J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

LIAPUNOV STABILITY AND CONTROLS*

SOLOMON LEFSCHETZ†

Let me consider first the elements of Liapunov's theory which are actually involved. I presume that you all know his basic stability definitions and I shall only describe the theorems required.

Let then

$$\dot{u} = U(u), \qquad U(0) = 0, \tag{1}$$

be an n-vector system where $U(u)$ is continuous for all u. Denote by $V(u)$ a positive definite Liapunov function for all u; $V(u)$ is of class C^1 for all u and $V(0) = 0$. As a consequence along the paths of (1), $\dot{V} = dV(u(t))/dt = (\partial V/\partial u) \cdot U$ for all solutions $u(t)$ of (1).

THEOREM 1. (*Liapunov*). *If* $-\dot{V} > 0$ *for all* $u \neq 0$ *then the origin is asymptotically stable in the large.*

THEOREM 2. (*Barbašin-Krasovskiĭ complement*). *If* $V(u) \to \infty$ *with* $\| u \|$ *then every solution* $u(t)$ *of* (1) *approaches* 0 *as* $t \to +\infty$.

These two propositions may be combined as the

L-B-K THEOREM. *If* V *and* $-\dot{V}$ *are positive definite for all* u *and* $V \to \infty$ *with* $\| u \|$, *then all solutions of* (1) *approach* 0 *as* $t \to \infty$.

AUXILIARY PROPERTY 3. (*Special case of Liapunov's instability theorem.*) *If there exists a Liapunov function* $V(u)$ *in the whole* u *space such that both* V *and* $\dot{V}$ *are positive definite then all solutions* $u(t)$ *are unbounded.*

This last property leads directly to Theorem 2. One closes the u-space at infinity by a point. Thus the space becomes a sphere with the origin as South pole S and infinity as North pole N. If V behaves as in Theorem 2, then $1/V = W$ behaves in accordance with Property 3 relative to N and yields Theorem 2.

After these preliminaries we are ready to discuss controls. A. Lurie has had the great merit to synthesize into a comparatively simple system a vast collection of controls. His system (1947) improved by Popov (1960) is

$$\begin{aligned} \dot{x} &= Ax - b\phi(\sigma), \\ \dot{\xi} &= \phi(\sigma), \\ \sigma &= c'x - \gamma. \end{aligned} \tag{2}$$

* Received by the editors October 7, 1964. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5-7, 1964.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island, and Princeton University, Princeton, New Jersey.

Here $\dot{x} = Ax$ is an n-vector equation in the deviation x from the normal, which one hopes to snuff out through imbedding in a larger $(n + 1)$-vector system control with parameter vectors b, c, scalar γ, scalar variable ξ, with a servo-characteristic $\phi(\sigma)$. Its admissible class is governed by: $\phi(\sigma)$ is continuous; $\phi(0) = 0$, $\sigma\phi(\sigma) > 0$; $\Phi(\sigma) = \int_0^\sigma \phi(\sigma)\, d\sigma$ diverges, as $|\sigma| \to \infty$. We also have the following general problem.

Problem of Lurie. To find necessary and sufficient conditions that the solutions of (2) all $\to 0$ as $t \to +\infty$, and this regardless of the choice of admissible $\phi(\sigma)$. This is *absolute stability.*

Now the requirements of design are merely to find *sufficient* conditions—*necessary* conditions being just a nice mathematical complement if one can find them also.

Several observations are now in order:

I. If $\gamma = 0$, the significant part of (2) reduces to

$$\dot{x} = Ax - b\phi(c'x), \tag{3}$$

and ξ ceases to play any role. The system (3) characterizes a *direct* control, (2) an *indirect* control. Practically, (2) is the more important scheme; but formally, each of the two types may be reduced to a special case of the other.

We will concentrate on (2) and so assume $\gamma \neq 0$.

II. Since $\gamma \neq 0$ one may replace ξ by σ in (2) and obtain the equivalent system,

$$\dot{x} = Ax - b\phi(\sigma), \qquad \dot{\sigma} = c'Ax - \rho\phi(\sigma), \qquad \rho = c'b + \gamma. \tag{4}$$

III. Absolute stability requires that for $\phi = \mu\sigma$, $\mu > 0$, the linearized system

$$\dot{x} = Ax - \mu b\sigma, \qquad \dot{\sigma} = c'Ax - \rho\mu\sigma, \tag{5}$$

be asymptotically stable for all $\mu > 0$, and in particular for $\mu > 0$ and small. This leads quite readily to the following two properties:

IV. No characteristic root of A may have positive real part.

V. $\gamma > 0$.

VI. Still playing with (5) one may show that:

(a) Zero is at most a simple characteristic root of A;

(b) $i\omega$ is at most a double root of A.

These two cases are the *critical* cases.

Stability. The whole trouble is due of course to the presence of the very general nonlinear characteristic $\phi(\sigma)$. At all events the only theory at our disposal is that of Liapunov. One must look therefore for a function $V(x, \sigma)$ with suitable behavior. The first general proposal (Lurie-Postnikov) was

a function

$$V(x, \sigma) = x'Bx + \Phi(\sigma), \tag{6}$$

where the first term is a quadratic form. At once

$$-\dot{V}(x, \sigma) = x'Cx + 2d'x\phi + \rho\phi^2, \tag{7}$$

$$A'B + BA = -C, \tag{8a}$$

$$d = Bb - \tfrac{1}{2}A'c. \tag{8b}$$

One will assume that A is stable (perhaps merely feebly). It turns out that the following are necessary and sufficient conditions to satisfy the L-B-K Theorem for all admissible ϕ, and hence sufficient to obtain absolute stability:

$$\text{existence of a matrix } C > 0, \tag{9a}$$

$$\rho > d'C^{-1}d \ (\Rightarrow \rho > 0). \tag{9b}$$

These are strong conditions on a whole matrix C.

The theorems of V. M. Popov. This was the situation until 1960, the date when Popov of Romania came out with two very strong theorems which we shall now discuss.

Set $A_z = zE - A$, so that $|A_z| = 0$ is the characteristic equation of A. Note in particular that since A is stable $|A_{i\omega}| \neq 0$ for all real ω; hence $A_{i\omega}^{-1}$ exists for all such ω.

We will state the two theorems of Popov; both refer to the system (2).

First Theorem of Popov. *A sufficient condition for absolute stability of the indirect control* (2) *is the existence of* $q \geqq 0$ *such that*

$$P(q, \omega) = q\gamma + \operatorname{Re}\{(1 + i\omega q)c'A_{i\omega}^{-1}b\} \geqq 0 \tag{10}$$

for all real ω.

Second Theorem of Popov. *A necessary condition to have absolute stability determined by means of a Liapunov function of the form "quadratic form in* x *and* σ *plus* $\beta\Phi(\sigma)$*" and the L-B-K Theorem, is that* (10) *hold for some* $q \geqq 0$ *and all real* ω.

Remarks on the first theorem. In the first place this theorem is already sufficient for the technical applications, all the more so since it only involves finding a real constant q and not the $n(n + 1)/2$ terms of a real matrix, namely C. This is already striking enough. Actually the discovery of q is quite simple. We have in fact $P(q, \omega) = S_1(\omega) - \omega q S_2(\omega)$, where S_1 and ωS_2 are real rational functions. The curve Γ represented parametrically by $y = S_1(\omega)$, $x = \omega S_2(\omega)$, is of the type known as *rational* and may be drawn with little difficulty. The inequality $P \geqq 0$ merely asserts this:

(a) if $q > 0$, the curve Γ is above the line $L: y - qx = 0$;

(b) if $q = 0$, the curve is above the x-axis;
(c) if $q = \infty$, the curve is to the left of the y-axis.
Whether (a), (b) or (c) holds may be readily verified.

It is also very remarkable that the first Popov theorem has been obtained without reference to Liapunov's general theory.

Remarks on the second theorem. The second theorem refers to a Liapunov function more complicated than the first,

$$V(x, \sigma) = x'Bx + \alpha(\sigma - c'x)^2 + \beta\Phi(\sigma). \tag{11}$$

Actually a priori V should have an additional term in $x\sigma$. However it is readily shown that for absolute stability this term must be absent. From (11) there follows

$$-\dot{V} = x'Cx + 2d'x\phi + \beta\rho\phi^2 + 2\alpha\gamma\sigma\phi, \quad d = Bb - \tfrac{1}{2}\beta A_c{}' - \alpha\gamma c. \tag{12}$$

The "ϵ-method" yields at once preliminary results. It demands that under the substitutions $x, \sigma, \phi \to \epsilon x, \sigma, \epsilon^3\phi$, V be positive. If $\alpha \neq 0$ its sign is that of $\alpha\sigma^2$ for ϵ small and $\alpha \neq 0$, and so $\alpha \geqq 0$. Similarly $\beta \geqq 0$ and also $\alpha + \beta > 0$. Hence the companion necessary conditions to (11) are

$$\alpha \geqq 0, \qquad \beta \geqq 0, \qquad \alpha + \beta > 0. \tag{11a}$$

This is assumed throughout.

Remarks on the two theorems. A widely open question is this: Can sufficiency of the first theorem be replaced by "necessary and sufficient condition", which might be strengthened by some simple condition such as (11a)? This problem is still unsolved.

Outline of proof of Popov's first theorem. We deal directly with system (2). After very extensive and complicated analytical work, resting above all upon Fourier transforms, it is shown as a consequence of the inequality (10) that $|\xi(t)|$ is bounded as $t \to +\infty$. From this point on, boundedness of $|\xi(t)|$ leads to that of $\|x(t)\|$, and more precisely to the proof that the solution $(x(t), \xi(t))$ of (2) is stable in the sense of Liapunov. Further analysis proves then that this solution approaches 0 whatever admissible ϕ, that is, absolute stability is achieved. (Details must be abandoned as they would require far more time than is reasonable.)

Outline of proof of the second theorem. It is convenient in (10) to replace q by $\beta/2\alpha$, so that (10) becomes

$$P(\alpha, \beta, \omega) = \beta\gamma + \operatorname{Re}\{(2\alpha\gamma + i\omega\beta)c'A_{i\omega}^{-1}b\} \geqq 0, \qquad \alpha \geqq 0, \qquad \beta \geqq 0, \qquad \alpha + \beta > 0. \tag{13}$$

This will enable us to consider in a very natural manner the value $q = \infty$.

We will now prove that *if* V, $-\dot{V}$ *given by* (11), (12) *are positive definite*

for all x, σ, *then* (13) *holds with the same* α, β *as in* (11), (12). This is the content of the second theorem.

A preliminary step is required. If $u'Fu$ is a real bilinear form and is positive whenever $u \neq 0$, we will write $F > 0$. Then the proof of the following is elementary.

Property $F > 0$ *for all real* u *is equivalent to*

$$\text{Re } v^*Fv > 0 \text{ for all complex } v \neq 0. \tag{14}$$

We have already seen that $V \geqq 0$ implies (11a). Write now

$$-\dot{V} = -2x'B(Ax - b\phi) + 2\gamma\phi(\sigma - c'x) - \beta\phi\{c'(Ax - b\phi) - \gamma\phi\} > 0.$$

If we apply (14), observe the relation $i\omega m(i\omega) = i\omega A_{i\omega}^{-1}b = AA_{i\omega}^{-1}b + b$, and substitute in $-\dot{V}$: $x = -m$, $\phi = \mu\sigma$, $\mu > 0$, $\sigma = 1/\mu$, then after some simplifications, we obtain

$$\frac{2\alpha\gamma}{\mu} + P(\alpha, \beta, \omega) > 0. \tag{15}$$

Since this must hold for all $\mu > 0$, (13) follows.

Complementary remark. Suppose that $\alpha = 0$; hence $\beta > 0$. Then one may divide V by β so that V becomes the function (6). The relation (15) yields then the stronger necessary condition

$$P(0, 1, \omega) > 0.$$

Some further recent results. The results of Popov stimulated a return, notably by Yacubovich and Kalman, to earlier more algebraic procedures, with strict adherence however to Liapunov's ideas.

Let the pair (A, b) $[(c', A)]$ be called *completely controllable* (denoted c.c.) [*completely observable* (denoted c.o.)] whenever the vectors A^hb, $0 \leqq h < n$, are linearly independent [whenever (A', c) is c.c.]. If both hold we say that (A, b, c) is c.c.o (Kalman). If it is not, one may replace (2) by a system of lower order with (A, b, c) c.c.o. In what follows we accept this preliminary reduction. We have then an important lemma.

LEMMA OF YACUBOVICH. *Let* A *be stable. Consider the system in the real matrix* B *and vector* q,

$$\begin{aligned} A'B + BA &= -\epsilon D - qq', \\ Bb - k &= \sqrt{\tau}q, \end{aligned} \tag{16}$$

where b, k *are real vectors,* $b \neq 0$; *with* $\epsilon > 0$, $\tau \geqq 0$ *real scalars and* D *a symmetric matrix* >0. *A necessary and sufficient condition for the existence of a solution* (B, q) *of* (16) *is that* ϵ *be small enough and that Kalman's*

relation,

$$\tau + 2 \operatorname{Re} (k' A_{i\omega}^{-1} b) > 0, \tag{17}$$

hold for every real ω.

By taking $\tau = \beta\gamma$, $k = \frac{1}{2}\beta A'c + \alpha\gamma c$, (17) reduces to $P(\alpha, \beta, \omega)$.

On the basis of the lemma, I proved the following theorem.

THEOREM. *The necessary and sufficient conditions to have* V *and* $-\dot{V}$ *of* (10), (11) *satisfy the L-B-K Theorem, and hence yield absolute stability, are* $P(\alpha, \beta, \omega) > 0$ *for all real* ω *plus* $\tau > 0$, *or else* $\tau = 0$, $d = 0$, $\alpha > 0$.

A stronger result was obtained by Kalman in replacing $\dot{V} < 0$ by $\dot{V} \leqq 0$, using a result of LaSalle, plus $P \geqq 0$, plus another quite complicated relation.

REFERENCES

M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Control Systems*, 1963; English transl., Holden-Day, San Francisco, 1963.

SOLOMON LEFSCHETZ, *Stability of Nonlinear Control Systems*, Academic Press, New York, 1965.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

AN EXISTENCE THEOREM IN PROBLEMS OF OPTIMAL CONTROL*

LAMBERTO CESARI†

1. Introduction. In the present paper we aim at a new existence theorem for the problem of Pontryagin in the mathematical theory of optimal control. Further existence theorems for the problems of Pontryagin and Lagrange will appear elsewhere.

As usual we denote by $I[x, u] = \int_{t_1}^{t_2} f_0(t, x, u)dt$ the "cost," that is, the functional which has to be minimized, by $u = u(t)$, $t_1 \leqq t \leqq t_2$, the "control function" with values $u(t) = (u_1, \cdots, u_m)$ in a given compact subset U of the m-dimensional Euclidean space E_m, and by $x = x(t) = (x_1, \cdots, x_n)$ a corresponding trajectory in the Euclidean space E_n, satisfying n given differential equations

$$\frac{dx_i}{dt} = f_i(t, x(t), u(t)), \qquad i = 1, \cdots, n,$$

and taking given initial and final values (see §2 for details and extensions). We shall denote by $f = f(t, x, u)$ the n-vector $f = (f_1, \cdots, f_n)$, and by $\tilde{f}$ the $(n + 1)$-vector $\tilde{f} = (f_0, f_1, \cdots, f_n)$. The subset U of E_m above may be fixed or, more generally, variable with t and x, and we write $U(t, x)$.

A number of important existence theorems have been proved so far. We mention here the existence theorems for the problem of "minimum time" ($f_0 = 1$) when f has the form $f = Ax + Bu$, A, B constant matrices, and U has the form $[|u_j| \leqq 1, j = 1, \cdots, m]$ by R. V. Gamkrelidze [3], successively extended by Pontryagin [10] to the case where U is a convex finite polyhedron in E_m. We mention the results of L. Markus and E. B. Lee [8] and the more general statements of A. F. Filippov [2] and E. Roxin [11]. L. Markus and E. B. Lee, and E. Roxin showed also how the lack of convexity conditions may easily lead to examples of Pontryagin's problem with no optimal solution. The most significant condition which is requested in Filippov's results is that the set $\tilde{Q}(t, x) = \tilde{f}(t, x, U(t, x))$ be convex, that is, for every (t, x), $\tilde{f}$ transforms $U(t, x)$ in a convex subset of the $(n + 1)$-dimensional $(x_0x_1 \cdots x_n)$-Euclidean space E_{n+1}.

In the calculus of variations for free problems (that is, when no differ-

* Received by the editors September 28, 1964, and in final revised form March 15, 1965. Presented in discussion at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan. This research was partially supported by the National Science Foundation under Grant G-57 at the University of Michigan.

ential equation is involved) the condition that f_0 be convex with respect to u is relevant for the existence of a minimizing solution, since then the cost is a lower semicontinuous functional. We have been exploring the possibility of using this condition in the general Pontryagin problem. Of course, this condition alone does not assure existence when a differential system is involved.

In the present paper we give an existence theorem by combining two conditions: (1) for each (t, x) the scalar f_0 is a convex function of u in the compact set $U(t, x)$; (2) for each (t, x), $Q(t, x) = f(t, x, U(t, x))$ is a convex subset of the n-dimensional Euclidean $(x_1 \cdots x_n)$-space E_n (while $\tilde{Q}(t, x)$ may not be convex in the $(n + 1)$-dimensional $(x_0 x_1 \cdots x_n)$-space E_{n+1}); (3) the "curvature" of f is small with respect to the "degree of convexity" of f_0 (see §2 for details).

The theorem of existence stated here (§4) combines Tonelli's direct method of the calculus of variations based on lower semicontinuity and convexity, with the typical reasoning of Filippov. The complete proofs are given in [1b]. Under conditions (1) and (2) new phenomena occur which are studied. Examples are shown to the effect that conditions (1) and (2) alone do not assure existence (§5).

Nevertheless, there is a situation where no convexity of f_0 or f is needed at all, as L. Neustadt [9] pointed out in a recent paper, namely, when f and f_0 are *linear* in x, that is, when $f_i = a_i(t)x + \varphi_i(u, t)$, $i = 0, 1, \cdots, n$, and U is compact.

2. Notations. Let t denote the independent variable, $t \in E_0 = [t_1 \leqq t \leqq T]$ let $x = (x_1, \cdots, x_n)$ be a vector representing the *state* of the system, $x \in E_n$. For each $(t, x) \in E_0 \times E_n$ let $U(t, x)$ be a set of vectors $u = (u_1, \cdots, u_n)$, or $U(t, x) \in E_n$. The set $U(t, x)$ may be fixed, or variable with t and x for $(t, x) \in E_0 \times E_n$. The variable vector $u \in U(t, x)$ represents the position of the *regulator*, and $U = U(t, x)$ is said to be the *control space*. Let $x_{10} = (x_{11}, \cdots, x_{n1})$ be a fixed point of E_n. Let $f_i(t, x, u)$, $i = 0, 1, \cdots, n$, be functions of t, x, u defined for $(t, x) \in E_0 \times E_n$, $u \in U(t, x)$, and let f and $\tilde{f}$ be the two vectors $f = (f_1, \cdots, f_n)$, $\tilde{f} = (f_0, f_1, \cdots, f_n)$. When needed, we shall use the auxiliary variable $x_0 \in E_1$, and then $\tilde{x}$ will denote the vector $\tilde{x} = (x_0, x_1, \cdots, x_n)$, $\tilde{x} \in E_{n+1} = E_1 \times E_n$.

Let $K = [x(t), u(t), t_1 \leqq t \leqq t_2]$ be the set of all *admissible strategies*, or *control functions* $u(t)$, and corresponding *trajectories* $x(t)$, that is, the set of all pairs of vector functions $u(t) = (u_1, \cdots, u_m)$, $x(t) = (x_1, \cdots, x_n)$, $t_1 \leqq t \leqq t_2$, with $t_1 \leqq t_2 \leqq T$, such that (1) $u(t)$ is measurable in $[t_1, t_2]$; (2) the differential system

$$\frac{dx_i}{dt} = f_i(t, x, u(t)), \quad i = 1, \cdots, n, \quad t_1 \leqq t \leqq t_2, \tag{1}$$

or

$$x' = f(t, x, u(t)), \qquad t_1 \leqq t \leqq t_2,$$

admits of the solution $x(t) = (x_1, \cdots, x_n)$, whose components $x_i(t)$ are absolutely continuous functions in $[t_1, t_2]$, satisfy (1) almost everywhere, and have initial values $x_i(t_1) = x_{i1}$, $i = 1, \cdots, n$, or $x(t_1) = x_{10} = (x_{11}, \cdots, x_{n1})$; (3) $u(t) \in U(t, x(t))$ for every $t_1 \leqq t \leqq t_2$.

The vector function $x(t)$, $t_1 \leqq t \leqq t_2$, is said to be a *trajectory* relative to the admissible control function $u(t)$ with initial point x_{10} and initial time t_1 (which are assumed to be fixed), and final point $x(t_2) = x_{20} = (x_{12}, \cdots, x_{n2})$ and final time t_2 both for the moment undetermined.

The points x_{20} which are final points of some trajectory $x(t)$ corresponding to some admissible control function $u(t)$ are said to be *accessible* from x_{10} (that is, starting from x_{10} at time t_1). Finally, we shall denote by M the set of all points (t, x, u) with $t \in E_0 = [t_1, T]$, $x \in E_n$, $u \in U(t, x)$.

If $b = (b_1, \cdots, b_m)$, $u = (u_1, \cdots, u_m)$ are any two vectors of the same dimension, we denote by $b \cdot u$ the usual inner product $b_1u_1 + \cdots + b_mu_m$. Therefore, each linear function in u can be written in the form $z(u) = r + b \cdot u$, $u \in E_m$, where r is a scalar, b a vector, (r, b constants), and we have also, for every $u_0 \in E_m$, $u_0 = (u_{01}, \cdots, u_{0m})$,

$$z(u) = r + b \cdot u = z(u_0) + b(u - u_0).$$

We shall denote by $|x| = (x \cdot x)^{1/2}$ the usual Euclidean norm of a vector x.

3. Hypotheses. (C) *Hypotheses of continuity and compactness.* (C_1) The functions f_i are continuous on M; (C_2) for every $t \in E_0$ and $x \in E_n$, the set $U(t, x)$ is compact; (C_3) the set $U(t, x)$, $(t, x) \in E_0 \times E_n$ is an upper semicontinuous function of (t, x), that is, we assume that, given $\epsilon > 0$, $t_0 \in E_0$, $x_0 \in E_n$, there exists a $\delta = \delta(\epsilon, t_0, x_0) > 0$ such that $U(t, x) \subset [U(t_0, x_0)]_\epsilon$ for every $(t, x) \in E_0 \times E_n$ with $|t - t_0| < \delta$, $|x - x_0| < \delta$, where U_ϵ denotes the closed neighborhood of U of radius ϵ in E_m; (C_4) there exists a constant $C > 0$ such that $x_1f_1 + \cdots + x_nf_n \leqq C(|x|^2 + 1)$ for every $t \in E_0$, $x \in E_n$, $u \in U(t, x)$.

The condition (C_4) assures that the trajectories $x(t)$, $t_1 \leqq t \leqq t_2$, with $x(t_1) = x_{10}$, $t_1 \leqq t_2 \leqq T$, (t_1, x_{10} fixed) belong to some bounded closed set D of the space E_n (see [2]). Indeed, if $z(t) = |x(t)|^2 + 1$, or $z(t) = x_1^2 + \cdots + x_n^2 + 1$, then, by (1) and ($C_4$), $z' = 2(x \cdot f) \leqq 2Cz$, hence $z(t) \leqq z(0) \exp[2C(t - t_1)] \leqq z(0) \exp[2C(T - t_1)]$. We can take for D a solid closed ball in E_n.

We shall denote by M_T the set of all (t, x, u) with $t \in E_0 = [t_1, T]$, $x \in D$, $u \in U(t, x)$; hence $M_T \subset M$. The hypotheses (C_2) and (C_3) together assure that the set M_T is compact (see [2]; for another proof, [1b]).

The condition (C_1) then assures that, for each $(t, x) \in E_0 \times D$, the set $Q(t, x) = f(t, x, U(t, x)) \subset E_n$ is compact, and $Q(t, x)$ describes a compact

set $\mathfrak{M} \subset E_n$ as (t, x) describes $E_0 \times D$, say $\mathfrak{M} = f(M_T)$. The functions $f_i(t, x, u)$, $i = 1, 2, \cdots, n$, are bounded in M_T, say $|f_i(t, x, u)| \leqq N$ for $(t, x, u) \in M_T$. As a consequence, the components $x_i(t)$, $i = 1, \cdots, n$, of the trajectories $x(t)$ are uniformly Lipschitzian with constant N, and are also bounded since they lie in D. We shall take N in such a way that also for f_0 we have $|f_0(t, x, u)| \leqq N$ for $(t, x, u) \in M_T$.

Finally, the components $u_j(t)$, $j = 1, \cdots, n$, of the strategies $u(t) \in K$, or control functions, are also uniformly bounded since $(t, x(t), u(t)) \in M_T$ and M_T is a bounded subset of the $(m + n + 1)$-dimensional (txu)-space E_{m+n+1}. We shall take N so that

$$
\begin{aligned}
|f_i(t, x(t), u(t))| &\leqq N, \qquad i = 0, 1, \cdots, n, \\
|u_j(t)| &\leqq N, \qquad j = 1, \cdots, m,
\end{aligned}
\tag{2}
$$

for every $t_1 \leqq t \leqq t_2 \leqq T$, and $u(t) \in K$.

For every $(t, x) \in E_0 \times D$ the set $U(t, x)$ is compact, but not necessarily convex. We shall denote by $U^*(t, x)$ the convex, closed and therefore compact hull of $U(t, x)$. We shall need the following further hypotheses.

(C_1') The functions $f_i(t, x, u)$, $i = 1, \cdots, n$, are defined and continuous on the compact set M_T^* of all (t, x, u) with $t \in E_0 = [t_1, T]$, $x \in D$, $u \in U^*(t, x)$.

If it happens that for each $(t, x) \in E_0 \times D$ the set $U(t, x)$ is convex, then condition (C_1') reduces to the condition (C_1).

(I) *Hypothesis of convexity.* For every $(t, x) \in E_0 \times D$ the set $Q(t, x) = f(t, x, U(t, x))$ is convex.

Under this hypothesis and (C) then $Q(t, x)$ is convex and compact. We shall now measure the degree of "convexity" of the scalar function f_0 and the "curvature" of the n-vector $f = (f_1, \cdots, f_n)$. This we do by means of the following hypotheses (or definition of convexity).

(α) *Hypothesis of convexity of* f_0. There is a nonnegative bounded and Borel measurable function $C = C(t, x, u)$, $(t, x, u) \in M_T^*$, with the following property: for each $\epsilon > 0$ and $(t_0, x_0, u_0) \in M_T^*$ there are $\delta = \delta(t_0, x_0, u_0, \epsilon) > 0$ and a linear function $z(u) = r + b \cdot u$ (also depending on t_0, x_0, u_0, ϵ) such that, for every $(t, x) \in E_0 \times D$ at a distance $\leqq \delta$ from (t_0, x_0) we have $f_0(t, x, u) \geqq z(u) + C|u - u_0|^2$ for each $u \in U^*(t, x)$; $f_0(t, x, u) \leqq z(u) + \epsilon$ for each $u \in U^*(t, x)$ with $|u - u_0| \leqq \delta$.

Condition (α) is certainly satisfied if, for each (t, x), the function f_0 is of class C^2 in $u = (u_1, \cdots, u_m)$, if the second partial derivatives f_{0hk} are continuous in M_T^*, and if the quadratic form

$$
\sum_{hk} f_{0hk}(t, x, u)\xi_h \xi_k, \qquad f_{0hk} = \frac{\partial^2 f_0}{\partial u_h \, \partial u_k},
$$

$$
h, k = 1, \cdots, m; \ \xi_1, \cdots, \xi_m \text{ real},
$$

is positive semidefinite (positive definite if we want $C > 0$). Condition (α) as given above is only a generalized form of this familiar condition, which does not require second order partial derivatives. This generalized form of stating convexity is often used in the calculus of variations.

(β) *Hypotheses of boundedness of the curvature of* f. There exists a nonnegative, bounded, Borel measurable function $D = D(t, x, u)$, $(t, x, u) \in M_T^*$, with the following property: for each $\epsilon > 0$, $(t_0, x_0) \in E_0 \times D$, $u_0 \in U^*(t_0, x_0)$, there are a $\delta = \delta(t_0, x_0, u_0) > 0$ and a linear function $Z(u) = R + Bu$ (R an n-vector, B an $n \times m$ matrix, δ, R, B depending on t_0, x_0, u_0, ϵ), such that for each $(t, x) \in E_0 \times D$ at a distance $\leqq \delta$ from (t_0, x_0) we have $|f(t, x, u) - Z(u)| \leqq \epsilon + D|u - u_0|^2$ for each $u \in U^*(t, x)$.

Condition (β) is certainly satisfied if, for each $(t, x) \in E_0 \times D$, $f(t, x, u)$ is of class C^2 in $u = (u_1, \cdots, u_m)$ with second order partial derivatives continuous in M_T^*, and

$$\sum_{i=1}^{n} \left| \sum_{h,k} f_{ihk}(t, x, u)\xi_h\xi_k \right| \leqq 2D(t, x, u)(\xi_1^2 + \cdots + \xi_m^2),$$

where $f_{ihk} = \partial^2 f_i / \partial u_h \partial u_k$, and $\sum_{hk}$ is taken for all $h, k = 1, \cdots, m$.

Finally, we require certain Lipschitz-type requirements on both the scalar function f_0 and the vector function $f = (f_1, \cdots, f_n)$.

(γ) *Lipschitz-type conditions for* f_0 *and* f. There are two functions $\Lambda(t, x, u)$, $L(t, x, u)$, $(t, x, u) \in M_T^*$, both nonnegative and Borel measurable, with the following properties:

(γ_1) $$|f_0(t, x, u) - f_0(t, x, u_0)| \leqq L(t, x, u_0)|u - u_0|$$

for each $(t, x) \in E_0 \times D$ and any two points $u, u_0 \in U^*(t, x)$;

(γ_2) if, for every $(t, x) \in E_0 \times D$, for every n-vector $z_0 = f(t, x, u_0)$, $u_0 \in U^*(t, x)$, and for every other n-vector $z \in f(t, x, U^*(t, x))$ we take $z = f(t, x, u)$, $u \in U^*(t, x)$ with $|u - u_0| =$ minimum, then we have

$$|u - u_0| \leqq \Lambda(t, x, u_0)|z - z_0|.$$

Condition (γ_2) obviously does not require the monotoneity of f in the vector u, a condition which would be impossible to verify if, for instance, $n < m$. Nevertheless, if $n \geqq m$, then condition (γ_2) is certainly verified if f is monotone in u and if

$$|u - u_0| \leqq \Lambda(t, x, u)|f(t, x, u) - f(t, x, u_0)|$$

for every $(t, x) \in E_0 \times D$ and any two $u, u_0 \in U^*(t, x)$. We say that $f(t, x, u)$ is monotone in u if $u, u_0 \in U^*(t, x)$, $u \neq u_0$, implies $f(t, x, u) \neq f(t, x, u_0)$.

4. Existence theorem for Pontryagin's problem.

THEOREM. *Under hypotheses* (C), (C$_1'$), (I), (α), (β), (γ), *let us assume that*

$$\Lambda(t, x, u)L(t, x, u)D(t, x, u) \leqq C(t, x, u), \qquad (t, x, u) \in M_T, \tag{2}$$

where the equality sign holds at most at points $(t, x, u) \in M_T^*$ *whose coordinates* t *lie in a subset of measure zero of* E_0. *Let us assume that the point* x_{20} *is accessible within the time* T *starting at the point* x_{10} *at the time* t_1. *Then there is an optimal solution* $u_0(t)$, $t_1 \leqq t \leqq t_2$, $u_0(t) \in K$, *for Pontryagin's problem*:

$$I = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt = \text{minimum}, u = (u_1, \cdots, u_m),$$

$$x'(t) = f(t, x(t), u(t)), t_1 \leqq t \leqq t_2, f = (f_1, \cdots, f_n), x = (x_1 \cdots, x_n),$$

$$(x(t), u(t)) \in K,$$

$$x(t_1) = x_{10}, x(t_2) = x_{20}, t_1 \leqq t_2 \leqq T, t_1, x_{10}, x_{20} \text{ fixed.}$$

We refer for the proof to [1b].

Remark 1. In the existence theorem above, hypothesis (2) simply states that the "curvature" of f is small with respect to the "convexity" of f_0. The comparison factor $L\Lambda$ depends "in a sense" on the "linear" parts only of f and f_0.

Remark 2. In the existence theorem above, the hypothesis that x_{20} is accessible can be replaced by the hypothesis that x_{20} is a point of accumulation of accessible points. The set of the points of accumulation is closed. This remark is essentially Roxin's (see [1b] for the proof).

Remark 3. In the existence theorem above, we can replace the fixed point-target x_{20} by a moving set-target $B(t)$ as in the paper by L. Markus and E. B. Lee [8]. We have only to require that $B(t)$ is a continuous set function. Precisely, we assume that $B(t)$, $t_1' \leqq t \leqq T$, is a variable subset of E_n such that (a) for each t the set $B(t)$ is closed, $t_1 \leqq t \leqq T$; (b) $B(t)$ varies with continuity, that is, we assume that, given $\epsilon > 0$, there is $\delta > 0$ such that t, $\bar{t} \in [t_1, T]$, $|t - \bar{t}| \leqq \delta$ imply $B(t) \subset [B(\bar{t})]_\epsilon$, $B(\bar{t}) \subset [B(t)]_\epsilon$ (see [1b] for the proof).

Remark 4. If $f_0 \geqq \nu$ for some constant $\nu > 0$, then the restriction $t_2 \leqq T < +\infty$ can be waived.

In the lines below we give examples of Pontryagin's problems where the existence theorem above can be applied, but not Filippov's theorem. Nevertheless, examples could be given where Filippov's theorem applies, but not the one above. The two theorems are independent. Also, we give an example where conditions (C), (C'), (I) are satisfied, as well as (α), (β), (γ) (thus certainly conditions 1 and 2 of the introduction), but not (2), and there is no optimal solution.

5. An example with no optimal solution. The following example shows that the condition represented by the logical union of (C), (I), and f_0 convex in u (thus, *both* the function f_0 is convex in u, and the subset $Q(t, x)$ of E_n is convex) is not sufficient to assure the existence of the absolute minimum of I.

Let us consider the differential system

$$x' = u(1 - v) + [2 - 2^{-1}(u - 1)^2]v,$$

$$y' = [2 - 2^{-1}(u - 1)^2](1 - v) + uv,$$

with $t_1 = 0$, initial point $(0, 0)$, fixed target $(0, 1)$, and fixed control space $U = [-1 \leqq u \leqq 1, -1 \leqq v \leqq 1]$. If

$$z_1 = f_1 = u(1 - v) + [2 - 2^{-1}(u - 1)^2]\, v,$$

$$z_2 = f_2 = [2 - 2^{-1}(u - 1)^2](1 - v) + uv,$$

we see that the segment $[v = 1, -1 \leqq u \leqq 1]$ is mapped by $f = (f_1, f_2)$ onto the arc of parabola $ABC = [z_1 = 2 - 2^{-1}(u - 1)^2, z_2 = u, -1 \leqq u \leqq 1]$, whose points $A = (0, -1)$, $B = (3/2, 0)$, $C = (2, 1)$ correspond to $u = -1, 0, 1$, respectively. The segment $[v = -1, -1 \leqq u \leqq 1]$ is mapped by f onto the arc $DEF = [z_1 = 2^{-1}u^2 + u - 3/2, z_2 = -u^2 + u + 3, -1 \leqq u \leqq 1]$, that is, of the parabola $(2z_1 + z_2)^2 - 6(z_1 - z_2) - 27 = 0$, whose points $D = (-2, 1)$, $E = (-3/2, 3)$, $F = (0, 3)$ correspond to $u = -1, 0, 1$, respectively. Each segment $[u = c, -1 \leqq v \leqq 1]$ is mapped by f onto the segment joining the points corresponding to $(c, 1)$ and $(c, -1)$ on the two parabolas. Thus, the image $Q = f(U)$ of U is the convex body $Q = (ABCFED)$ of the z_1z_2-plane. Let us take the cost functional

$$I = \int_{t_1}^{t_2} [x^2 + (y - t)^2 + (au + bv + c)^2]\, dt,$$

where $a = 2 + \sqrt{7} = 4.64575$, $b = -6(\sqrt{7} - \sqrt{3}) = -5.48220$, $c = -4 - \sqrt{3} + \sqrt{7} + \sqrt{21} = 1.49628$.

First we observe that the points $(\alpha_1, \beta_1) \in U$, $(\alpha_2, \beta_2) \in U$ with $\alpha_1 = 2 - \sqrt{3} = 0.26795$, $\beta_1 = 2^{-1}$, $\alpha_2 = 2 - \sqrt{7} = -0.64575$, $\beta_2 = 6^{-1}(1 - \sqrt{7}) = -0.27429$ are mapped by f into the points $(z_1 = 1, z_2 = 1)$, $(z_1 = -1, z_2 = 1)$. For $k = 1, 2, \cdots$, let $u_k(t), v_k(t), 0 \leqq t \leqq 1$, be defined by taking $u_k(t) = \alpha_1, v_k(t) = \beta_1$, or $u_k(t) = \alpha_2, v_k(t) = \beta_2$, according as t belongs to the intervals $k^{-1}(i - 1) \leqq t < k^{-1}(i - 1) + (2k)^{-1}$, or $k^{-1}(i - 1) + (2k)^{-1} \leqq t < k^{-1}i$, $i = 1, 2, \cdots, k$. Then the functions $x_k(t)$, $y_k(t)$, $0 \leqq t \leqq 1$, satisfy the differential equations $dx_k/dt = \pm 1$, $dy_k/dt = 1$, where we take $+1$, or -1, according as t belongs to one or the other of the two sets of intervals. Then $x_k(t) \to x_0(t) = 0$, $y_k(t) \to y_0(t) = t$ uniformly in $0 \leqq t \leqq 1$ as $k \to \infty$. If C_k, C_0 denote these trajectories, we say that $C_k \to C_0$.

The question as to whether C_0 is actually a trajectory, that is, whether there are admissible control functions $u_0(t)$, $v_0(t)$, $0 \leqq t \leqq 1$, whose corresponding trajectory is C_0, can be answered in the affirmative because of the convexity of Q. Actually, the point $(\alpha_0, \beta_0) \in U$ with $\alpha_0 = 2 - \sqrt{5} = -0.23607$, $\beta_0 = (11)^{-1}(4 - \sqrt{5}) = 0.16036$, is mapped by f into $(z_1 = 0, z_2 = 1)$, and thus $u_0(t) = \alpha_0$, $v_0(t) = \beta_0$, $0 \leqq t \leqq 1$, generate C_0. Now, according as t belongs to one or the other of the two systems of equal intervals above, we have respectively

$$au_k(t) + bv_k(t) + c = a\alpha_1 + b\beta_1 + c = 0,$$

$$au_k(t) + bv_k(t) + c = a\alpha_2 + b\beta_2 + c = 0,$$

and thus $au_k(t) + bv_k(t) + c = 0$ almost everywhere on $0 \leqq t \leqq 1$, $k = 1, 2, \cdots$. On the other hand $x_k(t) \to 0$, $y_k(t) \to t$ uniformly in $[0, 1]$, and thus, without computations, $I[C_k] \to 0$ as $k \to \infty$. Finally, $a\alpha_0 + b\beta_0 + c = -0.47957 \neq 0$, and

$$I[C_0] = \int_0^1 [0^2 + 0^2 + (a\alpha_0 + b\beta_0 + c)^2]\, dt > 0.$$

Let us prove that I has no absolute minimum in the class Ω of all trajectories satisfying the differential equations, boundary conditions and constraints above. Indeed, $I[C_k] \to 0$ shows that the infimum of $I[C]$ in Ω is zero, but this value cannot be attained by I in Ω. Indeed, $I[C] = 0$ implies $x = 0$, $y = t$, $au + bv + c = 0$ almost everywhere, and then the first two relations yield $u = \alpha_0$, $v = \beta_0$ almost everywhere, while $a\alpha_0 + b\beta_0 + c \neq 0$, a contradiction. Thus I cannot attain the value zero in Ω.

6. Examples where the existence theorem applies. I. Let $m = n = 1$. We consider the Pontryagin problem

$$I = \int_{t_1}^{t_2} (x^2 + u^2 + 1)\, dt = \text{minimum},$$

$$x' = Au + u^2, \qquad x, u \text{ scalars}, \qquad u \in U, \qquad U = [-1 \leqq u \leqq 1],$$

$$t_1 = 0, \qquad x(0) = 1, \qquad x(t_2) = 1, \qquad t_2 \text{ undetermined},$$

where A is a constant, $A > 4$. We have

$$f_0 = x^2 + u^2 + 1, \qquad f_{0u} = 2u, \qquad f_{0uu} = 2,$$

$$f = f_1 = Au + u^2, \qquad f_u = A + 2u, \qquad f_{uu} = 2.$$

We can take

$$C = 1, \qquad D = 1, \qquad \Lambda = (A - 2)^{-1}, \qquad L = 2,$$

since f is strictly increasing and $f_u = A + 2u \geqq A - 2 > 0$. Also, U is a segment and $Q = f(U)$ is also a segment (since $n = 1$), namely, the segment $[-A + 1 \leqq z \leqq A + 1]$ of the z-axis. Here U and Q are fixed sets, and they are compact and convex. The inequality $\Lambda LD < C$ reduces to the relation $A > 4$ which we required at the beginning. We have also $f_0 \geqq 1$, and the point $x = 1$ is certainly accessible from $x = 0$, since for $u = 1$ we have $x' = A + 1 > 5$ and $x = 1$ is reached in a time $t_2 < 1/5$. By our existence theorem, the problem above admits of an optimal solution. Let us note that $\tilde{f} = (f_0, f_1)$ and, for each x, the set $\tilde{Q} = \tilde{f}(U)$ is the arc of curve in the z_0z_1-plane:

$$z_0 = x^2 + u^2 + 1, \qquad z_1 = Au + u^2, \qquad -1 \leqq u \leqq 1,$$

and $\tilde{Q}$ is not a convex set. Thus Filippov's theorem cannot be applied.

II. Let $m = 2$, $n = 1$. Let us consider the Pontryagin problem

$$I = \int_{t_1}^{t_2} (u^2 + v^2 + x^2 + 1)\, dt = \text{minimum},$$

$$x' = u^2 + 2v^2 + Au + x, \qquad u, v, x \text{ scalars},$$

$$(u, v) \in U \equiv [|\, u \,| + |\, v \,| \leqq 1],$$

$$t_1 = 0, \qquad x(0) = 1, \qquad x(t_2) = 1, \qquad t_2 \text{ undetermined},$$

where A is a constant, $A > 10$. Let us note that U is convex and completely contained in the solid circle $u^2 + v^2 \leqq 1$. We have

$$f_0 = u^2 + v^2 + x^2 + 1, \qquad f_{0u} = 2u, \qquad f_{0v} = 2v,$$

$$|\operatorname{grad} f_0| = 2(u^2 + v^2)^{1/2} \leqq 2,$$

$$f_{0uu} = 2, \qquad f_{0vv} = 2, \qquad f_{0uv} = 0,$$

$$f_1 = u^2 + 2v^2 + Au + x, \qquad f_{1u} = 2u + A, \qquad f_{1v} = 4v,$$

$$|\operatorname{grad} f_1| \geqq A - 2,$$

$$f_{1uu} = 2, \qquad f_{1vv} = 4, \qquad f_{1uv} = 0.$$

We can take $C = 1$, $D = 2$, $L = 2$. We shall now find a value for Λ. First let us note that $f_1(u, v, x)$ has minimum and maximum values on U which are $-A + 1 + x$ and $A + 1 + x$, and these values are taken respectively at the vertices $(-1, 0)$, $(1, 0)$ of U. Let $z_0 = f_1(u_0, v_0, x)$ be a value taken by f_1 at some point, say $w_0 = (u_0, v_0) \in U$ (for some fixed x). Let z_1 be any other value also taken by f_1 on U (for the same x), and let $w_1 = (u_1, v_1)$ be the point of U closest to w_0 where $z_1 = f_1(u_1, v_1, x)$. Let S be the segment of U joining $w_0 = (u_0, v_0)$ to $(-1, 0)$ if $z_1 < z_0$, and to $(1, 0)$ if $z_1 > z_0$.

Thus f_1 takes on the value z_1 at some point $\bar{w} = (\bar{u}, \bar{v})$ of S, and $|w_1 - w_0| \leqq |\bar{w} - w_0|$. On the other hand, the directional cosines of S are $\alpha_1 \geqq 2^{-1/2}$, $\alpha_2 \leqq 2^{-1/2}$, and hence the directional derivative of f_1 along S in the direction of u increasing is

$$\frac{df_1}{ds} = f_{1u}\,\alpha_1 + f_{1v}\,\alpha_2 = (A + 2u)\alpha_1 + 4v\alpha_2 \geqq 2^{-1/2}(A + 2u - 4\,|\,v\,|).$$

The function $\zeta = 2u - 4\,|\,v\,|$ has the minimum value -4 on U, and this minimum is taken at the two points $(0, \pm 1)$. Thus $df_1/ds \geqq 2^{-1/2}(A - 4)$ along S, and

$$|w_1 - w_0| \leqq |\bar{w} - w_0| \leqq 2^{1/2}(A - 4)^{-1}\,|z_1 - z_0|.$$

We can take, therefore, $\Lambda = 2^{1/2}(A - 4)^{-1}$. The relation $\Lambda LD < C$ reduces now to $A > 4 + 4\sqrt{2}$ which is certainly satisfied since $A > 10$. Here $Q = f(U)$ is a segment since $n = 1$, and thus Q and U are both convex compact sets. Also, we have $f_0 \geqq 1$, and the point $x = 1$ is certainly accessible from $x = 0$, since, for $u = 1$, $v = 1$, we have $x' = A + 1 + x$, and the corresponding solution $x(t)$ with $x(0) = 0$ certainly goes beyond $x = 1$ in a finite time. By force of the existence theorem above, the problem II admits of an optimal solution.

Let us note that $\tilde{f} = (f_0, f_1)$, and, for each x, the set $\tilde{Q} = \tilde{f}(U)$ is the set of all (z_0, z_1) with $z_0 = f_0$, $z_1 = f_1$, $|u| + |v| \leqq 1$. Here z_0 can take its maximum value $x^2 + 2$ only at the four vertices $(\pm 1, 0)$, $(0, \pm 1)$ of U, and to these points there correspond points (z_0, z_1) with $z_0 = x^2 + 2$, $z_1 = A + 1 + x$, or $z_1 = -A + 1 + x$, or $z_1 = 2 + x$. Since these are the only points of $\tilde{Q}$ whose first coordinate z_0 is maximum, $z_0 = x^2 + 2$, it follows that $\tilde{Q}$ is not convex. Thus Filippov's theorem cannot be applied.

III. Let $m = 1$, $n = 2$. Let us consider the Pontryagin problem

$$I = \int_{t_1}^{t_2} (x^2 + y^2 + u^2 + 1)\,dt = \text{minimum},$$

$$x' = 2(u + 1)(u + 1 + A), \qquad y' = -1 + (u + 1)(u + 1 + A),$$

$$u \in U, \qquad U = [-1 \leqq u \leqq 1], \qquad x, y, u \text{ scalars},$$

$$t_1 = 0, \qquad x(0) = 0, \qquad y(0) = 0, \qquad x(t_2) = x_{12}, \qquad y(t_2) = y_{12},$$

where t_2 is undetermined, and A is a constant, $A > 2$. We have

$$f_0 = x^2 + y^2 + u^2 + 1, \qquad f_{0u} = 2u, \qquad f_{0uu} = 2,$$

$$f_1 = 2(u + 1)(u + 1 + A) = 2(u + 1)^2 + 2A(u + 1),$$

$$f_2 = -1 + (u + 1)(u + 1 + A) = (u + 1)^2 + A(u + 1) - 1,$$

and, if $f = (f_1, f_2)$, the set $Q = f(U)$ is a segment of the straight line $z_1 - 2z_2 = 2$. If $z(u)$ denotes $z(u) = (u+1)^2 + A(u+1)$, we see that $z_u = 2(u+1) + A \geqq A$, and $z_{uu} = 2$, thus $\operatorname{grad} f \geqq \sqrt{5}\, A$. We can take

$$C = 1, \qquad L = 2, \qquad \Lambda = (\sqrt{5}A)^{-1}, \qquad D = \sqrt{5}.$$

Thus, the inequality $\Lambda LD < C$ reduces to $A > 2$. We have also $f_0 \geqq 1$. By the existence theorem above we conclude that the problem III admits of an optimal solution for every point (x_{12}, y_{12}) which is accessible from $(0, 0)$.

Let us note that $\tilde{f} = (f_0, f_1, f_2)$ and that $\tilde{Q} = \tilde{f}(U)$ is the curve in $(z_0z_1z_2)$-space with $z_0 = f_0$, $z_1 = f_1$, $z_2 = f_2$, $-1 \leqq u \leqq 1$. Here the coordinate z_0 takes on its maximum value $x^2 + y^2 + 2$ (for fixed x, y) only at the endpoints of $\tilde{Q}$, that is, at the points $(x^2 + y^2 + 2, 0, -1)$, $(x^2 + y^2 + 2, 4(A+2), 4(A+2) - 1)$, and thus $\tilde{Q}$ is not a convex set. The theorem of Filippov cannot be applied.

IV. Let $m = n = 2$. We consider the Pontryagin problem

$$I = \int_0^{t_2} (x^2 + y^2 + u^2 + v^2 + 1)\, dt = \text{minimum},$$

$$x' = Au, \qquad y' = (1 - v)u^2 + Bv,$$

$$(u, v) \in U = [-1 \leqq u \leqq 1,\ 0 \leqq v \leqq 1],$$

$$x(0) = y(0) = 0, \qquad x(t_2) = 0, \qquad y(t_2) = 1,$$

where A, B are constants, $A \geqq 6$, $B \geqq 11$. We have

$$f_0 = x^2 + y^2 + u^2 + v^2 + 1, \qquad f = (f_1, f_2),$$

and we take

$$X = f_1 = Au, \qquad Y = f_2 = (1 - v)u^2 + Bv.$$

First the rectangle U is mapped by f onto the region $Q = f(U) = [-A \leqq X \leqq A,\ A^{-2}X^2 \leqq Y \leqq B]$, which is convex. Then we have

$$\begin{aligned} |f_0(u_1, v_1) - f_0(u_2, v_2)| &= |(u_1 - u_2)(u_1 + u_2) + (v_1 - v_2)(v_1 + v_2)| \\ &\leqq [(u_1 - u_2)^2 + (v_1 - v_2)^2]^{1/2}[(u_1 + u_2)^2 + (v_1 + v_2)^2]^{1/2} \\ &\leqq 2^{3/2}[(u_1 - u_2)^2 + (v_1 - v_2)^2]^{1/2} \end{aligned}$$

and thus $L = 2^{3/2}$. We have also

$$f_{0uu} = 2, \qquad f_{0uv} = 0, \qquad f_{0vv} = 2,$$

and therefore we can take $C = 1$. Then we have

$$X_1 - X_2 = A(u_1 - u_2),$$

$$\begin{aligned} Y_1 - Y_2 &= [(1 - v_1)u_1^2 + Bv_1] - [(1 - v_2)u_2^2 + Bv_2] \\ &= B(v_1 - v_2) + (1 - v_1)(u_1^2 - u_2^2) + u_2^2[(1 - v_1) - (1 - v_2)] \\ &= (B - u_2^2)(v_1 - v_2) + (1 - v_1)(u_1 - u_2)(u_1 + u_2). \end{aligned}$$

Since $B - u_2^2 \geqq B - 1$, we obtain

$$v_1 - v_2 = (B - u_2^2)^{-1}[(Y_1 - Y_2) - (1 - v_1)(u_1 - u_2)(u_1 + u_2)],$$

$$| v_1 - v_2 | \leqq (B - 1)^{-1}[| Y_1 - Y_2 | + 2A^{-1} | X_1 - X_2 |],$$

and hence

$$\begin{aligned} (u_1 - u_2)^2 + (v_1 - v_2)^2 &\leqq A^{-2}(X_1 - X_2)^2 \\ &\quad + (B - 1)^{-2}[(X_1 - X_2)^2 + (Y_1 - Y_2)^2](1 + 4A^{-2}) \\ &= [A^{-2} + (B - 1)^{-2}(1 + 4A^{-2})](X_1 - X_2)^2 \\ &\qquad + (B - 1)^{-2}(1 + 4A^{-2})(Y_1 - Y_2^2). \end{aligned}$$

Since $A \geqq 6$, $B \geqq 11$, we have

$$(B - 1)^{-2}(1 + 4A^{-2}) \leqq (100)^{-1}(1 + 4(36)^{-1}) = (90)^{-1} < 4 \cdot 9^{-2},$$

$$\begin{aligned} A^{-2} + (B - 1)^{-2}(1 + 4A^{-2}) &\leqq (36)^{-1} + (100)^{-1}(1 + 4(36)^{-1}) \\ &= (36)^{-1} + (90)^{-1} < 4 \cdot 9^{-2}, \end{aligned}$$

and finally

$$(u_1 - u_2)^2 + (v_1 - v_2)^2 \leqq 4 \cdot 9^{-2}[(X_1 - X_2)^2 + (Y_1 - Y_2)^2].$$

Thus, we can take $\Lambda = 2/9$. Next, $X_{uu} = X_{uv} = X_{vv} = 0$, $Y_{uu} = 2(1 - v)$, $Y_{uv} = -2u$, $Y_{vv} = 0$, $| Y_{uu} | \leqq 2$, $| Y_{uv} | \leqq 2$, and

$$| \xi^2 + 2\xi\eta | \leqq 2\xi^2 + \eta^2 \leqq 2(\xi^2 + \eta^2).$$

We can take $D = 1$. Finally,

$$\Lambda LD = \frac{2}{9} \cdot 2\sqrt{2} \cdot 1 = \frac{4\sqrt{2}}{9} < 1 = C.$$

The conditions of the existence theorem are satisfied, and thus problem IV admits of an optimal solution for every point (x_2, y_2) which is accessible. For instance, $x = 0$, $y = 1$ is accessible since, by taking $u = 0$, $v = 1$, we have $x' = 0$, $y' = B$, hence $x \equiv 0$, $y \equiv Bt$, and the point $x = 0$, $y = 1$ is certainly reached. Let us note that, in this example, f_2 is not a monotone function of u.

Let us note that $\tilde{f} = (f_0, f_1, f_2)$ and $\tilde{Q} = \tilde{f}(U)$ is a set of points (Z, X, Y)

with $Z = f_0$. The set $\tilde{Q}$ contains the points $(3 + x^2 + y^2, 1, B)$, $(3 + x^2 + y^2, -1, B)$ of maximum Z-coordinate, but no point of the segment which joins them. Thus $\tilde{Q}$ is not convex, and Filippov's theorem cannot be applied.

V. Let $m = 2$, $n = 1$. We consider the Pontryagin problem

$$I = \int_{t_1}^{t_2} (x^2 + u^2 + v^2 + 1)\, dt = \text{minimum},$$

$$x' = (1 - v^2)u^2 + Bv, \qquad (u, v) \in U = [-1 \leqq u \leqq 1, -1 \leqq v \leqq 1],$$

$$t_1 = 0, \qquad x(0) = y(0) = 0, \qquad x(t_2) = 0, \qquad y(t_2) = 1,$$

where t_2 is undetermined, and $B \geqq 8$ is a constant. We have $f_0 = x^2 + u^2 + v^2 + 1$, and we can take, as in problem IV, $C = 1$, $L = 2^{3/2}$. Let us take $X = f_1 = (1 - v^2)u^2 + Bv$. Then $X_v = B - 2vu^2 \geqq B - 2 > 0$, and hence $X_v > 0$, and X is increasing with v. For $v = -1$ we have $X = -B$, for $v = 1$ we have $x = B$, independently of u. Thus $Q = f(U)$ is the interval $Q = [-B \leqq X \leqq B]$. Given $X_0 = (1 - v_0^2)u_0^2 + Bv_0$ and any other $X \neq X_0$, $-B \leqq X \leqq B$, there is some v such that $X = (1 - v^2)u_0^2 + Bv$, and we have

$$X - X_0 = (v_0^2 - v^2)u_0^2 + B(v - v_0),$$

$$|X - X_0| = |v - v_0|\,[B - (v + v_0)u_0^2] \geqq |v - v_0|(B - 2),$$

$$|v - v_0| \leqq (B - 2)^{-1}|X - X_0|.$$

Thus, if $(\bar{u}, \bar{v})$ is the point of U with $X = (1 - \bar{v}^2)\bar{u}^2 + B\bar{v}$, at a minimum distance from (u_0, v_0), we have

$$[(\bar{u} - u_0)^2 + (\bar{v} - v_0)^2]^{1/2} \leqq |v - v_0| \leqq (B - 2)^{-1}|X - X_0|.$$

This proves that we can take $\Lambda = (B - 2)^{-1}$. We have now

$$f_{uu} = 2(1 - v^2), \qquad f_{uv} = -4uv, \qquad f_{vv} = -2u^2,$$

$$|f_{uu}| \leqq 2, \qquad |f_{uv}| \leqq 4, \qquad |f_{vv}| \leqq 2,$$

and we can take $D = 2$. Thus

$$\Lambda LD = (B - 2)^{-1} \cdot 2^{3/2} \cdot 2 \leqq \frac{4\sqrt{2}}{6} < 1 = C.$$

By force of the existence theorem we conclude that problem V has an optimal solution for every point (x_2, y_2) which is accessible. In particular, the point $(0, 1)$ is accessible. Indeed, by taking $v \equiv 0$, $u \equiv 0$, the differential system reduces to $x' = 1$, hence $x = t$, and the point $(0, 1)$ is reached at the time $t_2 = 1$. Here f_2 is not a monotone function of u. If we take $\tilde{f} = (f_0, f_1)$,

and $\tilde{Q} = \tilde{f}(U)$, we see, as in problem IV, that $\tilde{Q}$ is not convex, and Filippov's theorem does not apply.

VI. This example shows a problem (containing a parameter A) which has an optimal solution for $A > 0$ sufficiently large by force of the existence theorem of §4, while the same problem has no optimal solution for $A = 0$. Let $m = n = 2$. We consider the Pontryagin problem

$$I = \int_{t_1}^{t_2} [x^2 + (y - t)^2 + (au + bv + c)^2 + A(u^2 + v^2)]\, dt = \text{minimum},$$

$$x' = u(1 - v) + [2 - 2^{-1}(u - 1)^2]v + 4Au,$$

$$y' = [2 - 2^{-1}(u - 1)^2](1 - v) + uv + 4Av,$$

with $t_1 = 0$, initial point $(0, 0)$, fixed target $(0, 1)$, and fixed control space $U = [-1 \leqq u \leqq 1, -1 \leqq v \leqq 1]$, and where a, b, c are the same constants as in §5 reduced by the factor 10^{-1}, that is, $a = 10^{-1}(2 + \sqrt{7}) = 0.464575$, $b = -(3/5)(\sqrt{7} - \sqrt{3}) = -0.548220$, $c = 10^{-1}(-4 - \sqrt{3} + \sqrt{7} + \sqrt{21}) = 0.149628$. Thus, for $A = 0$, this problem reduces essentially to the one studied in §5 and has no optimal solution. Now we have $\tilde{f} = (f_0, f_1, f_2)$ with

$$f_0 = x^2 + (y - t)^2 + (au + bv + c)^2 + A(u^2 + v^2), \qquad A \geqq 0,$$

$$f_1 = u(1 - v) + [2 - 2^{-1}(u - 1)^2]v + 4Au,$$

$$f_2 = [2 - 2^{-1}(u - 1)^2](1 - v) + uv + 4Av.$$

Thus

$$f_{0u} = 2a(au + bv + c) + 2Au, \qquad f_{0v} = 2b(au + bv + c) + 2Av,$$

$$f_{0uu} = 2(a^2 + A), \qquad f_{0uv} = 2ab, \qquad f_{0vv} = 2(b^2 + B).$$

Since $-1 \leqq u \leqq 1$, $-1 \leqq v \leqq 1$, we have $|au + bv + c| \leqq a + |b| + c < 1.2$, $|f_{0u}|$, $|f_{0v}| < (1.2)(1.1) + 2A$, $(f_{0u}^2 + f_{0v}^2)^{1/2} < \sqrt{2}(1.32 + 2A)$, and we can take $L = \sqrt{2}(1.32 + 2A)$. For every u, v, u_0, v_0, we have

$$A(u^2 + v^2) + (au + bv + c)^2 = A(u_0^2 + v_0^2) + (au_0 + bv_0 + c)^2$$
$$+ 2A(u_0u + v_0v) + 2(au_0 + bv_0 + c)[a(u - u_0) + b(v - v_0)]$$
$$+ A[(u - u_0)^2 + (v - v_0)^2] + [a(u - u_0) + b(v - v_0)]^2,$$

where the last term is nonnegative. Hence,

$$A(u^2 + v^2) + (au + bv + c)^2 \geqq z(u_0, v_0, u, v) + A[(u - u_0)^2 + (v - v_0^2],$$

where z is a linear function of u and v. This proves that we can take $C = A$.

Now

$$f_{1u} = 1 - uv + 4A, \qquad f_{1v} = (2 - u) - 2^{-1}(u - 1)^2,$$

$$f_{1uu} = -v, \qquad f_{1uv} = -u, \qquad f_{1vv} = 0,$$

$$f_{2u} = (1 - u) + uv, \qquad f_{2v} = 2^{-1}(1 - u)^2 - (2 - u) + 4A,$$

$$f_{2uu} = -(1 - v), \qquad f_{2uv} = u, \qquad f_{2vv} = 0.$$

Then

$$|f_{1uu}| + 2|f_{1uv}| + |f_{1vv}| + |f_{2uu}| + 2|f_{2uv}| + |f_{2vv}|$$
$$\leqq 1 + 2 + 0 + 2 + 2 + 0 = 7,$$

and we can take $D = 7/2$. Finally,

$$f_{1u} \geqq 4A - 3, \qquad |f_{1v}| \leqq 3, \qquad |f_{2u}| \leqq 3, \qquad |f_{2v}| \geqq 4A - 3.$$

If (u_1, v_1), (u_2, v_2) are two points of U and (X_1, Y_1), (X_2, Y_2) the corresponding values of (f_1, f_2), we have, for $A > 3/2$,

$$|X_1 - X_2| \geqq (4A - 3)|u_1 - u_2| - 3|v_1 - v_2|,$$
$$|Y_1 - Y_2| \geqq (4A - 3)|v_1 - v_2| - 3|u_1 - u_2|,$$
$$[(X_1 - X_2)^2 + (Y_1 - Y_2)^2]^{1/2} \geqq \max(|X_1 - X_2|, |Y_1 - Y_2|)$$
$$\geqq 2^{-1}(|X_1 - X_2| + |Y_1 - Y_2|)$$
$$\geqq 2^{-1}(4A - 6)(|u_1 - u_2| + |v_1 - v_2|).$$

Thus

$$[(X_1 - X_2)^2 + (Y_1 - Y_2)^2]^{1/2} \geqq (2A - 3)[(u_1 - u_2)^2 + (v_1 - v_2)^2]^{1/2},$$

and we can take $\Lambda = (2A - 3)^{-1}$. Thus, for $A \geqq 7$, we have

$$\Lambda LD = (2A - 3)^{-1} \cdot \sqrt{2}(1.32 + 2A) \cdot (7/2)$$
$$= A \cdot (7\sqrt{2}/2)(1.32\,A^{-2} + 2\,A^{-1})(2 - 3\,A^{-1})^{-1}$$
$$\leqq A \cdot (7\sqrt{2}/2)(1.32 \cdot 7^{-2} + 2 \cdot 7^{-1})(2 - 3 \cdot 7^{-1})^{-1}$$
$$< (0.98482)A < A = C.$$

If f_{10}, f_{20} are the expressions of f_1 and f_2 at $A = 0$, and $\bar{f}_1 = u$, $\bar{f}_2 = v$, then both transformations $z_1 = f_{10}$, $z_2 = f_{20}$, and $z_1 = \bar{f}_1$, $z_2 = \bar{f}_2$, map U onto a convex set. Indeed, the first transformation was studied in §5, and the second one is affine. Thus the transformation $z_1 = f_1 = f_{10} + 4A\bar{f}_1$, $z_2 = f_2 = f_{20} + 4A\bar{f}_2$ also transforms U into a convex set. The point $(0, 1)$ is certainly accessible from $(0, 0)$. Indeed, for $v = 0$, we have $x' = (4A + 1)u$,

$y' = 2 - 2^{-1}(u-1)^2$; and hence for $v(t) = 0$, and $u(t) = 1/2$ if $0 \leqq t \leqq t_2/2$, $u(t) = -1/2$ if $t_2/2 < t \leqq t_2$, $t_2 = 8/11$, we have $x' = \pm(4A+1)/2$, and $y' = 15/8$, or $= 7/8$ respectively, and finally $x(t_2) = 0$, $y(t_2) = (15/8)(8/22) + (7/8)(8/22) = 1$. By force of the existence theorem the Pontryagin problem above has certainly an optimal solution for $A \geqq 7$, $T \geqq 8/11$.

REFERENCES

[1a] L. Cesari, *Semicontinuità e convessità nel calcolo delle variazioni*, Ann. Scuola Norm. Sup. Pisa, 18 (1964), pp. 389–423.

[1b] ———, *Un teorema di esistenza in problemi di controlli ottimi*, Ibid., 19 (1965), pp. 35–78.

[2] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. I. Mat. Meh., 2 (1959), pp. 25–32; English transl., this Journal, 1 (1962), pp. 76–84.

[3] R. V. Gamkrelidze, *The theory of time-optimal processes in linear systems*, Izv. Akad. Nauk SSSR, 22 (1958), pp. 449–474; English transl., Report 61–7, Department of Engineering, University of California, Los Angeles, 1961.

[4] L. M. Graves, *The existence of an extremum in problems of Mayer*, Trans. Amer. Math. Soc., 39 (1936), pp. 456–471.

[5a] J. P. LaSalle, *Time optimal control systems*, Proc. Nat. Acad. Sci. U. S. A., 45 (1959), pp. 573–577.

[5b] ———, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, Princeton, 1960, pp. 1–24.

[6] E. Magenes, *Sui teoremi di Tonelli per la semicontinuità nei problemi di Mayer e di Lagrange*, Ann. Scuola Norm. Sup. Pisa, 15 (1948), pp. 113–125.

[7] B. Manià, *Sui problemi di Lagrange e di Mayer*, Rend. Circ. Mat. Palermo, 58 (1934), pp. 285–310.

[8] L. Markus and E. B. Lee, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[9] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[10] L. S. Pontryagin, *Optimal control processes*, Uspehi Mat. Nauk, 14 (85), (1959), pp. 3–20; English transl., Automation Express, 2 (1959), pp. 26–30.

[11] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[12a] L. Tonelli, *Su la semicontinuità nei problemi di Mayer e di Lagrange*, Rend. Accad. Lincei, 24 (1936), pp. 339–404; Opere Scelte, Cremonese, Roma, 3 (1962), pp. 342–348.

[12b] ———, *Fondamenti di calcolo delle variazioni*, 2 vols., Zanichelli, Bologna, 1921–1923.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

ON VARIATIONAL THEORY AND OPTIMAL CONTROL THEORY*

MAGNUS R. HESTENES†

1. Introduction. During recent years the general problem of Bolza has become increasingly useful in optimal control theory. Unfortunately the classical formulation of the general problem of Bolza (see [1]) is not a convenient one for applications of this type. Consequently a new formulation has arisen which appears to be preferable to the classical one. The first formulation of this newer type, known to the author, is the one given by him [2] in 1949. The author was convinced at that time that the new formulation was superior to the old and has urged individuals to pursue this approach. It was not until the coming of the work of Pontryagin (see [3]) that the new formulation came to the fore. The maximum principle of Pontryagin embodies the first order necessary conditions for the problem of Bolza. It can be obtained from the theory of the problem of Bolza by translation. This was carried out in [2]. However, it is desirable to obtain this principle directly. This was done by Pontryagin, using a modification of a method devised by McShane [4]. The work of McShane is a significant extension of the works of Graves [5], Bliss [1], and Bolza. McShane was the first to establish first order necessary conditions without assumptions of normality. He did this by the use of a theorem of separation of cones, one of which degenerated to a halfline. Pontryagin used the same cones and made the important observation that these cones can be generated by limiting oneself to strong variations. McShane used both strong and weak variations. Since the cones used were closed, and since weak variations can be obtained as limits of strong variations, it is clear that the use of weak variations is unnecessary.

In the present paper we extend the methods of McShane and Pontryagin so as to enlarge the class of convenient applications. The results are not essentially new, since they have been embedded in the theory of calculus of variations and can be obtained from the results of McShane and Valentine [7]. However, it appears that the details of proof have been simplified.

The present paper is in part taken from the lecture notes of the author given at UCLA during the spring semester, 1963. We shall be concerned only with first order necessary conditions. The results here given have been established by Guinn [8] under weaker hypotheses. Sufficient conditions

* Received by the editors September 14, 1964, and in final revised form December 4, 1964. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mathematics, University of California, Los Angeles, California. The preparation of this paper was sponsored by the Office of Naval Research and the United States Army Research Office (Durham).

for local minima have been given by Pennisi [9] and in part by Mookini [10], using the method developed by McShane [11] and by the author [12].

2. Formulation of the problem. In the present paper an arc will be considered to be a system

$$x^i(t),\ u^k(t),\ b^\sigma,\quad t^1 \leqq t \leqq t^2;\ i = 1, \cdots, n;\ k = 1, \cdots, q;\ \sigma = 1, \cdots, r,$$

of n continuous functions $x^i(t)$, called *state functions*, q piecewise continuous functions $u^k(t)$, called *control functions*, and r constants b^σ, called *control parameters*. It will be convenient to designate an arc by the single symbol x. Thus, in vector notation,

$$x: \qquad x(t),\ u(t),\ b, \qquad t^1 \leqq t \leqq t^2.$$

The symbol x is therefore used in at least two senses, namely, to denote an arc as above and to designate a point $x = (x^1, \cdots, x^n)$. However, the context will make clear in which sense it is to be interpreted.

The problem to be considered is that of minimizing a function

$$I_0(x) = g_0(b) + \int_{t^1}^{t^2} L_0(t, x(t), u(t), b)\, dt,$$

in a class of arcs

$$x: \qquad x(t),\ u(t),\ b, \qquad t^1 \leqq t \leqq t^2,$$

satisfying a system of differential equations

$$\dot{x}^i(t) = f^i(t, x(t), u(t), b), \tag{2.1}$$

$$\begin{aligned} \varphi_\alpha(t, x(t), u(t), b) &\leqq 0, & 1 \leqq \alpha \leqq m', \\ \varphi_\alpha(t, x(t), u(t), b) &= 0, & m' < \alpha \leqq m, \end{aligned} \tag{2.2}$$

a set of initial and terminal conditions

$$t^s = T^s(b), \qquad x^i(t^s) = X^{is}(b), \qquad s = 1, 2, \tag{2.3}$$

and a set of isoperimetric relations

$$\begin{aligned} I_\gamma(x) &\leqq 0, & 1 \leqq \gamma \leqq p', \\ I_\gamma(x) &= 0, & p' < \gamma \leqq p, \end{aligned} \tag{2.4}$$

where

$$I_\gamma(x) = g_\gamma(b) + \int_{t^1}^{t^2} L_\gamma(t, x(t), u(t), b)\, dt. \tag{2.5}$$

This problem will be referred to as the *optimal control formulation of the problem of Bolza*. An equivalent problem formulated in a more classical

manner is that of minimizing a function

$$I_0(x) = g_0(b) + \int_{t^1}^{t^2} L_0(t, x(t), \dot{x}(t), b)\, dt$$

in a class of arcs

$$x: \qquad x(t), b, \qquad t^1 \leqq t \leqq t^2,$$

satisfying conditions of the form

$$\text{(2.6a)} \qquad \begin{aligned} \varphi_\alpha(t, x, \dot{x}, b) &\leqq 0, & 1 &\leqq \alpha \leqq m', \\ \varphi_\alpha(t, x, \dot{x}, b) &= 0, & m' &< \alpha \leqq m, \end{aligned}$$

$$\text{(2.6b)} \qquad t^s = T^s(b), \qquad x^i(t^s) = X^{is}(b), \qquad s = 1, 2,$$

$$\text{(2.6c)} \qquad \begin{aligned} I_\gamma(x) &\leqq 0, & 1 &\leqq \gamma \leqq p', \\ I_\gamma(x) &= 0, & p' &< \gamma \leqq p. \end{aligned}$$

Here

$$I_\gamma(x) = g_\gamma(b) + \int_{t^1}^{t^2} L_\gamma(t, x(t), \dot{x}(t), b)\, dt.$$

This problem will be called the *isoperimetric problem of Bolza with inequality constraints.*

The second problem is the special case of the first in which the differential equations (2.1) take the form $\dot{x}^i = u^i$. The first can be reduced to the second by introducing new state functions

$$x^{n+k}(t) = \int_{t^1}^{t} u^k(s)\, ds,$$

and adding the conditions

$$x^{n+k}(t^1) = 0, \qquad x^{n+k}(t^2) = b^{r+k},$$

where $b^{r+1}, \cdots, b^{r+q}$ are new control parameters. The details of this substitution will be left to the reader.

Inequality constraints can be replaced by equality constraints by a method used by Valentine [7]. However, this does not simplify our procedures and so we shall not pursue the matter. One can also eliminate isoperimetric conditions if one desires, but again this does not simplify our procedures. There are other modifications that can be made. For example, one can assume that b^σ does not appear in f^i, φ_α, L_γ, since these can be eliminated by introducing new state functions $x^{n+\sigma}(t)$ subject to the conditions

$$\dot{x}^{n+\sigma} = 0, \qquad x^{n+\sigma}(t^s) = b^\sigma, \qquad s = 1, 2.$$

Turning now to the control problem, it will be assumed that all functions used are of class C' on a region $\mathcal{R}$ in (t, x, u, b)-space. The class of all elements (t, x, u, b) in $\mathcal{R}$ satisfying the conditions

$$\begin{aligned} &\varphi_\alpha(t, x, u, b) \leqq 0, && 1 \leqq \alpha \leqq m', \\ &\varphi_\alpha(t, x, u, b) = 0, && m' < \alpha \leqq m, \end{aligned} \tag{2.7}$$

will be denoted by $\mathcal{R}_0$ and will be called the *class of admissible elements.* An arc

$$x: \qquad x(t), u(t), b, \qquad t^1 \leqq t \leqq t^2,$$

will be called *admissible* if its elements $(t, x(t), u(t), b)$ are in $\mathcal{R}_0$. The class of admissible arcs will be denoted by $\mathfrak{A}$. The class of admissible arcs satisfying the conditions (2.1), (2.3) and (2.4) will be denoted by $\mathfrak{B}$. We assume that we have given an admissible arc

$$x_0: \qquad x_0(t), u_0(t), b_0, \qquad t^1 \leqq t \leqq t^2,$$

in $\mathfrak{B}$ that minimizes $I_0(x)$ on $\mathfrak{B}$. In addition, we assume that the matrix

$$\left(\frac{\partial \varphi_\alpha}{\partial u^k} \quad \delta_{\alpha\beta}\varphi_\beta\right), \qquad \begin{aligned} &\alpha, \beta = 1, \cdots, m; \quad \beta \text{ not summed;} \\ &k = 1, \cdots, q; \quad \delta_{\alpha\alpha} = 1, \delta_{\alpha\beta} = 0 (\alpha \neq \beta); \end{aligned} \tag{2.8}$$

has rank m at each element $(t, x_0(t), u, b_0)$ in $\mathcal{R}_0$. Here α denotes the row index and k, β are column indices. The determinant (2.8) has rank m at an element $(\bar{t}, \bar{x}, \bar{u}, \bar{b})$ if and only if the matrix

$$\left(\frac{\partial \varphi_\alpha}{\partial u^k}\right), \qquad \alpha = \alpha_1, \cdots, \alpha_r,$$

has rank r, where $\alpha_1, \cdots, \alpha_r$ are the indices at which

$$\varphi_\alpha(\bar{t}, \bar{x}, \bar{u}, \bar{b}) = 0.$$

The problem at hand is to determine properties of x_0 which are consequences of the fact that x_0 minimizes I_0 on $\mathfrak{B}$.

3. First order necessary conditions and a maximum principle. The main theorem to be established in the present paper is the following. (Here and elsewhere a repeated index in a term denotes summation with respect to that index unless otherwise specified or implied.)

THEOREM 3.1. *Suppose that the arc*

$$x_0: \qquad x_0(t), u_0(t), b_0, \qquad t^1 \leqq t \leqq t^2$$

described above affords a minimum to I_0 on $\mathfrak{B}$. Then there exist multipliers

$$\lambda_0 \geqq 0, \lambda_\gamma, p_i(t), \mu_\alpha(t), \quad \gamma = 1, \cdots, p; i = 1, \cdots, n; \alpha = 1, \cdots, m;$$

not vanishing simultaneously on $t^1 \leqq t \leqq t^2$, *and functions*

$$H(t, x, u, b, p, \mu) = p_i f^i - \lambda_0 L_0 - \lambda_\gamma L_\gamma - \mu_\alpha \varphi_\alpha ,$$

$$G(b) = \lambda_0 g_0 + \lambda_\gamma g_\gamma ,$$

such that the following relations hold.

(i) *The multipliers* λ_γ *are constants and* $\lambda_\gamma \geqq 0$, $1 \leqq \gamma \leqq p'$, *with* $\lambda_\gamma = 0$ *in case* $I_\gamma(x_0) < 0$.

(ii) *The multipliers* $\mu_\alpha(t)$ *are piecewise continuous and are continuous at each point of continuity of* $u_0(t)$. *Moreover, for each* $\alpha \leqq m'$, *the relation* $\mu_\alpha(t) \geqq 0$ *holds and the equation*

$$\mu_\alpha(t)\varphi_\alpha(t, x_0(t), u_0(t), b_0) = 0, \qquad \alpha \text{ not summed}, \tag{3.1}$$

holds on $t^1 \leqq t \leqq t^2$.

(iii) *The multipliers* $p_i(t)$ *are continuous and have piecewise continuous derivatives. In fact there are constants* c_i, c *such that the relations*

$$p_i = -\int_{t^1}^{t} H_{x^i}\, ds + c_i , \qquad H = \int_{t^1}^{t} H_t\, ds + c, \qquad H_{u^k} = 0, \tag{3.2}$$

hold along x_0 *with* $p_i = p_i(t)$, $\mu_\alpha = \mu_\alpha(t)$.

(iv) *The transversality condition,*

$$dG + [-H\, dT^s + p_i(T^s)\, dX^{is}]_{s=1}^{s=2} - \int_{t^1}^{t^2} H_{b^\sigma}\, db^\sigma\, dt = 0, \tag{3.3}$$

is an identity in db^σ *on* x_0.

(v) *The inequality*

$$H(t, x_0(t), u, b_0, p(t), 0) \leqq H(t, x_0(t), u_0(t), b_0, p(t), 0) \tag{3.4}$$

holds whenever $(t, x_0(t), u, b_0)$ *is in* $\mathfrak{R}_0$.

The formula

$$H = \int_{t^1}^{t} H_t\, ds + c \tag{3.5}$$

is an abbreviation of the formula

$$H(t,x_0(t),u_0(t),b_0,p(t),\mu(t)) = \int_{t^1}^{t} H_t(s,x_0(s),u_0(s),b_0,p(s),\mu(s))\, ds + c.$$

It states that H is continuous along x_0 and has a piecewise continuous derivative given by H_t. A similar remark holds for the first integral in (3.2). Equations (3.2) are equivalent to the statements that p_i, H are continuous along x_0, and that on each subarc on which $u_0(t)$ is continuous we have

$$\frac{dp_i}{dt} = -H_{x^i} , \qquad \frac{dH}{dt} = H_t , \qquad H_{u^k} = 0.$$

The equations

$$\dot{p}_i = -H_{x^i}, \qquad \dot{x}^i = H_{p^i} = f^i, \qquad H_{u^k} = 0,$$

are the *Euler equations* for our problem and the inequality (3.4) is the *condition of Weierstrass.*

The relation $H_{u^k} = 0$ is a consequence of the condition of Weierstrass. In fact, it can be shown that (3.5) is a consequence of (3.4) and the relation

$$p_i = -\int_{t^1}^{t} H_{x^i}\, ds + c_i .$$

The maximum principle given in Theorem 3.1 is often called *Pontryagin's maximum principle.* It can also be found in [2].

In the inequality (3.4) we have set $\mu_\alpha = 0$. If one wishes to retain $\mu_\alpha = \mu_\alpha(t)$, the condition (3.4) may be stated as follows: At each element $(\bar{t}, \bar{x}, \bar{u}, b_0, \bar{p}, \bar{u})$ with

$$\bar{x} = x_0(\bar{t}), \qquad \bar{u} = u_0(\bar{t}), \qquad \bar{p} = p(\bar{t}), \qquad \bar{\mu} = \mu(\bar{t}),$$

the inequality

$$H(\bar{t}, \bar{x}, u, b_0, \bar{p}, \bar{\mu}) + \bar{\mu}_\alpha \varphi_\alpha(\bar{t}, \bar{x}, u, b_0) \leqq H(\bar{t}, \bar{x}, \bar{u}, b_0, \bar{p}, \bar{\mu})$$

holds whenever $(\bar{t}, \bar{x}, u, b_0)$ is in $\mathfrak{R}_0$. In obtaining this inequality we made use of (3.1).

The corresponding result for the isoperimetric problem of Bolza with inequality constraints is given in the following theorem. We assume, of course, that the matrix (2.8) with $u^k = \dot{x}^k$ has rank m along x_0. The class of admissible arcs satisfying (2.6) will be denoted by $\mathfrak{B}$ in this case also.

THEOREM 3.2. *Suppose that the arc*

$$x_0 : \qquad x_0(t), b_0, \qquad t^1 \leqq t \leqq t^2,$$

affords a minimum to I_0 on $\mathfrak{B}$. Then there exist multipliers

$$\lambda_0 \geqq 0, \lambda_\gamma, \mu_\alpha(t), \quad \gamma = 1, \cdots, p; \alpha = 1, \cdots, m;$$

not vanishing simultaneously on $t^1 \leqq t \leqq t^2$, and functions

$$F(t, x, \dot{x}, b, \mu) = \lambda_0 L_0 + \lambda_\gamma L_\gamma + \mu_\alpha \varphi_\alpha, \qquad G(b) = \lambda_0 g_0 + \lambda_\gamma g_\gamma,$$

such that the following relations hold.

(i) *The multipliers λ_γ are constants and $\lambda_\gamma \geqq 0$, $1 \leqq \gamma \leqq p'$, with $\lambda_\gamma = 0$ in case $I_\gamma(x_0) < 0$.*

(ii) *The multipliers $\mu_\alpha(t)$ are piecewise continuous and are continuous at each point of continuity of $\dot{x}_0(t)$. Moreover, $\mu_\alpha(t) \geqq 0$, $\alpha \leqq m'$, and the equation*

$$\mu_\alpha(t)\varphi_\alpha(t, x_0(t), \dot{x}_0(t), b_0) = 0, \qquad \alpha \text{ not summed},$$

holds on $t^1 \leqq t \leqq t^2$.

(iii) *There exist constants* c, c_1 *such that the relations,*

$$F - \dot{x}^i F_{\dot{x}^i} = \int_{t^1}^{t} F_t \, ds + c, \qquad F_{\dot{x}^i} = \int_{t^1}^{t} F_{x^i} \, ds + c_i ,$$

hold along x_0 *with* $\mu_\alpha = \mu_\alpha(t)$.

(iv) *The transversality condition,*

$$dG + [(F - \dot{x}^i F_{\dot{x}^i}) \, dT^s + F_{\dot{x}^i} \, dX^{is}]_{s=1}^{s=2} + \int_{t^1}^{t^2} F_{b^\sigma} \, db^\sigma \, dt = 0,$$

is an identity in db_σ *along* x_0 .

(v) *At each element* $(t, x, \dot{x}, b, \mu)$ *on* x_0 *one has*

$$E(t, x, \dot{x}, u, b, \mu) \geqq \mu_\alpha \varphi_\alpha(t, u, x, b),$$

whenever (t, x, u, b) *is in* $\mathfrak{R}_0$, *where* E *is the Weierstrass E-function*

$$E(t, x, \dot{x}, u, b, \mu) = F(t, x, u, b, \mu) - F(t, x, \dot{x}, b, \mu) - (u^i - \dot{x}^i) F_{\dot{x}^i}(t, x, \dot{x}, b, \mu).$$

This result is an easy consequence of Theorem 3.1. Since u plays the role of $\dot{x}$, we have

$$H = p_i u^i - F(t, x, u, b, \mu),$$

and hence

$$H_{u^i} = p_i - F_{\dot{x}^i} = 0, \qquad H_{x^i} = -F_{x^i} , \qquad H_t = -F_t , \qquad H_{b^\sigma} = -F_{b^\sigma}$$

along x_0 . Using these facts, one obtains Theorem 3.2 from Theorem 3.1. Theorem 3.1 can be obtained from Theorem 3.2 by the use of the transformation described in §2.

The proof of Theorem 3.1 will be simplified if we assume that the range $t^1 \leqq t \leqq t^2$ is fixed. To show that no generality is lost thereby, we replace the variable t by a state variable $x^0(t)$ subject to the condition

$$\dot{x}^0(t) = u^0(t) > 0, \qquad t^1 \leqq t \leqq t^2,$$
$$x^0(t^s) = T^s(b) = X^{0s}(b), \qquad s = 1, 2.$$

where

$$t^1 = T^1(b_0), \qquad t^2 = T^2(b_0).$$

Moreover, we set

$$x_0^0(t) = t, \qquad u_0^0(t) = 1,$$

and introduce the functions

$$\bar{L}_\rho = L_\rho(x^0, x, u, b)u^0, \qquad \bar{\varphi}_\alpha = \varphi_\alpha(x^0, x, u, b)u^0,$$
$$\bar{f}^0 = u^0, \qquad \bar{f}^i = f^i(x^0, x, u, b)u^0.$$

The problem then becomes that of minimizing

$$I_0(x) = g_0(b) + \int_{t^1}^{t^2} \bar{L}_0(x(t), u(t), b)\, dt,$$

in a class of arcs

$$x: \qquad x^j(t),\ u^k(t),\ b^\sigma, \quad t^1 \leqq t \leqq t^2; j = 0, 1, \cdots, n; k = 0, 1, \cdots, q;$$

satisfying the conditions

$$\begin{aligned} x^j &= \bar{f}^j(x, u, b), \\ \bar{\varphi}_\alpha &\leqq 0, & 1 \leqq \alpha \leqq m', \\ \bar{\varphi}_\alpha &= 0, & m' < \alpha \leqq m, \\ x^j(t^s) &= X^{js}(b), & s = 1, 2, \\ I_\gamma(x) &\leqq 0, & 1 \leqq \gamma \leqq p', \\ I_\gamma(x) &= 0, & p' < \gamma \leqq p, \end{aligned}$$

where

$$I_\gamma(x) = g_\gamma(b) + \int_{t^1}^{t^2} \bar{L}_\gamma(x(t), u(t), b)\, dt.$$

Since we have only altered the variable of integration, the arc

$$x_0: \qquad x_0{}^0 = t,\ x_0{}^i(t), \quad u_0{}^0(t) = 1,\ u_0{}^i(t), \quad b_0, \quad t^1 \leqq t \leqq t^2,$$

is a minimizing arc for the new problem. Suppose now that Theorem 3.1 has been established for this new problem except for the relation (3.5). We shall show that the theorem holds as stated for the original problem. Let

$$\bar{H} = p_j \bar{f}_j - \bar{L} = (p_0 + H)u^0,$$

where

$$\bar{L} = \lambda_\rho \bar{L}_\rho + \mu_\alpha \bar{\varphi}_\alpha .$$

Since

$$\bar{H}_{u^0} = p_0 + H = 0, \tag{3.6}$$

along x_0 we have

$$p_0(t) = -H(t, x_0(t), u_0(t), b_0).$$

Moreover, along x_0,

$$u^0 = 1, \qquad \bar{H}_{x^0} = H_t, \qquad \bar{H}_{x^i} = H_{x^i}, \qquad \bar{H}_{b^\sigma} = H_{b^\sigma}, \qquad \bar{H}_{u^k} = H_{u^k}.$$

It follows that (3.2) and (3.3) hold along x_0, as stated. Since $\bar{H} \equiv 0$ along x_0, the condition (3.4) for the transformed problem takes the form

$$(p_0(t) + H(t, x_0(t), u, b, 0))u^0 \leqq 0.$$

In view of the relation $u^0 > 0$, it follows from (3.6) and (3.1) that (3.4) holds for the original problem. We can accordingly assume that the interval $t^1 \leqq t \leqq t^2$ is fixed for all arcs in $\mathfrak{B}$, as was to be proved. We have shown also that (3.5) is a consequence of the remaining conditions for a minimizing arc, and a separate proof of this result need not be given.

4. An auxiliary lemma. In the course of our proof we shall need a property of the class $\mathfrak{R}_0$ of admissible elements (t, x, u, b). This property is described in the following lemma. In this lemma

$$x_0: \qquad x_0(t), u_0(t), b_0, \qquad t^1 \leqq t \leqq t^2,$$

is an admissible arc in $\mathfrak{B}$ such that the matrix (2.8) has rank m.

LEMMA 4.1. *There exists a function $U_0(t, x, b)$ defined over a neighborhood $\mathfrak{F}$ of those on x_0 such that (t, x, U_0, b) is in $\mathfrak{R}_0$ and*

$$U_0(t, x_0(t), b_0) = u_0(t), \qquad t^1 \leqq t \leqq t^2. \tag{4.1}$$

The function U_0 and its partial derivatives with respect to x^i and b^σ are continuous except at the values of t at which $u_0(t)$ is discontinuous. At a point $\bar{t}$ of discontinuity of $u_0(t)$, these functions have continuous left- and right-hand limits. Moreover, if we set

$$r_i^k(t) = \frac{\partial U_0^k}{\partial x^i}, \qquad s_\sigma^k(t) = \frac{\partial U_0^k}{\partial b^\sigma}, \tag{4.2}$$

along x_0, the relations

$$\varphi_{\alpha x^i} + \varphi_{\alpha u^k} r_i^k = 0, \qquad \varphi_{\alpha b^\sigma} + \varphi_{\alpha u^k} s_\sigma^k = 0, \tag{4.3}$$

hold on x_0 for each α and each value of t such that

$$\varphi_\alpha(t, x_0(t), u_0(t), b_0) = 0. \tag{4.4}$$

In the course of the proof of Lemma 4.1 we shall also prove the following result.

LEMMA 4.2. *Let $(\bar{t}, \bar{x}, \bar{u}, b_0)$ be an element in $\mathfrak{R}_0$ with $\bar{x} = x_0(\bar{t})$. There is a function $U(t, x, b)$ defined over a neighborhood $\mathfrak{N}$ of $(\bar{t}, \bar{x}, b_0)$ such that (t, x, U, b) is in $\mathfrak{R}_0$ and*

$$U(\bar{t}, \bar{x}, b_0) = \bar{u}.$$

The function U and its partial derivatives with respect to x^i and b^σ are continuous on this neighborhood.

Return now to the proof of Lemma 4.1. Consider first the case in which

$m' = 0$. Then $\mathfrak{R}_0$ is the set of points (t, x, u, b) in $\mathfrak{R}$ having

$$\varphi_\alpha(t, x, u, b) = 0, \qquad \alpha = 1, \cdots, m.$$

Moreover, the matrix (2.8) has rank m at a point (t, x, u, b) in $\mathfrak{R}_0$ if and only if the matrix

$$\left(\frac{\partial \varphi_\alpha}{\partial u^k}\right) \tag{4.5}$$

has rank m at this point.

Suppose next that $u_0(t)$ is continuous on $t^1 \leqq t \leqq t^2$. Then the functions

$$w_\alpha{}^k(t) = \frac{\partial \varphi_\alpha(t, x_0(t), u_0(t), b_0)}{\partial u^k}, \qquad \alpha = 1, \cdots, n; k = 1, \cdots, q;$$

are continuous on $t^1 \leqq t \leqq t^2$. Extend the functions $u_0{}^k(t)$, $w_\alpha{}^k(t)$ to be continuous on an extended interval $t^1 - \delta \leqq t \leqq t^2 + \delta$ for $\delta > 0$. Since the matrix (4.5) has rank m along x_0, it follows that the m-dimensional determinant

$$| w_\alpha{}^k(t) w_\beta{}^k(t) |, \qquad \alpha, \beta = 1, \cdots, m,$$

has rank m on $t^1 \leqq t \leqq t^2$. This determinant is the functional determinant with respect to z_β of the equations

$$\varphi_\alpha(t, x, u_0(t) + w_\beta(t) z_\beta, b) = 0$$

along the initial solution $(t, x, b, z) = (t, x_0(t), b_0, 0)$ on $t^1 \leqq t \leqq t^2$. By the implicit function theorem these equations have continuous solutions,

$$z_\beta = Z_\beta(t, x, b),$$

on a neighborhood $\mathfrak{F}$ of the elements (t, x, b) on x_0 such that

$$Z_\beta(t, x_0(t), b_0) = 0, \qquad t^1 \leqq t \leqq t^2,$$

and such that $Z_\beta(t, x, b)$ has continuous partial derivatives with respect to x^j and b^σ on $\mathfrak{F}$. The functions

$$U_0{}^k(t, x, b) = u_0{}^k(t) + w_\beta{}^k(t) Z_\beta(t, x, b)$$

have the properties described in the theorem, as one readily verifies.

The case $m' > 0$ can be reduced to the case $m' = 0$. This follows because a point (t, x, u, b) is in $\mathfrak{R}_0$ if and only if there exist numbers $v^1, \cdots, v^{m'}$ such that

$$\begin{aligned} \bar{\varphi}_\alpha &= \varphi_\alpha + (v^\alpha)^2 = 0, && \alpha \leqq m', \\ \bar{\varphi}_\alpha &= \varphi_\alpha = 0, && m' < \alpha \leqq m. \end{aligned} \tag{4.6}$$

The matrix (2.8) has rank m at (t, x, u, b) in $\mathfrak{R}_0$ if and only if the matrix

$$\left(\frac{\partial \bar{\varphi}_\alpha}{\partial u^k} \quad \frac{\partial \bar{\varphi}_\alpha}{\partial v^\beta}\right), \qquad \alpha = 1, \cdots, m; k = 1, \cdots, q'; \beta = 1, \cdots, m';$$

has rank m at the corresponding point (t, x, u, v, b) satisfying (4.6) with

$$v^\beta = [-\varphi_\beta(t, x, u, b)]^{1/2}, \qquad \beta = 1, \cdots, m'.$$

Considering $v^1, \cdots, v^{m'}$ to be additional u's, it follows from the result given in the last paragraph that equations (4.6) have continuous solutions

$$U_0^k(t, x, b), \; V_0^\beta(t, x, b), \quad k = 1, \cdots, q; \; \beta = 1, \cdots, m';$$

defined on a neighborhood $\mathfrak{F}$ of the values (t, x, b) on x_0 having continuous partial derivatives with respect to x^j and b^σ. Moreover (4.1) holds. We have, by (4.6),

$$\varphi_{\alpha x^j} + \varphi_{\alpha u^k} \frac{\partial U_0^k}{\partial x^j} + 2V_0^\alpha \frac{\partial V_0^\alpha}{\partial x^j} = 0,$$

and a similar formula for derivatives with respect to b^σ. Since $V_0^\alpha = 0$ when (4.4) holds, one obtains (4.3).

It remains to consider the case in which $u_0(t)$ has discontinuities at points $t_1, \cdots, t_{N-1}$. These points divide x_0 into subarcs $x_{01}, \cdots, x_{0N}$, on each of which $u_0(t)$ is continuous. Let $U_{0j}(t, x, b)$ be the functions with domains $\mathfrak{F}_j$ related to the arc x_{0j} as described above. Set

$$U_0(t, x, b) = U_{0j}(t, x, b), \quad t_{j-1} \leqq t \leqq t_j \; ; \; (t, x, b) \text{ in } \mathfrak{F}_j \; ;$$

where $t_0 = t^1 - \delta$, $t_N = t^2 + \delta$ and δ is a small positive constant. The function $U_0(t, x, b)$ so defined has the properties in the lemma.

Lemma 4.2 can be considered to be the special case of Lemma 4.1 in which x_0 degenerates to a point.

5. A reformulation of the problem. As remarked at the end of §3, we can assume that the functions $T^1(b)$, $T^2(b)$ are independent of b. This assumption simplifies some of the formulas given below and hence will be made throughout the remainder of this paper unless otherwise expressly stated. The interval $t^1 \leqq t \leqq t^2$ is then the same for all arcs in $\mathfrak{B}$.

Let $\mathfrak{C}$ be the class of all admissible arcs

$$x: \qquad x(t), u(t), b, \qquad t^1 \leqq t \leqq t^2,$$

on the fixed range $t^1 \leqq t \leqq t^2$ that satisfy the differential system

$$\dot{x}^i = f^i(t, x, u, b), \qquad x^i(t^1) = X^{i1}(b). \tag{5.1}$$

The problem at hand is then that of minimizing $I_0(x)$ on $\mathfrak{C}$ subject to the conditions

$$
\begin{aligned}
I_\gamma(x) &\leqq 0, && 1 \leqq \gamma \leqq p', \\
I_\gamma(x) &= 0, && p' < \gamma \leqq p, \\
x^i(t^2) &= X^{i2}(b).
\end{aligned} \tag{5.2}
$$

It will be convenient to introduce $p + n + 1$ new functions

$$
J_\rho(x) = G_\rho(b) + \int_{t^1}^{t^2} F_\rho(t, x, u, b)\, dt, \qquad \rho = 0, 1, \cdots, p + n,
$$

such that x_0 is a solution to our original problem if and only if x_0 minimizes $J_0(x)$ subject to the conditions

$$
\begin{aligned}
J_\rho(x) &\leqq 0, && 1 \leqq \rho \leqq p', \\
J_\rho(x) &= 0, && p' < \gamma \leqq p + n.
\end{aligned} \tag{5.3}
$$

These functions are to be chosen so that for a given function $U_0(t, x, b)$ related to x_0 as described in Lemma 4.1 we have

$$
\frac{\partial}{\partial x^j}\,[F_\rho(t, x, U_0(t, x, b), b] = 0 \tag{5.4}
$$

along x_0.

In order to construct the functions J_ρ, let

$$
r_j^{\,k}(t) = \frac{\partial U_0^{\,k}(t, x_0(t), b_0)}{\partial x^j}, \tag{5.5}
$$

and set

$$
\begin{aligned}
A_j^{\,i}(t) &= \frac{\partial}{\partial x^j}\,[f^i(t, x, U_0, b)] = f^i_{xj} + f^i_{uk} r_j^{\,k}, \\
B_{\gamma j}(t) &= \frac{\partial}{\partial x^j}\,[L_\gamma(t, x, U_0, b)] = L_{\gamma x j} + L_{\gamma u k} r_j^{\,k},
\end{aligned} \tag{5.6}
$$

where the right members are to be evaluated along x_0. For $\gamma = 0, 1, \cdots, p$, set

$$
\begin{aligned}
F_\gamma &= L_\gamma - B_{\gamma j} x^j + q_{\gamma i}(t)(f^i - A_j^{\,i} x^j), \\
G_\gamma &= g_\gamma - q_{\gamma i}(t^2) X^{i2} + q_{\gamma i}(t^1) X^{i1},
\end{aligned} \tag{5.7}
$$

where $q_{\gamma i}$ are solutions of the system

$$
\dot{q}_{\gamma j} + q_{\gamma i} A_j^{\,i} + B_{\gamma j} = 0, \qquad q_{\gamma j}(t^2) = 0. \tag{5.8}
$$

For an arc x in $\mathfrak{C}$ we have

$$
\frac{d}{dt}\,(q_{\gamma j} x^j) = -B_{\gamma j} x^j + q_{\gamma i}(f^i - A_j^{\,i} x^j).
$$

Using this fact together with the relation $q_{\gamma j}(t^2) = 0$, it is seen that on $\mathfrak{C}$,

$$J_\gamma(x) = I_\gamma(x) - [q_{\gamma i}(t^s)X^{is}]_{s=1}^{s=2} + \int_{t^1}^{t^2} \frac{d}{dt}(q_{\gamma i}x^i)\,dt = I_\gamma(x).$$

For $i = 1, \cdots, n$, we set

$$\begin{aligned} F_{p+i} &= P_{ij}(t)(f^j - A_h{}^j x^h), \\ G_{p+i} &= -[P_{ij}(t^s)X^{js}]_{s=1}^{s=2}, \end{aligned} \tag{5.9}$$

where $P_{ij}(t)$ are solutions of the system

$$\dot{P}_{ij} + P_{ih}A_j{}^h = 0, \qquad P_{ij}(t^2) = \delta_{ij}. \tag{5.10}$$

For an arc x in $\mathfrak{C}$ we have, by (5.7) and (5.1),

$$J_{p+1}(x) = G_{p+1} + \int_{t^1}^{t^2} \frac{d}{dt}(P_{ij}x^j)\,dt = x^i(t^2) - X^{i2}(b).$$

The conditions (5.3) are accordingly equivalent to the conditions (5.2) on $\mathfrak{C}$. It is a simple matter to verify that (5.4) holds.

An analogue of Theorem 3.1 for the reformulated problem is given in the following.

THEOREM 5.1. *Suppose that x_0 minimizes $J_0(x)$ on $\mathfrak{C}$ subject to (5.3). Then there exist constant multipliers $\lambda_0 \geqq 0, \lambda_1, \cdots, \lambda_{p+n}$, not all zero, and functions*

$$F(t, x, u, b) = \lambda_\rho F_\rho, \qquad \bar{G}(b) = \lambda_\rho G_\rho, \quad \rho = 0, 1, \cdots, p + n,$$

such that

(i) $\lambda_\gamma \geqq 0$, $1 \leqq \gamma \leqq p'$, *with* $\lambda_\gamma = 0$ *if* $J_\gamma(x_0) < 0$;

(ii) *the inequality*

$$F(t, x_0(t), u, b_0) \geqq F(t, x_0(t), u_0(t), b_0) \tag{5.11}$$

holds on $t^1 \leqq t \leqq t^2$ whenever $(t, x_0(t), u, b_0)$ is in $\mathfrak{R}_0$;

(iii) *the transversality condition*

$$d\bar{G} + \int_{t^1}^{t^2} (F_{u^k}s_\sigma{}^k + F_{b^\sigma})\,db^\sigma\,dt = 0 \tag{5.12}$$

is an identity in db^σ along x_0, where

$$s_\sigma{}^k = \frac{\partial U_0{}^k(t, x_0(t), b_0)}{\partial b^\sigma}. \tag{5.13}$$

The proof of Theorem 5.1 will be given in §7. As a further result we have

THEOREM 5.2. *Let F and $\bar{G}$ be related to x_0 as described in Theorem 5.1. There exist piecewise continuous multipliers $\mu_\alpha(t)$ that are continuous at each*

point of continuity of $u_0(t)$ *and are such that the equations*

$$F_{u^k} + \mu_\alpha(t)\varphi_{\alpha u^k} = 0 \tag{5.14}$$

hold along x_0 . *Moreover,* $\mu_\alpha(t) \geqq 0$, $\alpha \leqq m'$, *and*

$$\mu_\alpha(t)\varphi_\alpha = 0, \qquad \alpha \text{ not summed}, \tag{5.15}$$

along x_0 . *If we set*

$$\bar{F} = F + \mu_\alpha(t)\varphi_\alpha ,$$

the transversality condition 5.12 *takes the form,*

$$d\bar{G} + \int_{t^1}^{t^2} \bar{F}_{b^\sigma}\, db^\sigma\, dt = 0, \tag{5.16}$$

on x_0 *for all* db^σ.

In order to prove this result, let

$$f(t, u) = F(t, x_0(t), u, b_0), \qquad \bar{\varphi}_\alpha(t, u) = \varphi_\alpha(t, x_0(t), u, b_0).$$

By virtue of (5.11) the value $u = u_0(t)$ affords a minimum to $f(t, u)$ subject to the constraints

$$\bar{\varphi}_\alpha(t, u) \leqq 0, \qquad \alpha \leqq m',$$

$$\bar{\varphi}_\alpha(t, u) = 0, \qquad m' < \alpha \leqq m.$$

By virtue of the results to be given in the next section or by virtue of the multiplier rule for minimizing a function of a finite number of variables subject to constraints, it follows that for each t on $t^1 \leqq t \leqq t^2$ there exist multipliers $\mu_0 \geqq 0$, $\mu_\alpha(t)$, not all zero, such that

$$\mu_0 f_{u^k} + \mu_\alpha(t)\bar{\varphi}_{\alpha u^k} = 0 \tag{5.17}$$

at $u = u_0(t)$, and such that (5.15) holds, together with the relations $\mu_\alpha(t) \geqq 0$, $\alpha \leqq m'$. Since the matrix (2.8) has rank m on x_0 , the relations (5.17) and (5.15) cannot hold with $\mu_0 = 0$ unless $\mu_\alpha(t) = 0$, $\alpha = 1, \cdots, m$. This implies that $\mu_0 > 0$ and can be chosen to be unity. The multipliers $\mu_\alpha(t)$ are then unique. Using the fact that the matrix (2.8) has rank m on x_0 , it follows from (5.17) and (5.15) that $\mu_\alpha(t)$ is continuous at each point of continuity of $u_0(t)$. Setting $\bar{F} = F + \mu_\alpha(t)\varphi_\alpha$, it is seen that

$$\bar{F}_{u^k} = 0 \quad \text{along} \quad x_0 .$$

Using the relations (4.3) and (5.15), we conclude that

$$\int_{t^1}^{t^2} \mu_\alpha(t)\{\varphi_{\alpha u^k} s_\sigma{}^k + \varphi_{\alpha b^\sigma}\}\, dt = 0$$

along x_0 . Adding this result to (5.12), we obtain (5.16). This establishes Theorem 5.2.

By the use of Theorems 5.1 and 5.2 we can establish Theorem 3.1 as follows. Let λ_ρ be the multipliers given in Theorem 5.1 and set

$$p_i(t) = -\lambda_{p+j}P_{ij}(t) - \lambda_\beta q_{\beta i}(t), \quad \beta = 0, 1, \cdots, p.$$

In view of (5.8) and (5.10) the functions $p_i(t)$ satisfy the equations

$$\dot{p}_j + p_i A_j^{\ i} = \lambda_\beta B_{\beta j}.$$

Consequently, if we set

$$H(t, x, u, b, p, \mu) = p_j f^j - \lambda_\beta L_\beta - \mu_\alpha \varphi_\alpha,$$

we have, by (5.7) and (5.10),

$$H(t, x, u, b, p(t), \mu(t)) = -\dot{p}_j(t)x^j - F - \mu_\alpha\varphi_\alpha = -\dot{p}_j x^j - \bar{F}, \tag{5.18}$$

where F and $\bar{F}$ are the functions appearing in Theorems 5.1 and 5.2. By virtue of (5.11) we see that the inequality

$$H(t, x_0(t), u, b_0, p(t), 0) \leqq H(t, x_0(t), u_0(t), b_0, p(t), 0)$$

holds whenever $(t, x_0(t), u, b_0)$ is in $\mathfrak{R}_0$. Observe further that along x_0 with $\mu_\alpha = \mu_\alpha(t)$,

$$H_{u^k} = -\bar{F}_{u^k} = 0. \tag{5.19}$$

Using (4.3) and the fact that $\mu_\alpha(t) = 0$ if

$$\varphi_\alpha(t, x_0(t), u_0(t), b_0) < 0,$$

it follows that along x_0 we have

$$\mu_\alpha(t)\{\varphi_{\alpha x^i} + \varphi_{\alpha u^k} r_j^{\ k}\} = 0,$$

where $r_j^{\ k}$ is given by (5.5). Combining this result with (5.4) and (5.19), and setting $u = U_0(t, x, b)$ in (5.18), it is found by differentiation that, along x_0, one has

$$\dot{p}_i = -H_{x^i} - H_{u^k} r_i^{\ k} = -H_{x^i}.$$

Finally, by (5.7) and (5.9),

$$\bar{G} = G + [p_i(t^s)X^{is}]_{s=1}^{s=2},$$

where $G = \lambda_\beta g_\beta$. Inasmuch as

$$H_{b^\sigma} = -\bar{F}_{b^\sigma}$$

along x_0, (5.16) yields the transversality condition,

$$dG + [p_i(t^s)\, dX^{is}]_{s=1}^{s=2} - \int_{t^1}^{t^2} H_{b^\sigma}\, db^\sigma\, dt = 0,$$

along x_0 for all db^σ. Theorem 3.1 therefore holds as stated when t^1 and t^2 are fixed, as was to be proved.

6. A fundamental theorem. In the present section we digress for a moment and consider a set of real valued functions $J_\rho(x)$, $\rho = 0, 1, \cdots, p$,* defined on a space $\mathfrak{C}$ of elements x. The nature of the class $\mathfrak{C}$ is immaterial. However, we need the concept of one-sided derivatives of J_ρ at the initial point of a curve in $\mathfrak{C}$. Let x_0 be in $\mathfrak{C}$ and let

$$x(t), \qquad 0 \leqq t \leqq \delta,$$

be a curve in $\mathfrak{C}$ such that $x(0) = x_0$. The vector $k = (k^0, \cdots, k^p)$ defined by the formula

$$k^\rho = \frac{d}{dt} J_\rho(x(t)) \Big|_{t=0},$$

if it exists, can be considered to be a derivative of J_ρ, $\rho = 0, 1, \cdots, p$, at x_0. However, this definition of a derivative is not restrictive enough for our purposes. What we need is a concept that insures us that for any finite set of derivatives $k_1, \cdots, k_N$, every linear combination

$$k = k_1\alpha_1 + \cdots + k_N\alpha_N$$

with nonnegative coefficients is also a derivative. This can be accomplished by the definition of a derived set of vectors given in the next paragraph. It should be noted that it is not essential to generate the class of all derivatives of J_ρ at x_0, but to generate a class of derivatives large enough for our purposes.

Let x_0 be a point of $\mathfrak{C}$. A set K of vectors $k = (k^0, \cdots, k^p)$ in a Euclidean space $\mathfrak{E}^{p+1}$ will be called a *derived set of vectors for* J_ρ *at* x_0 *on* $\mathfrak{C}$ if, given any finite set of vectors $k_1, \cdots, k_N$ in K, there is a function

$$x(\epsilon) = x(\epsilon_1, \cdots, \epsilon_N),$$

defined on a set

$$0 \leqq \epsilon_j \leqq \delta, \qquad j = 1, \cdots, N; \delta > 0; \tag{6.1}$$

such that $x(0) = x_0$ and such that the functions

$$f_\rho(\epsilon) = J_\rho(x(\epsilon)) - J_\rho(x_0), \qquad \rho = 0, 1, \cdots, p,$$

are continuous on the set (6.1) and have

$$df_\rho = k_j{}^\rho \, d\epsilon_j$$

as their differentials at $\epsilon = 0$ on the set (6.1). It is readily verified that if K is a derived set for J_ρ at x_0, so also is the convex cone K^* generated by K.

* Here p plays the role of $p + n$ in the last section.

The cone K^* is of course the class of all vectors k of the form

$$k = k_1\alpha_1 + \cdots + k_N\alpha_N , \tag{6.2}$$

where $\alpha_j \geqq 0$, k_j is in K and N is arbitrary. Observe that if k is a vector of the form (6.2) and $x(\epsilon_1 , \cdots , \epsilon_N)$ is the function related to $k_1 , \cdots , k_N$ and x_0 as described above, then, upon setting $\epsilon_j(t) = \alpha_j t$, the curve

$$\bar{x}(t) = x(\epsilon(t)), \qquad 0 \leqq t \leqq t',$$

is in $\mathcal{C}$ if $t' < \delta/(\alpha_1 + \cdots + \alpha_N)$. Moreover,

$$k^\rho = \frac{d}{dt} J_\rho(\bar{x}(t)) \Big|_{t=0} . \tag{6.3}$$

The vectors k in K^* are therefore derivatives of J_ρ in the sense described at the beginning of this section. Since K^* is a convex cone we shall refer to it as the *derived cone* generated by K. It should be noted in passing that a vector k in the closure $K' = \bar{K}^*$ of K^* need not be a derivative in the sense described above.

We are now in position to state a fundamental theorem which is a generalized Lagrange multiplier rule. It is the basis of a large class of multiplier rules. In particular, it is the basis of the multiplier rule stated in the preceding pages.

THEOREM 6.1. *Suppose that K is a derived set for J_ρ at x_0 on $\mathcal{C}$. If x_0 minimizes $J_0(x)$ on $\mathcal{C}$ subject to the constraints*

$$\begin{aligned} J_\gamma(x) &\leqq 0, & 1 \leqq \gamma \leqq p', \\ J_\gamma(x) &= 0, & p' < \gamma \leqq p, \end{aligned} \tag{6.4}$$

then there exist multipliers $\lambda_0 \geqq 0$, $\lambda_1 , \cdots , \lambda_p$ not all zero, such that the inequality

$$L(k) = \lambda_\rho k^\rho \geqq 0 \tag{6.5}$$

holds for every vector k in K and hence for every vector in the closure K' of the convex cone generated by K. Moreover, $\lambda_\gamma \geqq 0$, $1 \leqq \gamma \leqq p'$, with $\lambda_\gamma = 0$ in case $J_\gamma(x_0) < 0$.

The fact that $\lambda_\gamma = 0$ when $J_\gamma(x_0) < 0$ signifies that J_γ plays no role in the multiplier rule. Intuitively this follows because when $J_\gamma(x_0) < 0$ the condition $J_\gamma(x) < 0$ is not a constraint locally.

The inequality (6.5) on K^* can be given the following interpretation in terms of the function $J = \lambda_\rho J_\rho$. Recall that to each vector k in K^* there is a curve

$$C: \qquad \bar{x}(t), \qquad 0 \leqq t \leqq t',$$

in $\mathcal{C}$ containing x_0 for $t = 0$ and such that k^ρ is the derivative of J_ρ at x_0

along C, as indicated by (6.3). From this fact we find that

$$\frac{d}{dt} J(\bar{x}(t)) \Big|_{t=0} = \lambda_\rho \frac{d}{dt} J_\rho(\bar{x}(t)) \Big|_{t=0} = L(k). \tag{6.6}$$

The inequality (6.5) therefore states that the derivatives of $J = \lambda_\rho J_\rho$ at x_0 along admissible curves C are nonnegative.

Turning now to the proof of Theorem 6.1, observe that it is sufficient to consider the case in which K coincides with its convex closure K^*. We shall accordingly assume that this is the case. Then K is a convex cone in $\mathcal{E}^{p+1}$. Let K^- be a second cone whose nonzero elements consisting of all vectors k are of the form

$$\begin{aligned} &k^0 < 0, \\ &k^\gamma < 0, \quad 1 \leqq \gamma \leqq p' \quad \text{and} \quad J_\gamma(x_0) = 0, \\ &k^\gamma = 0, \qquad\qquad\qquad p' < \gamma \leqq p. \end{aligned} \tag{6.7}$$

Let $K^+ = K - K^-$ be the set of all vectors of the form $k - k^-$, where k is in K and k^- is in K^-. The classes K^+ and K^- are convex cones in $\mathcal{E}^{p+1}$.

Theorem 6.1 will be established with the help of four lemmas, the first of which is the following.

LEMMA 6.1. *Suppose there exists a vector $k \neq 0$ in $\mathcal{E}^{p+1}$ which is not in K^+. Then the conclusions in Theorem* 6.1 *hold.*

Observe first that by virtue of convexity the interior of the cone K^+ is also the interior of its closure $\bar{K}^+$. Since K^+ is not all of $\mathcal{E}^{p+1}$, it follows that $\bar{K}^+$ cannot coincide with $\mathcal{E}^{p+1}$. If $\bar{k} = (-1, 0, \cdots, 0)$ is not in $\bar{K}^+$, choose $k_0 = \bar{k}$. Otherwise, let k_0 be any unit vector not in $\bar{K}^+$. Let k_1 be the vector in $\bar{K}^+$ nearest to k_0. If k is in $\bar{K}^+$, so also is $k_1 + tk$, $t \geqq 0$. The function

$$g(t) = \tfrac{1}{2} \mid k_1 + tk - k_0 \mid^2, \qquad t \geqq 0,$$

therefore has a minimum at $t = 0$. It follows that

$$g'(0) = (k_1 - k_0, k) \geqq 0, \tag{6.8}$$

where (h, k) is the usual inner product in $\mathcal{E}^{p+1}$. If we select $k = k_1$, then $g(t) \geqq g(0)$, $t \geqq -1$. In this event the value of $g'(0)$ is

$$(k_1 - k_0, k_1) = 0. \tag{6.9}$$

We shall show that the multipliers

$$\lambda_\rho = k_1^\rho - k_0^\rho$$

have the properties described in the theorem. By (6.8) we have

$$L(k) = \lambda_\rho k^\rho \geqq 0$$

on $\bar{K}^+$ and hence on K. The vector $\bar{k}_\gamma$, $\gamma = 0, 1, \cdots, p'$, defined by the relations

$$\bar{k}_\gamma{}^\gamma = 1, \qquad \bar{k}_\gamma{}^\rho = 0, \qquad \gamma \neq \rho,$$

is in $\bar{K}^+$ since $-\bar{k}_\gamma$ is in the closure $\bar{K}^-$ of K^-. Hence

$$L(\bar{k}_\gamma) = \lambda_\gamma \geqq 0, \qquad \gamma = 0, 1, \cdots, p'.$$

If $J_\gamma(x_0) < 0$, $1 \leqq \gamma \leqq p'$, then $-\bar{k}_\gamma$ is also in $\bar{K}^+$. Consequently $-\lambda_\gamma \geqq 0$. Hence $\lambda_\gamma = 0$ in this case. Similarly, if $\bar{k} = -\bar{k}_0$ is in $\bar{K}^+$, then $\lambda_0 = 0$. If $\bar{k}$ is not in K^+ then $k_0 = \bar{k}$, by virtue of our choice of k_0. In this event it follows from (6.9) that

$$0 < |k_1 - k_0|^2 = 2(k_1 - k_0, k_1) + |k_0|^2 - |k_1|^2 = 1 - |k_1|^2.$$

Hence $|k_1| < 1$ and $|k_1{}^0| < 1$. Consequently,

$$\lambda_0 = k_1{}^0 - k_0{}^0 = -k_1{}^0 + 1 > 0,$$

when $k_0 = \bar{k}$. This completes the proof of Lemma 6.1.

LEMMA 6.2. *If the cone $K^+ = K - K^-$ coincides with $\mathcal{E}^{p+1}$, there exist $N = p - p' + 1$ vectors $k_1, \cdots, k_N$ in K, whose sum*

$$\bar{k} = k_1 + \cdots + k_N$$

is in K^-, and such that the $(N-1) \times N$-dimensional matrix

$$(k_j{}^\gamma), \quad \gamma = p' + 1, \cdots, p; j = 1, \cdots, N; \tag{6.10}$$

has rank $N - 1$.

Since K^+ coincides with $\mathcal{E}^{p+1}$ we can select N vectors $k_1, \cdots, k_N$ such that the matrix (6.10) has rank $N - 1$ and such that

$$k_1{}^\gamma + \cdots + k_N{}^\gamma = 0, \quad \gamma = p' + 1, \cdots, p. \tag{6.11}$$

Inasmuch as k_j is in K^+, there are a vector k_j' in K and a vector $\bar{k}_j$ in K^- such that $k_j = k_j' - \bar{k}_j$. Since $\bar{k}_j{}^\gamma = 0$, $\gamma = p' + 1, \cdots, p$, it follows that $k_j' = k_j + \bar{k}_j$ has the properties assigned to k_j. The vectors $k_1, \cdots, k_N$ therefore can be chosen to be in K.

It remains to show that these vectors can be modified so that their sum $\bar{k}$ is in K^-. To this end observe first that there is a vector $k' \neq 0$ in K^- that is also in K. For, let $k \neq 0$ be a vector in K^-. Since k is also in K^+ we can choose k' in K and k'' in K^- such that $k = k' - k''$. The vector $k' = k + k''$ is a nonnull vector in K^- which is also in K. For a small positive number ϵ the vector $k^* = k' + \epsilon\bar{k}$ will be in K^- as well as in K. The vectors

$$k_j{}^* = \frac{1}{N} k' + \epsilon k_j$$

are in K and have their sum k^* in K^-. Since k' is in K^-, the matrix (6.10) with k_j replaced by $k_j{}^*$ is unaltered and hence has rank $N - 1$. This proves Lemma 6.2.

In view of Lemma 6.1 the proof of Theorem 6.1 will be complete when we have established the following.

LEMMA 6.3. *The cone $K^+ = K - K^-$ does not coincide with $\mathcal{E}^{p+1}$.*

Suppose that K^+ coincides with $\mathcal{E}^{p+1}$. Select $N = p - p' + 1$ vectors $k_1, \cdots, k_N$ in K having the properties described in Lemma 6.2. By virtue of the definition of a derived set there exists an N-parameter family

$$x(\epsilon_1, \cdots, \epsilon_N), \qquad 0 \leqq \epsilon_j \leqq \delta,$$

of points in $\mathcal{C}$ such that $x(0) = x_0$ and such that

$$f_\rho(\epsilon) = J_\rho(x(\epsilon)) - J_\rho(x_0), \qquad \rho = 0, 1, \cdots, p,$$

is continuous and has

$$df_\rho = k_j{}^\rho \, d\epsilon^j$$

as its differential at $\epsilon = 0$. Choose $\bar{h}^j = 1, j = 1, \cdots, N$. Then

$$k_j{}^\rho \bar{h}^j = \bar{k}^\rho,$$

where $\bar{k} = k_1 + \cdots + k_N$. Recall that $\bar{k}$ is in K^-. If δ' is sufficiently small, the quantities

$$\epsilon_j = t\bar{h}^j + ty^j, \qquad 0 \leqq t \leqq \delta', \ | y | \leqq 1,$$

will satisfy the relations $0 \leqq \epsilon_j \leqq \delta$. Set

$$F_\rho(y, t) = \frac{1}{t} f_\rho(t\bar{h} + ty), \qquad 0 < t \leqq \delta',$$

$$F_\rho(y, 0) = \bar{k}^\rho + k_j{}^\rho y^j, \qquad \rho = 0, 1, \cdots, p.$$

Since $f_\rho(\epsilon)$ has $k_j{}^\rho \, d\epsilon_j$ as its differential at $\epsilon = 0$, we have

$$\lim_{t=0^+} F_\rho(y, t) = F_\rho(y, 0), \tag{6.12}$$

uniformly with respect to y on the set $| y | \leqq 1$. Since $\bar{k}^\gamma = 0, \gamma > p'$, and the matrix (6.10) has rank $N - 1$, the functions $F_{p'+1}, \cdots, F_p$ satisfy the hypotheses of Lemma 6.4 given below. According to this lemma, the equations

$$F_\gamma(y, t) = 0, \qquad \gamma = p' + 1, \cdots, p,$$

have solutions

$$y(t), \qquad 0 \leqq t \leqq \delta''; \ \delta'' \leqq \delta';$$

such that

$$\lim_{t=0+} y(t) = y(0) = 0.$$

Set

$$\epsilon(t) = t\bar{h} + ty(t), \qquad x(t) = x(\epsilon(t)),$$
$$g_\rho(t) = f_\rho(\epsilon(t)) = J_\rho(x(t)) - J_\rho(x_0), \qquad \rho = 0, 1, \cdots, p.$$

By 6.12 we have

$$\lim_{t=0+} \frac{g_\rho(t)}{t} = \lim_{t=0+} F_\rho(y(t), t) = F_\rho(0, 0) = \bar{k}^\rho. \tag{6.13}$$

If γ is on the range $1 \leqq \gamma \leqq p'$ and $J_\gamma(x_0) = 0$, then $\bar{k}^\gamma < 0$ and, by (6.13),

$$g_\gamma(t) = J_\gamma(x(t)) < 0, \qquad 0 < t \leqq \delta'',$$

provided δ'' is sufficiently small. By continuity this is also true if $J_\gamma(x_0) < 0$, if δ'' is suitably chosen. Since $\bar{k}^0 < 0$, the number δ'' can be diminished still further so that

$$g_0(t) = J_0(x(t)) - J_0(x_0) < 0, \qquad 0 < t \leqq \delta''.$$

Finally, by construction,

$$0 = g_\gamma(t) = J_\gamma(x(t)) - J_\gamma(x_0) = J_\gamma(x(t)), \quad p' < \gamma \leqq p.$$

Hence, if t is on the range $0 < t \leqq \delta''$, the point $x(t)$ satisfies the constraints (6.4) and has $J_0(x(t)) < J_0(x_0)$, contrary to the minimizing property of x_0. The proof of Lemma 6.3 and hence of Theorem 6.1 will be complete when the following lemma has been established.

LEMMA 6.4. *Let $G_\alpha(y, t)$, $\alpha = 1, \cdots, m$, be continuous functions on the domain*

$$|y| \leqq r, \qquad 0 \leqq t \leqq \epsilon. \tag{6.14}$$

Suppose that

$$G_\alpha(y, 0) = a_{\alpha i} y^i, \quad \alpha = 1, \cdots, m; i = 1, \cdots, n;$$

where the matrix $(a_{\alpha i})$ has rank m. Let

$$N(t) = \max |G(y, t) - G(y, 0)| \quad \text{on} \quad |y| \leqq r.$$

There are a constant $M > 0$ and a function

$$y(t), \qquad 0 \leqq t \leqq \delta; \delta \leqq \epsilon;$$

such that $y(0) = 0$ and

$$(6.15) \qquad |y(t)| \leqq MN(t) \leqq r, \qquad G_\alpha(y(t), t) = 0, \qquad 0 \leqq t \leqq \delta.$$

Moreover

$$\lim_{t=0^+} y(t) = y(0) = 0.$$

In order to prove this result select constants $a_{\beta j}$, $\beta = m + 1, \cdots, n$; $j = 1, \cdots, n$, such that the matrix

$$A = (a_{ij}), \qquad i, j = 1, \cdots, n,$$

is nonsingular. Set

$$G_\beta(y, t) = a_{\beta j} y^j, \qquad \beta = m + 1, \cdots, n.$$

The enlarged set of functions $G_1, \cdots, G_n$ satisfies the hypothesis of the lemma. Moreover $N(t)$ is unaltered. Select M such that

$$|A^{-1}y| \leqq M|y|.$$

The function

$$F(y, t) = y - A^{-1}G(y, t) = A^{-1}(G(y, 0) - G(y, t))$$

satisfies the relation

$$|F(y, t)| \leqq MN(t)$$

on the set (6.14). Select δ such that

$$r(t) = MN(t) \leqq r, \qquad 0 \leqq t \leqq \delta.$$

For each t on the range $0 \leqq t \leqq \delta$ the transformation $x = F(y, t)$ maps the ball $|y| \leqq r(t)$ into itself. By the fixed-point theorem there is a point $y(t)$ such that

$$|y(t)| \leqq r(t), \qquad y(t) = F(y(t), t) = y(t) - A^{-1}G(y(t), t).$$

We have accordingly

$$A^{-1}G(y(t), t) = 0,$$

and hence also the relations (6.15). Since $N(t)$ is continuous and $N(0) = 0$, we have

$$\lim_{t=0^+} y(t) = y(0) = 0.$$

This proves Lemma 6.4 and completes the proof of Theorem 6.1.

7. Proof of Theorem 5.1. The proof of Theorem 5.1 will be based upon Theorem 6.1. Using the notations given in §5, let K be the class of all vectors $k = (k^0, \cdots, k^{p+n})$ of the form

$$k^\rho = F_\rho(t, x_0(t), u, b_0) - F_\rho(t, x_0(t), u_0(t), b_0), \qquad \rho = 0, 1, \cdots, p + n, \tag{7.1}$$

such that $(t, x_0(t), u, b_0)$ is in $\mathcal{R}_0$, and t is on the range $t^1 < t < t^2$ and distinct from the points of discontinuity of $u_0(t)$. Adjoin to K all vectors k of the form

$$k^\rho = \bar{k}_\sigma{}^\rho h^\sigma, \qquad h \text{ arbitrary}, \tag{7.2a}$$

where

$$\bar{k}_\sigma{}^\rho = G_{\rho b^\sigma} + \int_{t^1}^{t^2} \{F_{\rho u^k} s_\sigma{}^k + F_{\rho b^\sigma}\}\, dt, \tag{7.2b}$$

evaluated along x_0 with

$$s_\sigma{}^k = \frac{\partial U_0{}^k(t, x_0(t), b_0)}{\partial b^\sigma}.$$

It should be noted that $\bar{k}_\sigma{}^\rho$ is the derivative of

$$G_\rho(b) + \int_{t^1}^{t^2} F_\rho(t, x_0(t), U_0(t, x_0(t), b), b)\, dt,$$

with respect to b^σ at $b = b_0$.

We have the following.

LEMMA 7.1. *The class K is a derived set for J_ρ at x_0 on* $\mathcal{C}$.

Assume for a moment that Lemma 7.1 has been established. Then by Theorem 6.1 there exist multipliers $\lambda_0 \geqq 0, \lambda_1, \cdots, \lambda_{p+n}$, not all zero, such that

$$L(k) = \lambda_\rho k^\rho \geqq 0 \tag{7.3}$$

on the closure of K. Moreover $\lambda_\gamma \geqq 0$, $1 \leqq \gamma \leqq p'$, with $\lambda_\gamma = 0$ in case $J_\gamma(x_0) < 0$. Setting $F = \lambda_\rho F_\rho$, we see, by (7.1) and (7.3), that

$$F(t, x_0(t), u, b_0) \geqq F(t, x_0(t), u_0(t), b_0),$$

whenever $(t, x_0(t), u, b_0)$ is in $\mathcal{R}_0$, except possibly for a finite number of values of t on $t^1 \leqq t \leqq t^2$. By continuity considerations it holds at these values of t also. Using (7.2) and (7.3), we see that, along x_0,

$$dG + \int_{t^1}^{t^2} \{F_{u^k} s_\sigma{}^k + F_{b^\sigma}\}\, db^\sigma \geqq 0$$

holds for all db^σ. Since db^σ is arbitrary, the equality must hold. Theorem 5.1, therefore, will be proved when Lemma 7.1 has been established.

Turning now to the proof of Lemma 7.1, let $k_0, k_1, \cdots, k_N$ be $N + 1$ vectors in K. It is sufficient to consider the case when one of them, say k_0,

is of the form (7.2) and the remaining are of the form

$$k_j^{\rho} = F_{\rho}(t_j, x_0(t_j), u_j, b_0) - F_{\rho}(t_j, x_0(t_j), u_0(t_j), b_0), \qquad j = 1, \cdots, N. \tag{7.4}$$

We can assume that $t_1 \leqq t_2 \leqq \cdots \leqq t_N$. By virtue of Lemma 4.2 there is a continuous function $U_j(t, x, b)$ defined on a δ-neighborhood of $(t_j, x_0(t_j), b_0)$ such that $(t, x, U_j(t, x, b), b)$ is in $\mathcal{R}_0$,

$$U_j(t, x_0(t_j), b_0) = u_j,$$

and such that its partial derivatives with respect to x^j and b^{σ} are continuous on this set. We may choose δ independent of j and such that $t_i + N\delta$ is on $t^1 < t < t^2$ and does not exceed t_j when $t_i < t_j$. Moreover, δ can be chosen so that the function $U_0(t, x, b)$ of Lemma 4.1 used in §5 to define $J_{\rho}(x)$ is defined on the δ-neighborhood of the points (t, x, b) on x_0. Set

$$T_1 = t_1, \qquad T_j = t_j + \epsilon_1 + \cdots + \epsilon_{j-1}, \quad j = 1, \cdots, N,$$

where ϵ_j is restricted to the set

$$0 \leqq \epsilon_j \leqq \delta', \qquad 0 < \delta' < \delta, \tag{7.5}$$

where δ' is a positive constant. Let $M(\epsilon)$ be the complement on $t^1 \leqq t \leqq t^2$ of the set of nonoverlapping intervals

$$T_j \leqq T \leqq T_j + \epsilon_j, \qquad j = 1, \cdots, N.$$

Set

$$b^{\sigma}(\epsilon) = b_0{}^{\sigma} + \epsilon_0 h^{\sigma},$$

where h^{σ} are the numbers in (7.2) defining k_0 and $0 \leqq \epsilon_0 \leqq \delta'$. If δ' is sufficiently small the function $U(t, x, \epsilon)$ defined by the formulas

$$\begin{aligned} U(t, x, \epsilon) &= U_j(t, x, b(\epsilon)), \quad T_j \leqq t \leqq T_j + \epsilon_j ; j = 1, \cdots, N; \\ U(t, x, \epsilon) &= U_0(t, x, b(\epsilon)), \qquad\qquad t \text{ in } M(\epsilon)), \end{aligned} \tag{7.6}$$

is well defined on a neighborhood of the points (t, x) on x_0. The equations

$$\dot{x}^i = f^i(t, x, U(t, x, \epsilon), b(\epsilon)), \qquad x^i(t^1) = X^{i1}(b(\epsilon))$$

have a solution

$$x^i(t, \epsilon), \qquad t^1 \leqq t \leqq t^2,$$

for all ϵ on a set

$$0 \leqq \epsilon_j < \delta'', \qquad j = 0, 1, \cdots, N, \tag{7.7}$$

provided that δ'' is taken sufficiently small. The arc

$x(\epsilon)$: $\qquad x(t, \epsilon), u(t, \epsilon) = U(t, x(t, \epsilon), \epsilon), b(\epsilon), \qquad t^1 \leqq t \leqq t^2,$

is in $\mathcal{C}$ and $x(0) = x_0$. Moreover, the derivatives $x^i_{\epsilon_j}(t, \epsilon)$ are uniformly bounded piecewise continuous functions of t. The functions

$$f_\rho(\epsilon) = J_\rho(x(\epsilon)) - J_\rho(x_0),$$

are continuous and are in fact of class C' on the set (7.7). We shall show that at $\epsilon = 0$ the relation

$$\frac{\partial f_\rho}{\partial \epsilon_j} = k_j{}^\rho, \qquad j = 0, 1, \cdots, N,$$

holds. Setting $\epsilon_j = 0, j = 1, \cdots, N$, we see that

$$f_\rho(\epsilon) = -J_\rho(x_0) + G_\rho(b(\epsilon)) + \int_{t^1}^{t^2} F_\rho(t, x(t, \epsilon), U_0(t, x(t, \epsilon), b(\epsilon)), b(\epsilon))\, dt.$$

Taking the derivative with respect to ϵ_0 at $\epsilon_0 = 0$, and using the fact that

$$\frac{\partial F_\rho(t, x, U_0(t, x, b_0), b_0)}{\partial x^i} = 0 \tag{7.8}$$

along x_0, it is seen that, at $\epsilon = 0$,

$$\frac{\partial f_\rho}{\partial \epsilon_0} = \bar{k}_\sigma{}^\rho h^\sigma = k_0{}^\rho,$$

as desired. Setting $\epsilon_i = 0$, $i \neq j, j > 0$, we see that

$$f_\rho(\epsilon) = P_\rho(\epsilon) + Q_\rho(\epsilon),$$

where

$$P_\rho(\epsilon) = \int_{t_j}^{t_j+\epsilon_j} [F_\rho(t, x(t, \epsilon), U_j(t, x(t, \epsilon), b_0), b_0) - F(t, x_0(t), u_0(t), b_0)]\, dt,$$

$$Q_\rho(\epsilon) = \int_{M(\epsilon)} [F_\rho(t, x(t, \epsilon), U_0(t, x(t, \epsilon), b_0), b_0) - F_\rho(t, x_0(t), u_0(t), b_0)]\, dt.$$

We have, at $\epsilon = 0$,

$$\frac{\partial P_\rho}{\partial \epsilon_j} = \lim_{\epsilon_j = 0} \frac{P_\rho(\epsilon)}{\epsilon_j} = k_j{}^\rho, \qquad \frac{\partial Q_\rho}{\partial \epsilon_j} = 0,$$

the last equation holding by virtue of (7.8) and the boundedness of the derivative of $x(t, \epsilon)$ with respect to ϵ_j. Hence at $\epsilon = 0$,

$$\frac{\partial f_\rho}{\partial \epsilon_j} = k_j{}^\rho, \qquad \rho = 0, 1, \cdots, p$$

Since $f_\rho(\epsilon)$ is of class C', its differential at $\epsilon = 0$ is accordingly $k_j{}^\rho\, d\epsilon_j$.

It follows that K is a derived set for J_ρ at x_0 on $\mathcal{C}$, as was to be proved. This completes the proof of Theorem 5.1 and hence also of Theorem 3.1.

REFERENCES

[1] G. A. BLISS, *Lectures in the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[2] M. R. HESTENES, *A general problem in the calculus of variations with applications to paths of least time*, The RAND Corporation, RM-100, 1949; see also ASTIA Document No. 112382.

[3] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[4] E. J. MCSHANE, *On multipliers for Lagrange problems*, Amer. J. Math., 61 (1939), pp. 809–819.

[5] L. M. GRAVES, *On the Weierstrass condition for the problem of Bolza in the calculus of variations*, Ann. of Math., 33 (1932), pp. 747–752.

[6] L. D. BERKOVITZ, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[7] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, Dissertation, University of Chicago, Chicago, 1937.

[8] T. GUINN, *On first order necessary conditions for variational and optimal control problems*, Dissertation, University of California, Los Angeles, 1964.

[9] L. L. PENNISI, *An indirect sufficiency proof for the problem of Lagrange with differential inequalities as added side conditions*, Dissertation, University of Chicago, Chicago, 1952.

[10] E. H. MOOKINI, *Necessary conditions and sufficient conditions for the problem of Bolza in the calculus of variations*, Dissertation, University of California, Los Angeles, 1964.

[11] E. J. MCSHANE, *Sufficient conditions for a weak relative minimum in the problem of Bolza*, Trans. Amer. Math. Soc., 52 (1942), pp. 344–379.

[12] M. R. HESTENES, *An indirect sufficiency proof for the problem of Bolza in nonparametric form*, Trans. Amer. Math. Soc., 62 (1947), pp. 509–535.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

ON SOME DIFFERENTIAL GAMES*

L. S. PONTRYAGIN†

This paper briefly reports some work on differential games which has just been published in [1]. A particular case of such a game is the problem of the pursuit of one controlled object by another controlled object.

By a controlled object we shall mean one whose state at each instant of time is defined by a vector (call it x) in a certain vector space, and whose motion is described by the equation

$$\dot{x} = F(x, u), \qquad \dot{x} = \frac{dx}{dt}. \tag{1}$$

Here, u is the control parameter, a point in a certain manifold. If the initial value x_0 is given, then, by (1), we can determine the motion of the object, since the control u is known, that is, u is known as a function of the time. If x represents the state of a mechanical object, then some of the coordinates of the vector x represent the position of the object, and the others represent its velocity.

In the pursuit problem there are two objects, x and y. The motion of object y is described by the equation

$$\dot{y} = G(y, v), \tag{2}$$

where v is the control parameter. We shall assume that object x pursues object y, which is moving away from x. The pursuit is considered terminated when the geometric coordinates of the objects x and y coincide. The rules of pursuit are as follows: at each time t, the states x and y of both objects, and the value of the control parameter v of the pursued object, are assumed to be known. Our aim is to find a value of u (the control parameter of the pursuer) at the same time t, such that the pursuit will terminate in the shortest period of time.

It should be particularly emphasized that the future behavior of the pursued object is not assumed to be known. In actuality, for certain states x, y, it will be necessary to use not only the value of the parameter v, but also the values of a number of its derivatives. This means that we are

* Received by the editors October 19, 1964. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5-7, 1964. The original manuscript of this paper in Russian was translated into English by A. Naparstek.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† V. A. Steklov Mathematics Institute, Academy of Sciences of the USSR, Moscow, USSR, and Brown University, Providence, Rhode Island.

using some knowledge about the future behavior of object y. This premise may be given a sensible physical meaning, which I shall not dwell on here.

We assume that the parameters u and v belong to analytic manifolds, that the functions in (1) and (2) are analytic, and that the function $v(t)$ is piecewise analytic. It turns out that the function $u(t)$ obtained under these conditions is piecewise analytic as well.

In constructing a general theory, we had the following example as a model for it. In a Euclidean space E of arbitrary dimension, two points, ξ and η, undergo a motion as described by the equations

$$\ddot{\xi} + \alpha\dot{\xi} = \rho u, \tag{3}$$

$$\ddot{\eta} + \beta\dot{\eta} = \sigma v. \tag{4}$$

Here α, β, ρ, σ are positive numbers, and u and v are control vectors in E, which satisfy the conditions

$$|u| = 1, \qquad |v| = 1.$$

Thus, each of the points moves under the action of friction and a force of constant magnitude whose direction may vary.

In order to bring (3) and (4) to the form (1) and (2), it is necessary to set

$$x = (\xi, \dot{\xi}), \qquad y = (\eta, \dot{\eta}).$$

The pursuit is said to be terminated when $\xi = \eta$.

A general theory should answer the following two questions. Under what conditions can the pursuit be automatically terminated? What is the length of time required for its termination?

It has been found that the pursuit can definitely be terminated if the inequalities

$$\rho = \sigma, \qquad \frac{\rho}{\alpha} > \frac{\sigma}{\beta}, \tag{5}$$

hold. The time needed to end the pursuit, for a given initial position of the objects x, y, can be found to be the smallest positive root of a certain transcendental equation.

To simplify the calculations, we couple x and y into the single object $z = (x, y)$ and somewhat generalize the problem.

Let R^n denote n-dimensional Euclidean space. There is only one controlled object, whose state is defined by the vector z in R^n, and whose motion is described by the equation

$$\dot{z} = Z(z, u, v). \tag{6}$$

Here, u and v are the control parameters, which are points on analytic manifolds, the control $v(t)$ is assumed to be piecewise analytic, and Z is an

analytic function of all its arguments. In addition, we shall assume that

$$Z(z, u, v) = X(z, u) + Y(z, v). \tag{7}$$

In the space R^n, there is given an l-dimensional analytic manifold M^l, where $l < n$. The game is considered to be terminated when the point z reaches the manifold M^l. The rules of the game are as follows. At each time t, the position z of the object and the control v are assumed to be known. Our aim is to find a value for the control u, at the same time t, that will terminate the game in the shortest possible time.

Actually, for certain values of z, it will be necessary to use not only the value of the parameter v itself, but also the values of a number of its derivatives with respect to t, at the same instant of time.

One can attempt to solve this problem by Bellman's method of dynamic programming. In fact, it can be solved by this method if the corresponding Bellman function is single-valued and has continuous derivatives. However, the Bellman function for this problem is, as a rule, multiple-valued; it has branches which are analogous to the branches of an analytic function. For example, this holds true for the problem described by (3) and (4).

We solve this problem with the help of the maximum principle.

Along with the vector z, we introduce an auxiliary vector ψ in R^n and consider the function

$$\mathcal{H}(z, \psi, u, v) = \psi \cdot Z(z, u, v), \tag{8}$$

where on the right we have the scalar product of the vectors ψ and Z. As the function $\mathcal{H}$ is a Hamiltonian function, we consider the Hamiltonian system of equations

$$\frac{dz^i}{d\tau} = \frac{\partial \mathcal{H}}{\partial \psi_i}, \qquad \frac{d\psi_i}{d\tau} = -\frac{\partial \mathcal{H}}{\partial z^i}, \qquad i = 1, 2, \cdots, n. \tag{9}$$

This system is not complete, since, in addition to the vectors z and ψ, there are also the parameters u and v. We shall complete this system in a manner analogous to that used in the maximum principle. By virtue of (7), we have

$$\mathcal{H}(z, \psi, u, v) = \mathcal{H}_1(z, \psi, u) + \mathcal{H}_2(z, \psi, v), \tag{10}$$

in which

$$\mathcal{H}_1(z, \psi, u) = \psi \cdot X(z, u), \qquad \mathcal{H}_2(z, \psi, v) = \psi \cdot Y(z, v). \tag{11}$$

Let $M_1(z, \psi)$ be the maximum of the function $\mathcal{H}_1(z, \psi, u)$ for fixed values of z and ψ, and let $M_2(z, \psi)$ be the minimum of the function $\mathcal{H}_2(z, \psi, v)$ for fixed values of z and ψ. The additional relations for system (9) are the following:

$$\mathcal{H}_1(z, \psi, u) = M_1(z, \psi), \qquad \mathcal{H}_2(z, \psi, v) = M_2(z, \psi). \tag{12}$$

The system (9), (12) obtained in this manner is, generally speaking, complete. We shall assume that this system has the unique solution

$$z = z(\tau), \qquad \psi = \psi(\tau), \qquad u = u(\tau), \qquad v = v(\tau), \tag{13}$$

defined for all values $\tau \leqq 0$, with the initial conditions

$$z(0) = z_0, \qquad \psi(0) = \psi_0. \tag{14}$$

We regard as admissible only those initial conditions for which $z_0 \in M^l$, and for which ψ_0 is a unit vector ($|\psi_0| = 1$) orthogonal to the manifold M^l at the point z_0. The manifold N, consisting of all admissible pairs $(z_0, \psi_0) = \zeta$ of initial values, obviously has dimension $n - 1$. Each solution (13), with admissible initial values, that we have been considering depends, in reality, not only on the time τ, but also on the initial value ζ, so that we have

$$z = z(\tau, \zeta) = \omega(\tau, \zeta), \tag{15}$$

where $\tau < 0$, $\zeta \in N$.

The set of all pairs $(\tau, \zeta) = s$, where $\tau < 0$ and $\zeta \in N$, forms an n-dimensional analytic manifold S. The function ω (see (15)) defines an analytic mapping of the n-dimensional manifold S into the n-dimensional Euclidean space R^n. We denote the Jacobian at a point $s \in S$ of this mapping by $\mathfrak{D}(s)$:

$$\mathfrak{D}(s) = \det \frac{\partial \omega}{\partial s}. \tag{16}$$

The mapping ω is, as a rule, not one-to-one. In particular, it is not one-to-one for the example (3), (4). Thus, the inverse function $\omega^{-1}(z)$ is not single-valued. If, from the relation $\omega^{-1}(z) = s = (\tau, \zeta)$, one defines τ as a function of z,

$$\tau = \tau(z),$$

then we obtain the Bellman function. If it happens that the mapping ω is one-to-one and that the Jacobian $\mathfrak{D}(s)$ is everywhere different from zero, then the Bellman function $\tau(z)$ is single-valued, and with its help the problem can be solved by the method of dynamic programming.

The following theorem has been proved by the author under certain assumptions concerning the game which are not here formulated.

THEOREM. *Let $\hat{z}$ be the initial state of the object and let $\hat{s} = (\hat{\tau}, \hat{\zeta})$ be that pre-image of the point $\hat{z}$ under the mapping ω for which $|\tau|$ has the minimal value. Then, no matter how player v conducts himself, player u, by correct play, will certainly be able to terminate the game in a time not exceeding $|\hat{\tau}|$.*

REFERENCES

[1] L. S. PONTRYAGIN, *On some differential games*, Dokl. Akad. Nauk SSSR, 156 (1964), pp. 738–741.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

SOME GEOMETRICAL ASPECTS OF OPTIMAL PROCESSES*

G. LEITMANN†

Introduction. This paper is an account of preliminary results about some geometrical aspects of optimal processes. Related investigations into the geometry of optimally controlled systems can be found in the work of Halkin [1], Roxin [2], and Blaquiere [3].

Under the assumptions that the cost is additive and that the minimum value of the cost is a continuous function of the initial state, the existence of so-called limiting surfaces in cost-augmented state space is exhibited. Each member of this one-parameter family of surfaces is the locus of optimal trajectories, and bounds the region containing all trajectories which emanate from points on the surface. The existence of limiting surfaces is deduced without restriction to a particular system and cost functional.

For systems described by the usual set of differential state equations and an integral cost functional, further geometrical properties of limiting surfaces are found. While it is not the primary purpose of this investigation to give another derivation of the maximum principle [4], that principle is shown to be a consequence of the geometry of limiting surfaces. Furthermore, for regular optimal trajectories the connection with dynamic programming [5] is exhibited.

1. Notation and assumptions. Consider a system characterized by n variables $x_1, x_2, \cdots, x_n$. The *state* of the system may be thought of as a point

$$\mathfrak{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

in n-dimensional Euclidean space E^n.

At the outset of this discussion, we shall not specify the rules which govern the behavior of the system. Rather we shall assume that these rules are subject to some degree of change, and we shall make certain assumptions concerning the system's behavior.

Associated with any one of the given set of rules is a process which changes the state of the system from some point, $\mathfrak{x}$, along a path, π, in E^n. In other

* Received by the editors February 4, 1965, and in final revised form March 16, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mechanical Engineering, Division of Applied Mechanics, University of California, Berkeley, California. This paper is based on research supported, in part, by the National Science Foundation and reported in [6].

words, the state variables, x_j, $j = 1, 2, \cdots, n$, are functions of time,[1] $x_j(t)$. We shall assume that these paths are continuous, i.e., that the functions $x_j(t)$ are continuous on some time interval $[t_0, t_1]$. In the ensuing discussion we shall be concerned with the paths which can be generated by making all permissible changes in the rules which govern the dynamical behavior of the system.

Let p denote a path which emanates from a point $\mathfrak{x}$ at some time t and reaches a *prescribed* terminal point $\mathfrak{x}^1$ at some time t_1 ; neither the value of t, nor that of t_1 , is specified.[2] Let $\bar{p}$ denote a path which starts at point $\mathfrak{x}$ but does not reach $\mathfrak{x}^1$ in finite time; the terminal point of such a path will be denoted by $\bar{\mathfrak{x}}^1$.

Next consider a rule, or *functional*, which assigns a unique number to each process. We shall assume that this number, the value of the functional or the cost of the process, depends on the path, p or $\bar{p}$, and hence on the initial point, $\mathfrak{x}$, and on the terminal point, $\mathfrak{x}^1$ or $\bar{\mathfrak{x}}^1$. We shall denote the cost by $V(\mathfrak{x}; \mathfrak{x}^1, p)$ or $V(\mathfrak{x}; \bar{\mathfrak{x}}^1, \bar{p})$, respectively.

We shall assume further that there exist *optimal* rules, and we shall say that a rule is optimal if it results in a path, p^*, from $\mathfrak{x}$ to $\mathfrak{x}^1$, for which the cost takes on its *minimum* value. In other words,

$$V(\mathfrak{x}; \mathfrak{x}^1, p^*) \leqq V(\mathfrak{x}; \mathfrak{x}^1, p), \ \forall \text{ rules}. \tag{1}$$

Given a functional, and end states $\mathfrak{x}$ and $\mathfrak{x}^1$, the optimal rule and the corresponding optimal path need not be unique; many different rules may be optimal. However, it follows from definition (1) that the minimum value of the functional is unique; therefore, we shall write

$$V^*(\mathfrak{x}; \mathfrak{x}^1) \triangleq V(\mathfrak{x}; \mathfrak{x}^1, p^*).$$

In general, paths from $\mathfrak{x}$ to $\mathfrak{x}^1$ are either optimal or nonoptimal; the latter will be denoted by p', and so

$$\{\pi : \pi = p\} = \{\pi : \pi = p'\} \cup \{\pi : \pi = p^*\}.$$

Next, let us consider two subsets of state space E^n:
(i) the set E of all points $\mathfrak{x}$ from which the prescribed terminal point $\mathfrak{x}^1$ can be reached in finite time;
(ii) the set E^* of all points $\mathfrak{x}$ from which $\mathfrak{x}^1$ can be reached along an optimal path.
More concisely, we have

[1] In order to avoid cumbersome notation, we shall employ the same symbol for a function of time and for its value at a given time. The intended meaning should be evident from the context in which the symbol is used.

[2] The results can be extended readily to include the case of prescribed end manifolds and of prescribed end times.

$$E = \{\mathfrak{x}: \exists \pi = p\},$$
$$E^* = \{\mathfrak{x}: \exists \pi = p^*\}.$$

First of all, it is clear that $E^* \subset E$. Secondly, neither E nor E^* need be all of state space E^n. Thus, these sets may possess boundary points. We shall now make some assumptions:

(i)[3] At every point $\mathfrak{x}$ of boundary ∂E^*, there exists a sphere with center at $\mathfrak{x}$ and radius $r > 0$ such that every concentric sphere with radius ρ, $0 < \rho < r$, contains interior points (with respect to E^n) of E^*.

(ii)[4] $V^*(\mathfrak{x}; \mathfrak{x}^1)$ is defined and continuous on E^*.

(iii) Partial derivatives $\partial V^*(\mathfrak{x}; \mathfrak{x}^1)/\partial x_j$, $j = 1, 2, \cdots, n$, may have at most jump or infinite discontinuities.

(iv)[5] The value of the functional obeys the additivity property

$$V(\mathfrak{x}; \mathfrak{x}^1, p) = V(\mathfrak{x}; \mathfrak{x}^i, p^i) + V(\mathfrak{x}^i; \mathfrak{x}^1, p_i), \quad \forall \mathfrak{x}^i \in p, \quad \text{where} \quad p = p^i \cup p_i, \tag{2}$$

and

$$\lim_{\mathfrak{x}^i \to \mathfrak{x}^1} V(\mathfrak{x}^i; \mathfrak{x}^1, p_i) = 0.$$

2. Augmented state space and trajectories. Let us now introduce another variable, x_0, and consider an augmented state vector

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

which represents a point in $(n + 1)$-dimensional Euclidean space E^{n+1}.

Next we shall define trajectories, Γ or $\bar{\Gamma}$, in E^{n+1} by

$$\begin{aligned} \Gamma: &\quad x_0{}^i + V(\mathfrak{x}^i; \mathfrak{x}^1, p_i) = C, \quad \text{where} \quad p_i \subset p, \\ \bar{\Gamma}: &\quad x_0{}^i + V(\mathfrak{x}^i; \bar{\mathfrak{x}}^1, \bar{p}_i) = C, \quad \text{where} \quad \bar{p}_i \subset \bar{p}. \end{aligned} \tag{3}$$

Thus, a trajectory Γ, or $\bar{\Gamma}$, is a curve in E^{n+1} traced by a point $\mathbf{x}^i$ whose projection $\mathfrak{x}^i$ on E^n moves on a path p, or $\bar{p}$, respectively.

[3] This assumption precludes the possibility of portions of E^* consisting only of boundary points, e.g., isolated optimal paths.

[4] It can be shown that this assumption is valid for systems of the kind considered subsequently.

[5] Here, $\mathfrak{x}^i$ is regarded as the terminal point of path p^i. Subsequently, we shall restrict the analysis to functionals, namely, integrals, for which the additivity property is evidently valid.

The terminal point $\mathbf{x}^1$ of a trajectory Γ lies on a line X_0, which is parallel to the x_0-axis and intersects E^n in the prescribed terminal point $\mathfrak{x}^1$. We shall write $\Gamma = \Gamma'$ if $p = p'$, and $\Gamma = \Gamma^*$ if $p = p^*$.

3. Limiting surfaces Σ and optimal isocost surfaces S. In view of the assumption that $V^*(\mathfrak{x}; \mathfrak{x}^1)$ is defined and continuous on E^*, we may define an n-dimensional surface

$$\Sigma: \quad x_0 + V^*(\mathfrak{x}; \mathfrak{x}^1) = C. \tag{4}$$

Such a surface is composed of a single sheet, since $V^*(\mathfrak{x}; \mathfrak{x}^1)$ is a single-valued function of $\mathfrak{x}$. Since partial derivatives $\partial V^*(\mathfrak{x}; \mathfrak{x}^1)/\partial x_j$, $j = 1, 2, \cdots, n$, may have jump or infinite discontinuities, a Σ surface is piecewise regular.[6]

The function

$$x_0 = C - V^*(\mathfrak{x}; \mathfrak{x}^1)$$

corresponding to a given value of C vanishes on an $(n-1)$-dimensional surface

$$S: \quad V^*(\mathfrak{x}; \mathfrak{x}^1) = C. \tag{5}$$

As the value of C varies, (5) and (4) define two one-parameter families of surfaces: S surfaces in E^n and Σ surfaces in E^{n+1}, which we shall call *optimal isocost surfaces* and *limiting surfaces*, respectively. The first of these names is a consequence of the definitions of S and $V^*(\mathfrak{x}; \mathfrak{x}^1)$; namely, every point on a given S surface corresponds to the *same* minimum value of the functional. The reason for adopting the second name will become obvious in the next section.

4. Some properties of Σ surfaces. We shall now state some properties of Σ surfaces. First among these is the one embodied in the following lemma.

LEMMA 1. *Any optimal trajectory Γ^* which intersects line X_0 at a point $\mathbf{x}^1$ lies entirely in the Σ surface through $\mathbf{x}^1$.*

This lemma follows at once from the definition (1) of optimality together with additivity property (2).

In view of definition (4) of Σ surfaces, the members of the one-parameter family of these surfaces may be deduced from one another by translation parallel to the x_0-axis. Furthermore, these surfaces are ordered along the x_0-axis in the same way as the values of parameter C. Thus, one and only one Σ surface passes through a given point $\mathbf{x}^1$ in E^{n+1}. From this property, together with Lemma 1, we have:

LEMMA 2. *Any optimal trajectory Γ^* with a point on a given Σ surface lies*

[6] A regular point of a surface is one where the tangent plane of the surface is defined.

entirely on that Σ *surface; i.e., the* Σ *surfaces are the loci of all optimal trajectories.*

Next let us introduce some more nomenclature. A given Σ surface separates the domain in which it is defined in E^{n+1}—whose projection on E^n is E^*—into two open regions with respect to that Σ surface. We shall denote these regions by A/Σ ("above" Σ) and B/Σ ("below" Σ), respectively. For a given value of $\mathfrak{x}$, the terms "above" and "below," respectively, refer to larger and smaller values of x_0 than that of Σ. A point $\mathbf{x} \in A/\Sigma$ will be called an A-point relative to Σ, and a point $\mathbf{x} \in B/\Sigma$ a B-point relative to Σ.

From the definitions (3) and (4), respectively, of a nonoptimal trajectory Γ' and a Σ surface, together with Lemma 2, there follows:

LEMMA 3. *There exists no trajectory which starts on a given* Σ *surface and intersects line* X_0 *at a B-point relative to that* Σ *surface.*

Lemmas 2 and 3 lead quite readily to the following theorem and corollary.

THEOREM 1. *A trajectory (optimal or nonoptimal) whose initial point belongs to a given* Σ *surface cannot have a B-point relative to that* Σ *surface.*

COROLLARY 1.1. *A trajectory whose initial point is an A-point relative to a given* Σ *surface cannot have a B-point relative to that* Σ *surface, nor, indeed, a point in it.*

Theorem 1 embodies the limiting nature of the Σ surfaces; namely, *a given* Σ *surface bounds the domain of all trajectories which emanate from points on that* Σ *surface.* This property of Σ surfaces is fundamental in the subsequent discussion of optimal processes.

Thus far, we have not specified the rules which govern the behavior of the system. To deduce additional results, we shall now consider specific rules.

5. State equations and control. Henceforth, we shall restrict the analysis to systems for which the state variables satisfy state equations

$$\dot{x}_j = f_j(x_1, \cdots, x_n, u_1, \cdots, u_m), \quad j = 1, 2, \cdots, n, \tag{6}$$

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

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}$$

defines a point in m-dimensional Euclidean space E^m. Given the control as a function of time, $\mathbf{u} = \mathbf{u}(t)$, $t_0 \leqq t \leqq t_1$, equations (6) constitute a set of rules which govern the behavior of the system during interval $[t_0, t_1]$.

We shall restrict the values of the control to a prescribed[7] set U, i.e.,

[7] The analysis is easily extended to include the case of state-dependent set U.

$\mathbf{u} \in U \subset E^m$. Furthermore, we shall assume that the function $\mathbf{u}(t)$ is defined[8] and piecewise continuous on some interval $[t_0, t_1]$. A control satisfying these conditions will be termed admissible.

Concerning state equations (6), we shall assume that $f_j(\mathbf{x}, \mathbf{u})$ and $\partial f_j(\mathbf{x}, \mathbf{u})/\partial x_\alpha$, $\alpha, j = 1, 2, \cdots, n$, are defined and continuous on $E^n \times U$. Consequently, for given admissible control $\mathbf{u}(t)$, $t_0 \leqq t \leqq t_1$, and given initial conditions $\mathbf{x}(t_0) = \mathbf{x}^0$, there exists a unique continuous solution $\mathbf{x}(t)$ on $[t_0, t_1]$.

Note here that the interval $[t_0, t_1]$ is not specified. Rather, the initial state $\mathbf{x}^0$ and the terminal state $\mathbf{x}^1$ are prescribed. The subsequent results are readily extendable to include consideration of specified end times as well as of nonautonomous state equations, and of motion between end manifolds.

Next we shall consider a functional in integral form

$$\int_{t_0}^{t_1} f_0(x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t))\, dt, \tag{7}$$

where $\mathbf{u}(t)$ is a control which transfers the system from initial state $\mathbf{x}^0$ either to prescribed terminal state $\mathbf{x}^1$ or to some other terminal state $\bar{\mathbf{x}}^1$ in time $t_1 - t_0$.

Trajectories Γ or $\bar{\Gamma}$ are defined by

$$x_0(t) + \int_t^{t_1} f_0(x_1(s), \cdots, x_n(s), u_1(s), \cdots, u_m(s))\, ds = C,$$

so that

$$\dot{x}_0 = f_0(x_1, \cdots, x_n, u_1, \cdots, u_m), \tag{8}$$

where we take $x_0(t_0) = 0$, since the initial value of x_0 is of no consequence in determining the value of functional (7). Concerning $f_0(\mathbf{x}, \mathbf{u})$ and $\partial f_0(\mathbf{x}, \mathbf{u})/\partial x_j$, $j = 1, 2, \cdots, n$, we shall make the same assumptions of continuity as those made earlier for the functions $f_j(\mathbf{x}, \mathbf{u})$ and their partial derivatives.

Equations (6) and (8) constitute a set of $n + 1$ differential equations whose solutions define trajectories in E^{n+1}. We shall combine them into a single vector equation

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \tag{9}$$

where

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

[8] If $\mathbf{u}(t)$ is discontinuous at $t = t_c$, we shall take $\mathbf{u}(t_c) = \mathbf{u}(t_c - 0)$. Also, without loss of generality, we shall assume $\mathbf{u}(t)$ to be continuous at t_0 and t_1, i.e., $\mathbf{u}(t_0) = \mathbf{u}(t_0 + 0)$ and $\mathbf{u}(t_1) = \mathbf{u}(t_1 - 0)$.

An optimal control $\mathbf{u}^*(t)$, $t_0 \leqq t \leqq t_1$, results in transfer of the system from $\mathfrak{x}^0$ to $\mathfrak{x}^1$, while rendering the minimum value of x_0 at time t_1. The corresponding optimal solution of trajectory equations (9) will be denoted by $\mathbf{x}^*(t)$.

6. A linear transformation and tangent planes. Consider the homogeneous linear differential equations,

$$\dot{\eta}_j = \sum_{\alpha=0}^{n} \frac{\partial f_j(\mathbf{x}, \mathbf{u}^*(t))}{\partial x_\alpha}\bigg|_{\mathbf{x}=\mathbf{x}^*(t)} \eta_\alpha, \qquad j = 0, 1, \cdots, n, \tag{10}$$

whose solution, for given initial conditions $\eta_j(t') = \eta_j'$, $t_0 \leqq t' \leqq t_1$, defines an $(n+1)$-dimensional vector $\mathbf{n}(t)$, $t' \leqq t \leqq t_1$. Equations (10) define a nonsingular linear transformation, and hence a linear operator $A(t', t)$ such that

$$\mathbf{n}(t) = A(t', t)\mathbf{n}', \qquad t' \leqq t \leqq t_1. \tag{11}$$

Next consider an optimal trajectory Γ^*, generated by control $\mathbf{u}^*(t)$, $t_0 \leqq t \leqq t_1$, and let $\mathbf{x}^*(t')$ be a regular, interior point of a limiting surface Σ. A neighborhood in Σ of point $\mathbf{x}^*(t')$ is defined by

$$\Delta(\mathbf{x}^*(t')) \overset{\Delta}{=} \{\mathbf{x}:\mathbf{x} = \mathbf{x}(t', \epsilon)\}$$

and

$$\mathbf{x}(t', \epsilon) = \mathbf{x}^*(t') + \epsilon\mathbf{n}' + \mathbf{o}(\epsilon) \in \Sigma, \tag{12}$$

where $\mathbf{n}'$ belongs to the tangent plane $T(\mathbf{x}^*(t'))$ of $\Delta(\mathbf{x}^*(t'))$—and hence of Σ—at $\mathbf{x}^*(t')$, where ϵ is a parameter of first order smallness, and where $\lim_{\epsilon\to 0} \mathbf{o}(\epsilon)/\epsilon = 0$.

We shall be interested in the transform of $\Delta(\mathbf{x}^*(t'))$ along Γ^*, i.e., in

$$\Delta(\mathbf{x}^*(t'')) \overset{\Delta}{=} \{\mathbf{x}:\mathbf{x} = \mathbf{x}(t'', \epsilon)\}, \qquad t' \leqq t'' \leqq t_1,$$

where $\mathbf{x}(t, \epsilon)$ is the solution of trajectory equations (9) with control $\mathbf{u}^*(t)$, $t_0 \leqq t \leqq t_1$, and initial conditions (12). In other words,

$$\mathbf{x}(t'', \epsilon) = \mathbf{x}^*(t'') + \epsilon\mathbf{n}(t'') + \mathbf{o}(t'', \epsilon), \tag{13}$$

where

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

and $\mathbf{o}(t'', \epsilon)/\epsilon$ tends to zero uniformly with respect to t'', $t' \leqq t'' \leqq t_1$, as $\epsilon \to 0$.

From the properties of the linear transformation $A(t', t)$ and from (13), it follows that

$$T(\mathbf{x}^*(t'')) = A(t', t'')T(\mathbf{x}^*(t')),$$

where $T(\mathbf{x}^*(t''))$ is the tangent plane of $\Delta(\mathbf{x}^*(t''))$ at $\mathbf{x}^*(t'')$.

Recall now that points $\mathbf{x}(t'', \epsilon)$ belong to trajectories whose initial points lie in the limiting surface Σ. Thus, provided $\mathbf{x}^*(t'')$ is an interior point of Σ, it follows from Theorem 1 that there exists an $\alpha > 0$ such that points $\mathbf{x}(t'', \epsilon)$ belong to region $A/\Sigma \cup \Sigma$ for all $|\epsilon| < \alpha$. If, in addition, $\mathbf{x}^*(t'')$ is a regular point of Σ, then we conclude that $\Delta(\mathbf{x}^*(t''))$ is tangent to Σ at $\mathbf{x}^*(t'')$. Thus we have:

LEMMA 4. *If* $\mathbf{x}^*(t')$ *and* $\mathbf{x}^*(t'')$, $t'' \geqq t'$, *of optimal trajectory* Γ^* *on limiting surface* Σ *are regular interior points of* Σ, *then*

$$T(\mathbf{x}^*(t'')) = A(t', t'')T(\mathbf{x}^*(t')),$$

where $T(\mathbf{x}^*(t))$ *denotes the tangent plane of* Σ *at* $\mathbf{x}^*(t)$.

7. Regular interior points of a limiting surface. Suppose that $\mathbf{x}'$ is a regular interior point of limiting surface Σ. Let $\mathbf{n}(\mathbf{x}')$ denote the unit vector normal to Σ at $\mathbf{x}'$ and directed into region B/Σ, so that $n_0(\mathbf{x}') \leqq 0$.

As a consequence of Theorem 1, we have

$$\mathbf{n}(\mathbf{x}')\cdot\mathbf{f}(\mathbf{x}', \mathbf{u}) \leqq 0, \qquad \forall \mathbf{u} \in U. \tag{14}$$

Furthermore, if $\mathbf{x}' = \mathbf{x}^*(t')$ of optimal trajectory Γ^* generated by control $\mathbf{u}^*(t)$, $t_0 \leqq t \leqq t_1$, it follows from Lemma 2 that

$$\mathbf{n}(\mathbf{x}^*(t'))\cdot\mathbf{f}(\mathbf{x}^*(t'), \mathbf{u}^*(t')) = 0. \tag{15}$$

8. The maximum principle and the functional equation for regular optimal trajectories. Suppose that optimal trajectory Γ^* is *regular*, i.e., all points of Γ^* are regular interior points of the Σ surface on which Γ^* lies.

Consider the tangent plane $T(\mathbf{x}^0)$ of Σ at the initial point $\mathbf{x}^*(t_0) = \mathbf{x}^0$ of Γ^*. According to Lemma 4,

$$T(\mathbf{x}^*(t)) = A(t_0, t)T(\mathbf{x}^0),$$

where $T(\mathbf{x}^*(t))$ is the tangent plane of Σ at $\mathbf{x}^*(t)$. In other words, if any vector $\mathbf{n}^0 \in T(\mathbf{x}^0)$ is transformed according to

$$\mathbf{n}(t) = A(t_0, t)\mathbf{n}^0,$$

then

$$\mathbf{n}(t) \in T(\mathbf{x}^*(t)). \tag{16}$$

The equations adjoint to variational equations (10) are

$$\dot{\lambda}_j = -\sum_{\alpha=0}^{n} \frac{\partial f_\alpha(\mathbf{x}, \mathbf{u}^*(t))}{\partial x_j}\bigg|_{\mathbf{x}=\mathbf{x}^*(t)} \lambda_\alpha, \qquad j = 0, 1, \cdots, n. \tag{17}$$

For given initial condition $\boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}^0$, the solution $\boldsymbol{\lambda}(t)$ of (17) is unique and continuous on $[t_0, t_1]$. It follows from (10) and (17) that

$$\boldsymbol{\lambda}(t)\cdot\mathbf{n}(t) = \text{const.}, \qquad t_0 \leqq t \leqq t_1. \tag{18}$$

Let us now choose the following initial conditions:

(i) $$\mathbf{n}(t_0) = \mathbf{n}^0 \in T(\mathbf{x}^0);$$

(ii) $\boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}^0 \perp T(\mathbf{x}^0)$ and directed into B/Σ; i.e., $\boldsymbol{\lambda}^0$ is codirectional with $\mathbf{n}(\mathbf{x}^0)$, so that $\lambda_0(t_0) \leqq 0$.
Consequently, we have

$$\boldsymbol{\lambda}^0 \cdot \mathbf{n}^0 = 0,$$

and hence, by (18), that

$$\boldsymbol{\lambda}(t) \cdot \mathbf{n}(t) = 0, \qquad t_0 \leqq t \leqq t_1 . \tag{19}$$

In view of (16) and (19), we conclude that vectors $\boldsymbol{\lambda}(t)$ and $\mathbf{n}(\mathbf{x}^*(t))$ are codirectional on $[t_0 , t_1]$. It follows from (14) and (15) that

$$\boldsymbol{\lambda}(t) \cdot \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}) \leqq 0, \ \forall \mathbf{u} \in U, \tag{20}$$

and

$$\boldsymbol{\lambda}(t) \cdot \mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)) = 0 \tag{21}$$

for all t on $[t_0 , t_1]$.

Furthermore, since the right-hand side of adjoint equations (17) is independent of x_0, it follows that

$$\lambda_0(t) = \text{const.} \leqq 0 \tag{22}$$

Conditions (20)–(22) embody the maximum principle of Pontryagin for the case of regular optimal trajectories.

Let us recall now the equation of the Σ surfaces, namely,

$$\Phi(\mathbf{x}) \overset{\Delta}{=} x_0 + V^*(\mathfrak{x}; \mathfrak{x}^1) = C,$$

and let us consider grad $\Phi(\mathbf{x})$. This vector is defined at a point $\mathbf{x}^*(t')$ of optimal trajectory Γ^* on limiting surface Σ, provided:
(i) $\mathbf{x}^*(t')$ is a regular interior point of Σ, and
(ii) $\left.\dfrac{\partial V^*(\mathfrak{x}; \mathfrak{x}^1)}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}^*(t)}, j = 1, 2, \cdots, n$, are finite.

If conditions (i) and (ii) are satisfied, then grad $\Phi(\mathbf{x}^*(t'))$ and $\boldsymbol{\lambda}(t')$ are collinear. Furthermore, their projections on the x_0-axis are constant. Hence, we have

$$\boldsymbol{\lambda}(t') = \lambda_0(t') \text{ grad } \Phi(\mathbf{x}^*(t')). \tag{23}$$

Since $\lambda_0(t') \neq 0$ at a point where condition (ii) is met, we may put $\lambda_0(t) \equiv -1$. Thus, if Γ^* is regular and condition (ii) is met at one point of Γ^*, it follows from (20)–(22) with (23) that

$$\min_{\mathbf{u} \in U} \left[f_0(\mathbf{x}^*(t), \mathbf{u}) + \sum_{j=1}^{n} f_j(\mathbf{x}^*(t), \mathbf{u}) \left.\frac{\partial V^*(\mathfrak{x}; \mathfrak{x}^1)}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}^*(t)} \right] = 0$$

for all $t \in [t_0, t_1]$. This equation is one version of Bellman's functional equation.

9. Nonregular interior points of a limiting surface. Thus far we have restricted the discussion to regular interior points of Σ surfaces. Next we shall consider nonregular interior points of limiting surfaces.

In view of our assumptions on $V^*(\mathfrak{x}; \mathfrak{x}^1)$ and its partial derivatives, a Σ surface is *piecewise regular*; i.e., it may consist of intersecting subsurfaces whose interior points are regular points of Σ. Suppose that subsurfaces Σ_β, $\beta = 1, 2, \cdots, \gamma$, intersect in a manifold

$$M^d \triangleq \Sigma_1 \cap \Sigma_2 \cap \cdots \cap \Sigma_\gamma ,$$

where superscript d denotes the dimension of the intersection manifold. Also, let $T_\beta(\mathbf{x})$ denote the tangent plane of Σ_β at point $\mathbf{x}$.

For the ensuing discussion, we shall assume that manifold M^d is *nondegenerate*; i.e., at a point $\mathbf{x} \in M^d$:

(i) all tangent planes $T_\beta(\mathbf{x})$ possess linearly independent normals, if $\gamma \leqq n + 1$; and

(ii) $n + 1$ of the tangent planes $T_\beta(\mathbf{x})$ possess linearly independent normals, if $\gamma > n + 1$.

One consequence of the assumed nondegeneracy is that

$$d = n + 1 - \gamma \quad \text{for} \quad \gamma \leqq n + 1,$$

$$d = 0 \quad \text{for} \quad \gamma \geqq n + 1.$$

In other words, for $\gamma \geqq n + 1$ the intersection manifold M^d reduces to a point in E^{n+1}.

Next we consider the n-dimensional surface $\mathcal{S}(\mathbf{x})$, composed of the portion of $\bigcup_{\beta=1}^{\gamma} T_\beta(\mathbf{x})$ which adheres to Σ at point $\mathbf{x}$. Since Σ is single-sheeted, so is $\mathcal{S}(\mathbf{x})$. Hence, $\mathcal{S}(\mathbf{x})$ divides E^{n+1} into two open regions. We shall characterize these regions by means of rays L_- and L_+ at $\mathbf{x}$. L_- and L_+ are parallel to the x_0-axis; L_- points into the negative x_0 direction, and L_+ into the positive x_0 direction. The region containing L_- will be denoted by $B/\mathcal{S}(\mathbf{x})$, the other by $A/\mathcal{S}(\mathbf{x})$. If L_- lies in $\mathcal{S}(\mathbf{x})$, we utilize L_+ in a similar fashion. If both L_- and L_+ lie in $\mathcal{S}(\mathbf{x})$, we define these regions at point $\mathbf{x} + \Delta\mathbf{x} \in \Sigma$, and let $|\Delta\mathbf{x}| \to 0$ along a continuous curve in Σ; this can always be done in view of the continuity of $V^*(\mathfrak{x}; \mathfrak{x}^1)$, which rules out the possibility of every such curve belonging to an x_0-cylindrical sheet.

Of course, the definition of the surface $\mathcal{S}(\mathbf{x})$ is valid at a regular point $\mathbf{x}$ of Σ. In fact, in that case, the surface $\mathcal{S}(\mathbf{x})$ reduces to the tangent plane $T(\mathbf{x})$ of Σ.

Let us consider a vector $\mathbf{n}'$ at the point $\mathbf{x}^*(t')$ of optimal trajectory Γ^*, and its transform $\mathbf{n}'' = A(t', t'')\mathbf{n}'$ at $\mathbf{x}^*(t'')$, $t'' > t'$. Then, by means of Corollary 1.1, we can deduce the next two lemmas.

LEMMA 5. *If*

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

then

$$\mathbf{n}'' \in [A/\mathcal{S}(\mathbf{x}^*(t''))] \cup \mathcal{S}(\mathbf{x}^*(t'')).$$

LEMMA 6. *If*

$$\mathbf{n}'' \in [B/\mathcal{S}(\mathbf{x}^*(t''))] \cup \mathcal{S}(\mathbf{x}^*(t'')),$$

then

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

A direct consequence of these lemmas is the following.

LEMMA 7. *A point* $\mathbf{x}^*(t')$, $t_0 < t' < t_1$, *of an optimal trajectory* Γ^* *cannot be an isolated nonregular point, i.e., one such that* $\mathbf{x}^*(t' \pm \Delta t)$ *are regular points of* Σ *as* $\Delta t \to 0$.

10. Attractive and repulsive manifolds. If $\mathbf{x}$ is a nonregular interior point of Σ, i.e., $\mathbf{x} \in M^d$, $d < n$, then there exist three possibilities:
(i) $B/\mathcal{S}(\mathbf{x})$ is separable;
(ii) $A/\mathcal{S}(\mathbf{x})$ is separable;
(iii) neither $B/\mathcal{S}(\mathbf{x})$ nor $A/\mathcal{S}(\mathbf{x})$ is separable.
Here we shall restrict the discussion to cases (i) and (ii) only. We shall denote the separable region, an open cone, by $\mathcal{C}_t(\mathbf{x})$ and the corresponding separating hyperplane by $\mathcal{I}(\mathbf{x})$.

Of course, at a *regular* interior point $\mathbf{x}$ of Σ, both $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$ are separable with $\mathcal{S}(\mathbf{x}) \equiv T(\mathbf{x}) \equiv \mathcal{I}(\mathbf{x})$. We shall distinguish between *nonregular* interior points of Σ, where $\mathcal{C}_t(\mathbf{x}) \equiv A/\mathcal{S}(\mathbf{x})$ and $\mathcal{C}_t(\mathbf{x}) \equiv B/\mathcal{S}(\mathbf{x})$, respectively. By means of Lemmas 5 and 6 one can establish:

LEMMA 8. *An optimal trajectory* Γ^* *cannot join nonregular interior points* $\mathbf{x}' = \mathbf{x}^*(t')$ *and* $\mathbf{x}'' = \mathbf{x}^*(t'')$, $t'' > t'$, *if* $\mathcal{C}_t(\mathbf{x}') \equiv B/\mathcal{S}(\mathbf{x}')$ *and* $\mathcal{C}_t(\mathbf{x}'') \equiv A/\mathcal{S}(\mathbf{x}'')$.

If, in addition, there exists a separating hyperplane different from every one of the tangent planes $T_\beta(\mathbf{x})$, $\beta = 1, 2, \cdots, \gamma$, we have:

LEMMA 9. *An optimal trajectory* Γ^* *cannot join points* $\mathbf{x}' = \mathbf{x}^*(t')$ *and* $\mathbf{x}'' = \mathbf{x}^*(t'')$, $t'' > t'$, *if* $\mathbf{x}'$ *is a nonregular interior point where* $\mathcal{C}_t(\mathbf{x}') \equiv B/\mathcal{S}(\mathbf{x}')$ *and* $\mathbf{x}''$ *is a regular interior point; or if* $\mathbf{x}'$ *is a regular interior point and* $\mathbf{x}''$ *is a nonregular interior point where* $\mathcal{C}_t(\mathbf{x}'') \equiv A/\mathcal{S}(\mathbf{x}'')$.

Consequently, we shall employ the terms *attractive manifold* and *repulsive manifold*, if $\mathbf{x} \in M^d$, $d < n$, and $\mathcal{C}_t(\mathbf{x}) \equiv B/\mathcal{S}(\mathbf{x})$ or $\mathcal{C}_t(\mathbf{x}) \equiv A/\mathcal{S}(\mathbf{x})$, respectively.

It can be shown that an optimal trajectory Γ^* cannot join points $\mathbf{x}^*(t')$ and $\mathbf{x}^*(t'')$, $t'' > t'$, belonging to *attractive* manifolds of dimensions

$n + 1 - \gamma'$ and $n + 1 - \gamma''$, respectively, unless $\gamma' \leqq \gamma''$. The converse result holds, if both points belong to *repulsive* manifolds.

11. Convex cone of f vectors. We shall define the unit normal $\mathbf{n}_\beta(\mathbf{x})$ of subsurface Σ_β at the point $\mathbf{x} \in M^d$ by considering the normal $\mathbf{n}(\mathbf{x} + \Delta\mathbf{x})$ at a regular interior point of Σ, and letting $|\Delta\mathbf{x}| \to 0$ along a continuous curve in Σ_β.

From (14), together with the continuity of $\mathbf{f}(\mathbf{x}, \mathbf{u})$ for given $\mathbf{u}$, it follows that

$$\mathbf{n}_\beta(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}) \leqq 0, \qquad \forall \mathbf{u} \in U, \qquad \beta = 1, 2, \cdots, \gamma. \tag{24}$$

Let us recall now that the surface $\mathcal{S}(\mathbf{x})$ is generated by vectors $\mathbf{t}$ tangent to Σ at $\mathbf{x}$. We shall denote the surface generated by vectors $-\mathbf{t}$ by $\mathcal{S}_-(\mathbf{x})$, and the corresponding separable open cone by $\mathcal{C}_{t-}(\mathbf{x})$. Condition (24), together with the properties of intersection manifolds, leads to the following.

LEMMA 10. *If* $\mathbf{x}$ *belongs to an attractive manifold, then*

$$\mathbf{f}(\mathbf{x}, \mathbf{u}) \in \mathcal{C}_{t-}(\mathbf{x}) \cup \mathcal{S}_-(\mathbf{x}), \qquad \forall \mathbf{u} \in U.$$

If $\mathbf{x}$ *belongs to a repulsive manifold, then*

$$\mathbf{f}(\mathbf{x}, \mathbf{u}) \in \mathcal{C}_t(\mathbf{x}) \cup \mathcal{S}(\mathbf{x}), \qquad \forall \mathbf{u} \in U.$$

Let us now define the γ-dimensional convex cone $\mathcal{C}_n(\mathbf{x})$ of vectors $\mathbf{N}(\mathbf{x}) = \sum_{\beta=1}^{\gamma} \alpha_\beta \mathbf{n}_\beta(\mathbf{x})$, where α_β, $\beta = 1, 2, \cdots, \gamma \leqq n$, are nonnegative constants, not all of which are zero. From condition (24), it follows at once that

$$\mathbf{N}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}) \leqq 0, \qquad \forall u \in U, \qquad \forall N(\mathbf{x}) \in \mathcal{C}_n(\mathbf{x}). \tag{25}$$

Furthermore, since the hyperplane $T^{(n+1-\gamma)}(\mathbf{x}) = \bigcap_{\beta=1}^{\gamma} T_\beta(\mathbf{x})$ is tangent to Σ at $\mathbf{x}$, it follows from (15) that

$$N(\mathbf{x}^*(t')) \cdot \mathbf{f}(\mathbf{x}^*(t'), \mathbf{u}^*(t')) = 0, \qquad \forall N(\mathbf{x}^*(t')) \in \mathcal{C}_n(\mathbf{x}^*(t')), \tag{26}$$

for all $t' \in [t_0, t_1]$.

12. Boundary points of Σ. If subset E^* of E possesses a boundary ∂E^*, then the domain on which limiting surfaces are defined in E^{n+1} is bounded by an x_0-cylindrical surface, $\mathcal{B}$.

We shall assume here that $E^* \equiv E$; namely, prescribed terminal point $\mathfrak{x}^1$ cannot be reached from a point from which it cannot be reached along an optimal path.

Let $I/\mathcal{B}$ denote the open set of points interior to the domain of definition of Σ surfaces, and let $O/\mathcal{B}$ denote the open set of points exterior to that domain. Then we have at once the following results.

LEMMA 11. *If* $\mathbf{x}(t')$ *and* $\mathbf{x}(t'')$, $t'' > t'$, *are points of a trajectory, and* $\mathbf{x}(t') \in O/\mathcal{B}$, *then* $\mathbf{x}(t'') \notin I/\mathcal{B}$.

LEMMA 12. *If point* $\mathbf{x}^*(t')$ *of optimal trajectory* Γ^* *belongs to boundary* $\mathfrak{B}$, *then* Γ^* *lies in* $\mathfrak{B}$ *for all* t, $t' \leqq t \leqq t_1$.

If $\mathbf{x}'$ is a regular point of boundary $\mathfrak{B}$, and $\mathbf{n}(\mathbf{x}')$ denotes the unit vector normal to $\mathfrak{B}$ at $\mathbf{x}'$ and directed into region $I/\mathfrak{B}$, then it follows from Lemma 11 that

$$\mathbf{n}(\mathbf{x}')\cdot\mathbf{f}(\mathbf{x}', \mathbf{u}) \leqq 0, \qquad \forall \mathbf{u} \in U. \tag{27}$$

And, as a consequence of Lemma 12, we have

$$\mathbf{n}(\mathbf{x}^*(t'))\cdot\mathbf{f}(\mathbf{x}^*(t'), \mathbf{u}^*(t')) = 0. \tag{28}$$

It is noteworthy that the salient properties of the boundary surface $\mathfrak{B}$ correspond to those of a limiting surface Σ provided we invoke the correspondence of regions $O/\mathfrak{B}$ and $I/\mathfrak{B}$ to A/Σ and B/Σ, respectively. In particular, we note that

(i) Lemma 11 corresponds to Corollary 1.1;

(ii) conditions (27) and (28) correspond to (14) and (15), respectively;

(iii) Lemma 12 corresponds to Lemma 2.

In fact, if for a given Σ surface, we disregard the points of $I/\mathfrak{B} \cup \mathfrak{B}$ which belong to A/Σ, the intersection $\Sigma \cap \mathfrak{B}$ has the salient features of an *attractive manifold.*

13. The maximum principle. Upon invoking the various lemmas stated in the preceding sections, it is again possible to arrive at Pontryagin's maximum principle. Now, however, we no longer have an adjoint vector $\lambda(t)$ which is normal to a Σ surface, but rather one which is normal to a separating hyperplane $\mathfrak{I}(\mathbf{x})$ of the separable cone $\mathfrak{C}_t(\mathbf{x})$. Of course, at a regular point of Σ this distinction disappears. Nonetheless, it is an important distinction at nonregular points, since it invalidates[9] relation (23) between $\lambda(t)$ and grad $\Phi(\mathbf{x}^*(t))$.

REFERENCES

[1] H. HALKIN, *The principle of optimal evolution*, Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 284 f.

[2] E. ROXIN, *A geometric interpretation of Pontryagin's maximum principle*, Ibid., pp. 303f.

[3] A. BLAQUIERE, *Sur la théorie de la commande optimale*, Course at the Faculty of Sciences, University of Paris, 1963.

[4] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[5] R. E. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[6] A. BLAQUIERE AND G. LEITMANN, *On the geometry of optimal processes—part I*, I. E. R. Report AM 64-10, University of California, Berkeley.

[9] A simple example which demonstrates the inapplicability of (23) at nonregular points is the well-known problem of time optimal transfer from (x_1^0, x_2^0) to (0, 0) for $\dot{x}_1 = x_2$, $\dot{x}_2 = u$, $|u| \leqq 1$.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

ON A CERTAIN PROBLEM FOR PARABOLIC DIFFERENTIAL EQUATIONS CONNECTED WITH OPTIMAL PURSUIT*

E. F. MISHCHENKO†

Let there be in n-dimensional space two moving points, one of which is controlled according to the law

$$\dot{z} = f(x, u), \tag{1}$$

where u is the control parameter, $| u | \leqq 1$, and the other is a random point of the Markov type.

We denote by $p(\sigma, x, \tau, y)$ the probability density of the random point being at time τ in the position y if at time σ it is in the position x.

Suppose that moving together with z is a small neighborhood Σ_z of z, where Σ_z is bounded by a piecewise-smooth surface, for example, a sphere S_z of radius ϵ and center at the point z.

It is required to calculate the probability of the following event: that in the time interval $\sigma \leqq t \leqq \tau$ the random point will be covered by the neighborhood Σ_z ; that is, the random point crosses the surface S_z .

Since the probability $\phi(\sigma, x, \tau)$ which is sought is a functional of the control $u(t)$, the problem reduces to an application of the maximum principle as soon as this functional has been calculated.

It is clear, however, that in certain instances there may be interest in the problem of the pursuit of the random point by the controlled point in the sense that only a certain number of the coordinates of the controlled point come close to the corresponding coordinates of the random point. Thus, we arrive at the natural generalization of the above-formulated problem, namely:

Suppose that in the space $(z^1, z^2, \cdots, z^n)$ there moves a k-dimensional manifold M, changing its form and position according to the law

$$M = M_s . \tag{2}$$

(We shall consider the manifold M_s to be twice continuously differentiable). Suppose also that moving together with M is its n-dimensional ϵ-neighborhood $U(M)$. It is required to calculate the probability of the following

* Received by the editors December 16, 1964. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964. The original manuscript of this paper in Russian was translated into English by A. Naparstek.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† V. A. Steklov Mathematics Institute, Academy of Sciences of the USSR, Moscow, USSR, and the University of Michigan, Ann Arbor, Michigan.

event: that the random Markov point falls in the neighborhood $U(M)$ in the time interval $\sigma \leqq s \leqq \tau$.

As A. N. Kolmogorov has proved, the function $p(\sigma, x, \tau, y)$ satisfies the parabolic differential equation

$$\frac{\partial p}{\partial \sigma} + \sum_{i,j=1}^{n} a^{ij}(\sigma, x) \frac{\partial^2 p}{\partial x^i \partial x^j} + \sum_{j=1}^{n} b^j(\sigma, x) \frac{\partial p}{\partial x^j} = 0. \tag{3}$$

It is easy to see that the probability $\phi(\sigma, x, \tau)$ being sought also satisfies the same differential equation, where $\phi(\tau, x, \tau) = 0$, and $\phi(\sigma, x, \tau) = 1$ whenever x belongs to the boundary $U(M_\sigma)$ of the manifold M_σ. We denote this boundary by $V(M_\sigma)$.

Thus, we need to solve the differential equation

$$\frac{\partial \phi}{\partial \sigma} + \sum_{i,j=1}^{n} a^{ij}(\sigma, x) \frac{\partial^2 \phi}{\partial x^i \partial x^j} + \sum_{j=1}^{n} b^i(\sigma, x) \frac{\partial \phi}{\partial x^i} = 0 \tag{4}$$

under the conditions

$$\begin{aligned} &\text{(a)} && \phi(\tau, x, \tau) = 0, \\ &\text{(b)} && \phi(\sigma, x, \tau)\,|_{x \in V(M_\sigma)} = 1. \end{aligned} \tag{5}$$

It appears that the result of solving this equation up to an accuracy greater than a high order of ϵ may be written in the form of an explicit formula. In order to do this, several helpful constructions are necessary.

Through each point m_s of the manifold M_s we draw the tangent plane $P(m_s)$. Then we choose n linearly independent vectors $e_1, e_2, \cdots, e_n$, leaving the point m_s such that

(a) $e_1, e_2, \cdots, e_k$ are contained in $P(M_s)$, and

(b) in the coordinate system $\xi^1, \xi^2, \cdots, \xi^n$, referring to the basis $e_1, e_2, \cdots, e_n$, the differential operator

$$a^{ij}(s, m_s) \frac{\partial^2}{\partial x^i \partial x^j}$$

may be written in the form of the Laplace operator

$$\sum_{\nu=1}^{n} \frac{\partial^2}{(\partial \xi^\nu)^2}.$$

We denote by $Q(m_s)$ the subspace spanned by the vectors $e_{k+1}, \cdots, e_n$ which is associated with $P(m_s)$. The collection of points in the subspace $Q(m_s)$ which are separated (in the metric of R^n) from the plane $P(m_s)$ by a distance ϵ is an ellipsoid E_{m_s}. Let its equation in the coordinates ξ be

$$\sum_{i,j=1}^{n} c_{ij} \xi^i \xi^j = \epsilon^2. \tag{6}$$

It is obvious that we ehav

$$V(M_s) = E_{m_s} \times M_s \tag{7}$$

with an accuracy greater than a high order of ϵ. Further, we denote by $w(\xi^{k+1}, \cdots, \xi^n)$ the harmonic function vanishing as $|\xi| \to \infty$ and equal to one on the ellipsoid E'_{m_s}, distinguished in $Q(m_s)$ by the equation

$$\sum_{i,j=1}^{n} c_{ij}\xi^i\xi^j = 1.$$

It is known that w can be represented in the form

$$w = \frac{\alpha(m_s)}{\rho^{n-k-2}} + \pi(\xi^{k+1}, \cdots, \xi^n), \tag{8}$$

where $\rho^2 = (\xi^{k+1})^2 + \cdots + (\xi^n)^2$, $\alpha(m_s)$ is uniquely defined by the dimensions of the ellipsoid E'_{m_s}, and π is the double layer potential created by E'_{m_s} at the point $(\xi^{k+1}, \cdots, \xi^n)$. Upon differentiating the right and left side of (8) in the ρ direction and after taking the integral over the surface E'_{m_s}, we easily see that

$$\int_{E'_{m_s}} \frac{\partial w}{\partial \rho}\, dE'_{m_s} = \frac{4\pi^{(n-k)/2}}{\Gamma\left(\frac{n-k}{2} - 1\right)} \alpha(m_s) = \beta(m_s), \tag{9}$$

in which Γ is Euler's gamma function.

We can now formulate the following proposition:

The solution of (4) *under the conditions* (5) *can be represented in the form*

$$\phi(\sigma, x, \tau) = \epsilon^{n-k-2} \int_\sigma^\tau ds \int_{M_s} p(\sigma, x, s, m_s)\beta(m_s)\, dM_s + \omega(\sigma, x, \tau, \epsilon), \tag{10}$$

where ω *has a magnitude of order* ϵ^{n-k-1} *for any point* x *which is separated from the manifold* M_σ *by a finite distance independent of* ϵ.

In (10) the interior integration is carried out over the entire manifold M_s, in which the element of volume is indexed at each point by the reference frame $e_1, e_2, \cdots, e_k$. It is easy to see that this definition of volume depends only on the coefficients a^{ij} of (4) and does not depend on the admissible arbitrariness in the choice of the reference frame $e_1, \cdots, e_k$.

The schema of proof of the formulated proposition follows.

The function

$$\Phi(\sigma, x, \tau) = \epsilon^{n-k-2} \int_\sigma^\tau ds \int_{M_s} p(\sigma, x, s, m_s)\beta(m_s)\, dM_s \tag{11}$$

is the solution of (4) in the exterior of the manifold M_σ and satisfies (5a). But it does not satisfy the boundary condition (5b), the second of condi-

tions (5). It appears, however, that one can construct an n-dimensional elliptical neighborhood of the manifold M_σ, on the boundary of which the values of the solutions $\Phi(\sigma, x, \tau)$ and $\phi(\sigma, x, \tau)$ actually coincide. Let us construct this neighborhood.

For this, we shall take in each subspace $Q(m_\sigma)$ the ellipsoid $E^*_{m_\sigma}$ distinguished by the equation

$$\rho = \epsilon, \tag{12}$$

and we shall set

$$V^*(M_\sigma) = E^*_{m_\sigma} \times M_\sigma . \tag{13}$$

The surface V^* is, indeed, the boundary of the required neighborhood of the manifold M_σ. We emphasize that, generally speaking, V^* does not coincide with V.

Now, by using known asymptotic expansions for the function $p(\sigma, x, s, y)$ and by performing elementary, although tedious, calculations, we find that for $x_0 \in V^*(M_\sigma)$,

$$\Phi(\sigma, x_0, \tau) = \alpha(m_{0\sigma}) + \omega_1(\sigma, x_0, \tau, \epsilon), \tag{14}$$

where ω_1 has the order $O(1)$ for $\tau - \sigma \leqq \epsilon$, and vanishes for $\epsilon \to 0$ when $\tau - \sigma > \epsilon$. Here, $m_{0\sigma}$ denotes the projection of the point x_0 onto the manifold M_σ in the direction determined by $Q(m_s)$.

On the other hand, by using methods which are natural analogues of the method in [2], we can write down the solution $\phi(\sigma, x, \tau)$ in a certain special form, in which it is immediately possible to observe that

$$\phi(\sigma, x_0, \tau) = \alpha(m_{0\sigma}) + \omega_2(\sigma, x_0, \tau, \epsilon), \tag{15}$$

and moreover, ω_2 has the same asymptotic character with respect to ϵ as does ω_1. Upon equating relations (14) and (15), it is not difficult to extract (10).

In the case when the manifold M_s is simply a point $z(s)$, and Σ_z its spherical neighborhood of radius ϵ, (10) turns out much simpler. Namely,

$$\phi(\sigma, x, \tau) = \epsilon^{n-2} \int_\sigma^\tau p(\sigma, x, s, z(s))\beta(s)\, ds + o(\epsilon^{n-2}), \tag{16}$$

where

$$\beta(s) = \int_{A_s\Sigma} \frac{\partial w(s, \xi)}{\partial n}\, d\Sigma.$$

Here, A_s denotes the linear transformation $\xi = A_s\bar{\xi}$ which converts the differential form

$$\sum_{i,j} a^i\,(s, z(s))\, \frac{\partial^2}{\partial \xi^i \partial \xi^j}$$

to the Laplace operation, and $w(s, \xi)$ is the harmonic function satisfying the conditions

$$w(s, \xi) = 1, \qquad \xi \in A_s\Sigma,$$
$$w(s, \xi) \to 0, \qquad |\xi| \to \infty.$$

In conclusion, we note that in case $n - k = 2$, formulas (10) and (16) are no longer valid; but instead of these, the following two formulas, respectively, are valid:

$$\phi(\sigma, x, \tau) = \frac{2\pi}{|\log \epsilon|} \int_\sigma^\tau ds \int_{M_s} p(\sigma, x, s, m_s)\, dM_s,$$

and

$$\phi(\sigma, x, \tau) = \frac{2\pi}{|\log \epsilon|} \int_\sigma^\tau p(\sigma, x, s, z(s))\, ds.$$

REFERENCES

[1] E. F. Mishchenko and L. S. Pontryagin, *A statistical optimal control problem*, Izv. Akad. Nauk SSSR, Ser. Mat., 25(1961), pp. 477–498.

[2] A. N. Kolmogorov, E. F. Mishchenko, and L. S. Pontryagin, *A probability problem in optimal control*, Dokl. Akad. Nauk SSSR, 145(1962), pp. 993–995; Soviet Math. Dokl., 3(1962), pp. 1143–1145.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

A DOUBLY SINGULAR PROBLEM IN OPTIMAL INTERPLANETARY GUIDANCE*

J. V. BREAKWELL†

The problem we wish to describe is a doubly singular problem, i.e., a problem singular in 2 controls, which has recently confronted Frank Tung and the author in connection with "minimum effort" interplanetary guidance.

The problem has the following background. A vehicle launched from a low-altitude earth orbit towards a planet will require several trajectory corrections to insure arrival in the close vicinity of the planet. It is not difficult to schedule several correction times so that the total amount of fuel necessary for implementing these corrections is but a small fraction of the fuel necessary to launch the vehicle away from the earth orbit. This is true, at any rate, if launching accuracy comes up to present-day standards, and if subsequent measurements for trajectory determination, which we shall suppose are made by on-board optical instruments, approach the accuracy of engineering estimates thereof. It is possible [1], [2], indeed, to optimize the number and timing of the corrections so as to minimize the average total velocity correction consistent with a specified reasonable terminal accuracy. This optimum timing will depend considerably on the frequency (not necessarily constant) as well as the accuracy of the measurements, i.e., on the (variable) "information rate".

The average total velocity correction depends partly on launching accuracy and partly on the information rate history, especially near arrival at a planet, but is not greatly affected by a reduction in information rate over the long mid-course phase. Furthermore, whether on-board measurements of the planet against a star background are made photographically by astronauts or by powered star- and planet-trackers, there are good reasons for reducing the total number of measurements to be made to a number very much smaller than the number possible by measuring throughout at a maximum rate, say once a minute, even though the average total velocity correction would thereby be somewhat increased. We are led, thus, to formulate the following:

PROBLEM. *Optimize the variable observation rate as well as the correction schedule so as to achieve a desired terminal accuracy with a maximum value*

* Received by the editors December 14, 1964, and in revised form March 11, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Aeronautics and Astronautics, Stanford University, Stanford, California.

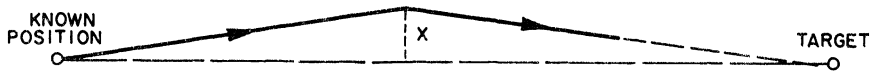

FIG. 1

of a specified linear combination of total number of observations and average total velocity correction.

For mathematical simplicity we shall idealize the observation, whether at maximum rate or a lower rate, as a continuous process corrupted by white noise. We shall also ignore, for mathematical simplicity, any errors in executing the indicated trajectory corrections, and we may add that engineering estimates indicate that these are relatively insignificant. We shall furthermore confine attention to an essentially one-dimensional control problem in which we assume that we are approaching the vicinity of the target planet with a constant velocity vector interrupted by velocity impulses perpendicular to the nominal straight line approach to the target, or else by a continuous acceleration $\ddot{x}$ in this perpendicular direction. (See Fig. 1.)

We shall make the further important simplifying assumption that the control acceleration, whether impulsive or not, is *linear* in the predicted miss, that is,

$$\ddot{x}(t) = -u(t)\hat{x}_f(t), \tag{1}$$

where

$$\hat{x}_f(t) = \hat{x}(t) + (T - t)\hat{\dot{x}}(t),$$

$\hat{x}$ and $\hat{\dot{x}}$ being the best linear estimates of the instantaneous state components x and $\dot{x}$ based on assumed noise levels in the (unbiased) continuous measurements of x and on an assumed variance of an (unbiased) initial transverse velocity error $\dot{x}(0)$, $x(0)$ being assumed negligible, and where $T - t$ is the time-to-go. Note that $\hat{x}_f(t)$ is the best estimate of

$$x_f(t) = x(t) + (T - t)\dot{x}(t),$$

which is the miss in the absence of further control. In the case of impulsive corrections, $u(t)$ consists of δ-functions.

Now it may be shown that the variance $q(t)$ of the error $(\hat{x}_f(t) - x_f(t))$ in the predicted miss decreases according to the information rate, independently of $u(t)$:

$$\dot{q}(t) = -r(t)a(t)q^2(t), \tag{2}$$

where $r(t)$ is the observation rate and $a(t)$ is a measure of the geometrical effectiveness of the measurements and increases markedly as $t \to T$ in the case of angular measurements of the target's instantaneous direction.

On the other hand, if $p(t)$ denotes the variance of the predicted miss $\hat{x}_f(t)$ and $s(t)$ the variance of the miss $x_f(t)$, then

$$\dot{s}(t) = -2\tau u(t)p(t),$$

where τ is (time-to-go) $T - t$, and

$$s(t) = p(t) + q(t),$$

since the error in $x_f(t)$ is known to be independent of $\hat{x}_f(t)$. It follows that

$$\dot{p}(t) = -2\tau u(t)p(t) + r(t)a(t)q^2(t). \tag{3}$$

Note that $p(0) = 0$, since the unbiased initial predicted miss vanishes, while $q(0) = s(0) = T^2 \operatorname{cov}(\dot{x}(0))$. Furthermore, the average total velocity correction is

$$\int_0^T E\{|u(t)\hat{x}_f(t)|\}\,dt = \sqrt{\frac{2}{\pi}} \int_0^T u(t)\sqrt{p(t)}\,dt,$$

assuming $u(t)$ is nonnegative, and the total number of observations is represented by $\int_0^T r(t)\,dt$, while the mean-squared terminal miss is $s(T) = p(T) + q(T)$.

We thus seek to determine the "control variables" $u(t)$ and $r(t)$, subject to the inequalities

$0 \leqq r \leqq R$ (the maximum observation rate),

$0 \leqq u \leqq \infty$ (it is simpler to ignore any finite upper limit),

so as to minimize

$$\int_0^T [2u(t)\sqrt{p(t)} + \alpha r(t)]\,dt,$$

α being a specified constant, subject to a given sum $s(T)$ of the final values of the "state variables" p and q whose known initial values are 0 and $T^2 \operatorname{cov}(\dot{x}(0))$ and which satisfy the differential constraints (2) and (3), where $a(t)$ is a known function. We assume that $s(T) < s(0)$, so that a negative $u(t)$ would not be helpful!

Note that this is a doubly singular problem in that both control variables occur only linearly in the appropriate Hamiltonian. We do not know the solution to this problem but we wish to present a conjecture as to its nature. Before doing so, let us describe the solution of the much simpler problem [3] in which the observation rate $r(t)$, and hence also $q(t)$, is prescribed; it is desired to minimize

$$\int_0^T 2u(t)\sqrt{p(t)}\,dt$$

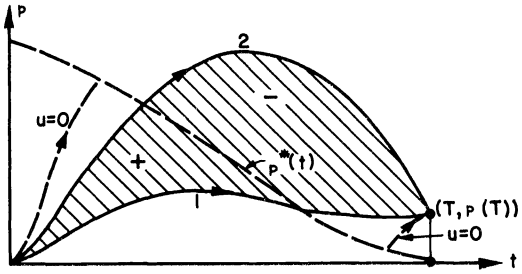

FIG. 2

subject effectively to a specified $p(T)$. With the aid of (3) we may write this cost as

$$\int_0^T \frac{b(t)\,dt - dp}{\tau\sqrt{p}},$$

where $b(t)$ denotes the known function $r(t)a(t)q^2(t)$. The difference in cost associated with 2 different strategies $u(t)$ leading to the specified $p(T)$ is thus expressed as a line integral around a closed curve in the (t, p)-plane, which, according to Green's Theorem, is

$$\iint \left[\frac{\partial}{\partial t}\left(-\frac{1}{\tau\sqrt{p}}\right) - \frac{\partial}{\partial p}\left(\frac{b(t)}{\tau\sqrt{p}}\right)\right] dt\,dp,$$

evaluated over the enclosed area.

But (see Fig. 2) the integrand of the double integral is easily found to be positive or negative, respectively, below or above a "critical curve"

$$p^*(t) = \tfrac{1}{2}\tau b(t),$$

which is continuous whenever the observation rate $r(t)$ is continuous. Since

$$-\infty \leqq \dot{p}(t) \leqq b(t),$$

where $b(t)$ is the upper limit corresponding to $u = 0$, we see that the optimal strategy consists of a period of no control, while p rises from 0 to the critical curve, followed by a period of *continuous* (non-impulsive) control as long as $r(t)$ and hence $b(t)$ is continuous, and provided that $\dot{p}^*(t)$ does not exceed $b(t)$, followed finally by a period of no control just before arriving near the target. Note that an instantaneous drop in observation rate and hence in $p^*(t)$ requires an impulsive correction, but any sharp rise in observation rate produces a critical curve which cannot be followed. In case, for example, $r(t)$ is zero between t_A and t_B, it is fairly easy to see that the optimal strategy corresponds to the p and s histories shown in Fig. 3. The drop in observation rate is followed immediately by an impulsive partial cor-

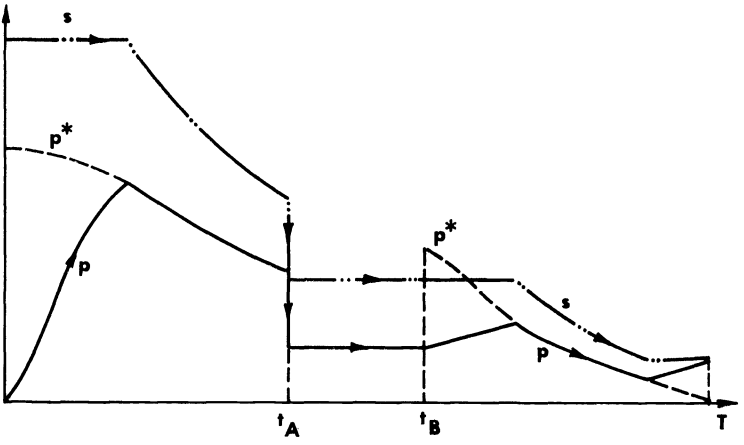

FIG. 3

rection, followed by a period of no control lasting until some time after the rise in observation rate when p rises again to critical level p^*.

It should be stressed that p and s are mean-squared quantities whose optimal histories correspond to an optimal choice of $u(t)$. A typical history of the random process $|\hat{x}_f(t)|$ is not necessarily monotonic prior to control turn-on and its value at turn-on is not generally at some preassigned critical level. Neither is its final value specified.

Coming back now to our problem of optimizing both the observation rate $r(t)$ and the trajectory control gain $u(t)$, we *conjecture* that the optimal $r(t)$ is "bang-bang", i.e., that it consists of one or more periods of observation at maximum rate R, separated by periods of no observation $r = 0$. If so, any period of observation must be followed immediately by an impulsive correction and then a period of no control lasting into the next observation period.

If this conjecture is true, we will now show that the solution can be determined numerically without an undue amount of iteration. To establish the conjecture, however, it would be necessary to rule out the possibility of doubly-singular arcs, involving intermediate levels of both $r(t)$ and $u(t)$. This we have not so far been able to do, but it is appropriate to suggest here that Kelley's recent work [4] on tests for singular extremals might be applicable.

To investigate the computation of the solution if the conjecture is true, let $\lambda(t)$ and $\mu(t)$ be adjoint variables corresponding to $p(t)$ and $q(t)$. The Hamiltonian to be minimized is thus

$$H = \lambda(raq^2 - 2\tau up) - \mu raq^2 + 2u\sqrt{p} + \alpha r.$$

The rates of change of the adjoint variables are:

$$\dot{\lambda} = -\frac{\partial H}{\partial p} = 2\lambda\tau u - \frac{u}{\sqrt{p}}, \qquad \dot{\mu} = -\frac{\partial H}{\partial q} = 2(\mu - \lambda)raq.$$

The terminal constraint on $p + q$ requires that λ, μ satisfy the end-constraint:

$$\lambda(T) = \mu(T).$$

The minimization of H with respect to u shows that

$$\lambda\tau\sqrt{p} = 1$$

during any control period ($u(t) > 0$), and it is easy to show that an impulsive correction preserves the product $\lambda\sqrt{p}$, the instantaneous drop in p being matched by a rise in λ.

The minimization of H with respect to r shows that $r = 0$ or $r = R$ according to the sign (+ or −) of the following switching function:

$$F = \alpha - (\mu - \lambda)aq^2.$$

But we find that

$$\frac{d}{dt}[(\mu - \lambda)q^2] = -\frac{u}{\sqrt{p}}q^2.$$

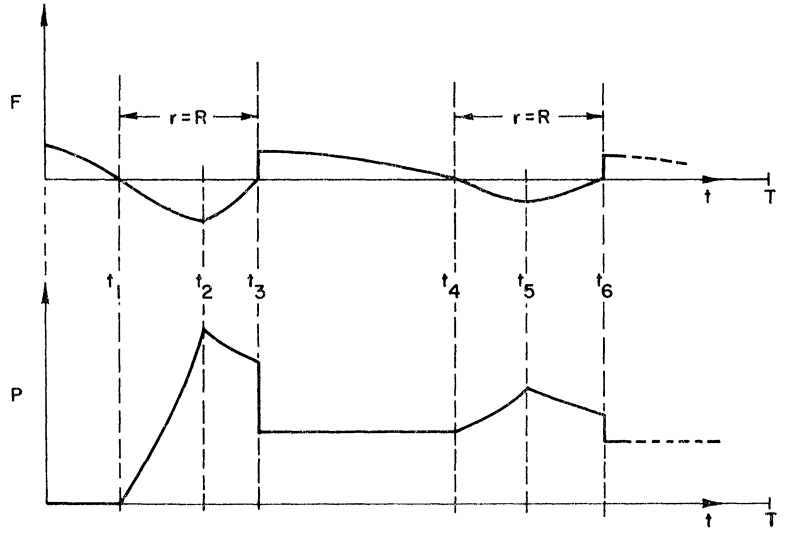

FIG. 4

Assuming that $a(0) = 0$ and that $a(t)$ is an increasing function, the computation may proceed (see Fig. 4) as follows:
guess an initial positive value for $(\mu - \lambda)$;
keep $r = 0$ and $q = q(0)$ until time t_1 when $a(t)$ reaches the value $\alpha/\{[\mu(0) - \lambda(0)]q^2(0)\}$;
now keep $r = R$ and $u = 0$ and compute $q(t)$ and $p(t) = s(0) - q(t)$ until time t_2 when $p(t)$ reaches the critical $p^*(t) = \frac{1}{2}\tau Raq^2$;
compute $\lambda(t_2) = 1/\tau_2\sqrt{p(t_2)}$ and $\mu(t_2)$ such that $[\mu(t_2) - \lambda(t_2)]q^2(t_2) = [\mu(0) - \lambda(0)]q^2(0)$;
now keep $p(t) = p^*(t)$, $\lambda(t) = 1/\tau\sqrt{p^*(t)}$, compute q and $u/\sqrt{p}$, the known time derivative of $\lambda(t)$, and by numerical integration compute $(\mu - \lambda)q^2$, until time t_3 when F again reaches zero;
apply an impulsive drop in p to be determined later, together with the corresponding impulsive rise in λ, F becoming again positive;
keep $r = 0$ and q constant until time t_4 when F is again zero;
now keep $r = R$ and $u = 0$ until time t_5 when $p(t)$ again reaches a critical level $p^*(t)$, at which time require that λ, which is unchanged since the impulse, be equal to $1/\tau_5\sqrt{p^*(t_5)}$;
this last requirement is used, in an iterative loop, to determine the amount of the impulse at t_3 ;
the computation then continues on to a time t_6 at which F again reaches zero, to be followed by another impulse whose amount will be determined by a later iterative loop, etc.;
each of the times t_3 , t_6 , $\cdots$ may be also considered as a time of final observation cut-off to be followed immediately by an impulse whose magnitude is such that λ jumps to the value μ, in order that $\lambda(T)$ shall equal $\mu(T)$.

This computation thus provides a finite number of terminal variances $s(T)$ for every guess of the single quantity $\mu(0) - \lambda(0)$, and the optimal strategy for any $s(T) < q(0)$ is thus quite easily obtained.

REFERENCES

[1] D. F. Lawden, *Optimal programme for correctional manoeuvres*, Astronaut. Acta, 4 (1960), pp. 106–123.

[2] J. V. Breakwell, *Fuel requirements for crude interplanetary guidance*, Advances in Astronaut. Sci., 5 (1960), pp. 53–65.

[3] J. V. Breakwell and C. T. Striebel, *Minimum effort control in interplanetary guidance*, this Journal, to appear.

[4] H. J. Kelley, *A second variation test for singular extremals*, presented at Control Optimization Symposium, Monterey, California, 1964.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

CONTROLLABILITY OF NONLINEAR PROCESSES*

LAWRENCE MARKUS†

Introduction. Control theory, as formulated within the framework of ordinary differential equations, has two general approaches: the qualitative theory of controllability, and the quantitative theory of optimal control. These two aspects of control theory are unified by the study of the geometric properties of the set of attainability. In this paper we present some new results and examples which illuminate the significance of the set of attainability in both the qualitative and quantitative facets of control theory.

1. Controllability and regulation. Consider a nonlinear autonomous process described by a differential system in R^n (real n-space):

$$\text{S)} \qquad \dot{x} = f(x, u),$$

where the state vector x is in R^n, the control vector u is in some nonempty restraint set $\Omega \subset R^m$ at each time t, and the coefficient function $f(x, u) \in C^1$ in R^{n+m}. We seek to steer or control the initial state $x_0 \in R^n$ to some final target x_1 (usually $x_1 = 0$ is the origin, but the target could be all R^n). The class Δ of admissible controllers consists of bounded measurable functions $u(t) \subset \Omega$ on various finite time durations $0 \leqq t \leqq t_1$, each of which steers the response $x(t)$ of

$$\dot{x} = f(x, u(t)), \qquad x(0) = x_0,$$

to $x(t_1)$ in the prescribed target. In optimal control theory a cost functional $C(u)$ is defined and we seek to minimize $C(u)$ among all admissible controllers; however we shall suppress this concept when investigating qualitative control theory.

The first two theorems on controllability are easy generalizations of well-known [2] global stability theory, and the third theorem is a nonlinear analogue of an important linear controllability criterion [5].

THEOREM 1. *Consider the control process in* R^n,

$$\text{S)} \qquad \dot{x} = f(x, u),$$

$f \in C^1$ *in* R^{n+m}, *and restraint* $u(t) \subset \Omega \subset R^m$. *Assume there exist a scalar*

* Received by the editors March 22, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† School of Mathematics, Institute of Technology, University of Minnesota, Minneapolis, Minnesota. This research was supported by NONR 3776–(00).

function $V(x)$ and an m-vector function $U(x)$ in C^1 in R^n such that:

(i) $V(x) \geqq 0$, *with equality if and only if* $x = 0$;

(ii) $\lim_{|x|\to\infty} V(x) = +\infty$;

(iii) $U(x) \subset \Omega$;

(iv) $\dfrac{\partial V(x)}{\partial x^i} f^i(x, U(x)) < 0$ *for* $x \neq 0$.

Then each initial state $x_0 \in R^n$ can be controlled to an arbitrarily small neighborhood of the origin.

Proof. Define the response $x(t)$ from $x_0 \neq 0$ by the solution of the differential system

$$\dot{x} = f(x, U(x)), \qquad x(0) = x_0 .$$

Let the control be $u(t) = U(x(t))$, so $u(t) \subset \Omega$. Along the solution $x(t)$ consider $V(x(t))$ and compute

$$\frac{dV}{dt} = \frac{\partial V}{\partial x^i}(x(t))\, f^i(x(t), U(x(t))) < 0,$$

as long as $x(t) \neq 0$ (if $x(t_1) = 0$, then the origin target has been reached)· Thus $x(t)$ exists for all $0 \leqq t < \infty$ and lies within the compact set in R^n defined by $V(x) \leqq V(x_0)$. By the usual technique of A. M. Lyapunov we find that $\lim_{t\to+\infty} V(x(t)) = 0$ and so $\lim_{t\to+\infty} x(t) = 0$, as required.

Example. Consider the regulation towards zero of the angular velocity $\omega = (\omega_1, \omega_2, \omega_3)$ of a rigid body rotating in inertial space. The Euler equations of motion are

$$\begin{aligned} I_1\dot{\omega}_1 &= (I_2 - I_3)\omega_2\omega_3 + u_1(t), \\ I_2\dot{\omega}_2 &= (I_3 - I_1)\omega_3\omega_1 + u_2(t), \\ I_3\dot{\omega}_3 &= (I_1 - I_2)\omega_1\omega_2 + u_3(t). \end{aligned}$$

Here I_1, I_2, I_3 are the positive constant principal moments of inertia of the body and the control vector $u(t) = (u_1(t), u_2(t), u_3(t))$ satisfies the restraint $|u_i(t)| \leqq 1$ for $i = 1, 2, 3$. Define the Lyapunov function

$$V(\omega_1, \omega_2, \omega_3) = \tfrac{1}{2}[I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2],$$

and take

$$U(\omega) = -\epsilon \operatorname{grad} V(\omega),$$

where the constant $\epsilon > 0$ is chosen so small that $|\epsilon I_i\omega_i| \leqq 1$ for $i = 1, 2, 3$, within the solid ellipsoid $V(\omega) \leqq V(\omega(0))$. Then the initial state $\omega(0)$ can be steered towards $\omega = 0$.

THEOREM 2. *Consider the control process in R^n,*

$$\text{S)} \qquad \dot{x} = f(x, u),$$

with $f(x, u) \in C^1$ *in* R^{n+m} *and with restraint* $u(t) \subset \Omega \subset R^m$. *Assume*:

(i) $f(0, 0) = 0$,

(ii) *there exists an m-vector function* $U(x)$ *in* C^1 *in* R^n *with* $U(x) \subset \Omega$ *and* $U(0) = 0$,

(iii) *every eigenvalue* $\lambda(x)$ *of* $J(x) + J^T(x)$ *satisfies* $\lambda(x) < -\epsilon < 0$ *for all* $x \in R^n$ *and some* $\epsilon > 0$, *where*

$$J(x) = \frac{\partial f}{\partial x}(x, U(x)) + \frac{\partial f}{\partial u}(x, U(x)) \frac{\partial U}{\partial x}.$$

Then each initial state $x_0 \in R^n$ *can be controlled to an arbitrarily small neighborhood of the origin.*

Proof. By a result of Krasovskii [2, 8] every solution of the autonomous system $\dot{x} = f(x, U(x))$ exists on $0 \leqq t < \infty$ and approaches the origin as t increases. Let the required response be the solution $x(t)$ with $x(0) = x_0$. Define the control $u(t) = U(x(t))$ so $u(t) \subset \Omega$. Thus the control $u(t)$ yields the response $x(t)$ by the process $\mathcal{S}$ and $\lim_{t\to\infty} x(t) = 0$.

The above two theorems assert that the response $x(t)$ approaches the origin as t increases. In order to complete the regulation of an initial state x_0 to the exact origin $x_1 = 0$ we need a local controllability result.

Define the domain $\mathcal{C}$ of null controllability for the process

$$\mathcal{S}) \qquad \dot{x} = f(x, u),$$

with $f \in C^1$ in R^{n+m}, to be all those initial states $x_0 \in R^n$ which can be steered to the origin in a finite time, by admissible controllers. It is clear that $\mathcal{C}$ is a connected set, and also $\mathcal{C}$ is open in R^n if and only if $\mathcal{C}$ contains a neighborhood of the origin.

Consider the linear process in R^n,

$$\mathcal{L}) \qquad \dot{x} = Ax + Bu,$$

with restraint $\Omega \subset R^m$ such that the convex hull $H(\Omega)$ contains $u = 0$ in its interior. Then the domain of null controllability for $\mathcal{L}$ is open in R^n if and only if the controllability condition obtains (see [5, 9]):

$$\text{rank } [B, AB, A^2B, \cdots, A^{n-1}B] = n.$$

The next theorem asserts that the above algebraic controllability condition (at the origin) implies the geometric controllability of the nonlinear process $\mathcal{S}$. For the linear process $\mathcal{L}$, the "bang-bang" principle states that responses to controllers in $H(\Omega)$ can also be attained by controllers restricted to Ω. We shall prove an analogue of this bang-bang principle for the nonlinear system $\mathcal{S}$ near the origin.

Example. Consider the nonlinear scalar process in R^1,

$$\dot{x} = u + u^2,$$

with restraint $\Omega: |u| = 2$. Then the domain of null controllability $\mathcal{C}$ is the half-closed interval $x \leqq 0$. Yet the linear approximation near $x = 0$, $u = 0$ is $\dot{x} = u$ which is controllable and $H(\Omega)$ contains $u = 0$ in its interior. This example shows that the set Ω must be sufficiently small if a suitable nonlinear bang-bang principle is to hold. Note that the set of attainability from the origin, for the nonlinear process with control restraint Ω, lies in the half-axis $x > 0$. For this same process, with the restraint $|u| \leqq 2$, the set of attainability is a segment containing both positive and negative values of x.

We first demonstrate a special version of the theorem of A. Lyapunov on vector measures [1, 3, 4]. Consider a compact time interval $\mathcal{I}: 0 \leqq t \leqq T$ and consider the σ-algebra $\mathcal{B}$ of all Lebesgue measurable subsets of $\mathcal{I}$ (modulo null sets). Let μ be the usual Lebesgue measure on $\mathcal{B}$ so $\{\mathcal{I}, \mathcal{B}, \mu\}$ is a measure σ-algebra. On $\mathcal{B}$ we define a metric by the distance

$$\rho(E, F) = \mu(E \Delta F) = \mu((E \cup F) - (E \cap F)),$$

for sets $E, F \in \mathcal{B}$ (see [4]). It is known [3] that in $\mathcal{B}$ (or in any nonatomic σ-subalgebra $\mathcal{A} \subset \mathcal{B}$ with the induced measure and metric), there exists a topological image of a segment $0 \leqq \alpha \leqq 1$ by sets $D_\alpha \in \mathcal{B}$ with $\mu(D_\alpha) = \alpha\mu(\mathcal{I})$; and $D_{\alpha_1} \subset D_{\alpha_2}$ if and only if $\alpha_1 \leqq \alpha_2$.

Notation. Let $\{\mathcal{I}, \mathcal{A}, \mu\}$ be a measure σ-algebra with the usual metric. A k-partition of $\mathcal{I}$ is a collection of k sets $A_1, \cdots, A_k$ in $\mathcal{A}$ with $A_1 \cup A_2 \cup \cdots \cup A_k = \mathcal{I}$ and $A_i \cap A_j = \varnothing$ for $i \neq j$. Consider the product metric on the k-fold product of $\mathcal{B}$ with itself and so define a topology on the space $\mathcal{P}_k$ of all k-partitions of $\mathcal{I}$. Let S be a topological space and define a continuous family of k-partitions of $\mathcal{I}$ to be a continuous map of S into $\mathcal{P}_k$.

LEMMA. *Let $h_1(t), \cdots, h_k(t)$ be integrable n-vector functions on the finite real interval $\mathcal{I}: 0 \leqq t \leqq T$. Let S be the $(k-1)$-simplex with barycentric coordinates*

$$\alpha = (\alpha_1, \cdots, \alpha_k), \qquad \alpha_i \geqq 0, \qquad \sum_{i=1}^{k} \alpha_i = 1.$$

Then there exists a continuous family of k-partitions of $\mathcal{I}$ in $\mathcal{B}$,

$$\alpha \to \{A_1(\alpha), \cdots, A_k(\alpha)\},$$

such that the integrable function

$$h(t, \alpha) = \begin{cases} h_1(t) & \text{for } t \in A_1(\alpha), \\ \vdots & \\ h_k(t) & \text{for } t \in A_k(\alpha), \end{cases}$$

satisfies the convexity condition

$$\int_0^T h(t, \alpha)\, dt = \alpha_1 \int_0^T h_1(t)\, dt + \cdots + \alpha_k \int_0^T h_k(t)\, dt.$$

Proof. There exists a nonatomic σ-subalgebra $\mathfrak{A} \subset \mathfrak{B}$ such that the kn-vector $h^* = (h_1, \cdots, h_k)$ satisfies the identity [1, 3]

$$\int_D h^* \, dt = \frac{\mu(D)}{\mu(\mathcal{S})} \int_{\mathcal{S}} h^* \, dt$$

for every $D \in \mathfrak{A}$.

Now let D_α be a topological image of the segment $0 \leqq \alpha \leqq 1$ into $\mathfrak{A}$ such that $\mu(D_\alpha) = \alpha\mu(\mathcal{S})$; and $D_{\alpha_1} \subset D_{\alpha_2}$ if and only if $\alpha_1 \leqq \alpha_2$. For each point $\alpha = (\alpha_1, \cdots, \alpha_k)$ of S we define the k-partition of $\mathcal{S}$ in $\mathfrak{A}$ by

$$A_1(\alpha) = D_{\alpha_1}, \quad \text{so} \quad \mu(A_1) = \alpha_1\mu(\mathcal{S}),$$

$$A_2(\alpha) = D_{\alpha_1+\alpha_2} - D_{\alpha_1}, \quad \text{so} \quad \mu(A_2) = (\alpha_1 + \alpha_2)\mu(\mathcal{S}) - \alpha_1\mu(\mathcal{S}) = \alpha_2\mu(\mathcal{S}),$$

$$A_3(\alpha) = D_{\alpha_1+\alpha_2+\alpha_3} - D_{\alpha_1+\alpha_2}, \quad \text{so} \quad \mu(A_3) = \alpha_3\mu(\mathcal{S}),$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$A_k(\alpha) = D_1 - D_{1-\alpha_k}, \qquad\qquad \text{so} \quad \mu(A_k) = \alpha_k\mu(\mathcal{S}).$$

Then it is easy to verify that $\alpha \to \{A_1(\alpha), \cdots, A_k(\alpha)\}$ is a continuous family of k-partitions of $\mathcal{S}$ in $\mathfrak{A}$.

Therefore

$$\int_{A_i(\alpha)} h_i \, dt = \alpha_i \int_{\mathcal{S}} h_i \, dt, \qquad\qquad i = 1, \cdots, k.$$

Thus

$$\int_0^T h(t, \alpha) \, dt = \alpha_1 \int_0^T h_1(t) \, dt + \cdots + \alpha_k \int_0^T h_k(t) \, dt,$$

as required.

THEOREM 3. *Consider the control process in* R^n,

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

with $f \in C^1$ *in* R^{n+m} *and restraint* $u(t) \subset \Omega \subset R^m$. *Assume*:

(i) $f(0, 0) = 0$,

(ii) Ω *contains* $m + 1$ *vectors* $u_1, u_2, \cdots, u_{m+1}$ *which span an* m*-simplex with* $u = 0$ *in its interior, and* Ω *also contains* $\epsilon u_1, \epsilon u_2, \cdots, \epsilon u_{m+1}$ *for certain arbitrarily small* $\epsilon > 0$,

(iii) rank $[B, AB, A^2B, \cdots, A^{n-1}B] = n$, *where*

$$A = \frac{\partial f}{\partial x}(0, 0), \qquad B = \frac{\partial f}{\partial u}(0, 0).$$

Then the domain $\mathfrak{C}$ *of null controllability of* S *is an open neighborhood of the origin in* R^n.

Proof. For each small bound

$$|u(t)| \leqq \epsilon < \epsilon_0 < 1$$

on $0 \leqq t \leqq 1$, the response $x(t)$ of

$\mathcal{S}$) $$\dot{x} = f(x, u(t)), \qquad x(0) = 0,$$

and $x_L(t)$ of the linear approximating system

$\mathcal{L}$) $$\dot{x} = Ax + Bu(t), \qquad x_L(0) = 0,$$

are defined on $0 \leqq t \leqq 1$ and there satisfy a corresponding bound

$$|x(t)| + |x_L(t)| \leqq c(\epsilon) < 1$$

where $\lim_{\epsilon \to 0} c(\epsilon) = 0$. Here the norm of a vector or matrix is the sum of the absolute values of all the components.

The restraint set Ω contains the vertices $\bar{u}_1, \cdots, \bar{u}_{m+1}$ (which we can assume to have norms less than $\epsilon_0 > 0$) of the m-simplex $\bar{W}$ and so the convex hull $H(\Omega)$ contains all of $\bar{W}$. For the linear process $\mathcal{L}$ the set of attainability K at $t = 1$, for solution initiating at the origin, with controllers in $\bar{W}$, is a convex set which contains $x = 0$ in its interior. By the linear bang-bang principle every point of K can be attained by responses of $\mathcal{L}$ to controllers which assume only the $m + 1$ values at the vertices of $\bar{W}$. Let $\bar{u}_1(t), \cdots, \bar{u}_{n+1}(t)$ be such controllers whose corresponding linear responses

$$\bar{x}_{Li}(t) = e^{At} \int_0^t e^{-As} B \bar{u}_i(s)\, ds, \qquad i = 1, \cdots, n + 1,$$

determine the vertices $\bar{x}_{L1}(1), \cdots, \bar{x}_{L,n+1}(1)$ of an n-simplex $\bar{S}$ centered at $x = 0$. Denote the inscribed and circumscribed radii of $\bar{S}$ by $c_1 > 0$ and $c_2 > 0$, respectively.

Take barycentric coordinates $\alpha = (\alpha_1, \cdots, \alpha_{n+1})$ in $\bar{S}$ and use the lemma to obtain a continuous family of $(n + 1)$-partitions of the time interval $\mathcal{I} = [0 \leqq t \leqq 1]$ for the functions

$$h_i(t) = e^{-At} B \bar{u}_i(t), \qquad i = 1, \cdots, n + 1,$$

so that $h(t, \alpha) = h_i(t)$ for $t \in A_i(\alpha)$, $i = 1, \cdots, n + 1$, satisfies the convexity condition

$$\int_0^1 h(t, \alpha)\, dt = \alpha_1 \int_0^1 h_1(t)\, dt + \cdots + \alpha_{n+1} \int_0^1 h_{n+1}(t)\, dt.$$

But this means that the controller family $\bar{u}(t, \alpha) = \bar{u}_i(t)$ for $t \in A_i(\alpha)$, $i = 1, \cdots, n + 1$, determines linear responses $\bar{x}_L(t, \alpha)$ with

$$\bar{x}_L(1, \alpha) = \alpha_1 \bar{x}_{L1}(1) + \cdots + \alpha_{n+1} \bar{x}_{L,n+1}(1).$$

Therefore the map of $\bar{S}$ into R^n defined by the linear responses

$$\alpha \to \bar{x}_L(1, \alpha)$$

is the identity map on $\bar{S}$.

Now repeat this entire construction with W, having vertices $u_1 = \epsilon\bar{u}_1$, $\cdots$, $u_{m+1} = \epsilon\bar{u}_{m+1}$, replacing $\bar{W}$ as the restraint simplex, for a suitably small $\epsilon > 0$. We use the control family $u(t, \alpha) = \epsilon\bar{u}(t, \alpha)$ to obtain the linear responses $x_L(t, \alpha) = \epsilon\bar{x}_L(t, \alpha)$. Then, if α are the barycentric coordinates of $S = \epsilon\bar{S}$ we find that $\alpha \to x_L(1, \alpha)$ is the identity map of S onto itself.

We compare the map $\alpha \to x_L(1, \alpha)$ with the map by the nonlinear responses $x(t, \alpha)$ of $\mathcal{S}$ with controllers $u(t, \alpha)$. Clearly the map $\alpha \to x(1, \alpha)$ is continuous on S and we show that it approximates the identity map closely on the boundary of S.

For the required estimates we restrict $\epsilon > 0$ so

$$| u(t, \alpha) | < \epsilon < \epsilon_0$$

and

$$| x(t, \alpha) | + | x_L(t, \alpha) | < c(\epsilon),$$

on $0 \leqq t \leqq 1$, lie in a region wherein

$$| f(x_L(t, \alpha), u(t, \alpha)) - Ax_L(t, \alpha) - Bu(t, \alpha) | \leqq c_3 | x_L(t, \alpha) | + c_3 | u(t, \alpha) |$$

and

$$\left| \frac{\partial f}{\partial x}(x, u) \right| \leqq | A | + 1,$$

where $c_3 = c_3(\epsilon)$ is determined (explicitly, below) in terms of the constants $| A |, | B |, c_1, c_2$, and ϵ. Let us bow our heads and compute

$$\begin{aligned} | x(t, \alpha) - x_L(t, \alpha) | &\leqq \int_0^t | f(x(s, \alpha), u(s, \alpha)) - Ax_L(s, \alpha) - Bu(s, \alpha) | \, ds \\ &\leqq \int_0^t | f(x(s, \alpha), u(s, \alpha)) - f(x_L(s, \alpha), u(s, \alpha)) | \, ds \\ &\quad + \int_0^t | f(x_L(s, \alpha), u(s, \alpha)) - Ax_L(s, \alpha) - Bu(s, \alpha) | \, ds, \end{aligned}$$

so

$$| x(t, \alpha) - x_L(t, \alpha) | \leqq (| A | + 1) \int_0^t | x(s, \alpha) - x_L(s, \alpha) | \, ds + \int_0^t 2c_3 \epsilon \, ds.$$

But this implies that (using standard inequalities)

$$|x(1, \alpha) - x_L(1, \alpha)| \leqq 2c_3 \epsilon \frac{e^{|A|+1} - 1}{|A| + 1}.$$

Now we can choose $\epsilon > 0$ so small that

$$c_3(\epsilon) = \frac{c_1}{4}\left[\frac{e^{|A|+1} - 1}{|A| + 1}\right]^{-1},$$

and then

$$|x(1, \alpha) - x_L(1, \alpha)| \leqq \frac{c_1}{2}\epsilon.$$

Thus the Euclidean norm of $x(1, \alpha) - x_L(1, \alpha)$ is less than $(c_1/2)\epsilon$ for $\alpha \in S$.

But for α on the boundary of S the Euclidean norm of $x_L(1, \alpha)$ is not less than ϵc_1. By a simple index argument (or else the Brouwer fixed-point theorem), we conclude that the image of S, by the nonlinear response $x(1, \alpha)$, covers an open ball neighborhood $\mathcal{C}_0$ of the origin in R^n.

Consider this entire construction for the system

$\hat{\mathcal{S}}$) $$\dot{x} = \hat{f}(x, u) = -f(x, u),$$

which satisfies the same hypotheses as does $\mathcal{S}$. Let $\hat{\mathcal{C}}_0$ be the corresponding open ball neighborhood of the origin covered by responses of $\hat{\mathcal{S}}$. If $\hat{u}(t)$ steers $\hat{x}(t)$ from $\hat{x}(0) = 0$ to some point $\hat{x}(1)$ in $\hat{\mathcal{C}}_0$, then $u(t) = \hat{u}(1 - t)$ steers $x(t) = \hat{x}(1 - t)$ by $\mathcal{S}$ from $x(0) = \hat{x}(1)$ to $x(1) = 0$. Thus the domain of null controllability $\mathcal{C}$ for $\mathcal{S}$ contains the open set $\hat{\mathcal{C}}_0$ and $\mathcal{C}$ is an open neighborhood of the origin in R^n.

Remark. Theorem 3 holds if $\epsilon u_1, \cdots, \epsilon u_{m+1}$ lie in Ω only for certain suitably small $\epsilon > 0$; thus Ω could be a finite point set. Of course, if Ω contains $u = 0$ in its interior, then the hypotheses of the theorem are verified. It should further be noted that the small simplices in Ω need not be strictly homothetic to a fixed simplex as long as these very small simplices do not become too flattened. The simplex in Ω could be replaced by any convex polytope about $u = 0$.

The following example shows that the algebraic condition of controllability is not necessary for the geometric property of controllability for nonlinear processes.

Example. In R^2 consider the process

$\mathcal{S}$) $$\dot{x} = y^3, \qquad \dot{y} = -x + u,$$

with restraint $\Omega: -1 \leqq u \leqq 1$ in R^1. For $u \equiv 1$ we have the family of responses along the curves

$$\frac{y^4}{4} + \frac{(x - 1)^2}{2} = \text{const.},$$

and for $u \equiv -1$ we have the responses along the curves

$$\frac{y^4}{4} + \frac{(x+1)^2}{2} = \text{const.}$$

Now the topological map of R^2 onto itself defined by

$$x \to \xi = x, \quad y \to \eta = y^2 \operatorname{sgn} y,$$

carries the above families of ovals onto the extremal solution curves for the linear controllable system

$$\dot{\xi} = \eta, \quad \dot{\eta} = -\xi + u, \qquad -1 \leqq u \leqq 1.$$

Thus every initial state (x, y) in R^2 can be steered to the origin by solutions of $\mathcal{S}$ with controllers in Ω. However the controllability matrix for $\mathcal{S}$ at the origin yields

$$[B, AB], \quad \text{where} \quad A = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

which has rank one.

2. Geometry of optimal control. The existence of an optimal controller for a control process

$$\mathcal{S}) \qquad \dot{x} = f(x, u),$$

with $f \in C^1$ in R^{n+m}, is usually presented as a consequence of the closedness of the appropriate set of attainability [7]. The following example fails to have a closed set of attainability and the time-optimal controller fails to exist.

Example. Consider first the process in R^3,

$$\mathcal{S}) \qquad \dot{x} = \sin 2\pi u, \qquad \dot{y} = \cos 2\pi u, \qquad \dot{z} = -1,$$

with restraint $\Omega: -1 \leqq u \leqq 1$ in R^1. We wish to steer $(0, 0, 1)$ to $(0, 0, 0)$ in minimal cost

$$C(u) = \int_0^{t_1} (x^2 + y^2 + 1)\, dt.$$

Since $z(t) = 1 - t$, we can assume $t_1 = 1$ for all admissible controllers $u(t) \subset \Omega$ on $0 \leqq t \leqq 1$.

For each integer $k = 1, 2, 3, \cdots$, define a piecewise continuous controller $u^{(k)}(t)$ on $0 \leqq t \leqq 1$ so that

$$\sin 2\pi u^{(k)}(t) = \sin 2\pi k t$$

and

$$\cos 2\pi u^{(k)}(t) = \cos 2\pi kt.$$

The corresponding responses to $\mathcal{S}$ are

$$x^{(k)}(t) = \frac{1 - \cos 2\pi kt}{2\pi k}, \quad y^{(k)}(t) = \frac{\sin 2\pi kt}{2\pi k}, \quad z^{(k)}(t) = 1 - t,$$

which steer $(0, 0, 1)$ to the origin $(0, 0, 0)$. The cost is easily computed to be

$$C(u^{(k)}) = 1 + \frac{1}{2\pi^2 k^2}.$$

Thus the infimum of all costs is $m = 1$; yet this cannot be attained unless

$$\int_0^1 [x(t)^2 + y(t)^2]\, dt = 0$$

or $x(t) \equiv y(t) \equiv 0$. But this is impossible since $\dot{x}^2 + \dot{y}^2 = 1$ (almost always). Thus no optimal controller exists for the process $\mathcal{S}$. We note that the set of attainability K for the extended process

$$\dot{x} = \sin 2\pi u, \qquad \dot{y} = \cos 2\pi u, \qquad \dot{z} = -1, \qquad \dot{\tau} = x^2 + y^2 + 1,$$

from the initial state $(0, 0, 1, 0)$, is not closed at time $t = 1$, since K does not contain the limit point $(0, 0, 0, 1)$.

We can further modify the example to yield a time-optimal control problem which has no optimal controller. Consider in R^3,

$$\hat{\mathcal{S}}) \qquad \frac{dx}{d\tau} = \frac{\sin 2\pi u}{x^2 + y^2 + 1}, \quad \frac{dy}{d\tau} = \frac{\cos 2\pi u}{x^2 + y^2 + 1}, \quad \frac{dz}{d\tau} = \frac{-1}{x^2 + y^2 + 1},$$

with restraint $\Omega\colon -1 \leqq u \leqq 1$ in R^1. We wish to steer $(0, 0, 1)$ to $(0, 0, 0)$ in minimal time $\tau^* > 0$. It is easy to see that the infimum of all times τ_1 required to steer $(0, 0, 1)$ to $(0, 0, 0)$ by $\hat{\mathcal{S}}$ is $m = 1$; yet this cannot be achieved and no optimal controller exists for $\hat{\mathcal{S}}$.

Now consider a control process in R^n,

$$\mathcal{S}) \qquad \dot{x} = f(x, u),$$

with $f \in C^1$ in R^{n+m} and restraint set $\Omega \subset R^m$. Fix an initial state $x_0 \in R^n$ and consider all bounded measurable controllers $u(t)$ on $0 \leqq t \leqq t_1$ which produce responses $x(t)$ on $0 \leqq t \leqq t_1$, $x(0) = x_0$, by the process $\mathcal{S}$. The set of all endpoints $x(t_1)$ of all such responses is the set of attainability $K(t_1)$ for $\mathcal{S}$ from x_0 by admissible controllers $u(t) \subset \Omega$ on $0 \leqq t \leqq t_1$.

A controller $\bar{u}(t)$ on $0 \leqq t \leqq t_1$ with response $\bar{x}(t)$ leading to the boundary of $K(t_1)$ must necessarily satisfy the maximal principle [3]. That is, there exists a nontrivial solution $\bar{\eta}(t)$ of the adjoint variational system

$$\dot{\eta} = -\eta \frac{\partial f}{\partial x}(\bar{x}(t), \bar{u}(t))$$

such that

$$\bar{\eta}(t)f(\bar{x}(t), \bar{u}(t)) = \max_{u \in \Omega} \bar{\eta}(t)f(\bar{x}(t), u)$$

almost always on $0 \leqq t \leqq t_1$. In particular, the optimal controller $u^*(t)$ on $0 \leqq t \leqq t^*$ (for the augmented system where the cost is treated as a new spatial coordinate) must steer to the boundary of $K(t^*)$ and hence must satisfy the corresponding maximal principle.

For linear processes

$$\mathfrak{L}) \qquad \dot{x} = Ax + Bu,$$

the maximal principle is both necessary and sufficient that $\bar{u}(t)$ on $0 \leqq t \leqq t_1$ steers the response $\bar{x}(t)$ to the boundary of $K(t_1)$. However for nonlinear processes the maximal principle for $\bar{u}(t)$ does not imply that $\bar{x}(t_1)$ lies on the boundary of $K(t_1)$. The following example illustrates this phenomenon for nonlinear systems.

Example. Consider in R^2,

$$\dot{x} = yu - xv, \qquad \dot{y} = -xu - yv,$$

with control restraints $|u| \leqq 1$, $|v| \leqq 1$ in R^2. In polar coordinates the differential system becomes

$$\dot{r} = -rv(t), \qquad \dot{\varphi} = -u(t).$$

Take the initial point $r_0 = 1$, $\varphi_0 = 0$, and study controllers on the duration $0 \leqq t \leqq \pi$. The control functions $u(t)$ and $v(t)$ enter independently in the angular and radial velocities. Thus it is easy to compute the set of attainability $K(\pi)$ as the annular ring

$$K(\pi): \qquad e^{-\pi} \leqq r \leqq e^{\pi}, \qquad 0 \leqq \varphi \leqq 2\pi.$$

Here $K(\pi)$ is compact but it is not convex, nor even simply connected. The controllers $u(t) \equiv +1$, $v(t) \equiv 0$, and also $u(t) \equiv -1$, $v(t) \equiv 0$, each satisfy the maximal principle but they steer responses to the point $r = 1$, $\varphi = \pi$, which lies interior to $K(\pi)$.

We can modify this example to obtain a more complicated geometry for $K(\pi)$. Consider the process in R^2 described in polar coordinates by the system

$$\dot{r} = -rv(t)h(\varphi),$$

$$\dot{\varphi} = -u(t)\left[1 - \left(\frac{R - r}{2R}\right)^4 \left(\sin^2 \frac{1}{R - r}\right) h(\pi - \varphi)\right],$$

where $h(\varphi)$ with period 2π is a function in C^∞ satisfying the conditions

$$h(\varphi) = h(-\varphi), \qquad 0 \leqq h(\varphi) \leqq 1,$$

$$h(\varphi) = 0 \quad \text{on} \quad \frac{\pi}{2} \leqq \varphi \leqq \pi,$$

$$h(\varphi) \equiv 1 \quad \text{for} \quad \varphi \quad \text{near} \quad 0.$$

Also the constant R is specified by

$$R = \exp \int_0^{\pi/2} h(\varphi)\, d\varphi.$$

The restraints are $|u| \leqq 1$, $|v| \leqq 1$, with initial point $r_0 = 1$, $\varphi_0 = 0$, as before.

At $t = \pi/2$ the set $K(\pi/2)$ meets the ray $\varphi = \pi/2$ only for $u(t) \equiv -1$, so $\dot{\varphi} = 1$. In this case $\dot{r} = -rv(t)h(t)$ and the segment $\varphi = \pi/2$, $1/R \leqq r \leqq R$ is an edge of $K(\pi/2)$. Similarly the segment $\varphi = -\pi/2$, $1/R \leqq r \leqq R$ in $K(\pi/2)$ can be attained only for $u(t) \equiv +1$. Thus $K(\pi/2)$ is a half-annular ring, with radial width $(e^{\pi/2} - e^{-\pi/2})$ at $\varphi = 0$ and tapering to a width of $(R - 1/R)$ at $\varphi = \pm\pi/2$.

Now consider $K(\pi)$ for this system. Only some of the points on the rays $\varphi = \pm\pi/2$ at time $t = \pi/2$ can reach the ray $\varphi = \pi$ at time $t = \pi$. In the left half plane the differential system is

$$\dot{r} = 0, \qquad \dot{\varphi} = -u(t)\left[1 - \left(\frac{R - r}{2R}\right)^4 \left(\sin^2 \frac{1}{R - r}\right) h(\pi - \varphi)\right].$$

Thus the intersection of $K(\pi)$ with the ray $\varphi = \pi$ occurs only at radii satisfying

$$(R - r)^4 \sin^2 \frac{1}{R - r} = 0,$$

that is, at a countable set of points accumulating at $\varphi = \pi$, $r = R$.

This analysis shows that $K(\pi)$ consists of an annular-like region, with radial width tapering to a minimum of $(R - 1/R)$ at $\varphi = \pi$, and with an infinite number of disjoint open regions excised along the ray $\varphi = \pi$. Thus $K(\pi)$ is infinitely connected and its boundary cannot be described by a finite number of simple closed continuous curves.

REFERENCES

[1] D. Blackwell, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.

[2] W. Hahn, *Theory and Application of Lyapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[3] H. Halkin, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 1964, pp. 1–82.

[4] P. R. HALMOS, *The range of a vector measure*, Bull. Amer. Math. Soc., 54 (1948), pp. 416–421.

[5] R. E. KALMAN, Y. C. HO, AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[6] J. P. LASALLE, *The time optimal control problem*, Contributions to Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.

[7] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[8] L. MARKUS AND H. YAMABE, *Global stability criteria for differential systems*, Osaka Math. J., 12 (1960), pp. 305–317.

[9] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

MINIMAX PROBLEMS AND UNILATERAL CURVES IN THE CALCULUS OF VARIATIONS*

J. WARGA†

1. Introduction. Let

$$\dot{x} = \frac{dx}{dt} = g(x, t, \rho(t)) \qquad \text{a.e. in } T = [t_0, t_1], \tag{1.1}$$
$$x(t_0) \in B_0, \qquad x(t_1) \in B_1,$$

be a system of ordinary differential equations and boundary conditions, where $x = (x^1, \cdots, x^n) \in E_n$, $g(x, t, \rho)$ is a function from $E_n \times T \times R$ to E_n, B_0 and B_1 are specified closed sets in E_n, and the "control" function $\rho(t)$ has its range in a specified set R. The problem of determining a control function $\rho(t)$ and an initial point $x(t_0)$ that yield the minimum of $x^1(t_1)$, subject to (1.1), is one of the first and most intensively studied subjects of the mathematical control theory and serves as a point of departure for related investigations.

We are concerned here with two generalizations of this problem which have been considered in recent years. *Unilateral problems* arise when the integral curves $x(t)$ are restricted to some preassigned set. Pioneering work in this field was done by Gamkrelidze [2], [3] who considered the additional side condition $a(x(t)) \leqq 0$, $t \in T$. Gamkrelidze succeeded in deriving necessary conditions for minimum based on the a priori assumption that there exists a minimizing unilateral curve with a finite number of "corners" and satisfying certain "regularity" conditions. Similar results were later derived by Berkovitz [1] who applied results of the classical calculus of variations.

Minimax problems are another generalization of the fundamental problem of the mathematical control theory. The *pursuit problem*, investigated by Kelendzheridze [4], [5], [6], belongs to that class. This problem deals with the control of a pursuing point ξ and a pursued point η; these points satisfy the relations

$$\dot{\xi} = h_1(\xi, u), \qquad \dot{\eta} = h_2(\eta, v) \qquad \text{a.e. in } T, \tag{1.2}$$
$$\xi(0) = \xi_0, \qquad \eta(0) = \eta_0,$$

* Received by the editors March 29, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Research and Advanced Development Division, AVCO Corporation, 201 Lowell Street, Wilmington, Massachusetts.

where $u(t)\colon T \to U$ and $v(t)\colon T \to V$ are "conflicting" controls. It is assumed that for every choice of an admissible control $v(t)$ there exists a control $u(t)$ such that the corresponding solutions $\xi(t)$ and $\eta(t)$ of (1.2) have a point in common, say $\xi(\bar{t}) = \eta(\bar{t})$. Let $\bar{t}_{u,v}$ represent the smallest value of $\bar{t}$ such that $\xi(\bar{t}) = \eta(\bar{t})$. Then the solution of the pursuit problem consists in determining $t^* = \max_v \min_u \bar{t}_{u,v}$ and controls $\bar{u}(t)$ and $\bar{v}(t)$ such that $t^* = \min_u \bar{t}_{u,\bar{v}} = \bar{t}_{\bar{u},\bar{v}}$. Kelendzheridze [4] derived necessary conditions for minimax in the special case where the pursuing point ξ obeys a linear law $\dot{\xi} = h_1(\xi, u) = A\xi + Bu + c$. His attempt [5] to generalize these results without the linearity assumption seems to be affected by an apparent logical mistake (in incorrectly applying a dynamic programming principle to the problem). His results, in fact, cannot be generalized to the nonlinear case as can be demonstrated by a rather simple counterexample (see §4).

Certain general results applicable to unilateral and minimax problems, and obtained during 1962 and 1963, are described in references [9], [10] and [11]. In [7], the unilateral problems were considered without making any a priori assumptions about the minimizing curves. It was shown there that such problems admit a "relaxed" solution and a procedure was indicated for uniformly approximating this relaxed solution with solutions of (1.1). In [9], necessary conditions for relaxed minimum are derived for problems in which the minimizing curves are restricted by the single inequality $a(x(t)) \leqq 0$, $t \in T$. These conditions generalize Gamkrelidze's results; they reduce to a form equivalent to Gamkrelidze's when certain a priori verifiable relations are satisfied (see Appendix A, A2.8). In [11], these results are generalized to problems in which the unilateral curves are restricted by the simultaneous inequalities $a^k(x(t)) \leqq 0$, $t \in T$, $k = 1, \cdots, m$.

The minimax problems studied in [10] can be described by the system

$$\begin{aligned} &\dot{x} = g(x, t, p, \rho(t)) \qquad \text{a.e. in } T, \\ &x(t_0, p) = b_0 \in B_0, \qquad x(t_1, p) \in B_1, \qquad p \in P. \end{aligned} \tag{1.3}$$

Here p is a parameter with values in a metric space P and $x(t, p)$ represents a solution for a fixed p. The problem consists in determining a control $\rho(t)$ that minimizes $\max_{p \in P} x^1(t, p)$. It is proved that all such problems belonging to a rather general class admit an optimal "relaxed" (or "generalized") control and that this relaxed control can be "simulated" by functions $\rho(t)$ from T to R. Furthermore, necessary conditions for relaxed minimax are derived; these conditions generalize the Weierstrass E-condition and the transversality conditions.

In §§2 and 3 we shall describe and discuss some typical problems, or classes of problems, to which the above results are applicable. In §4 we

shall discuss the pursuit problem. Appendices A and B contain the assumptions and the statements of the basic theorems proven in [11] and [10], respectively.

2. A generalized unilateral problem. As in §1, we consider the system

$$\begin{aligned} &\dot{x} = g(x, t, \rho) \qquad \text{a.e. in } T = [t_0, t_1], \\ &x(t_0) \in B_0, \qquad x(t_1) \in B_1, \\ &a^k(x(t)) \leqq 0, \qquad t \in T, \qquad k = 1, \cdots, m. \end{aligned} \tag{2.1}$$

The previously mentioned unilateral problems consist in minimizing $x^1(t_1)$ (or some other function of $x(t_1)$). We may also consider a related minimax problem.

Let $\phi(x)$ be a scalar function defined on some open set V containing $A = \{x \in E_n \mid a^k(x) \leqq 0, k = 1, \cdots, m\}$, and assume that $\phi(x)$ has continuous first- and second-order partial derivatives on V. The problem of minimizing $\max_{t \in T} \phi(x(t))$, subject to (2.1), can be easily reduced to an ordinary unilateral problem. Indeed, we can adjoin to the system (2.1) the relations

$$\begin{aligned} &\dot{s} = 0 \qquad \text{a.e. in } T, \\ &a^{m+1}(x(t)) = \phi(x(t)) - s \leqq 0, \quad t \in T. \end{aligned}$$

Since $s(t_1) = s$ is an upper bound of $\phi(x(t))$ on T, the greatest lower bound of $s(t_1)$ over all admissible choices of controls $\rho(t)$ and of boundary conditions will be identical with $\inf \max_{t \in T} \phi(x(t))$.

A very simple problem of this kind is discussed in [9, p. 437]. Assume that a train has to cover a unit distance in a unit time between two consecutive stops, and assume further that its acceleration can be varied at will between $-\alpha$ and $+\alpha$ (where $\alpha > 4$). The problem consists in modulating the acceleration in such a manner that the maximum speed attained is as low as possible.

More complicated problems of this kind arise in the study of the modulation of the lift and drag controls of a space vehicle reentering the Earth atmosphere. In a simple model of such reentry, the state of the vehicle is represented by its distance h to the center of the Earth and by the magnitude v and the inclination angle γ of its velocity vector. The instantaneous heat flux due to atmospheric heating is assumed to be a specified function of v and h. In one application, the initial conditions are specified and it is desired to modulate the lift and drag coefficients of the vehicle in such a manner as to minimize the maximum (with respect to time) of the heat flux. In another application, the initial values of h and v are specified and it is desired to determine the extreme values of the initial angle γ for which

the scalar acceleration can be kept during the descent below a preassigned limit. The first of these two related applications is a minimax problem of the form min $\max_{t\in T}\phi(x(t))$, and the second one is a conventional unilateral problem.

The solutions to both of these problems exhibit a property common to many unilateral problems. The minimizing curves are uniquely determined during a certain portion of the time interval $[t_0, t_1]$, but they can be rather arbitrarily chosen during the remaining time. We can illustrate this type of behavior with a very simple example.

Let a point (t, x) in a plane be initially at $(0, a)$, where $1 < a < \sqrt{2}$. Assume that dx/dt can be arbitrarily chosen between -1 and $+1$ for all t. The point (t, x) must remain on the outside, or on the boundary, of the unit circle with its center at $(0, 0)$. The problem consists in determining $x(t)$, subject to the above restrictions, so as to minimize $x(2)$.

This minimum can be achieved as follows: we choose any curve $(t, x(t))$ joining $(0, a)$ to $(\sqrt{2}/2, \sqrt{2}/2)$, remaining outside, or on the boundary, of the unit circle, and whose slope dx/dt remains between -1 and 1; for $t > \sqrt{2}/2$, we choose the straight line $x(t) = \sqrt{2} - t$.

3. Minimax problems. Next we consider minimax problems described by system (1.3). Such problems may arise, in particular, when certain parameters are not completely specified, or when they are subject to unpredictable variations from case to case. This is indeed the situation in most practical applications.

Consider, as an example, a chemical reaction of fixed duration which is used to produce a particular chemical substance. The reaction is described by a system of ordinary differential equations involving time dependent concentrations of reactants and their derivatives, as well as the time dependent fuel flow and certain fuel parameters. The reaction is controlled by varying, with time, the flow of the fuel into the furnace. The fuel parameters are known within certain limits only. A maximin problem will arise if a decision is made to control the fuel flow in such a manner as to maximize the guaranteed yield (for all possible values of the fuel parameters) of the desired chemical substance.

Another application is of the following type: we wish to minimize $y^1(t_1)$ among all the solutions of the system $\dot{y} = g(y, t, p_0, \rho)$, $y(t_0) \in B_0$, $y(t_1) \in B_1{}^*$, where p_0 is a fixed value of some parameter and $\rho(t)\colon T \to R$ and $y(t_0)$ are further restricted by the condition $x(t_1, p) \in B_1$, $p \in P$. Here $x(t, p)$ is a solution of the system $\dot{x} = g(x, t, p, \rho)$ and $x(t_0, p) = y(t_0)$. Such a problem would arise if we had a good estimate of the parameter p_0 but wanted to insure that the system will reach the "safe" region B_1 in case the parameters undergo an unforeseen change.

Certain variants of the problems described by (1.3) can be reduced to the same form as (1.3), or can be handled by the same methods. Thus, the initial condition $x(t_0, p) = b_0$ (b_0 independent of p) can be replaced by the condition $x(t_0, p) = b_0(p)$, where $b_0(p)$ is a specified function, or by the condition $x(t_0, p) \in B_0$. Problems of variable duration, or time minimizing problems, can also be reduced to the standard form.

The necessary conditions for minimax of Appendix B become greatly simplified when the minimax problem satisfies two special conditions: (a) the set B_1 is the entire Euclidean space E_n, that is, $x(t_1, p)$ is unrestricted; and (b) for every admissible choice of the relaxed control $\sigma(t)$: $T \to S$ and of the initial point b_0 there exists a unique point $\bar{p}(\sigma, b_0)$ in P that maximizes $x^1(t_1, p)$. Assume that these two conditions are satisfied, let $\bar{\sigma}(t)$ and $\bar{b}_0$ minimize $\max_{p \in P} x^1(t_1, p)$, and let $p^* = \bar{p}(\bar{\sigma}, \bar{b}_0)$. Then the necessary conditions for minimax of Appendix B are identical with the customary necessary conditions ([8, Theorem 6.1, p. 142], essentially the Weierstrass E-condition and the transversality conditions) that would be obtained by assuming that $\bar{\sigma}(t)$ and $\bar{b}_0$ minimize $x^1(t_1 ; p^*)$. Formally, these conditions can be derived *as if* p^* yielded $\max_p x^1(t_1 ; p, \bar{\sigma}(t), \bar{b}_0)$ and $\bar{\sigma}$ and $\bar{b}_0$ yielded $\min_{\sigma, b_0} x^1(t_1 ; p^*, \sigma, b_0)$, where $x(t; p, \sigma, b_0)$ is the relaxed solution of (1.3) corresponding to the parameter p, the initial point b_0, and the relaxed control σ: $T \to S$.

4. The pursuit problem. Necessary conditions for optimal pursuit were derived by Kelendzheridze in [4] for the case $h_1(\xi, u) = A\xi + Bu + c$, and were conjectured by him in [5] for nonlinear $h_1(\xi, u)$. These necessary conditions are quite strong; they imply that an optimal trajectory of the pursuing point ξ is an extremal of the system $\dot{\xi} = h_1(\xi, u)$, and an optimal trajectory of the pursued point η is an extremal of the system $\dot{\eta} = h_2(\eta, v)$.

The first assertion (that ξ follows an extremal) can be easily seen to hold for a wide class of pursuit problems, even if $h_1(\xi, u)$ is nonlinear. Indeed, the control v of the pursued object η becomes known to the pursuer as soon as it is chosen. Consequently, the trajectory $\bar{\eta}(t)$ that will be followed by η is known to the pursuer ξ from the very beginning, and the problem of determining u becomes a "standard" control problem of a minimum-time transfer of $(0, \xi_0)$ to the set $\{(t, \bar{\eta}(t)) \mid t \geqq 0\}$. It is well known that for a rather general class of problems such a minimum-time transfer requires that ξ follow an extremal of the system $\dot{\xi} = h_1(\xi, u)$.

The situation is quite different for the pursued object η. Having chosen an escape trajectory, which then becomes immediately known to the pursuer, η must face pursuit along any trajectory available to the pursuer. In some cases (as, e.g., when $h_1(\xi, u)$ is linear) only one pursuit trajectory can lead to a minimum-time capture of η for any given choice of the escape

trajectory. Having "communicated" his strategy to the pursuer, the pursued can in turn predict the counterstrategy of the opponent. Under such circumstances, escape along an extremal is often the best policy for the pursued object, as has been shown by Kelendzheridze for linear $h_1(\xi, u)$. The reader may observe the similarity between this result and the necessary conditions for minimax as described in the last paragraph of §3.

In many other cases, however, this result is not applicable. We shall describe a simple counterexample: let ξ and η be points in a plane, and let $\xi(0) = (2, 0)$ and $\eta(0) = (0, 0)$. Let $p(r)$ be a differentiable (even analytic) increasing function of a nonnegative variable r such that $dp(0)/dr = 0$, $p(0) = 1.01$, $p(1.9) = 1.02$, and $p(2) = 100$. Let admissible controls $u(t) = (u^1(t), u^2(t))$ and $v(t) = (v^1(t), v^2(t))$ be such that $|u(t)|^2 = (u^1(t))^2 + (u^2(t))^2 \leqq 1$ and $|v(t)|^2 = (v^1(t))^2 + (v^2(t))^2 \leqq 1$. Finally, let

$$\dot{\xi} = p(|\xi|)u, \quad \dot{\eta} = v.$$

We can easily verify that all of Kelendzheridze's assumptions are satisfied, including the existence of minimax solutions.

If η follows an extremal of the system $\dot{\eta} = v$, then its trajectory is a straight line, and $|\eta(t)| = t$. Let ϕ_0 be the angle that the trajectory of η forms with the positive η^1-axis. Then η can be captured at a time $t_1 \leqq 1.03$. Indeed, ξ can first move at its maximum speed along the circle with center $(0, 0)$ and radius 2 and it will reach the point $s = (2 \cos \phi_0, 2 \sin \phi_0)$ at $t' = .02\phi_0 \leqq .02\pi$. Then ξ can move radially to meet η and it will capture η before

$$t = \frac{2}{2.01} + \frac{1.01}{2.01}\frac{2\pi}{100} \sim 1.03.$$

If, on the other hand, η remains at $(0, 0)$ for all t ($|u(t)| \equiv 0$), then it will escape capture until $t_1^* > 1.86$, since $|\dot{\xi}| = p(|\xi|) \leqq 1.02$ for $|\xi| \leqq 1.9$.

This example shows that escape along an extremal is not always a good policy. A man standing in the middle of a circular marsh and trying to escape a distant car will do better by staying inside the marsh than by running away to firm ground.

Necessary conditions for an optimal escape strategy must, therefore, be much more complicated in general for nonlinear $h_1(\xi, u)$ than suggested by Kelendzheridze's results. We conjecture that the latter results still apply, however, in many cases where the "attainable set" of ξ is simply connected for all t. (The attainable set of ξ at time $\bar{t}$ is the set $\{\xi(\bar{t}; u) \mid u(t): T \to U\}$, where $\xi(t, u)$ is the solution of the system $\dot{\xi} = h_1(\xi, u)$, $\xi(0) = \xi_0$, corresponding to a given control $u(t)$.)

APPENDIX A

Statement of the problem and assumptions. Let R be a compact Hausdorff space, E_n the Euclidean n-space, T the closed interval $[t_0, t_1]$ of the real axis, V an open set in E_n, and B_0 and B_1 closed sets in V. We are also given a function $g(x, t, \rho) = (g^1(x, t, \rho), \cdots, g^n(x, t, \rho))$ from $V \times T \times R$ to E_n and a function $a(x) = (a^1(x), \cdots, a^m(x))$ from V to E_m.

Let $G(x, t) = \{g(x, t, \rho) \mid \rho \in R\}$, $x \in V$, $t \in T$, and let $F(x, t)$ be the convex closure of $G(x, t)$.

1. *Definition.* We define an *original admissible curve with respect to* $a(x)$ as any absolutely continuous function $x(t)$ from T to V such that, for some function $\rho(t)$ from T to R,

$$\frac{dx(t)}{dt} = \dot{x}(t) = g(x(t), t, \rho(t)) \qquad \text{a.e. in } T \tag{1.1}$$

or, equivalently,

$$\dot{x}(t) \in G(x(t), t) \qquad \text{a.e. in } T, \tag{1.1 Original}$$

and

$$x(t_0) \in B_0, \qquad x(t_1) \in B_1, \tag{1.2}$$

$$a^k(x(t)) \leqq 0, \quad k = 1, 2, \cdots, m, \quad t \in T. \tag{1.3}$$

We similarly define a *relaxed admissible curve with respect to* $a(x)$ except that (1.1), respectively (1.1 Original), is replaced by

$$\dot{x}(t) \in F(x(t), t) \qquad \text{a.e. in } T. \tag{1.1 Relaxed}$$

An *original* (respectively, *relaxed*) *minimizing curve with respect to* $a(x)$ is a curve that minimizes the value $x^1(t_1)$ among all original (respectively, relaxed) admissible curves with respect to $a(x)$.

We now state our basic assumptions.

2. *Assumptions.* There exist a finite or denumerable collection of disjoint (Lebesgue) measurable subsets T_r, $r = 1, 2, \cdots$, of T whose union T' has measure $|T| = t_1 - t_0$, positive constants c_1 and ϵ_1, a function $\epsilon(h)$, $h > 0$, converging to 0 as $h \to +0$, and a compact set $D \subset V$ such that the following five conditions are satisfied.

(2.1) The functions

$$g^i(x, t, \rho) \quad \text{and} \quad \frac{\partial g^i(x, t, \rho)}{\partial x^j}, \qquad i, j = 1, \cdots, n,$$

exist over $V \times T' \times R$, and over that set they are continuous functions of

(x, t), uniformly in ρ, and continuous functions of ρ for each (x, t); furthermore,

$$\| g(x, t, \rho) - g(x, t', \rho) \| \leqq \epsilon(| t - t' |),$$

provided t and t' belong to the same set T_r, where

$$\| g \| = \| (g^1, \cdots, g^n) \| = \sum_{i=1}^{n} | g^i |.$$

(2.2) $\| g(x, t, \rho) \| \leqq c_1$ and $\| g_x(x, t, \rho) \| \leqq c_1$ on $V \times T' \times R$;

here g_x is the matrix $(\partial g^i/\partial x^j)$, $i, j = 1, \cdots, n$, and

$$\| g_x \| = \sum_{i,j=1}^{n} \left| \frac{\partial g^i}{\partial x^j} \right|.$$

(2.3) The functions

$$a^k(x), \quad \frac{\partial a^k(x)}{\partial x^i}, \quad \frac{\partial^2 a^k(x)}{\partial x^i \partial x^j}, \qquad k = 1, \cdots, m, \quad i, j = 1, \cdots, n,$$

exist and are continuous on V; furthermore, $\| a^k \| \leqq c_1$, $\| a_x{}^k \| \leqq c_1$, and $\| a_x{}^k g \| \leqq c_1$, $k = 1, \cdots, m$. Here $a_x{}^k$ is the gradient of a^k,

$$\| a_x{}^k \| = \sum_{j=1}^{n} \left| \frac{\partial a^k}{\partial x^j} \right|, \quad \text{and} \quad a_x{}^k g = \sum_{j=1}^{n} g^j \frac{\partial a^k}{\partial x^j}.$$

(2.4) There exists at least one relaxed admissible curve with respect to $a(x)$.

(2.5) All relaxed admissible curves with respect to $(a^1(x) - \epsilon_1, \cdots, a^m(x) - \epsilon_1)$ are contained in D.

3. *Definition.* A function $f(x, t, \sigma)$ from $V \times T \times S$ to E_n is a *proper representation of* $F(x, t)$ if

(3.1) $$F(x, t) = \{f(x, t, \sigma) \mid \sigma \in S\}, \qquad x \in V, t \in T;$$

(3.2) for every absolutely continuous curve $x(t)$ satisfying (1.1 Relaxed) there exists a function $\sigma(t)$ from T to S such that

$$\dot{x}(t) = f(x(t), t, \sigma(t)) \qquad \text{a.e. in } T$$

and $f(x, t, \sigma(\tau))$ is, for all $x \in V$ and almost all $t \in T$, a (Lebesgue) measurable function of τ on T;

(3.3) $$f^i(x, t, \sigma) \quad \text{and} \quad \frac{\partial f^i(x, t, \sigma)}{\partial x^j}, \qquad i, j = 1, \cdots, n,$$

exist and are continuous functions of (x, t) on $V \times T'$ for every σ in S;

(3.4) $\| f(x, t, \sigma) \| \leqq c_1$ and $\| f_x(x, t, \sigma) \| \leqq c_1$ on $V \times T' \times S$;

(3.5) the set

$$H(x, t, \alpha) = \{(f(x, t, \sigma), f_x{}^T(x, t, \sigma)\alpha) \mid \sigma \in S\}$$

in $E_n \times E_n$ is compact and convex for every $(x, t, \alpha) \in V \times T' \times E_n$. (Here $f_x{}^T$ is the transpose of the matrix f_x.)

4. *Definition.* Let $B \subset E_n$. We shall say that $(C, c(\xi))$ is a *proper representation of B at x* if

(4.1) C is a compact and convex set in some Euclidean space;

(4.2) $c(\xi)$ is a continuous and continuously differentiable function from C to B;

(4.3) $x = c(\xi)$ for some $\xi \in C$.

All of the conditions stated in Definitions 3 and 4 are directly verifiable, except for (3.2). We indicate, therefore, two methods of constructing proper representations of $F(x, t)$.

5. *The Filippov representation.* Let S be a compact set in some Euclidean space, and let $f(x, t, \sigma)$ be continuous on $V \times T' \times S$ and satisfy (3.1), (3.3), (3.4), and (3.5). Then (3.2) follows from a lemma of Filippov.

6. *The Young representation.* Let S be the class of probability measures defined on the Borel subsets of R, and let $f(x, t, \sigma) = \int_R g(x, t, \rho)\, d\sigma$. Then (3.1)–(3.4) follow from Assumption 2 and from [7, Theorem 4.1, p. 124]. Condition (3.5) is easily verified, since S is a convex set and $f(x, t, \sigma)$ is linear in σ.

Existence of a minimizing curve. Necessary conditions for minimum.

THEOREM A. *Let Assumption 2 be satisfied. Then there exists a curve $x(t)$ which is a relaxed minimizing curve with respect to $a(x)$, and this curve can be uniformly approximated by solutions of the differential equations* (1.1).

Let $f(x, t, \sigma)$ be a proper representation of $F(x, t)$, let $(C_i, c_i(\xi_i))$ be a proper representation of B_i at $x(t_i)$, $i = 0, 1$, and let

$$Z^k = \{t \in T \mid a^k(x(t)) = 0\}, \qquad k = 1, \cdots, m,$$

$$Z = \bigcup_{l=1}^{m} Z^l,$$

and

$$K(t) = \{k \mid a^k(x(t)) = 0\}, \quad t \in T.$$

Finally, let $\delta_i = (\delta_i{}^1, \delta_i{}^2, \cdots, \delta_i{}^n)$, $i = 1, \cdots, n$, where $\delta_i{}^j = 0$, $i \neq j$, and $\delta_i{}^i = 1$. Then either

(A.1) *there exist a point $\xi_1{}^*$ in C_1 and numbers γ^a, γ^k, $k \in K(t_1)$, such that $c_1(\xi_1{}^*) = x(t_1)$, $\gamma^a \geqq 0$, $\gamma^k \geqq 0$, $k \in K(t_1)$,*

$$\gamma^a + \sum_{l \in K(t_1)} \gamma^l \neq 0,$$

and

$$(\gamma^a\delta_1 + \sum_{l\in K(t_1)} \gamma^l a_x{}^l(x(t_1)))\cdot c_{1,\xi}(\xi_1{}^*)\xi_1{}^* = \min_{\xi_1\in C_1} (\gamma^a\delta_1 + \sum_{l\in K(t_1)} \gamma^l a_x{}^l(x(t_1)))\cdot c_{1,\xi}(\xi_1{}^*)\xi_1,$$

where $c^i_{1,\xi}$ *is the gradient of* $c_1{}^i$ *and* $c_{1,\xi} = (c^1_{1,\xi}, \cdots, c^n_{1,\xi})$, *or*

(A.2) *there exist a function* $\sigma(t)$ *from* T *to* S, *a function* $\mu(t) = (\mu^1(t), \cdots, \mu^m(t))$ *from* T *to* E_m, *a function* $z(t)$ *from* T *to* E_n, *a closed subset* M *of* Z, *points* $\xi_0{}^* \in C_0$ *and* $\xi_1{}^* \in C_1$, *and a nonnegative number* γ^1 *such that all of the following hold.*

(A.2.1) $\mu^k(t) \geqq 0$, $k = 1, \cdots, m$, *and* $\| z(t) \| + \| \mu(t) \| > 0$, $t \in T$, *where*

$$\| \mu(t) \| = \sum_{l=1}^{m} | \mu^l(t) |.$$

(A.2.2) $z(t)$ *is absolutely continuous on every closed subinterval of* $T - M$.

(A.2.3) *For every* k, $k = 1, \cdots, m$, $\mu^k(t)$ *is nonincreasing on every subinterval of* $T - M$, $\mu^k(t)$ *is constant on every subinterval of* $T - M - Z^k$, *and* $\mu^k(t_1)a^k(x(t_1)) = 0$.

(A.2.4)
$$\dot{x}(t) = f(x(t), t, \sigma(t)) \qquad \text{a.e. } in\ T, \text{ and}$$
$$\dot{z}(t) = -f_x{}^T(x(t), t, \sigma(t))z(t) - \sum_{l=1}^{m} \mu^l(t) b_x{}^l(x(t), t, \sigma(t)) \qquad \text{a.e. } in\ T - M,$$

where

$$b^k(x, t, \sigma) = a_x{}^k(x)\cdot f(x, t, \sigma), \qquad b_x{}^k = \left(\frac{\partial b^k}{\partial x^1}, \cdots, \frac{\partial b^k}{\partial x^n}\right),$$

$f_x{}^T$ *denotes the transpose of the matrix* f_x,

$$z(t) = O = (0, \cdots, 0), \qquad \mu^k(t) = 0, \quad k \notin K(t), \quad t \in M,$$
$$z(t - 0) = \lim_{\tau\to t-} z(\tau) = O \text{ and } \mu^k(t - 0) = 0, k \notin K(t),$$

if $t \in M$ *and* t *is the right endpoint of some open subinterval of* $T - M$.

(A.2.5) *The Weierstrass E-condition*:

$$v(t)\cdot f(x(t), t, \sigma(t)) = \min_{\sigma\in S} v(t)\cdot f(x(t), t, \sigma) \qquad \text{a.e. } in\ T,$$

where

$$v(t) = z(t) + \sum_{l=1}^{m} \mu^l(t) a_x{}^l(x(t)).$$

(A.2.6) *Support (transversality) conditions*:

$$c_1(\xi_1^*) = x(t_1), \qquad c_0(\xi_0^*) = x(t_0).$$

(A.2.6.1) $$v(t_0)\cdot c_{0,\xi}(\xi_0^*)\xi_0^* = \min_{\xi_0 \in C_0} v(t_0)\cdot c_{0,\xi}(\xi_0^*)\xi_0 .$$

(A.2.6.2) $$(\gamma^1\delta_1 - z(t_1))\cdot c_{1,\xi}(\xi_1^*)\xi_1^* = \min_{\xi_1 \in C_1} (\gamma^1\delta_1 - z(t_1))\cdot c_{1,\xi}(\xi_1^*)\xi_1 .$$

(A.2.7) *There exists a point* t_0^* *in* T, $t_0^* < t_1$, *such that* $\| v(t) \| \neq 0$, $t_0^* < t \leqq t_1$, *and*
either $t_0^* = t_0$,
or $t_0^* \in Z$ *and*

$$z(t_0^*) = -\sum_{l=1}^{m} \bar{\gamma}^l a_x{}^l(x(t_0^*))$$

for some numbers $\bar{\gamma}^k \geqq \mu^k(t_0^*)$, $k \in K(t_0^*)$, *and* $\bar{\gamma}^k = \mu^k(t_0^*)$, $k \notin K(t_0^*)$, *or* $t_0^* \in Z$ *and*

$$\| \sum_{l \in K(t_0^*)} \bar{\gamma}^l a_x{}^l(x(t_0^*)) \| = 0$$

for some numbers $\bar{\gamma}^k$, $k \in K(t_0^*)$, *such that* $\bar{\gamma}^k \geqq 0$ *and* $\sum_{l \in K(t_0^*)} \bar{\gamma}^l = 1$.

(A.2.8) *If there exists a negative number* β *such that, for every subset* K *of* $\{1, 2, \cdots, m\}$, *the relations* $x \in V$; $t \in T$; $a^k(x) < 0$, $k \notin K$; $a^k(x) = 0$, $k \in K$; $\gamma^k \geqq 0$, $k \in K$; *and* $\sum_{l \in K} \gamma^l = 1$ *imply*

$$\min_{\sigma \in S} \sum_{l \in K} \gamma^l a_x{}^l(x)\cdot f(x, t, \sigma) < \beta,$$

then the set M *is empty or contains the single point* t_0.

APPENDIX B

Definitions and assumptions. Let R be a compact Hausdorff space, E_n the Euclidean n-space, T the closed interval $[t_0, t_1]$ of the real axis, V an open set in E_n, B_0 and B_1 compact sets in V and P a compact set in some metric space. We are given the function $g(x, t, p, \rho) = (g^1(x, t, p, \rho), \cdots, g^n(x, t, p, \rho))$ from $V \times T \times P \times R$ to E_n, and it is continuous on R for every $(x, t, p) \in V \times T \times P$.

1. *Definition.* We shall refer to a function $\rho(t)$ from T to R as an *original control*, and we shall say that it is an *admissible original control* if there exist a point b_0 in B_0 and a function $x(t, p)$ from $T \times P$ to V that is absolutely continuous on T for every p in P and such that, for every p in P,

(1.1) $$\frac{dx(t, p)}{dt} = \dot{x}(t, p) = g(x(t, p), t, p, \rho(t)) \qquad \text{a.e. in } T,$$

(1.2) $$x(t_0, p) = b_0 ,$$

(1.3) $$x(t_1, p) \in B_1 .$$

We shall say that $\rho(t)$ is a *minimizing original control* if $x^0 = \max_{p\in P} x^1(t_1, p)$ exists and $\rho(t)$ minimizes x^0 among all admissible original controls.

We next define relaxed (or generalized) controls. This concept, patterned after Young's definition of generalized curves, is introduced to simulate "limits" of rapidly oscillating original controls.

Let measurable sets in R be the Borel sets, and let S be the class of probability measures over R. Then $\sigma \in S$ if σ is a completely additive, nonnegative set function defined on Borel sets, with $\sigma(R) = 1$. Let

$$(1.4)\quad f(x, t, p, \sigma) = \int_R g(x,t,p,\rho)\, d\sigma, \quad \text{for} \quad (x,t,p,\sigma) \in V \times T \times P \times S.$$

2. *Definition.* We shall refer to a function $\sigma(t)$ from T to S as a *relaxed control* and we shall say that it is an *admissible relaxed control* if there exists a point b_0 in B_0 and a function $x(t, p)$ from $T \times P$ to V that is absolutely continuous on T for every p in P and such that, for every p in P,

$$(2.1)\qquad \frac{dx(t, p)}{dt} = \dot{x}(t, p) = f(x(t, p), p, \sigma(t)) \qquad \text{a.e. in } T,$$

$$(2.2)\qquad x(t_0, p) = b_0,$$

$$(2.3)\qquad x(t_1, p) \in B_1.$$

A function $\sigma(t)$ from T to S is a *minimizing relaxed control* if $x^0 = \max_{p\in P} x^1(t_1, p)$ exists and $\sigma(t)$ minimizes x^0 among all admissible relaxed controls.

We observe that every original control is also a relaxed control.

3. *Assumption.* There exist a finite or denumerable collection of disjoint measurable subsets T_r, $r = 1, 2, \cdots$, of T such that $T' = \cup T_r$ has measure $t_1 - t_0$, a positive constant c, a function $\epsilon(h)$, $h > 0$, converging to 0 as $h \to 0+$, and a compact set $D \subset V$ such that the following six conditions are satisfied.

(3.1) The functions

$$g^i(x, t, p, \rho) \text{ and } \frac{\partial g^i(x, t, p, \rho)}{\partial x^j}, \qquad i, j = 1, \cdots, n,$$

exist over $V \times T' \times P \times R$, and over that set they are continuous functions of (x, t, p) uniformly in ρ, are uniformly continuous in p, and are continuous in ρ for each (x, t, p). Furthermore, $| g(x, t, p, \rho) - g(x, t', p, \rho) | \leqq \epsilon(| t - t' |)$, provided t and t' belong to the same set T_r, $r = 1, 2, \cdots$. (Here $| g |$ represents the euclidean length of g.)

$$(3.2)\quad | g(x, t, p, \rho) | \leqq c \text{ and } | g_x(x, t, p, \rho) | \leqq c \text{ on } V \times T' \times P \times R,$$

where g_x is the matrix $(\partial g^i/\partial x^j)$, $i, j = 1, \cdots, n$, and

$$| g_x | = \sum_{i,j=1}^{n} \left| \frac{\partial g^i}{\partial x^j} \right| .$$

(3.3) There exists at least one admissible relaxed control.

(3.4) If $x(t)$ is an absolutely continuous function from T to V such that

$$\dot{x}(t) = f(x(t), t, p, \sigma(t)) \qquad \text{a.e. in } T,$$

$$x(t_0) \in B_0 , \quad x(t_1) \in B_1 ,$$

for some $\sigma(t)$ from T to S and some p in P, then $x(t) \in D$ for $t \in T$.

(3.5) $B_0 = c_0(C_0)$ and $B_1 = c_1(C_1)$, where C_0 and C_1 are compact, convex Euclidean sets and $c_0(\xi_0)$ and $c_1(\xi_1)$ are continuously differentiable homeomorphic mappings of C_0 and C_1 onto B_0 and B_1, respectively.

(3.6) The set C_1 (hence also B_1) is of dimension n, and the matrix $c_{1,\xi} = (\partial c_1{}^i/\partial \xi^j)$, $i, j = 1, \cdots, n$, is nonsingular over C_1.

Existence of minimax. Approximations with original controls. Necessary conditions.

THEOREM B. *Let Assumption* 3 *be satisfied, and let* $f(x, t, p, \sigma)$ *be defined as in* (1.4). *Then the following conclusions hold.*

(B.1) *There exist a minimizing relaxed control* $\sigma(t)$, *an associated point* $b_0 \in B_0$, *and a function* $x(t, p)$ *satisfying Definition* 2. *The vector function* $f(x, t, p, \sigma(t))$ *and the matrix function* $f_x(x, t, p, \sigma(t))$ *are measurable on* T *for every* $(x, p) \in V \times P$.

(B.2) *There exist a sequence* $\rho_1(t), \rho_2(t), \cdots$, *of original controls and a sequence of functions* $x_1(t, p), x_2(t, p), \cdots$, *from* $T \times P$ *to* V, *and absolutely continuous on* T *for every* p *in* P, *such that*

$$\frac{dx_s(t, p)}{dt} = g(x_s(t, p), t, p, \rho_s(t)) \quad \text{a.e. } in\ T, \qquad p \in P, s = 1, 2, \cdots ,$$

$\lim_{s\to\infty} x_s(t, p) = x(t, p)$ *uniformly on* $T \times P$, *and* $g(x, t, p, \rho_s(t))$ *is a measurable function of* t *for all* $(x, p) \in V \times P$, $s = 1, 2, \cdots$.

(B.3) *There exist* (a) *a measurable subset* T^* *of* T *of measure* $t_1 - t_0$, (b) *a nonnegative regular measure* ω *defined on Borel subsets of* P, (c) ω*-integrable functions* $\psi^0(p)$ *and* $\psi(p) = (\psi^1(p), \cdots, \psi^n(p))$, (d) *a continuous function* $\bar{\xi}(p)$ *from* P *to* C_1, (e) *a point* $\bar{\xi} \in C_0$, (f) *continuous functions* $h_j(t, p)$, $j = 1, \cdots, n$, *from* $T \times P$ *to* E_n, *absolutely continuous on* T *for each* p *in* P, *such that*

(B.3.1) $\omega(P) > 0$; $\omega(U) = 0$, *where*

$$U = \{p \in P \mid x(t_1, p) \in \text{interior of } B_1 \text{ and } x^1(t_1, p) < \max_{p' \in P} x^1(t_1, p')\};$$

$$\sum_{j=0}^{n} |\psi^j(p)| = 1 \qquad \text{a.e. } \textit{with respect to } \omega;$$

$\psi^0(p) \geqq 0$ *on* P *and* $\psi^0(p) = 0$ a.e. *in* Z *with respect to* ω, *where*

$$Z = \{p \in P \mid x^1(t_1, p) < \max_{p' \in P} x^1(t_1, p')\};$$

(B.3.2)
$$\frac{dx(t,p)}{dt} = f(x(t,p), t, p, \sigma(t)) \quad \textit{on} \quad T^* \times P,$$

$$\frac{dh_j(t,p)}{dt} = -f_x{}^T(x(t,p), t, p, \sigma(t))h_j(t,p) \quad \textit{on} \quad T^* \times P, \qquad j = 1, \cdots, n,$$

$$x(t_0, p) = c_0(\xi) = b_0 \in B_0,$$

$$x(t_1, p) = c_1(\bar{\xi}(p)) \in B_1 \quad \textit{for} \quad p \in P,$$

$$h_j(t_0, p) = \delta_j, \qquad p \in P, \qquad j = 1, \cdots, n,$$

where $f_x{}^T$ *is the transpose of the matrix* $f_x = (\partial f^i/\partial x^j)$, $i, j = 1, \cdots, n$, *and* δ_j *is the* jth *column of the unit matrix of order* n;

(B.3.3) *the Weierstrass E-condition holds*:

$$\sum_{j=1}^{n} \int_P \psi^j(p)h_j(t,p) \cdot f(x(t,p), t, p, \sigma(t))d\omega = \min_{\sigma \in S} \sum_{j=1}^{n} \int_P \psi^j(p)h_j(t,p) \cdot f(x(t,p), t, p, \sigma)\, d\omega \quad \textit{on} \quad T^*;$$

(B.3.4) *the support (transversality) conditions hold*:

$$\int_P \psi(p)\, d\omega \cdot c_{0,\xi}(\bar{\xi})\bar{\xi} = \min_{\xi \in C_0} \int_P \psi(p)\, d\omega \cdot c_{0,\xi}(\bar{\xi})\xi,$$

$$[\psi^0(p)\delta_1 - \sum_{j=1}^{n} \psi^j(p)h_j(t_1, p)] \cdot c_{1,\xi}(\bar{\xi}(p))\bar{\xi}(p) = \min_{\xi \in c_1} [\psi^0(p)\delta_1 - \sum_{j=1}^{n} \psi^j(p)h_j(t_1, p)] \cdot c_{1,\xi}(\bar{\xi}(p))\xi$$

a.e. *with respect to* ω, *where* $c_{0,\xi}$ *is the matrix* $(\partial c_0{}^i/\partial \xi_0{}^j)$, $i = 1, \cdots, n$; $j = 1, \cdots, l$; $l =$ *dimension of* C_0, *and* $c_{1,\xi} = (\partial c_1{}^i/\partial \xi^j)$, $i, j = 1, \cdots, n$.

REFERENCES

[1] L. D. Berkovitz, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.

[2] R. V. Gamkrelidze, *Optimum-rate processes with bounded phase coordinates*, Dokl. Akad. Nauk SSSR, 125 (1959), pp. 475–478.

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

[4] D. L. KELENDZHERIDZE, *On the theory of optimal pursuit*, Dokl. Akad. Nauk SSSR, 138 (1961), pp. 529–532; Soviet Math. Dokl., 2 (1961), pp. 654–656.

[5] ———, *On a problem of optimum tracking*, Avtomat. i Telemeh., 23 (1962), pp. 1008–1013; Automat. Remote Control, 23 (1963), pp. 942–947.

[6] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[7] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[8] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129–145.

[9] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[10] ———, *On a class of minimax problems in the calculus of variations*, Michigan Math. J., to appear.

11] ———, *Unilateral variational problems with several inequalities*, Ibid., to appear.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

ON SOME EXTREMAL PROBLEMS IN THE THEORY OF DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO THE THEORY OF OPTIMAL CONTROL*

R. V. GAMKRELIDZE†

1. Introduction. The development in the last decade of optimal control theory which culminated with the Pontryagin maximum principle [1] has brought about a renewed interest in the calculus of variations, and particularly in such classical phases thereof as the theory of the first variation and first order necessary conditions. It turned out that the classical Lagrange problem in the calculus of variations could be looked upon as a special case of the optimal control problem, and that the necessary condition for optimality, the Pontryagin maximum principle, obtained for the latter problem contained all the classical first order necessary conditions (the Euler-Lagrange equations, the Lagrange multiplier rule, Legendre's necessary condition, and the Weierstrass inequality [1, Chap. 5]). It should also be pointed out that many of the results recently obtained for the optimal control problem were actually contained in earlier works of Valentine [2] (see also Berkovitz [3], [4] and Hestenes [5]).

It is now possible to unify many of the results within a single general framework, and it is to this that the present paper is devoted. The central idea of this unification is intimately connected with the concepts of generalized curves, relaxed variational problems, and "chattering" controls considered respectively by Young [6], Warga [7] and the author [8].

In §§2, 3 of this paper we shall formulate a general extremal problem in the theory of differential equations and derive necessary conditions for extremality. In §4 we shall show that the conventional optimal control problem and the classical problems from the calculus of variations are special cases of our formulated general extremal problem, and that the maximum principle is implied by the necessary conditions derived in §3. (It should be mentioned that some ideas contained in the work of Halkin [9] are quite close to the considerations contained in §4.)

2. Formulation of the extremal problem. Let us consider a family F whose elements $f(x, t)$ are n-dimensional, vector-valued functions defined for $x \in G$ and $t \in I$, where G is a given region (open set) in R^n (an n-di-

* Received by the editors April 7, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5-7, 1964.

† The V. A. Steklov Mathematics Institute, USSR Academy of Sciences, Moscow, USSR, and Department of Mathematics, University of Michigan, Ann Arbor, Michigan.

mensional linear vector space) and I is a preassigned bounded, open interval. We shall suppose that the functions $f \in F$ satisfy the following conditions:

1. Each f is measurable[1] in t over I for every fixed $x \in G$, and is of class C^1 with respect to x in G.

2. For every $f \in F$ and compact subset X of G there exists a function $m(t)$, integrable over I and possibly depending on f and X, such that

$$|f(x, t)| \leqq m(t), \qquad |f_x(x, t)| \leqq m(t), \quad \text{for all} \quad x \in X \quad \text{and} \quad t \in I.$$

Here, vertical bars denote any vector norm in a finite dimensional linear vector space, and f_x is the $n \times n$ Jacobian matrix derived from f.

We shall now introduce the concept of extremality. Let us consider the vector differential equation

$$\dot{x} = \bar{f}(x, t), \tag{2.1}$$

where $\bar{f}$ is a fixed element of F, and let

$$z(t), \qquad t_1 \leqq t \leqq t_2, \qquad [t_1, t_2] \subset I, \tag{2.2}$$

be a solution of this differential equation (i.e., $z(t)$ is an absolutely continuous function that satisfies (2.1) for almost all $t \in [t_1, t_2]$) with boundary values

$$z(t_1) = z_1, \qquad z(t_2) = z_2. \tag{2.3}$$

The point $(t_1, t_2, z_1, z_2) \in R^{2n+2}$ will be denoted by q_z. Analogously, if we have a differential equation

$$\dot{x} = f(x, t), \tag{2.4}$$

where f is an arbitrary element of F, and if

$$x(t), \qquad \tau_1 \leqq t \leqq \tau_2, \qquad [\tau_1, \tau_2] \subset I,$$

is an arbitrary solution of (2.4), we shall denote by q_x the point $(\tau_1, \tau_2, x(\tau_1), x(\tau_2)) \in R^{2n+2}$. The set of all such q_x (for all possible solutions x of all possible equations of the form (2.4) with $f \in F$) will be denoted by Q. Clearly, $Q \subset R^{2n+2}$.

Let us consider a differentiable manifold N in R^{2n+2} with boundary M. For every point $q \in M$, let $N_T(q)$ and $M_T(q)$ denote, respectively, the tangent halfplane to N and tangent plane to M at the point q.

DEFINITION 2.1. The solution (2.2) of (2.1) will be called *an F, N extremal* if $q_z \in M$ and if there exists a neighborhood U of q_z such that

$$U \cap N \cap Q \subset M.$$

[1] In this paper, measurability is to be understood in the following sense: a function is measurable if the preimage of every Borel set is a Borel set.

In other words, $z(t)$ is an F, N, extremal if Q intersects N near q_z only along the boundary M of N, without penetrating the interior of N.

Of course, the necessary conditions for extremality in such a general formulation (without assuming any additional properties for F) may have to be quite trivial and uninteresting. To obtain meaningful conditions, we must confine ourselves to particular classes F. For this reason, we shall restrict ourselves to quasiconvex families F, whose definition we shall now present. The notion of quasiconvexity encompasses (as we shall see in §4) almost all of the extremal problems involving the minimization of integral type functionals which arise in the classical calculus of variations and in the theory of optimal control. Further, it enables us to derive, in a very simple manner, a necessary condition for extremality which contains the Pontryagin maximum principle, and the classical first order conditions in the calculus of variations, as special cases.

Let P^r denote the set of all vectors $\alpha = (\alpha_1, \cdots, \alpha_r) \in R^r$ such that $\alpha_i \geqq 0$ for each i, and $\sum_{i=1}^r \alpha_i = 1$. Let $[F]$ denote the convex hull of the family F, i.e.,

$$[F] = \Big\{ h(x,t) : h(x,t) = \sum_{i=1}^{r} \alpha_i f_i(x,t), \quad \text{where} \quad (\alpha_1, \cdots, \alpha_r) \in P^r, f_i \in F \quad \text{for each } i, \quad r > 0 \quad \text{arbitrary} \Big\}.$$

DEFINITION 2.2. The family of functions F will be called *quasiconvex* if it satisfies the hypotheses set forth at the beginning of this section and if, for every compact set $X \subset G$, every finite collection $f_1, \cdots, f_r$ of elements of F, and every $\epsilon > 0$, there exist functions $f_\alpha \in F$, defined for every $\alpha \in P^r$ (and depending on X, the f_i and ϵ), such that the functions

$$g(x,t;\alpha) = \sum_{i=1}^{r} \alpha_i f_i(x,t) - f_\alpha(x,t)$$

satisfy the following conditions[2]

(2.5) 1. $|g(x,t;\alpha)| < \bar{m}(t), |g_x(x,t;\alpha)| < \bar{m}(t)$ for all $x \in X$, $t \in I$, and $\alpha \in P^r$,

where $\bar{m}(t)$ is some function integrable over I and possibly depending on X and the f_i (but not on ϵ);

(2.6) 2. $\left| \int_{\tau_1}^{\tau_2} g(x,t;\alpha)\, dt \right| < \epsilon$

for every $x \in X$, $\alpha \in P^r$, $\tau_1 \in I$ and $\tau_2 \in I$;

[2] The author would like to thank L. Neustadt for drawing his attention to the necessity of including conditions 1 and 3 in the definition of quasiconvexity.

3. for every sequence $\{\alpha^i\}$ with $\alpha^i \in P^r$, which converges to some $\bar{\alpha} \in P^r$, $g(x, t; \alpha^i)$ converges in measure (as a function of t on I) to $g(x, t; \bar{\alpha})$, for every $x \in X$.

It is clear that if the family F is quasiconvex, then it is dense in its convex hull $[F]$ with respect to the topology induced by (2.6). The concept of quasiconvexity is, as we shall see in §4, intimately connected with the notion of "chattering controls".

To formulate the necessary conditions, we shall introduce some notation. It is convenient to consider that the vectors $f(x, t)$ of F are column vectors. The symbol ψ will always denote an n-dimensional row vector. For each element $f(x, t)$ from our class F, we can form the product $\psi f(x, t) = H(\psi, x, t)$. Clearly, H is a scalar-valued function which is linear in ψ, belongs to the class C^1 with respect to x, is measurable with respect to t, and depends on the choice of f in F. We shall call H the *Hamiltonian function* of our problem.

Let the solution (2.2) of (2.1) be an F, N extremal. Let $\bar{H}(\psi, x, t) = \psi \bar{f}(x, t)$. Then we have the following necessary conditions for extremality.

THEOREM 2.1. *Let the solution $z(t)$, $t_1 \leqq t \leqq t_2$, of* (2.1) *be an F, N extremal, and suppose that F is a quasiconvex family of functions. Then there exists a nonzero, absolutely continuous vector-valued function $\psi(t)$, $t_1 \leqq t \leqq t_2$, such that $z(t)$, $\psi(t)$, $t_1 \leqq t \leqq t_2$, satisfy the following Hamiltonian system of equations for almost all t, $t_1 \leqq t \leqq t_2$:*

$$\dot{z}(t) = \frac{\partial \bar{H}(\psi(t), z(t), t)}{\partial \psi} = \bar{f}(z(t), t),$$

$$\dot{\psi}(t) = -\frac{\partial \bar{H}(\psi(t), z(t), t)}{\partial x} = -\psi(t)\bar{f}_x(z(t), t), \tag{2.7}$$

and such that the inequality

$$\int_{t_1}^{t_2} \bar{H}(\psi(t), z(t), t)\, dt \geqq \int_{t_1}^{t_2} H(\psi(t), z(t), t)\, dt = \int_{t_1}^{t_2} \psi(t) f(z(t), t)\, dt \tag{2.8}$$

holds for every element $f(x, t) \in F$. Further, if $\bar{f}(z(t), t)$, considered as a function of t, is continuous at the end points t_1 and t_2, then the $(2 + 2n)$-dimensional row vector

$$(\bar{H}(\psi(t_1), z(t_1), t_1), -\bar{H}(\psi(t_2), z(t_2), t_2), -\psi(t_1), \psi(t_2)) \tag{2.9}$$

is orthogonal to M at $q_z = (t_1, t_2, z(t_1), z(t_2))$ (the transversality condition).

3. Proof of the necessary condition. This section is devoted to the proof of Theorem 2.1.

Let us consider the convex family of functions $[F] - \bar{f}$, which is the convex hull of the family $F - \bar{f} = \{f - \bar{f} : f \in F\}$. We shall denote the elements of $[F] - \bar{f}$ by δf.

If $m^*(t)$ is any nonnegative, real-valued, integrable function on I, we shall say that the family F is *m^*-quasiconvex* if F satisfies the conditions of Definition 2.2, but with (2.6) replaced by the condition:

$$\left| \int_{\tau_1}^{\tau_2} g(x(t), t; \alpha)\, dt \right| < \epsilon$$

for every $\alpha \in P^r$, $\tau_1 \in I$, $\tau_2 \in I$, and absolutely continuous function $x(t)$ from I to X that satisfies the inequality

$$|\dot{x}(t)| \leqq m^*(t)$$

for almost all $t \in I$.

Let us show that F is quasiconvex if and only if F is m^*-quasiconvex. It is obvious that m^*-quasiconvexity implies quasiconvexity. Conversely, suppose that F is quasiconvex, and let a compact set $X \subset G$, elements $f_i \in F$, and a number $\epsilon > 0$ be given. Since $m^*(t)$ is integrable, there is a subdivision of I defined by points s_i, $i = 0, 1, \cdots, k$, $s_i \leqq s_{i+1}$, such that

$$\int_{s_i}^{s_{i+1}} m^*(t)\, dt < \epsilon \left(2 \int_I \bar{m}(t)\, dt\right)^{-1}, \qquad i = 0, \cdots, k-1, \tag{3.1}$$

where $\bar{m}(t)$ is the function arising in Definition 2.2. Let $x(t)$ be an absolutely continuous function from I to X such that $|\dot{x}(t)| \leqq m^*(t)$ almost everywhere in I. It then follows from (3.1) that

$$|x(t) - x(s_i)| < \epsilon \left(2 \int_I \bar{m}(t)\, dt\right)^{-1} \tag{3.2}$$

for all $t \in [s_i, s_{i+1}]$ and all $i = 0, \cdots, k-1$.

Since F is quasiconvex, for every $\alpha \in P^r$ there exists a function $f_\alpha(x, t) \in F$ such that the functions $g(x, t; \alpha) = \sum_{i=1}^r \alpha_i f_i(x, t) - f_\alpha(x, t)$ satisfy (2.6) with ϵ replaced by $\epsilon/(2k)$, (2.5), and condition 3 in Definition 2.2. Let $\tau_1 \in I$, $\tau_2 \in I$. Adjoining the points τ_1 and τ_2 to the subdivision points s_i and/or reindexing, if necessary, we shall suppose that $\tau_1 = s_0$ and $\tau_2 = s_k$. Clearly, for every $\alpha \in P^r$,

$$\begin{aligned} \int_{\tau_1}^{\tau_2} g(x(t), t; \alpha)\, dt &= \sum_{i=0}^{k-1} \int_{s_i}^{s_{i+1}} g(x(s_i), s; \alpha)\, ds \\ &\quad + \sum_{i=0}^{k-1} \int_{s_i}^{s_{i+1}} [g(x(s), s; \alpha) - (g(x(s_i), s; \alpha)]\, ds. \end{aligned} \tag{3.3}$$

Now, by hypothesis (see (2.6) with τ_1 and τ_2 replaced by s_i and s_{i+1}, respectively),

$$\left| \int_{s_i}^{s_{i+1}} g(x(s_i), s; \alpha)\, ds \right| < \frac{\epsilon}{2k}, \qquad \text{for } i = 0, \cdots, k-1. \tag{3.4}$$

Also (see (3.2) and (2.5)),

$$\left|\sum_{i=0}^{k-1}\int_{s_i}^{s_{i+1}}[g(x(s),s;\alpha)-g(x(s_i),s;\alpha)]\,ds\right|$$

$$\leqq \sum_{i=0}^{k-1}\int_{s_i}^{s_{i+1}}[\max_{x\in X}|g_x(x,s;\alpha)|]\,|x(s)-x(s_i)|\,ds$$

(3.5)

$$< \frac{\epsilon}{2\int_I \bar{m}(t)\,dt}\sum_{i=0}^{k-1}\int_{s_i}^{s_{i+1}}\bar{m}(s)\,ds = \frac{\epsilon}{2}\frac{\int_{\tau_1}^{\tau_2}\bar{m}(s)\,ds}{\int_I \bar{m}(t)\,dt}$$

$$\leqq \frac{\epsilon}{2}.$$

Hence, combining (3.3)–(3.5), we obtain

$$\left|\int_{\tau_1}^{\tau_2} g(x(t),t;\alpha)\,dt\right| < \epsilon,$$

which proves that F is m^*-quasiconvex. Note that we have shown that if F is quasiconvex, then F is m^*-quasiconvex for every nonnegative integrable function $m^*(t)$, and that the function $\bar{m}(t)$ in (2.5) may be chosen independently of m^*.

We now turn to the proof of Theorem 2.1. Let $X \subset G$ be a fixed, compact set such that each point $z(t)$, $t_1 \leqq t \leqq t_2$, is an interior point of X. Let $\delta f = \sum_{i=1}^{r}\alpha_i f_i - \tilde{f}$, where $(\alpha_1, \cdots, \alpha_r) \in P^r$ and the $f_i \in F$, be an arbitrary element of $[F] - \tilde{f}$. Let $\bar{m}(t)$ be the integrable function, for the set X and the elements $\tilde{f}, f_1, \cdots, f_r$ of F, that arises in Definition 2.2. Let $m(t)$ be a function integrable over I such that $|f_i(x,t)| < m(t)$ and $|\tilde{f}(x,t)| < m(t)$ for each i, $x \in X$ and $t \in I$ (such a function exists by the definition of quasiconvexity), and let $m^*(t) = m(t) + \bar{m}(t)$.

Since F is quasiconvex, and consequently m^*-quasiconvex, there exists, for every $\epsilon \in [0, 1]$, a function $g_\epsilon(x,t)$ from $G \times I$ to R^n, in class C^1 with respect to x, and depending on δf and ϵ, such that

$$(\tilde{f} + \epsilon\,\delta f + g_\epsilon) \in F, \tag{3.6}$$

$$|g_\epsilon(x,t)| < \bar{m}(t), \quad \left|\frac{\partial g_\epsilon(x,t)}{\partial x}\right| < \bar{m}(t) \tag{3.7}$$

for every $t \in I$ and $x \in X$,

$$\left|\int_{\tau_1}^{\tau_2} g_\epsilon(x(t),t)\,dt\right| < \epsilon^2, \tag{3.8}$$

for every solution $x(t)$ of (3.9) (see below) sufficiently near $z(t)$, and every interval $[\tau_1, \tau_2] \subset I$ on which $x(t)$ is defined.

Henceforth, let us suppose that, for every $\delta f \in ([F] - f)$ and every $\epsilon \in [0, 1]$, such a function g_ϵ has been chosen (of course, this choice is generally not unique). Below we shall specifiy the g_ϵ more precisely for certain $\delta f \in [F] - \bar{f}$.

Let us "perturb" the differential equation (2.1). Namely, let us consider a fixed element $\delta f \in [F] - \bar{f}$, and for every $\epsilon \in [0, 1]$, let $g_\epsilon(x, t)$ be the function chosen as indicated above such that (3.6)–(3.8) hold. We then consider the "perturbed" equations:

$$\dot{x} = \bar{f}(x, t) + \epsilon\, \delta f(x, t) + g_\epsilon(x, t). \tag{3.9}$$

Further, let us consider the following linear variational equation of (2.1) along the solution $z(t)$:

$$\delta\dot{z}(t) = \frac{\partial \bar{f}(z(t), t)}{\partial x}\, \delta z(t) + \delta f(z(t), t). \tag{3.10}$$

The functions $\partial \bar{f}(z(t), t)/\partial x$ and $\delta f(z(t), t)$ are defined for $t \in [t_1, t_2]$, the interval on which $z(t)$ is defined. However, it is clear that, for arbitrary δt_1 and δt_2, we can consider these functions to be defined for $t_1 + \epsilon\, \delta t_1 \leqq t \leqq t_2 + \epsilon\, \delta t_2$, whenever $\epsilon > 0$ is sufficiently small, since the solution $z(t)$ of (2.1) can always be extended beyond t_1 and t_2, if necessary.

It is now not difficult to show (using standard arguments) that, if $\epsilon > 0$ is sufficiently small, then the solution $x(t)$ of (3.9), satisfying the initial condition

$$x(T) = z(T) + \epsilon\, \delta w,$$

where T is an arbitrary number in $[t_1, t_2]$ and δw is an arbitrary fixed vector in R^n, exists for $t_1 + \epsilon\, \delta t_1 \leqq t \leqq t_2 + \epsilon\, \delta t_2$, and has the form

$$x(t) = z(t) + \epsilon\, \delta z(t) + o(\epsilon), \qquad t_1 + \epsilon\, \delta t_1 \leqq t \leqq t_2 + \epsilon\, \delta t_2, \tag{3.11}$$

where $\delta z(t)$ is the solution of (3.10) with the initial value $\delta z(T) = \delta w$, and $o(\epsilon)/\epsilon \to 0$ as $\epsilon \to 0$ uniformly in t, $t_1 \leqq t \leqq t_2$. Let $\Phi(t)$ be a nonsingular matrix function that satisfies the equation

$$\dot{\Phi}(t) = \frac{\partial \bar{f}(z(t), t)}{\partial x}\, \Phi(t).$$

Then, $\delta z(t)$ is given by the formula

$$\delta z(t) = \Phi(t) \left[\Phi^{-1}(T) \delta w + \int_T^t \Phi^{-1}(s) \delta f(z(s), s))\, ds \right]. \tag{3.12}$$

At the endpoints $t_i + \epsilon\, \delta t_i$, $i = 1$ or 2, we have

$$x(t_i + \epsilon\, \delta t_i) = z(t_i + \epsilon\, \delta t_i) + \epsilon\, \delta z(t_i) + o(\epsilon).$$

By hypothesis, $\bar{f}(z(t), t)$ is continuous at the points t_i. Therefore,

$$z(t_i + \epsilon\,\delta t_i) = z(t_i) + \epsilon\,\delta t_i \bar{f}(z(t_i), t_i) + o(\epsilon),$$

and

$$x(t_i + \epsilon\,\delta t_i) = z_i + \epsilon[\delta z_i + \delta t_i \bar{f}_i] + o(\epsilon) = z_i + \epsilon\,\delta x_i + o(\epsilon), \tag{3.13}$$

where

$$\begin{aligned} &z_i = z(t_i), \qquad \delta z_i = \delta z(t_i), \qquad \bar{f}_i = \bar{f}(z(t_i), t_i), \\ &\delta x_i = \delta z_i + \delta t_i \bar{f}_i, \quad \text{and} \quad o(\epsilon)/\epsilon \to 0 \quad \text{as} \quad \epsilon \to 0. \end{aligned} \tag{3.14}$$

It follows from (3.13) that the points q_x for the solutions (3.11) defined in §2 have the form

$$\begin{aligned} q_x &= (t_1 + \epsilon\,\delta t_1, t_2 + \epsilon\,\delta t_2, z_1 + \epsilon[\delta z_1 + \delta t_1 \bar{f}_1] + o(\epsilon), \\ &\qquad z_2 + \epsilon[\delta z_2 + \delta t_2 \bar{f}_2] + o(\epsilon)) \\ &= q_z + \epsilon(\delta t_1, \delta t_2, \delta z_1 + \delta t_1 \bar{f}_1, \delta z_2 + \delta t_2 \bar{f}_2) + o(\epsilon). \end{aligned}$$

Or,

$$q_x = q_z + \epsilon(\delta t_1, \delta t_2, \delta x_1, \delta x_2) + o(\epsilon). \tag{3.15}$$

In the sequel, the number $T \in [t_1, t_2]$ will be assumed to be fixed.

Let us consider the convex set $A = R^1 \times R^1 \times R^n \times ([F] - \bar{f})$, i.e.,

$$A = \{(\delta t_1, \delta t_2, \delta w, \delta f) : \delta t_1 \in R^1, \delta t_2 \in R^1, \delta w \in R^n, \delta f \in ([F] - \bar{f})\}.$$

For every finite subset $\{\delta f_1, \cdots, \delta f_m\}$ of $[F] - \bar{f}$, we define the convex set $A(\delta f_1, \cdots, \delta f_m)$ as follows:

$$\begin{aligned} A(\delta f_1, \cdots, \delta f_m) = \{(\delta t_1, \delta t_2, \delta w, \delta f) : \delta t_1 \in R^1, \\ \delta t_2 \in R^1, \delta w \in R^n, \delta f \in [\delta f_1, \cdots, \delta f_m]\}, \end{aligned}$$

where $[\delta f_1, \cdots, \delta f_m]$ denotes the convex hull of $\{\delta f_1, \cdots, \delta f_m\}$. It is evident that we can identify $A(\delta f_1, \cdots, \delta f_m)$ with $R^1 \times R^1 \times R^n \times P^m$, i.e., with a convex subset of R^{2+n+m}. Later, when referring to a topology on $A(\delta f_1, \cdots, \delta f_m)$ (for example, when we discuss continuous functions on these sets), we shall mean the conventional Euclidean topology in R^{2+n+m}.

Formulas (3.12), (3.14), and (3.15) give rise to a mapping h (depending on $\epsilon \geqq 0$) from A to R^{2+2n} of the form

$$\begin{aligned} h((\delta t_1, \delta t_2, \delta w, \delta f), \epsilon) &= \frac{q_x - q_z}{\epsilon} \\ &= L(\delta t_1, \delta t_2, \delta w, \delta f) + \varphi(\delta t_1, \delta t_2, \delta w, \delta f, \epsilon), \end{aligned} \tag{3.16}$$

where L, the linear part of h, is given by

$$L(\delta t_1, \delta t_2, \delta w, \delta f) = (\delta t_1, \delta t_2, \delta x_1, \delta x_2). \tag{3.17}$$

Let $\delta f_1, \cdots, \delta f_m$ be fixed elements of $([F] - \bar{f})$, so that there exist functions $f_i \in F$, $i = 1, \cdots, r$, and vectors $\alpha^i = (\alpha_{i1}, \cdots, \alpha_{ir}) \in P^r$, $i = 1, \cdots, m$, such that $\delta f_i = \sum_{j=1}^{r} \alpha_{ij} f_j - \bar{f}$. For every $\beta = (\beta_1, \cdots, \beta_m) \in P^m$, let δf^β be the function in $[\delta f_1, \cdots, \delta f_m]$ defined by the relation

$$\delta f^\beta = \sum_{i=1}^{m} \beta_i \delta f_i = \sum_{j=1}^{r} \left(\sum_{i=1}^{m} \beta_i \alpha_{ij} \right) f_j - \bar{f}.$$

Then, because of the quasiconvexity of F, for every fixed $\epsilon \in (0, 1)$ and $\beta \in P^m$, there exists a function $g_\epsilon(x, t; \beta)$ from $G \times I$ to R^n, in class C^1 with respect to x, and depending on ϵ, such that (see (3.6)–(3.8))

1. $[\bar{f}(x, t) + \epsilon \delta f^\beta(x, t) + g_\epsilon(x, t; \beta)] \in F$;

2. $|g_\epsilon(x, t; \beta)| < \bar{m}(t)$, $\left| \dfrac{\partial g_\epsilon(x, t; \beta)}{\partial x} \right| < \bar{m}(t)$ for every $t \in I$ and $x \in X$, where $\bar{m}(t)$ is a function integrable over I that is independent of β and ϵ;

3. $$\left| \int_{\tau_1}^{\tau_2} g_\epsilon(x(t), t; \beta)\, dt \right| < \epsilon^2$$

for every solution of the equation

$$\dot{x} = \bar{f}(x, t) + \epsilon \delta f^\beta(x, t) + g_\epsilon(x, t; \beta)$$

sufficiently near $z(t)$, and every interval $[\tau_1, \tau_2] \subset I$ on which $x(t)$ is defined; and, in addition (see condition 3 of Definition 2.2),

4. for every sequence $\{\beta^j\}$, with $\beta^j \in P^m$ for every j, which converges to some $\bar{\beta} \in P^m$, $g_\epsilon(x, t; \beta^j)$ converges in measure on I to $g_\epsilon(x, t; \bar{\beta})$ for every fixed $x \in X$ and $\epsilon \in [0, 1]$.

For any preassigned functions $\delta f_1, \cdots, \delta f_m$ in $([F] - \bar{f})$, we may always suppose that in choosing the functions g_ϵ (for every $\delta f \in [F] - \bar{f}$ and every $\epsilon \in [0, 1]$) in such a way that (3.6)–(3.8) hold, we select functions $g_\epsilon(x, t; \beta)$ as described above for every $\delta f \in [\delta f_1, \cdots, \delta f_m]$. Then, it can be readily shown that, for every compact subset of $A(\delta f_1, \cdots, \delta f_m)$ there is a positive number ϵ_0 such that the mapping h (see (3.16)) is defined and continuous on this subset for every fixed $\epsilon < \epsilon_0$. We emphasize that h is not necessarily continuous as a function of ϵ, because g need not be continuous in ϵ. It is easily seen that the linear map L is continuous on $A(\delta f_1, \cdots, \delta f_m)$, so that φ is also continuous on this set for every fixed $\epsilon < \epsilon_0$. We have shown (see (3.15)) that $\varphi(\delta t_1, \delta t_2, \delta w, \delta f, \epsilon) \to 0$, as $\epsilon \to 0$, for fixed δt_i, δw, and δf. By examining the estimates that led to (3.11), it is easily seen that $\varphi \to 0$ uniformly in any preassigned compact subset of $A(\delta f_1, \cdots, \delta f_m)$ as $\epsilon \to 0$.

Let

$$K = L(A).$$

Evidently, K is a convex set in R^{2+2n} containing the origin.

We return to the manifold $N \subset R^{2+2n}$ with boundary M. Let us denote $N_T(q_z)$ and $M_T(q_z)$ simply by N_T and M_T, respectively. Clearly, M_T is the edge of N_T. Since N is differentiable, there exists a homeomorphism $\bar{h}$, from a neighborhood U_{q_z} of q_z in N_T onto a neighborhood of q_z in N, of the form

$$\bar{h}(y) = y + \bar{\varphi}(y), \tag{3.18}$$

where

$$\lim_{\substack{y \to q_z \\ y \in U_{q_z}}} \frac{\bar{\varphi}(y)}{|\, y - q_z \,|} = 0. \tag{3.19}$$

We shall now show that the convex sets K and $(N_T - q_z)$ in R^{2n+2}, which have the origin as a common point, can be separated, i.e., that there exists a hyperplane Π of dimension $(2 + 2n - 1)$ through 0, such that K is contained in one of the halfspaces defined by Π, and $(N_T - q_z)$ is contained in the other.

Suppose the contrary. Then the carrier planes of $(N_T - q_z)$ and of K are in general position, and there exists a point $\hat{q}$ which is a (relative) interior point of both $(N_T - q_z)$ and K. Let $U_{\hat{q}}$ be a bounded neighborhood of $\hat{q}$, in the carrier plane of $(N_T - q_z)$, which is contained in the interior of $(N_T - q_z)$. Let m be the dimension of K, $0 \leqq m \leqq 2 + 2n$. Then there exists a simplex K_m of dimension m such that $\hat{q} \in$ interior of K_m, and $K_m \subset$ interior of K. Consequently, there exist elements $\delta f_i \in ([F] - \bar{f})$, $i = 1, \cdots, m + 1$, and an m-simplex $S_m \subset A(\delta f_1, \cdots, \delta f_m, \delta f_{m+1})$ such that $L(S_m) = K_m$. Note that S_m is compact.

Let us choose the functions g_ϵ that correspond to elements $\delta f \in [\delta f_1, \cdots, \delta f_{m+1}]$ in the manner indicated above. Then the restriction of h on S_m, for fixed ϵ (see (3.16)), is defined for all $\epsilon > 0$ sufficiently small, and is continuous on S_m for each such ϵ. Further, if $\epsilon \geqq 0$ is sufficiently small, the set $(q_z + \epsilon U_{\hat{q}})$ is contained in U_{q_z}. Thus, for $\epsilon > 0$ small enough, we can define the following continuous mapping from $S_m \times U_{\hat{q}}$ into R^{2+2n}: if $\xi \in S_m$, $\eta \in U_{\hat{q}}$, let

$$\begin{aligned} \gamma(\xi, \eta; \epsilon) &= h(\xi, \epsilon) - \frac{1}{\epsilon} \bar{h}(q_z + \epsilon \eta) + \frac{1}{\epsilon} q_z \\ &= L(\xi) - \eta + \varphi(\xi, \epsilon) - \frac{1}{\epsilon} \bar{\varphi}(q_z + \epsilon \eta). \end{aligned}$$

Since the carrier planes of K and of $(N_T - q_z)$ are in general position, and $\hat{q} \in L(S_m) \cap U_{\hat{q}}$, the image of $S_m \times U_{\hat{q}}$ under the mapping $L(\xi) - \eta$

(the linear portion of the mapping $\gamma(\xi, \eta; \epsilon)$) contains a neighborhood of the origin.

Since $\varphi(\xi, \epsilon) \to 0$ as $\epsilon \to 0$ and $\frac{1}{\epsilon}\bar{\varphi}(q_z + \epsilon\eta) \to 0$ as $\epsilon \to 0$ (see (3.19)) uniformly in $(\xi, \eta) \in S_m \times U_{\hat{q}}$, it follows from the Brouwer fixed point theorem that, for all sufficiently small $\epsilon > 0$, the image of $S_m \times U_{\hat{q}}$ under the mapping $\gamma(\xi, \eta; \epsilon)$ will contain the origin; i.e., for every $\epsilon > 0$ sufficiently small, there exist points $\xi \in S_m$, $\eta \in U_{\hat{q}}$ such that

$$L(\xi) + \varphi(\xi, \epsilon) = \eta + \frac{1}{\epsilon}\bar{\varphi}(q_z + \epsilon\eta),$$

or (see (3.16) and (3.18)) there is a point $q_x \in Q$ such that

$$q_x = \bar{h}(q_z + \epsilon\eta).$$

Now $\eta \in U_{\hat{q}} \subset$ interior of $(N_T - q_z)$, so that $(q_z + \epsilon\eta) \in$ interior of N_T. Since $\bar{h}$ is a homeomorphism, this means that there exist points $q_x \in Q \cap$ (interior of N) arbitrarily close to q_z, which contradicts the fact that $z(t)$, $t_1 \leqq t \leqq t_2$, is an F, N extremal.

Thus, let Π be the plane that separates K and $(N_T - q_z)$ in R^{2+2n}, and let $p \neq 0$ be the normal to Π such that (considering p to be a row vector)

(3.20) $\quad p\zeta \leqq 0 \leqq p\eta \quad$ for every $\quad \zeta \in K \quad$ and $\quad \eta \in (N_T - q_z)$.

Since $(M_T - q_z) \subset (N_T - q_z)$, and $(M_T - q_z)$ is a plane through the origin, it follows from (3.20) that $(M_T - q_z) \subset \Pi$. Therefore, p is orthogonal to M_T, i.e., to M at q_z. Let $p = (\chi_1, \chi_2, \rho_1, \rho_2)$ where χ_1 and χ_2 are scalars, ρ_1 and ρ_2 are n-vectors. It follows from (3.20) and the definition of K that

$$\chi_1\delta t_1 + \chi_2\delta t_2 + \rho_1\delta x_1 + \rho_2\delta x_2 \leqq 0 \tag{3.21}$$

for every vector $\zeta = (\delta t_1, \delta t_2, \delta x_1, \delta x_2) \in K$, i.e., for arbitrary scalars δt_1, δt_2, and vectors (see (3.12) and (3.14)) δx_i of the form

$$\delta x_i = \delta t_i \bar{f}_i + \Phi(t_i)\left[\Phi^{-1}(T)\delta w + \int_T^{t_i} \Phi^{-1}(s)\delta f(z(s), s)\, ds\right],$$

where δw is an arbitrary vector in R^n and δf is arbitrary in $([F] - \bar{f})$. Hence, (3.21) implies that

$$\sum_{i=1}^{2}\left\{(\chi_i + \rho_i\bar{f}_i)\delta t_i + \rho_i\Phi(t_i)\left[\Phi^{-1}(T)\delta w + \int_T^{t_i} \Phi^{-1}\delta f\, ds\right]\right\} \leqq 0$$

for arbitrary $\delta w \in R^n$, $\delta f \in [F] - \bar{f}$, and real numbers δt_1, δt_2. Consequently (since $0 \in [F] - \bar{f}$),

$$\chi_i + \rho_i\bar{f}_i = 0, \qquad i = 1 \text{ or } 2, \tag{3.22}$$

$$\rho_1\Phi(t_1) = -\rho_2\Phi(t_2). \tag{3.23}$$

Therefore, also for every $\delta f \in [F] - \bar{f}$,

$$\sum_{i=1}^{2} \rho_i \Phi(t_i) \int_T^{t_i} \Phi^{-1}\delta f \, ds = \rho_2 \Phi(t_2) \int_{t_1}^{t_2} \Phi^{-1}\delta f \, ds \leqq 0. \tag{3.24}$$

Let

$$\psi(t) = \rho_2\Phi(t_2)\Phi^{-1}(t).$$

Then, by virtue of the definition of $\Phi(t)$, the row vector ψ satisfies the following differential equation:

$$\dot{\psi}(t) = -\psi(t)\bar{f}_x(z(t), t), \tag{3.25}$$

and (3.24) can be rewritten in the form

$$\int_{t_1}^{t_2} \psi(s)\delta f(z(s), s) \, ds \leqq 0. \tag{3.26}$$

Since (3.26) must hold for every $\delta f \in (F - \bar{f}) \subset ([F] - \bar{f})$, (3.25) and (3.26) imply (2.7) and (2.8).

Note that $\psi(t) \not\equiv 0$. For if $\psi(t) \equiv 0$, then $\rho_2 = 0$, which by (3.22) and (3.23) implies that $\rho_1 = 0$ and $\chi_1 = \chi_2 = 0$, i.e., that $p = 0$, which is a contradiction.

Since p is orthogonal to M at q_z, the proof of the theorem will be complete if we can show that p coincides with the vector (2.9). But this is an immediate consequence of (3.22), (3.23) and the definition of $\psi(t)$.

Note that (3.20) implies that p is directed into the halfspace defined by Π that contains $(N_T - q_z)$.

4. Applications to control theory and the calculus of variations.

Let y be an $(n - 1)$-dimensional column phase vector which satisfies the differential equation:

$$\dot{y} = Y(y, u, t), \tag{4.1}$$

where the function Y is defined on $G \times U \times I$, G being an open set in R^{n-1}, U an arbitrary (but fixed) set in R^r, and I a bounded, open time interval. Let us assume that Y is of class C^1 with respect to y, and measurable[3] in (u, t) for every fixed $y \in G$. We shall consider differential equations (4.1) where, for u, we substitute a so-called "control" function $u(t)$. We shall confine ourselves here to control functions defined on I which are measurable and essentially bounded, and whose range is contained in U; and we shall denote this class of functions by Ω. We shall also suppose

[3] See footnote 1.

that, for every function $u(t) \in \Omega$, and every compact subset X of G, there exists a function $\bar{m}(t)$, integrable over I and possibly depending on X and $u(t)$, such that

$$| Y(y, u(t), t) | \leqq \bar{m}(t), \qquad | Y_y(y, u(t), t) | \leqq \bar{m}(t)$$

for every $y \in X$ and $t \in I$.

Let us suppose that we are interested in solutions $y(t)$, $\tau_1 \leqq t \leqq \tau_2$, $(\tau_1, \tau_2 \in I)$, of (4.1), with $u = u(t)$ and $u(t) \in \Omega$, that satisfy the boundary conditions

$$\varphi_i(\tau_1, \tau_2, y(\tau_1), y(\tau_2)) = 0, \qquad i = 1, \cdots, k, \qquad k \leqq 2n, \tag{4.2}$$

where $\varphi_1, \cdots, \varphi_k$ are preassigned differentiable functions from R^{2n} to R^1. Finally, let us suppose that there is given a functional

$$J = \int_{\tau_1}^{\tau_2} L(y(t), u(t), t)\, dt, \tag{4.3}$$

where L is a scalar-valued function satisfying the same hypotheses as Y.

Let $\bar{\Omega}$ denote the class of all 4-tuples $(\bar{u}, y, \tau_1, \tau_2)$ such that $\bar{u} \in \Omega$, τ_1 and τ_2 are real numbers in I with $\tau_1 \leqq \tau_2$, and y is a function from $[\tau_1, \tau_2]$ to R^{n-1} that is a solution of (4.1) with $u = \bar{u}(t)$ satisfying the boundary conditions (4.2). Then the standard optimal control problem consists in finding an element $(u^*, y^*, t_1, t_2) \in \bar{\Omega}$ that minimizes (in $\bar{\Omega}$) the value of the functional J given by (4.3).

We shall show that if (u^*, y^*, t_1, t_2) is a solution of this problem, then it is possible to construct a quasiconvex family F of functions $f(x, t)$ and a manifold N with boundary M in R^{2+2n} (see §2) such that the n-vector valued function

$$x^*(t) = \begin{pmatrix} y_0^*(t) \\ y^*(t) \end{pmatrix}, \qquad t_1 \leqq t \leqq t_2,$$

where

$$y_0^*(t) = \int_{t_1}^{t} L(y^*(s), u^*(s), s)\, ds,$$

is an F, N extremal. Indeed, let

$$\tilde{Y}(x, u, t) = \begin{pmatrix} L(y, u, t) \\ Y(y, u, t) \end{pmatrix},$$

$$F = \{f(x, t) : f(x, t) = \tilde{Y}(x, u(t), t),\ u(t) \in \Omega\}, \tag{4.4}$$

where $x = (x_0, y) \in R^1 \times G$ and $t \in I$; and let R^{2+2n} be the space of parameters $(\tau_1, \tau_2, \eta_0', \eta_1, \eta_0'', \eta_2)$, where τ_1, τ_2, η_0' and η_0'' are scalars,

and η_1 and η_2 are vectors in R^{n-1}. The manifold N in R^{2+2n} consists of all points $(\tau_1, \tau_2, \eta_0', \eta_1, \eta_0'', \eta_2)$, where τ_1, τ_2, η_1 and η_2 are near t_1, t_2, $y^*(t_1)$, and $y^*(t_2)$, respectively, that satisfy the relations

$$\varphi_i(\tau_1, \tau_2, \eta_1, \eta_2) = 0, \qquad i = 1, \cdots, k,$$
$$\eta_0' = 0, \qquad \eta_0'' \leqq y_0^*(t_2);$$

and M consists of the points of N for which the last inequality is replaced by an equality. It is now evident that the function $x^*(t)$, defined above, is an F, N extremal.

Sometimes (e.g., in the minimum-time problem), the optimal control problem may instead be formulated as follows. Find an element $(u^*, y^*, t_1, t_2) \in \bar{\Omega}$ that minimizes (in $\bar{\Omega}$) the function $\varphi_0(\tau_1, \tau_2, y(\tau_1), y(\tau_2))$, where φ_0 is a preassigned differentiable function from R^{2n} to R^1 (in this case, we require that $k < 2n$ in (4.2)). Let

$$F' = \{f(x, t) : f(x, t) = Y(x, u(t), t), u(t) \in \Omega\}, \tag{4.4'}$$

where x and the values of f are both in R^{n-1}. If (u^*, y^*, t_1, t_2) is a solution of this problem, let the manifold N in R^{2n} consist of all points $(\tau_1, \tau_2, \eta_1, \eta_2)$, with τ_i a real number near t_i and η_i a vector in R^{n-1} near $y^*(t_i)$, $i = 1$ or 2, that satisfy the relations

$$\varphi_i(\tau_1, \tau_2, \eta_1, \eta_2) = 0, \qquad i = 1, \cdots, k,$$
$$\varphi_0(\tau_1, \tau_2, \eta_1, \eta_2) \leqq \varphi_0(t_1, t_2, y^*(t_1), y^*(t_2));$$

and let M consist of those points of N for which the last inequality is replaced by an equality. Evidently, the function $y^*(t)$ is then an F', N extremal.

In order to be able to apply the necessary conditions stated in Theorem 2.1 to the two optimal control problems described above, we need only verify that the families F and F' given by (4.4) and (4.4') are quasiconvex. But this is an almost immediate consequence of the following fundamental lemma.

LEMMA 4.1. *Let $f_i(x, t)$, $i = 1, \cdots, m$, be functions from $X \times I$ to R^n, where X is a compact metric space and I is a finite, open time interval. The functions f_i are assumed to be measurable in t over I for every fixed $x \in X$, of class C^r, $r \geqq 0$, with respect to $x \in X$, and to be dominated, together with their first r partial derivatives with respect to x, by a function $\bar{m}(t)$ integrable over I:*

$$|f_i(x, t)| \leqq \bar{m}(t), \quad \left|\frac{\partial^j f_i(x, t)}{\partial x^j}\right| \leqq \bar{m}(t), \quad \text{for all } x \in X, t \in I, \tag{4.5}$$
$$j = 1, \cdots, r, \text{ and } i = 1, \cdots m.$$

Further, let $p_i(t)$, $i = 1, \cdots, m$, *be nonnegative real-valued measurable functions which, for almost all* $t \in I$, *satisfy the relation*:

$$\sum_{i=1}^{m} p_i(t) = 1. \tag{4.6}$$

Then, for every $\epsilon > 0$, *it is possible to subdivide* I *into sufficiently small subintervals* E_j, $j = \pm 1, \pm 2, \cdots$, *and to assign to each* E_j *one of the functions* $f_1(x, t), \cdots, f_m(x, t)$, *which we shall denote by* f_{E_j}, *such that the function* $f(x, t)$, *defined by the relation*

$$f(x, t) = f_{E_j}(x, t) \quad \text{for} \quad t \in E_j, \qquad j = \pm 1, \pm 2, \cdots,$$

satisfies the inequality

$$\left| \int_{t_1}^{t_2} \left[\sum_{i=1}^{m} p_i(t) f_i(x, t) - f(x, t) \right] dt \right| < \epsilon$$

for every t_1, t_2 *in* I, *and* $x \in X$. *Consequently,* $f(x, t)$ *is of class* C^r *with respect to* x, *measurable in* t *for every fixed* $x \in X$, *and*

$$|f(x, t)| \leqq \bar{m}(t), \quad |\partial^j f(x, t)/\partial x^j| \leqq \bar{m}(t) \quad \text{for every} \quad t \in I, \ x \in X, \quad \text{and} \quad j = 1, \cdots, r.$$

Remark. In order that the function $f(x, t)$ be well defined, it is clear that the intervals E_j must be mutually disjoint. We must keep this in mind in the proof of the lemma.

Proof. Let us subdivide I into mutually disjoint subintervals I_α, $\alpha = \pm 1, \pm 2, \cdots$, and further subdivide each I_α into m mutually disjoint intervals E_{α_i}, $i = 1, \cdots, m$, such that

$$\text{meas } (E_{\alpha_i}) = \int_{I_\alpha} p_i(t)\, dt. \tag{4.7}$$

Further, let us define the function $f(x, t)$ by means of the relation

$$f(x, t) = f_i(x, t) \quad \text{for} \quad t \in E_{\alpha_i}; \quad i = 1, \cdots, m; \alpha = \pm 1, \pm 2, \cdots. \tag{4.8}$$

We shall now show that, for every $\bar{\epsilon} > 0$, there exists a $\delta > 0$ such that if

$$\max_{\alpha} (\text{meas } I_\alpha) < \delta,$$

then

$$\left| \int_{t_1}^{t_2} \left[\sum_{i=1}^{m} p_i(t) f_i(x, t) - f(x, t) \right] dt \right| \leqq \bar{\epsilon} \tag{4.9}$$

for every $t_1 \in I$, $t_2 \in I$, and $x \in X$.

Note. It will be clear from the proof that the intervals E_{α_i} defined above

may be replaced by arbitrary, measurable, mutually disjoint sets satisfying the relations

$$\bigcup_{i=1}^{m} E_{\alpha i} = I_\alpha\,,\ \alpha = \pm 1, \pm 2, \cdots;\ \text{meas}\,(E_{\alpha i}) = \int_{I_\alpha} p_i(t)\,dt.$$

Let $\bar{\epsilon} > 0$ be given, and let $\epsilon = \bar{\epsilon}/2(m + 2 + \text{meas}\, I)$. Let $g_i = g_i(x, t)$, $i = 1, \cdots, m$, denote continuous functions defined on $X \times I$ satisfying the inequalities

$$(4.10)\quad \int_I |\, f_i(x,t) - g_i(x, t)\,|\,dt \leqq \epsilon \quad \text{for every} \quad x \in X \quad \text{and} \quad i = 1, \cdots, m.$$

(The existence of such functions will be demonstrated later.) Let us suppose that the subdivision of I into the I_α is sufficiently fine that

$$(4.11)\quad |\, g_i(x, t') - g_i(x, t'')\,| \leqq \epsilon \quad \text{for every} \quad x \in X,\ i = 1, \cdots, m,$$

whenever t' and t'' both belong to the same I_α, and that

$$(4.12)\qquad \int_{I_\alpha} \bar{m}(t)\,dt \leqq \epsilon \quad \text{for every} \quad \alpha = \pm 1, \pm 2, \cdots,$$

where $\bar{m}(t)$ is the function introduced in (4.5). Such a subdivision exists because of the continuity of the functions g_i on $X \times I$, and the absolute continuity of the indefinite integral of $\bar{m}(t)$.

Let us now estimate the expression

$$\sum_{\alpha=k_1}^{k_2} \left| \int_{I_\alpha} \left[\sum_{i=1}^{m} p_i(t) f_i(x, t) - f(x, t) \right] dt \right|,$$

where k_1 and k_2, $k_1 < k_2$, are arbitrary integers and x is an arbitrary element of X. We shall make use of the equality

$$\int_{I_\alpha} f(x, t)\,dt = \sum_{i=1}^{m} \int_{E_{\alpha i}} f_i(x, t)\,dt,$$

which follows from the definition (4.8). We have

$$\left| \int_{I_\alpha} \left(\sum_{i=1}^{m} p_i f_i - f \right) dt \right| = \left| \int_{I_\alpha} \sum_i p_i f_i\,dt - \sum_i \int_{E_{\alpha i}} f_i\,dt \right|$$

$$= \left| \int_{I_\alpha} \sum_i p_i (f_i - g_i)\,dt + \int_{I_\alpha} \sum_i p_i g_i\,dt - \sum_i \int_{E_{\alpha i}} g_i\,dt + \sum_i \int_{E_{\alpha i}} (g_i - f_i)\,dt \right|$$

$$\leqq \sum_i \int_{I_\alpha} p_i\,|f_i - g_i|\,dt + \sum_i \int_{E_{\alpha i}} |f_i - g_i|\,dt + \left| \int_{I_\alpha} \sum_i p_i g_i\,dt - \sum_i \int_{E_{\alpha i}} g_i\,dt \right|.$$

By virtue of (4.6) and (4.10), we have

$$\sum_{\alpha=k_1}^{k_2}\left(\sum_{i=1}^{m}\int_{I_\alpha} p_i\,|f_i-g_i|\,dt+\sum_{i=1}^{m}\int_{E_{\alpha i}}|f_i-g_i|\,dt\right)$$

$$\leqq \sum_{\alpha=k_1}^{k_2}\sum_{i=1}^{m}\left(\int_{I_\alpha}|f_i-g_i|\,dt+\int_{E_{\alpha i}}|f_i-g_i|\,dt\right)$$

$$\leqq 2\sum_{\alpha=k_1}^{k_2}\sum_{i=1}^{m}\int_{I_\alpha}|f_i-g_i|\,dt\leqq 2\sum_{i=1}^{m}\int_{I}|f_i-g_i|\,dt\leqq 2m\epsilon.$$

Consequently,

$$\sum_{\alpha=k_1}^{k_2}\left|\int_{I_\alpha}\left(\sum_{i=1}^{m}p_i(t)f_i(x,t)-f(x,t)\right)dt\right| \tag{4.13}$$

$$\leqq 2m\epsilon+\sum_{\alpha=k_1}^{k_2}\left|\int_{I_\alpha}\sum_{i=1}^{m}p_ig_i\,dt-\sum_{i=1}^{m}\int_{E_{\alpha i}}g_i\,dt\right|.$$

We now estimate the second term in the right-hand side of (4.13). Let t_α be an arbitrary point in I_α, $\alpha=\pm1,\pm2,\cdots$. Denote $g_i(x,t_\alpha)$ by $g_{\alpha i}=g_{\alpha i}(x)$. It follows from (4.7) that

$$\int_{I_\alpha}\sum_{i=1}^{m}p_i(t)g_{\alpha i}(x)\,dt=\sum_{i=1}^{m}g_{\alpha i}\int_{I_\alpha}p_i(t)\,dt$$

$$=\sum_{i=1}^{m}g_{\alpha i}\,\text{meas}\,(E_{\alpha i})=\sum_{i=1}^{m}\int_{E_{\alpha i}}g_{\alpha i}(x)\,dt.$$

Consequently (see (4.11), (4.7), and (4.6)),

$$\left|\int_{I_\alpha}\sum_{i=1}^{m}p_ig_i\,dt-\sum_{i=1}^{m}\int_{E_{\alpha i}}g_i\,dt\right|$$

$$=\left|\int_{I_\alpha}\sum_{i}p_i(g_i-g_{\alpha i})\,dt+\sum_{i}\int_{E_{\alpha i}}(g_{\alpha i}-g_i)\,dt\right|$$

$$\leqq\sum_{i}\int_{I_\alpha}p_i\,|g_i-g_{\alpha i}|\,dt+\sum_{i}\int_{E_{\alpha i}}|g_{\alpha i}-g_i|\,dt$$

$$\leqq\epsilon\left(\int_{I_\alpha}\sum_{i}p_i\,dt+\sum_{i}\int_{E_{\alpha i}}dt\right)\leqq 2\epsilon\,\text{meas}\,I_\alpha.$$

Thus, (4.13) can be written in the form

$$\sum_{\alpha=k_1}^{k_2}\left|\int_{I_\alpha}\left(\sum_{i=1}^{m}p_i(t)f_i(x,t)-f(x,t)\right)dt\right|\leqq 2m\epsilon+\sum_{\alpha=k_1}^{k_2}2\epsilon\,\text{meas}\,I_\alpha \tag{4.14}$$

$$\leqq 2m\epsilon+2\epsilon\,\text{meas}\,I=(2m+2\,\text{meas}\,I)\epsilon.$$

For every t_1 and t_2 in I, there exist integers k_1 and k_2 such that

$$\int_{t_1}^{t_2}\left[\sum_{i=1}^{m}p_i(t)f_i(x,t)-f(x,t)\right]dt=\sum_{\alpha=k_1}^{k_2}\int_{I_\alpha}\left[\sum_{i=1}^{m}p_if_i-f\right]dt$$

$$+ \int_{K_1} (\sum_i p_i f_i - f)\, dt + \int_{K_2} (\sum_i p_i f_i - f)\, dt,$$

where $K_1 \subset I_{\beta_1}$, $K_2 \subset I_{\beta_2}$ for some β_1, β_2. Consequently (see (4.5), (4.14), and (4.12)), we obtain the inequality

$$\left| \int_{t_1}^{t_2} (\sum_i p_i f_i - f)\, dt \right| \leqq \sum_{\alpha=k_1}^{k_2} \left| \int_{I_\alpha} (\sum_i p_i f_i - f)\, dt \right|$$

$$+ \int_{I_{\beta_1}} (\sum_i p_i |f_i| + |f|)\, dt + \int_{I_{\beta_2}} (\sum_i p_i |f_i| + |f|)\, dt$$

$$\leqq 2(m + \text{meas } I)\epsilon + 2 \int_{I_{\beta_1}} \bar{m}(t)\, dt$$

$$+ 2 \int_{I_{\beta_2}} \bar{m}(t)\, dt \leqq 2(m + \text{meas } I + 2)\epsilon = \bar{\epsilon},$$

which is the desired inequality (4.9).

It only remains to prove that there exist continuous functions $g_i(x, t)$ satisfying (4.10). Let us fix the index i. Without loss of generality, we shall let $i = 1$. Let us cover X by a finite number of open sets U_j, $j = 1, \cdots, s$, such that

$$\int_I |f_1(x', t) - f_1(x'', t)|\, dt \leqq \epsilon/2 \tag{4.15}$$

whenever x' and x'' belong to the same U_j. Such a covering always exists because the function $\bar{F}(x', x'')$ defined by

$$\bar{F}(x', x'') = \int_I |f_1(x', t) - f_1(x'', t)|\, dt$$

is continuous on the compact set $X \times X$ (this follows from the continuity of f_1 in x and from (4.5)). Further, $\bar{F}(x, x) = 0$ for all $x \in X$, i.e., $\bar{F}$ vanishes on the diagonal in $X \times X$. Finally, for every neighborhood D of the diagonal, there exists a number $\rho > 0$ such that $(x', x'') \in D$ whenever the distance between x' and x'' is less than ρ, which implies the validity of (4.15).

For each $j = 1, \cdots, s$, let x_j be a fixed point of U_j. Then, as is well known, there exist continuous functions $h_j(t)$, $j = 1, \cdots, s$, defined for $t \in I$, such that

$$\int_I |f_1(x_j, t) - h_j(t)|\, dt \leqq \frac{\epsilon}{2}, \qquad j = 1, \cdots, s. \tag{4.16}$$

Now let the functions $\varphi_1(x), \cdots, \varphi_s(x)$ form the partition of unity on X with respect to the covering $\{U_j\}$, i.e., the φ_j are nonnegative, $\varphi_j(x) = 0$ for $x \notin U_j$, and $\sum_{j=1}^{s} \varphi_j(x) \equiv 1$.

We shall prove that if $g_1(x, t)$ is given by the equality

$$g_1(x, t) = \sum_{j=1}^{s} \varphi_j(x) h_j(t),$$

then $g_1(x, t)$ provides the desired continuous approximation to $f_1(x, t)$. In fact,

$$\begin{aligned} &\int_I |f_1(x, t) - g_1(x, t)| \, dt \\ &\leqq \int_I \left| \sum_{j=1}^{s} \varphi_j(x)[f_1(x, t) - f_1(x_j, t)] + \sum_{j=1}^{s} \varphi_j(x)[f_1(x_j, t) - h_j(t)] \right| dt \\ &\leqq \sum_{j=1}^{s} \varphi_j(x) \int_I |f_1(x, t) - f_1(x_j, t)| \, dt \\ &\qquad + \sum_{j=1}^{s} \varphi_j(x) \int_I |f_1(x_j, t) - h_j(t)| \, dt. \end{aligned} \tag{4.17}$$

But, for each $x \in X$,

$$\varphi_j(x) \int_I |f_1(x, t) - f_1(x_j, t)| \, dt \leqq \frac{\epsilon}{2} \varphi_j(x), \qquad j = 1, \cdots, s. \tag{4.18}$$

Indeed, (4.18) is a consequence of (4.15) if $x \in U_j$, and, if $x \notin U_j$, both sides of (4.18) vanish. Hence, using (4.16), we can rewrite (4.17) in the form

$$\int_I |f_1(x, t) - g_1(x, t)| \, dt \leqq \sum_{j=1}^{s} \frac{\epsilon}{2} \varphi_j(x) + \sum_{j=1}^{s} \frac{\epsilon}{2} \varphi_j(x) = \epsilon,$$

which was to be proved. This completes the proof of the lemma.

In order to show that the family F given by (4.4) is quasiconvex, let X be any compact subset of G, let $f_i(x, t) = \tilde{Y}(x, u_i(t), t)$, where $u_i(t) \in \Omega$, $i = 1, \cdots, m$, and let $p_i(t) \equiv \alpha_i$, $i = 1, \cdots, m$, where $(\alpha_1, \cdots, \alpha_m) \in P^m$. If we now apply Lemma 4.1, noting that the sets I_α may be chosen independently of the functions $p_i(t)$ (see (4.11) and (4.12)), and that the intervals E_{α_i} (for every $\alpha = \pm 1, \pm 2, \cdots$) may be selected in such a way that E_{α_i} is to the left of, and adjoins, the interval $E_{\alpha_{i+1}}$, for $i = 1, \cdots, m - 1$, we conclude almost immediately that the family F is quasiconvex. Note that the function $f(x, t)$ defined by (4.8) in this case also has the form $f(x, t) = \tilde{Y}(x, u(t), t)$, where $u(t) \in \Omega$ is given by

$$u(t) = u_i(t) \quad \text{for} \quad t \in E_{\alpha_i}, \quad i = 1, \cdots, m; \ \alpha = \pm 1, \pm 2, \cdots. \tag{4.19}$$

The control $u(t)$ defined by (4.19) is sometimes referred to as a "chattering control".

In precisely the same way it can be shown that the family F' defined by (4.4′) is also quasiconvex. Thus, Theorem 2.1 may be applied to both of the optimal control problems discussed above.

All of the previous statements carry over to the case where the set U is replaced by a family of subsets $U(t)$ of R^r, defined for every $t \in I$, and the class Ω is redefined as the set of all measurable, essentially bounded functions on I satisfying the relation $u(t) \in U(t)$ for almost all $t \in I$.

If the set U is fixed, and if the functions L and Y are, in addition, continuous with respect to both u and t, we can show that Theorem 2.1 implies the Pontryagin maximum principle as stated in [1]. Namely, suppose that $\tilde{f}(x, t) = \tilde{Y}(x, u^*(t), t)$ where $u^*(t) \in \Omega$. Then we shall show that (2.8) implies that

$$\psi(t)\tilde{Y}(z(t), u^*(t), t) = \sup_{v \in U} \psi(t)\tilde{Y}(z(t), v, t) \quad \text{for almost all } t \in [t_1, t_2], \tag{4.20}$$

which is here equivalent to the maximum principle. Indeed, suppose the contrary, so that if

$$J = \{t: \psi(t)\tilde{Y}(z(t), u^*(t), t) < \sup_{v \in U} \psi(t)\tilde{Y}(z(t), v, t),\ t \in [t_1, t_2]\},$$

then meas $(J) > 0$. Let $\bar{t} \in J$ be a regular point for $u^*(t)$ (see [1, pp. 76–77]), i.e.,

$$\lim_{\text{meas } E \to 0} \frac{\text{meas } (u^{-1}(O) \cap E)}{\text{meas } E} = 1, \tag{4.21}$$

where E is an arbitrary interval which contains $\bar{t}$, for every neighborhood O of $u^*(\bar{t})$. Such a point $\bar{t}$ exists because almost every point of I is a regular point for $u^*(t)$. Since $\bar{t} \in J$, there is a point $v^* \in U$ such that

$$\psi(\bar{t})\tilde{Y}(z(\bar{t}), u^*(\bar{t}), \bar{t}) < \psi(\bar{t})\tilde{Y}(z(\bar{t}), v^*, \bar{t}). \tag{4.22}$$

Relations (4.21) and (4.22), together with the continuity of the functions ψ, $\tilde{Y}$, and z, imply that if

$$J^* = \{t: \psi(t)\tilde{Y}(z(t), u^*(t), t) < \psi(t)\tilde{Y}(z(t), v^*, t),\ t \in [t_1, t_2]\},$$

then meas $(J^*) > 0$. Let us define $\tilde{u}(t)$ as follows:

$$\tilde{u}(t) = \begin{cases} u^*(t) & \text{if } t \notin J^*, t \in I, \\ v^* & \text{if } t \in J^* \cap I. \end{cases}$$

Clearly, $\tilde{u}(t) \in \Omega$, so that $\tilde{Y}(x, \tilde{u}(t), t) \in F$. But then

$$\int_{t_1}^{t_2} \psi(t)\tilde{Y}(z(t), u^*(t), t)\, dt < \int_{t_1}^{t_2} \psi(t)\tilde{Y}(z(t), \tilde{u}(t), t)\, dt,$$

which contradicts (2.8), and thereby proves the validity of (4.20).

Since it was shown in [1, Chap. 5] that all of the first order necessary conditions of the classical calculus of variations follow from the Pontryagin maximum principle, these conditions are also a special case of Theorem 2.1.

Finally, let us demonstrate how Theorem 2.1 may be applied to an optimal control problem which does not directly fall within the framework of the Pontryagin maximum principle. Namely, let (4.1) have the form

$$\dot{y} = A(t)y + B(t)u, \tag{4.23}$$

let $U = R^r$, and let $\varphi_0(\tau_1, \tau_2, y(\tau_1), y(\tau_2)) = \tau_2 - \tau_1$. Let us suppose that A and B are measurable, essentially bounded functions of t over I. Then it is clear that Y satisfies the hypotheses described at the beginning of this section. Let us then consider the second type of optimal control problem described above, with the following additional constraint imposed on the members of $\bar{\Omega}$:

$$\int_I |\, \bar{u}(t)\,|^2\, dt \leqq 1, \tag{4.24}$$

where $|\, u\,|$ denotes the ordinary Euclidean norm in R^r. In more conventional terminology, we have a minimum-time problem for systems described by the linear equation (4.23) under the constraint (4.24). Here, instead of using (4.4′), we define the family $\tilde{F}$ as follows:

$$\tilde{F} = \left\{ f(x, t) : f(x, t) = A(t)x + B(t)u(t), u(t) \in \Omega, \int_I |\, u(t)\,|^2\, dt \leqq 1 \right\}.$$

It is easily seen that $\tilde{F}$ is convex, and a fortiori, quasiconvex. Applying Theorem 2.1 in the manner indicated above, we conclude, on the basis of (2.8), that (if $u^*(t)$ is the sought optimal control)

$$\int_{t_1}^{t_2} \psi(t)B(t)u^*(t)\, dt \geqq \int_{t_1}^{t_2} \psi(t)B(t)u(t)\, dt$$

for every function $u(t) \in \Omega$ such that $\int_I |\, u\,|^2\, dt \leqq 1$. But this is equivalent to the relation

$$u^*(t) = B'(t)\psi'(t) \left\{ \int_{t_1}^{t_2} |\, B'(s)\psi'(s)\,|^2\, ds \right\}^{-1/2} \qquad \text{for } t_1 \leqq t \leqq t_2,$$

where the prime denotes transpose. Also, note that by virtue of (4.23),

(2.7) has the form

$$\dot{\psi} = -\psi A(t).$$

5. Generalizations. It is clear that the point q_z defined in §2 may instead be defined as the point $(t_1, t_2, t_3, \cdots, t_m, z_1, z_2, z_3, \cdots, z_m) \in R^{m(n+1)}$, where $z(t_i) = z_i$ for $i = 1, \cdots, m$, and $t_i \in [t_1, t_2]$ for $i = 3, \cdots, m$. It is then possible to define an F, N extremal in an entirely analogous manner (N is now a differentiable manifold in $R^{m(n+1)}$ with boundary M), and, with only slightly modified arguments, derive a theorem which generalizes Theorem 2.1.

The next order of generalization leads to the so-called problem of "restricted phase coordinates" (which has been treated, for example in [1, Chap. 6], [4], and [11]), for which, in place of $R^{m(n+1)}$, one must consider an appropriate function space. This problem will be treated in future papers.

6. Acknowledgment. This paper was written while the author was visiting the Department of Mathematics at the University of Michigan. The author would like to express his gratitude to Professors George Hay and Lamberto Cesari for their immeasurable aid and cooperation and for having made the author's visit possible. Further, the friendly scientific cooperation afforded by Professor Cesari contributed to the content of this paper. Finally, I would like to thank my friend and colleague, Professor Lucien Neustadt, for his great help during the preparation of this work[4], portions of which he is partially responsible for.

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Fizmatgiz, Moscow, 1961; English transl., John Wiley, New York, 1962.

[2] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, Contributions to the Calculus of Variations 1933–1937, University of Chicago Press, Chicago, 1937, pp. 407–448.

[3] L. D. BERKOVITZ, *Variational methods of control and programming*, J. Math. Anal. Appl., 3(1961), pp. 145–169.

[4] ———, *On control problems with bounded state variables*, Ibid., 5(1962), pp. 488–498.

[5] M. R. HESTENES, *On variational theory and optimal control theory*, this Journal, 3(1965), pp. 23–48.

[6] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. et Lettres Varsovie, Cl. III, 30(1937), pp. 212–234.

[4] L. Neustadt's contribution to this work was supported in part by the United States Air Force through the Air Force Office of Scientific Research under Contract No. AF 49(638)-1318.

[7] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4(1962), pp. 111–128.

[8] R. V. GAMKRELIDZE, *On sliding optimal regimes*, Dokl. Akad. Nauk SSSR, 143(1962), pp. 1243–1245; Soviet Math. Dokl., 3(1962), pp. 390–395.

[9] H. HALKIN, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12(1964), pp. 1–82.

[10] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Mosk. Univ., Ser. Mat., Meh., Astron., Fiz., Him., 2(1959), pp. 25–32; English transl., this Journal, 1(1962), pp. 76–84.

[11] J. WARGA, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112(1964), pp. 432–455.

[12] R. E. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[13] G. LEITMANN, ed., *Optimization Techniques*, Academic Press, New York, 1962.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

EXTREMAL PROBLEMS IN AERODYNAMICS*

ANGELO MIELE†

1. Introduction. The determination of optimum aerodynamic shapes has interested the scientific community for centuries. Historically speaking, the first problem of this kind was the study by Sir Isaac Newton of the body of revolution having minimum drag for a given length and diameter. Not only did Newton employ an analytical technique analogous to the modern calculus of variations, but he also postulated a law of resistance which has been recognized to be a good approximation to that of a hypersonic inviscid flow. In the early part of this century, the use of advanced mathematical techniques in the analysis of subsonic and supersonic flows stimulated a renewed interest in optimization problems. In particular, Munk determined the lift distribution which minimizes the induced drag of a subsonic wing having a given span and lift; furthermore, Von Kármán determined the shape of the slender forebody of revolution which minimizes the pressure drag in linearized supersonic flow for a given length and diameter. In more recent times, the advent of jet and rocket engines as aircraft propulsion systems and the parallel increase in flight velocities and altitudes have made it necessary to extend the optimization of aerodynamic shapes to a wider range of Mach and Reynolds numbers, thereby including the hypersonic and free-molecular flow regimes.

Since the distributions of pressure and skin-friction coefficients depend on the flow regime, a single optimum body does not exist; rather, a succession of optimum configurations exists, that is, one for each flow regime and each set of free-stream conditions [1]. In addition, the optimum geometry depends on the quantity being extremized (aerodynamic drag, lift-to-drag ratio, surface-integrated heat transfer rate, sonic boom of an aircraft, thrust of a nozzle) as well as on the constraints employed in the optimization process, whether geometric quantities (length, thickness, volume, wetted area, planform area, frontal area) or aerodynamic quantities (lift, bending moment, pitching moment, position of the center of pressure).

In §2 the physical models of interest in the theory of optimum aerodynamic shapes are reviewed. The corresponding mathematical models are illustrated in §3 for problems involving one independent variable and in §4

* Received by the editors April 9, 1965, and in revised form April 14, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mechanical Engineering, Rice University, Houston, Texas. This work was supported by the Office of Scientific Research, Office of Aerospace Research, United States Air Force under Grant No. AF-AFOSR-828-65.

for problems involving two independent variables. Finally, new trends in the theory of optimum aerodynamic shapes as well as certain problems of interest in the immediate future are outlined in §5.

2. Physical models. In this section, some of the physical models of current interest in the theory of optimum aerodynamic shapes are reviewed.

2.1. Linearized supersonic flow. For relatively slender shapes in flight at Mach numbers not too close to unity and yet not too large with respect to unity, the small perturbation theory can be employed when estimating the aerodynamic forces acting on a body. In other words, the nonlinear set of equations governing the motion can be replaced by one which is linear.

Because of the linearity, the method of superposition can be employed, and general analytical solutions can be derived for the aerodynamic forces acting on either a two-dimensional shape or an axisymmetric shape whose contour is arbitrarily prescribed. For a two-dimensional shape, the local pressure coefficient has the form $C_p \sim \dot{y}$; that is, it is proportional to the inclination of the tangent to a surface element with respect to the free-stream direction*. On the other hand, for an axisymmetric shape, the pressure coefficient no longer depends on the local slope of a surface element, but it is governed by the geometry of the entire body portion preceding that element.

2.2. Nonlinearized supersonic flow. Whenever the combination of thickness ratio and Mach number is such that the linearization process is not permissible, a more precise approach to the determination of the fluid properties is necessary. In this connection, one can employ a pressure coefficient derived from second or higher order approximations to the equations of motion or, where possible, one can use the complete set of equations.

If expansion processes are considered (this is the case with a rocket nozzle), the pressure coefficient has the form $C_p = C_p(w)$; that is, it depends only on the local velocity w. On the other hand, if compression processes are studied (this is the case with a forebody), the pressure coefficient has the form $C_p = C_p(w, p_0)$; that is, it depends on the local values of both the velocity w and the stagnation pressure p_0. In turn, the local values of the velocity and the stagnation pressure depend on the geometry of the entire body portion preceding a given surface element and must be determined by solving the partial differential equations governing the flow field within a certain region of interest.

2.3. Newtonian hypersonic flow. Whenever the free-stream Mach number is sufficiently large with respect to unity, the shock wave generated by

* The symbol x denotes a coordinate in the undisturbed flow direction, y a coordinate perpendicular to x, and $\dot{y}$ the derivative dy/dx.

the body lies so close to the body that it can be regarded to be identical with it. Consequently, the pressure distribution can be determined with the assumption that the particles striking the body conserve the tangential component of their velocity but lose the normal component.

For both two-dimensional and axisymmetric shapes, the local pressure coefficient is given by the *sine-squared law* $C_p = 2 \sin^2 \theta$, where θ denotes the inclination of the tangent to a surface element with respect to the free-stream direction. This law is also valid for a three-dimensional shape, providing θ is interpreted as the angle which the free-stream velocity forms with the particular tangent which is coplanar with said velocity and the normal to the surface element under consideration.

2.4. Newton-Busemann hypersonic flow. A basic hypothesis of the Newtonian flow model is that the pressure at a point immediately behind the shock wave is identical with the pressure at the corresponding point of the body. Even if one admits that the layer of gas between the shock wave and the body is infinitely thin, the equality of the pressures is justified only if the gas particles—after crossing the shock wave—move along rectilinear paths; this is precisely the case for a wedge or a cone. On the other hand, if the body surface is either convex or concave, the gas particles in the thin layer between the shock wave and the body move along curvilinear paths, that is, they are subjected to centripetal accelerations. Therefore, the actual pressure on the body is lower than that predicted with the Newtonian theory for convex bodies, but higher for concave bodies.

The resulting pressure correction was first calculated by Busemann; hence, this flow model is called the Newton-Busemann model and, while more complicated than the Newtonian model, it is still relatively simple for analytical purposes. The reason is that, if the slender body approximation is made, the local pressure coefficient has the functional form $C_p = C_p(y, \dot{y}, \ddot{y})$; that is, it depends only on the geometric properties of a surface element and is independent of the configuration of the body portion preceding that element.

2.5. Free-molecular flow. In the previous sections, it was tacitly assumed that the gas is a continuum, that is, the mean free path is small with respect to a characteristic dimension of the body. Whenever the mean free path is large with respect to a characteristic dimension of the body, the nature of the flow is free-molecular. The incident molecules are undisturbed by the presence of the vehicle, that is, the incoming and reflected flows are transparent to each other. For analytical purposes, two idealized models have been employed thus far, and are now illustrated.

In the *specular reflection model*, the molecules hitting the surface are reflected optically, which means that the tangential velocity component is

unchanged while the normal velocity is reversed. Under the hypersonic approximation (that is, if the square of the normal component of the speed ratio is much greater than one), the pressure coefficient is given by $C_p = 4 \sin^2 \theta$ and the skin-friction coefficient is $C_f = 0$.

In the *diffuse reflection* model, the molecules hitting the surface are first absorbed and then reemitted with a Maxwellian velocity distribution corresponding to an equilibrium temperature intermediate between that of the incoming flow and that of the solid surface; while the pressure and skin-friction coefficients are considerably more complicated than in the specular model, they still have the form $C_p = C_p(\theta)$ and $C_f = C_f(\theta)$; in other words, they depend only on the orientation of a surface element with respect to the free-stream direction and are independent of the geometry of the body portion preceding that element.

3. Extremal problems in one independent variable. In the theory of optimum aerodynamic shapes, certain functional forms involving one independent variable and one or several dependent variables are of frequent interest.

3.1. One dependent variable. A rather general problem in one independent variable occurs whenever a functional is to be extremized with respect to the class of arcs $y(x)$ which satisfy a set of isoperimetric constraints. If the functional and the isoperimetric constraints involve the sum of a line integral and a function of the end coordinates, the variational problem is represented by the relations

$$\text{(a)} \qquad I = \int_{x_i}^{x_f} f(x, y, \dot{y})\, dx + g(x_i, y_i, x_f, y_f),$$

$$\text{(b)} \qquad K_j = \int_{x_i}^{x_f} \varphi_j(x, y, \dot{y})\, dx + \gamma_j(x_i, y_i, x_f, y_f), \qquad j = 1, \cdots, p. \tag{1}$$

In these relations, x denotes the independent variable, y the dependent variable, and $\dot{y}$ the derivative dy/dx; the subscripts i, f stand for the initial and final points, respectively; the symbols f, g, φ_j, γ_j denote arbitrarily specified functions of the arguments within the parentheses; I is the quantity being extremized, and the symbols K_j denote some prescribed constants. The problem is to find, within the class of arcs $y(x)$ which satisfy the isoperimetric constraints (1b) and certain prescribed end conditions, that particular arc which minimizes the functional (1a).

Variational problems of this type arise whenever two requirements are met. First of all, the configuration must have special geometric properties so that the body is described by a single curve; this is precisely the case with

a two-dimensional wing, a body of revolution, a conical body, and—more generally—a homothetic body whose longitudinal contour or transversal contour is arbitrarily prescribed. Second, the flow regime must be such that the pressure and skin-friction coefficients are functions of, at most, the local coordinates and the slope of the contour; this situation occurs in linearized supersonic flow, Newtonian hypersonic flow, and free-molecular flow.

Since problems of type (1) are relatively simple from an analytical point of view, they have been treated extensively in the literature. In particular, two-dimensional wings in linearized supersonic flow have been considered in [2]–[4]; two-dimensional, axisymmetric, and conical bodies in Newtonian hypersonic flow have been treated in [5]–[15]; and axisymmetric bodies in free-molecular flow have been investigated in [16]–[21]. In a typical case, I is the aerodynamic drag. The constants K_j are: (a) the enclosed area, the moment of inertia of the contour, and the moment of inertia of the enclosed area of a two-dimensional wing; (b) the wetted area and the volume of an axisymmetric body; or (c) the base area of a conical fuselage.

3.2. Several dependent variables. In the previous problem, there are one independent variable and one dependent variable. An important generalization arises whenever the functional being extremized involves several dependent variables y_k, $k = 1, \cdots, n$, and they are required to satisfy a set of isoperimetric constraints and differential constraints. This problem, called the *Bolza problem*, is the most general problem of the calculus of variations in one independent variable and is represented by the relations

$$I = \int_{x_i}^{x_f} f(x, y_k, \dot{y}_k)\, dx + g(x_i, y_{ki}, x_f, y_{kf}),$$

$$K_j = \int_{x_i}^{x_f} \varphi_j(x, y_k, \dot{y}_k)\, dx + \gamma_j(x_i, y_{ki}, x_f, y_{kf}), \qquad j = 1, \cdots, p, \tag{2}$$

$$\psi_j(x, y_k, \dot{y}_k) = 0, \qquad j = 1, \cdots, q,$$

which reduce to those characteristic of a *Lagrange problem* for $g = \gamma_j = 0$ and those characteristic of a *Mayer problem* for $f = \varphi_j = 0$.

Problems of type (2) arise in the study of two-dimensional or axisymmetric bodies in nonlinearized supersonic flow, providing the aerodynamic forces and the geometric constraints can be expressed as one-dimensional integrals to be evaluated along the same reference line, whether the contour of the body or a characteristic line of the flow field.

As an example of isentropic flow, consider the shock-free, supersonic expansion of a gas in a two-dimensional or axisymmetric nozzle of given length. In this problem, the thrust, the mass flow, and the length can be expressed as integrals of quantities evaluated along the left-going characteristic line joining the axis of symmetry with the final point. The minimal

problem is a Bolza problem, with this understanding: the quantity I is the thrust; the constants K_j are the mass flow and the length; and the constraints $\psi_j = 0$ are the differential equations to be satisfied along a characteristic line, namely, the direction and compatibility conditions [22]–[27].

If compression processes are considered, that is, if shock waves are present in the flow field, a more complicated situation arises. Nevertheless, if the properties of the stream function are exploited, the optimum two-dimensional or axisymmetric forebody of given length can be studied within the framework of the calculus of variations in one independent variable. Once more, the drag, the mass flow, and the length can be expressed as integrals of quantities evaluated along the right-going characteristic line joining the shock wave with the final point. Therefore, this is a Bolza problem, with this understanding: the quantity I is the drag; the constants K_j are the mass flow and the length; and the constraints $\psi_j = 0$ are the differential equations to be satisfied along a characteristic line, namely the direction and compatibility conditions as well as the equation defining the stream function distribution [28]–[29].

Problems of type (2) also arise in the study of two-dimensional or axisymmetric bodies in Newtonian hypersonic flow, Newton-Busemann hypersonic flow, and free-molecular flow whenever an inequality constraint is imposed on the configuration and/or derivatives of higher order than the first are present. At first glance, these problems do not seem to be covered by the Bolza formulation since only first order derivatives are present in (2) and inequality constraints are not even mentioned. However, by the judicious use of auxiliary variables, each problem can be converted into a Bolza problem.

As a first example, the slope of a configuration in Newtonian hypersonic flow may be required to be nonnegative everywhere; that is, the inequality constraint $\dot{y} \geqq 0$ is to be accounted for. This inequality constraint can be converted into a differential constraint if the real auxiliary variable u defined by the relationship $\dot{y} - u^2 = 0$ is introduced [9]–[11].

As a second example, it is known that the aerodynamic drag of a slender, two-dimensional or axisymmetric body in Newton-Busemann hypersonic flow depends functionally not only on the ordinate and the slope but also on the curvature $\ddot{y}$ (see [30], [31]). In order to convert the associated variational problem into a Bolza problem, one has to introduce the auxiliary variable u defined by the differential constraint $\dot{y} - u = 0$; because of the relationship $\ddot{y} = \dot{u}$, the aerodynamic drag can then be expressed in terms of y, u, $\dot{u}$ (see [32]).

As a third example, the pressure coefficient of a slender, two-dimensional or axisymmetric body in Newton-Busemann hypersonic flow may be required to be nonnegative everywhere. The conversion of the inequality

constraint $C_p(y, \dot{y}, \ddot{y}) \geqq 0$ into the corresponding equality constraint necessary for the Bolza formulation is performed by introducing the auxiliary variables u and v defined by the differential constraints $\dot{y} - u = 0$ and $C_p(y, u, \dot{u}) - v^2 = 0$ (see [32]).

As a fourth example, a configuration in diffuse, free-molecular flow may be designed in such a way that each molecule strikes the body only once. This is the same as stating that the body is convex; that is, the inequality $\ddot{y} \leqq 0$ must be satisfied everywhere. The conversion of this inequality constraint into the corresponding equality constraint necessary for the Bolza formulation is performed by introducing the auxiliary variables u and v defined by the differential constraints $\dot{y} - u = 0$ and $\dot{u} + v^2 = 0$ (see [1, Chap. 28]).

4. Extremal problems in two independent variables. In the theory of optimum aerodynamic shapes, certain functional forms involving two independent variables and one or several dependent variables are of considerable interest.

4.1. One dependent variable. A rather general problem in two independent variables occurs whenever a functional is to be extremized with respect to the class of surfaces $z(x, y)$ which satisfy a set of isoperimetric constraints. If the functional and the isoperimetric constraints involve the sum of a surface integral and a line integral evaluated along the boundary of the surface, the variational problem is represented by the relations

$$\text{(a)} \qquad I = \iint_S f(x, y, z, z_x, z_y)\, dx\, dy + \oint_B g(x, y, z, \dot{y}, \dot{z})\, dx,$$

(3)

$$\text{(b)} \qquad K_j = \iint_S \varphi_j(x, y, z, z_x, z_y)\, dx\, dy + \oint_B \gamma_j(x, y, z, \dot{y}, \dot{z})\, dx, \qquad j = 1, \cdots, p.$$

In the surface integrals, x and y are the independent variables; z denotes the dependent variable, z_x the derivative $\partial z/\partial x$, and z_y the derivative $\partial z/\partial y$; and the symbol S denotes the domain of integration in the xy-plane. In the line integrals, x is the independent variable; y and z denote the dependent variables; $\dot{y}$ and $\dot{z}$ denote the derivatives dy/dx and dz/dx; and the symbol B denotes the boundary of the domain S. Also, the symbols f, g, φ_j, γ_j denote arbitrarily specified functions of the arguments within the parentheses; I is the quantity being extremized, and K_j denote some prescribed constants. The problem is to find, within the class of surfaces $z(x, y)$ which satisfy the isoperimetric constraints (3b) and certain prescribed boundary conditions, that particular surface which minimizes the functional (3a).

Variational problems of this type arise whenever the flow regime is such that the pressure and skin-friction coefficients are functions of, at most, the local coordinates and the slopes of the surface defining the body. This situation occurs in certain problems of linearized supersonic flow, Newtonian hypersonic flow, and free-molecular flow.

As an example, certain minimal properties of three-dimensional wings in linearized supersonic flow can be studied within the frame of the problems described by (3) if use is made of the reverse flow theorems concerning a thin, cambered wing in linearized supersonic flow [33], [34]. In a typical case, I is the aerodynamic drag and the constants K_j are the prescribed values of the lift, the bending moment, and the pitching moment.

As another example, the minimum drag problem of a wing or a fuselage in Newtonian hypersonic flow is a problem described by (3) (see [35], [36]). Typically, I is the aerodynamic drag, and the constant K represents the volume enclosed by a wing of given planform or a fuselage of given base shape.

4.2. Several dependent variables. In the previous problem, there are two independent variables and one dependent variable. An important generalization arises whenever the functional being extremized involves several dependent variables z_k, $k = 1, \cdots, n$, and they are required to satisfy a set of isoperimetric constraints and differential constraints within the domain of integration S and along the boundary B. This problem, called the *Bolza problem*, is the most general problem of the calculus of variations in two independent variables and is represented by the relations

$$
\begin{aligned}
I &= \iint_S f(x, y, z_k, z_{kx}, z_{ky})\, dx\, dy + \oint_B g(x, y, z_k, \dot{y}, \dot{z}_k)\, dx, \\
K_j &= \iint_S \varphi_j(x, y, z_k, z_{kx}, z_{ky})\, dx\, dy + \oint_B \gamma_j(x, y, z_k, \dot{y}, \dot{z}_k)\, dx, \qquad j = 1, \cdots, p, \\
\psi_j(x, y, z_k, z_{kx}, z_{ky}) &= 0 \quad \text{within} \quad S, \qquad j = 1, \cdots, q, \\
\chi_j(x, y, z_k, \dot{y}, \dot{z}_k) &= 0 \quad \text{along} \quad B, \qquad j = 1, \cdots, r,
\end{aligned}
\tag{4}
$$

which reduce to those characteristic of a *Lagrange problem* for $g = \gamma_j = 0$ and to those characteristic of a *Mayer problem* for $f = \varphi_j = 0$.

Problems of type (4) may arise in the study of axisymmetric bodies in linearized or nonlinearized supersonic flow, whenever constraints are imposed not only on the length and the diameter but also on integrated quantities such as the wetted area or the volume.

As an example, consider the shock-free expansion of a gas in an axisymmetric nozzle in nonlinearized supersonic flow, and assume that a general-

ized isoperimetric constraint is imposed on the contour—so as to treat the cases of given wetted area, weight, or linear combination of the wetted area and the weight simultaneously. Specifically, one deals with the flow properties in a region S limited by a boundary B formed by the nozzle contour, the right-going characteristic through the initial point, and the left-going characteristic through the final point. After the thrust and the generalized isoperimetric constraint are expressed as integrals of quantities evaluated along the nozzle contour, the minimal problem can be treated as a Bolza problem, with this understanding: the quantity I is the thrust; the constant K is the value prescribed for the generalized isoperimetric constraint; the constraints $\psi_j = 0$ are the irrotationality condition and the continuity equation to be satisfied at every point of the region S; and the constraints $\chi_j = 0$ are the tangency condition along the nozzle contour as well as the direction and compatibility conditions along the remainder of the contour B (see [37]).

As another example, consider the problem of finding the axisymmetric body, forebody, or ducted forebody which minimizes the drag in linearized supersonic flow for given constraints imposed on the length, the coordinates of some intermediate point, and the volume. Although a rather different formulation has been employed in the literature (see [38]–[50]; see also [1, Chap. 7]), these problems can be studied as problems of the Bolza type. Specifically, one deals with the flow properties in a region S limited by a boundary B formed by the body contour, the left-going characteristic through the initial point, and the right-going characteristic through the final point. After the drag and the volume are expressed as integrals of quantities evaluated along the body contour, the minimal problem can be treated as a Bolza problem, with this understanding: the quantity I is the drag; the constant K is the value prescribed for the volume; the constraints $\psi_j = 0$ are the irrotationality condition and the continuity equation to be satisfied at every point of the region S; and the constraints $\chi_j = 0$ are the tangency conditions along the body contour as well as the direction and compatibility conditions along the remainder of the contour B.

7. Engineering trends and unsolved problems. Despite the variety of the results already obtained, the theory of optimum aerodynamic shapes is only at its beginning. There are interesting and useful variational problems in one independent variable yet to be solved in every flow regime. An analogous remark is even more appropriate for variational problems involving two, three, or four independent variables, since these problems have only occasionally been treated in the literature.

Among the engineering problems which deserve to be investigated in the near future, the following should be mentioned.

In supersonic flow:
(a) the determination of the axisymmetric closed body, forebody, or ducted forebody which minimizes the total drag (the sum of the pressure drag and the friction drag) for a given volume; and
(b) the determination of three-dimensional wings, fuselages, and wing-fuselage combinations which minimize the total drag under the condition that the lift is given, the volume is given, and the boom intensity on the ground does not exceed a prescribed limit.

In hypersonic flow:
(a) the determination of the axisymmetric body which minimizes the surface-integrated heat transfer rate; and
(b) the determination of three-dimensional wings, fuselages, and wing-fuselage combinations which minimize the total drag for given conditions imposed on the lift and the volume.

In free-molecular flow:
(a) the determination of three-dimensional shapes having minimum drag for a given volume.

Mathematically speaking, some of these problems can be studied within the framework of the Bolza problem in one independent variable; on the other hand, more complex problems require an extension of the existing methodology to the cases where the independent variables are two, three, or four (see [51]–[53]; see also [1, Chap. 4]). These multi-dimensional problems occur whenever certain simplifying circumstances (e.g., two-dimensional flow, axisymmetric flow, steady flow) are not invoked. Since only a number of them are amenable to analytical solutions, it is necessary to develop numerical techniques in order to solve the associated boundary value problems.

The vista is expanding so rapidly on this promising application that it is not difficult to predict that—providing sufficient research effort is expended in this area and providing the present rate of progress is maintained in the design of digital computing machines—the calculus of variations approach will become a fundamental instrument in the design of optimum aerodynamic configurations.

REFERENCES

[1] A. Miele, ed., *Theory of Optimum Aerodynamic Shapes*, Academic Press, New York, 1965.
[2] G. Drougge, *Wing sections with minimum drag at supersonic speeds*, The Aeronautical Research Institute of Sweden, Report No. 26, 1949.
[3] D. R. Chapman, *Airfoil profiles for minimum pressure drag at supersonic velocities—general analysis with application to linearized supersonic flow*, NACA, Report No. 1063, 1952.
[4] A. Miele and R. E. Pritchard, *The effect of friction on optimum two-dimensional*

wings in linearized supersonic flow, Boeing Scientific Research Laboratories, Flight Sciences Laboratory, TR No. 78, 1963.

[5] A. J. Eggers, Jr., M. M. Resnikoff, and D. H. Dennis, *Bodies of revolution having minimum drag at high supersonic airspeeds*, NACA, Report No. 1306, 1957.

[6] E. Large, *Nose shape for minimum drag in hypersonic flow*, J. Aerospace Sci., 29 (1962), pp. 98–99.

[7] H. Kennet, *The effect of skin friction on optimum minimum-drag shapes in hypersonic flow*, Ibid., 29 (1962), pp. 1486–1487.

[8] A. Miele, *Slender shapes of minimum drag in Newtonian flow*, Z. Flugwiss., 11 (1963), pp. 203–210.

[9] D. G. Hull, *On slender bodies of minimum drag in Newtonian flow*, Boeing Scientific Research Laboratories, Flight Sciences Laboratory, TR No. 67, 1963.

[10] A. Miele, R. E. Pritchard, and D. G. Hull, *General theory of optimum hypersonic slender bodies including frictional effects*, J. Astronaut. Sci., 10 (1963), pp. 41–54.

[11] A. Miele and J. D. Cole, *Optimum slender bodies in hypersonic flow with a variable friction coefficient*, AIAA J., 1 (1963), pp. 2289–2293.

[12] G. G. Chernyi and A. L. Gonor, *The determination of body shapes of minimum drag using the Newton and the Busemann pressure laws*, Presented at the Symposium on Extremal Problems in Aerodynamics, Boeing Scientific Research Laboratories, Seattle, 1962.

[13] A. L. Gonor, *On three-dimensional bodies of minimum drag at hypersonic speeds*, J. Appl. Math. Mech., 27 (1963), pp. 273–280.

[14] A. Miele and G. R. Saaris, *On the optimum transversal contour of a body at hypersonic speeds*, Astronautica Acta, 9 (1963), pp. 184–198.

[15] R. Bellman, *On a variational problem of Miele*, Ibid., 9 (1963), pp. 199–200.

[16] W. J. Carter, *Optimum nose shapes for missiles in the superaerodynamic region*, J. Aero. Sci., 24 (1957), pp. 527–532.

[17] I. D. Chang, *On optimum nose shapes for missiles in the superaerodynamic region*, Ibid., 25 (1958), pp. 57–58.

[18] H. S. Tan, *On optimum nose curves for missiles in the superaerodynamic regime*, Ibid., 25 (1958), pp. 56–57.

[19] ———, *On optimum nose curves for superaerodynamic missiles*, Ibid., 25 (1958), pp. 263–264.

[20] ———, *On a special Bolza variational problem and the minimization of superaerodynamic hypersonic nose drag*, Quart. Appl. Math., 17 (1959), pp. 311–314.

[21] ———, *Nose drag in free-molecule flow and its minimization*, J. Aero. Sci., 26 (1959), pp. 360–366.

[22] K. G. Guderley and E. Hantsch, *Best shapes for axisymmetric supersonic jet nozzles*, Z. Flugwiss., 3 (1955), pp. 305–313.

[23] G. V. R. Rao, *Exhaust nozzle contour for optimum thrust*, Jet Propulsion, 28 (1958), pp. 377–382.

[24] ———, *Spike nozzle contour for optimum thrust*, Planetary and Space Sci., 4 (1961), pp. 92–101.

[25] K. G. Guderley, *On Rao's method for the computation of exhaust nozzles*, Z. Flugwiss., 7 (1959), pp. 345–350.

[26] L. E. Sternin, *The boundary of the region of existence of optimal nozzles free of shock waves*, Soviet Physics Dokl., 6 (1962), pp. 574–575.

[27] V. M. Borisov and Iu. D. Shmyglevskii, *The formulation of variational problems of gas dynamics*, J. Appl. Math. Mech., 27 (1963), pp. 269–272.

[28] K. G. GUDERLEY, J. V. ARMITAGE, AND E. M. VALENTINE, *Nose and inlet shapes of minimum drag in supersonic flow*, Aeronautical Research Laboratories, USAF, Report No. ARL 62-342, 1962.

[29] IU. D. SHMYGLEVSKII, *On a class of bodies of revolution with minimum wave resistance*, J. Appl. Math. Mech., 24 (1960), pp. 1390-1396.

[30] W. D. HAYES AND R. F. PROBSTEIN, *Hypersonic Flow Theory*, Academic Press, New York, 1959.

[31] J. D. COLE, *Newtonian slender-body theory of flow around minimum drag shapes*, Presented at the Symposium on Extremal Problems in Aerodynamics, Boeing Scientific Research Laboratories, Seattle, 1962.

[32] A. MIELE, *A study of slender shapes of minimum drag using the Newton-Busemann pressure coefficient law*, AIAA J., 1 (1963), pp. 168-178.

[33] R. T. JONES, *The spanwise distribution of lift for minimum induced drag of wings having a given lift and a given bending moment*, NACA, TN No. 2249, 1950.

[34] ———, *The minimum drag of thin wings in frictionless flow*, J. Aero. Sci., 18 (1951), pp. 75-81.

[35] T. STRAND, *Wings and bodies of revolution of minimum drag in Newtonian flow*, Convair, Report No. ZA-303, 1958.

[36] A. TOOMRE, *Zero-lift minimum drag hypersonic bodies*, Grumman Aircraft Engineering Corporation, Report No. RE-110, 1959.

[37] K. G. GUDERLEY AND J. V. ARMITAGE, *A general method for the determination of best supersonic rocket nozzles*, Presented at the Symposium on Extremal Problems in Aerodynamics, Boeing Scientific Research Laboratories, Seattle, 1962.

[38] T. VON KÁRMÁN, *The problem of resistance in compressible fluids*, Convegno di Scienze Fisiche, Matematiche e Naturali sul tema: Le Alte Velocitá in Aviazione, Reale Accademia d'Italia, Roma, 1935.

[39] C. FERRARI, *On the determination of the projectile of minimum wave drag, Parts 1 and 2*, Atti Accad. Sci. Torino, 74 (1939), pp. 675-693; 75 (1939), pp. 61-96.

[40] M. J. LIGHTHILL, *Supersonic flow past bodies of revolution*, ARC, RM No. 2003, 1945.

[41] W. R. SEARS, *On projectiles of minimum wave drag*, Quart. Appl. Math., 4 (1947), pp. 361-366.

[42] W. HAACK, *Projectile shapes for smallest wave drag*, Graduate Division of Applied Mathematics, Brown University, Translation No. A9-T-3, 1948.

[43] C. FERRARI, *On the problem of the fuselage and the ogive of minimum wave drag*, Atti Accad. Sci. Torino, 84 (1950), pp. 1-18.

[44] M. C. ADAMS, *Determination of shapes of boattail bodies of revolution for minimum wave drag*, NACA, TN No. 2550, 1951.

[45] C. FERRARI, *On the determination of the external form of the axisymmetric duct of minimum drag in linearized supersonic flow for given conditions imposed on the meridian contour*, Mem. Accad. Sci. Torino, 1 (1955), pp. 83-138.

[46] H. M. PARKER, *Minimum-drag ducted and pointed bodies of revolution based on linearized supersonic theory*, NACA, Report No. 1213, 1955.

[47] ———, *Minimum-drag ducted and closed three-point body of revolution based on linearized supersonic theory*, NACA, TN No. 3704, 1956.

[48] K. C. HARDER AND C. RENNEMANN, JR., *On boattail bodies of revolution having minimum wave drag*, NACA, Report No. 1271, 1956.

[49] M. A. HEASLET, *The minimization of wave drag for wings and bodies with given base area or volume*, NACA, TN No. 3289, 1957.

[50] M. A. Heaslet and F. B. Fuller, *Drag minimization for wings and bodies in supersonic flow*, NACA, Report No. 1385, 1958.

[51] A. I. Egorov, *On optimal control processes in distributed objects*, J. Appl. Math. Mech., 27 (1963), pp. 1045–1058.

[52] K. A. Lur'e, *The Mayer-Bolza problem for multiple integrals and the optimization of the performance of systems with distributed parameters*, Ibid., 27 (1963), pp. 1284–1299.

[53] A. I. Egorov, *Optimal control processes in certain systems with distributed parameters*, Automat. Remote Control, 25 (1964), pp. 557–567.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

CONVEX PROGRAMMING AND OPTIMAL CONTROL*

A. A. GOLDSTEIN†

Abstract. The use of convex programming to attack problems of optimal control is not new, but it is becoming of increasing interest. Techniques of steepest descent and gradient projection have been used by Balakrishnan [1], Goldstein [2], [3], Neustadt [4], and Neustadt and Paiewonsky [5]. For the case of unbounded fuel-optimal linear controls Neustadt [4] has shown that the problem may be cast into the form of an infinite linear program. More recently Dantzig [7] and Van Slyke [8] have obtained results in this direction for bounded linear controls. This paper will be concerned with the case of fuel-optimal linear controls. This problem will be reduced to the case of minimizing a convex function on E_n, and techniques of infinite convex programming will be applied. In the important case when the thrust magnitude is constrained, the convex function is continuously differentiable, and techniques of steepest descent may be applied. This approach has already been suggested by Neustadt and Paiewonsky [5].

1. We denote by $H(A)$ the convex hull of A and by δA the boundary of A. Let A be a compact convex subset of E_{n+1} and let φ denote the *support function* for A, i.e., $\varphi(\xi) = \max\{[a, \xi]: a \in A\}$. This function is defined everywhere except at the origin and is convex, continuous, and positively homogeneous. Let $\alpha(\xi)$ denote the *support set for A at ξ*, i.e., all the members of δA such that $[a, \xi] = \varphi(\xi)$ for all $a \in \alpha(\xi)$. Clearly $\alpha(\lambda\xi) = \alpha(\xi)$ if $\lambda > 0$ and $\alpha(\xi)$ is compact and convex. If A is *strictly convex*, i.e., δA contains no line segments, then $\alpha(\xi)$ is a singleton and the map $\alpha: \xi \to \alpha(\xi)$ will be called the *support mapping* for A. With strict convexity it is easy to prove that the support function is differentiable and its gradient is the support mapping, i.e., $\nabla\varphi(\xi) = \alpha(\xi)$. Furthermore α is continuous. For the sake of completeness we shall prove these statements when needed in the sequel. We now turn our attention to the following problems of mathematical programming which arise from the linear problems of fuel-optimal controls. The connection between the abstract geometrical setting presented here and the problem of synthesis of fuel optimal controls may be found in Meditch and Neustadt [6].

Let θ denote the $(n + 1)$-vector $(0, \cdots, 0, 1)$, and let K denote the hyperplane $\{x \in E_{n+1} : [\theta, x] = 0\}$. The projection of x on K will be denoted by $\underline{x}$. Thus $\underline{x} = (x_1, \cdots, x_n, 0)$. Similarly if $S \subset E_{n+1}$, the projection of S on K will be denoted by $\underline{S}$. Let J denote the hyperplane $\{(\gamma_1, \gamma_2, \cdots, \gamma_n, -1): \gamma \in E_n\}$. In the following problems the set A is described indirectly as follows. *For each $\xi \in J$ a point of $\alpha(\xi)$ is given.*

* Received by the editors April 27, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mathematics, University of Washington, Seattle, Washington.

Thus a solution of the linear programming problem of maximizing $[\xi, \cdot]$ on A is available. This solution is constructed with the aid of the "maximum principle" of optimal control.

Problem 1. Let A be a compact convex subset of E_{n+1} with $a^0 \in \delta A$ and $\underline{a}^0$ an interior point of $\underline{A}$. Assume that $a^0 \in \alpha(\hat{\xi})$ for some $\hat{\xi}$ in J. Given $\underline{a}^0$, find $\hat{\xi}$.

Problem 2. As the above problem with the added hypothesis that A is strictly convex. Find $\hat{\xi}$ so that $a^0 = \alpha(\hat{\xi})$.

The algorithms to be presented generate sequences or subsequences $\{\xi^k\}$ converging to $\hat{\xi}$. We turn our attention first to Problem 1.

2. The following lemma [9, p. 257] will be helpful for the discussion of Problem 1.

LEMMA. *Let Ω be a compact subset of E_n and b a continuous real valued function on Ω. Let $F(x) = \max\{[\mathfrak{A}, x] - b(\mathfrak{A}) : \mathfrak{A} \in \Omega\}$. If there exists $x^0 \in E_n$ such that $F(x^0) \leqq F(x)$ for all x, then there exists a set $\{\mathfrak{A}^0, \cdots, \mathfrak{A}^k\} \subset \Omega$ with $k \leqq n$ such that*

$$\text{(i)} \quad \min_x \max_{0 \leqq i \leqq k} \{[\mathfrak{A}^i, x] - b(\mathfrak{A}^i)\} = F(x^0) = [\mathfrak{A}^i, x^0] - b(\mathfrak{A}^i), \quad 0 \leqq i \leqq k,$$

and

$$\text{(ii)} \quad 0 \in H\{\mathfrak{A}^0, \mathfrak{A}^1, \cdots, \mathfrak{A}^k\}.$$

Let $f(\xi) = \varphi(\xi) - [a^0, \xi] = \max_{a \in A}[a - a_0, \xi]$. We are interested in the problem of minimizing f on the hyperplane J because of the following.

THEOREM. *A point $\hat{\xi} \in J$ minimizes f on J if and only if $a^0 \in \alpha(\hat{\xi})$.*

Proof. Assume $a^0 \in \alpha(\hat{\xi})$. Clearly $f(\xi) \geqq 0$ and $f(\hat{\xi}) = 0$.

Assume $\hat{\xi}$ minimizes f on J. Let $\Omega = A - a^0$. Then $f(\xi) = \max\{[\mathfrak{A}, \xi] : \mathfrak{A} \in \Omega\}$. The above lemma applies showing that there exists a set $\{\mathfrak{A}^1, \mathfrak{A}^2, \cdots, \mathfrak{A}^k\}$ in Ω with $k \leqq n + 2$, $f(\hat{\xi}) = [\mathfrak{A}^i, \hat{\xi}]$, $1 \leqq i \leqq k$, and 0 belongs to $H\{\mathfrak{A}^1, \cdots, \mathfrak{A}^k\}$. Thus $0 = \sum_{i=1}^k \lambda^i(a^i - a^0)$, $\lambda^i \geqq 0$, and $\sum \lambda^i = 1$. We conclude that $f(\hat{\xi}) = 0$ and that furthermore a^0 belongs to $H\{a^1, a^2, \cdots, a^k\}$. Since $f(\hat{\xi}) = \max_{a \in A}[a - a^0, \hat{\xi}]$, $a^i \in \alpha(\hat{\xi})$, $1 \leqq i \leqq k$, while since $\alpha(\hat{\xi})$ is convex, $a^0 \in \alpha(\hat{\xi})$.

Let $A^0 = \{\alpha(\xi) : \xi \in J\}$. A^0 is that piece of the boundary of A such that the normals to A^0 all have negative $(n+1)$st components.

LEMMA. *The set of ordered pairs $\{(\underline{a}, a_{n+1}) : a \in A^0\}$ is a continuous convex function.*

Proof. If the above set is not a function, then for some $\underline{a}$ there are distinct numbers a_{n+1} and a'_{n+1}. Assume $a_{n+1} > a'_{n+1}$. Let $a' = (\underline{a}, a'_{n+1})$ and $(\underline{a}, a_{n+1}) = a$. There exists $\xi \in J$ such that $a \in \alpha(\xi)$ and $[a, \xi] \geqq [b, \xi]$ for all $b \in A$. If $b = a'$ then $a_{n+1} \leqq a'_{n+1}$, a contradiction. That this function is convex and continuous may be proved by a similar argument using supporting hyperplanes.

3. The above function will be named a_{n+1} ; the values are $a_{n+1}(\underline{a})$. Let $b(\underline{a} - \underline{a}^0) = a_{n+1}(\underline{a})$. The restriction of f to J may be then written:

$$\xi \to f(\xi) = \max\{[\mathfrak{a}, \underline{\xi}] - b(\mathfrak{a}) : \mathfrak{a} \in \underline{A} - \underline{a}^0\} + a^0_{n+1}.$$

Let $F(\xi) = f(\xi) - a^0_{n+1}$. Since F and f have the same extremals we shall minimize F to obtain a point $\hat{\xi}$ such that $a^0 \in \alpha(\hat{\xi})$. Let $\Omega = \underline{A} - \underline{a}^0$. Observe that $0 \in \text{int}(\Omega)$ by the hypothesis of Problem 1. The following algorithm will be employed [9, pp. 260, 263].

Algorithm. Let Ω denote a bounded subset of E_n and $\mathfrak{a}$ and b bounded functions on Ω. Set $F(x) = \sup\{[\mathfrak{a}, x] - b(\mathfrak{a}) : \mathfrak{a} \in \Omega\}$. Assume there exists a subset Ω_0 of Ω such that $0 \in \text{int}(H\{\mathfrak{a} : \mathfrak{a} \in \Omega_0\})$. At the mth step of the algorithm a set Ω_m is available. Choose x^m to minimize $F^m(x) = \sup\{[\mathfrak{a}, x] - b(\mathfrak{a}) : \alpha \in \Omega_m\}$. Select* $\mathfrak{a}' \in \Omega$ to maximize $[\mathfrak{a}, x^m] - b(\mathfrak{a})$ with a tolerance of $1/m$. Set $\Omega_{m+1} = \Omega_m \cup \{\mathfrak{a}'\}$.

THEOREM. *The algorithm is effective in the sense that:*

(i) $$F^m(x^m) \uparrow \inf F(x),$$

(ii) *the sequence* $\{x^m\}$ *has cluster points each of which minimize* F.

Proof. See [9, p. 263].

To apply the algorithm to minimize the above function F two ingredients are still necessary: the first, a method of determining x^m, and the second, a way of finding Ω_0. The determination of x^m can be made via the algorithms of [10] or via linear programming. The set Ω_0 can sometimes be determined by trial or inspection. Generally it is most easily obtained by a starting procedure, for example, [9, p. 261]. In the starting procedure the set Ω is augmented by $n + 1$ points $\mathfrak{a}^0, \cdots, \mathfrak{a}^n$ which contain 0 in the interior of their convex hull. In the situation of minimizing F these points can be taken to be the vertices of any simplex containing $\underline{a}^0$ in its interior. It is not necessary that these vertices belong to the set $\underline{A}$. Numbers $b(A^i)$, $0 \leqq i \leqq n$, are also chosen, hopefully sufficiently large so that if $F(\bar{x}) = \inf F(x)$, then $[\mathfrak{a}^i, \bar{x}] - b(\mathfrak{a}^i) < F(\bar{x})$.

If this condition is satisfied, the solution of the augmented system will be the same as for the original system. Since $F(x^m) \geqq \inf F(x) \geqq F^m(x^m)$, computations are terminated when $F(x^m) - F^m(x^m)$ is reasonably small. If, however $[\mathfrak{a}^i, x^m] - b(\mathfrak{a}^i) \geqq F^m(x^m)$ for some i, $b(\mathfrak{a}^i)$ was not chosen sufficiently large. The process must be repeated, for example, replacing $b(\mathfrak{a}^i)$ by $10\, b(\mathfrak{a}^i)$. Since $\inf F(x)$ exists it is clear that eventually $b(\mathfrak{a}^i)$ will be sufficiently large so that the augmented equations do not influence the value of $\inf F(x)$ or the point where the infimum is achieved.

4. We now turn our attention to Problem 2. We first observe that if A

* The formula $\varphi(x) - [\underline{a}^0, \underline{x}]$ may be used to calculate $F(x)$, and $\mathfrak{a}' = \underline{a}(x^m) - \underline{a}^0$, where $a(x^m)$ is any point of $\alpha(x^m)$.

is strictly convex, $\alpha(\xi)$ is a singleton. It follows that if $\hat{\xi}$ minimizes F, then $\hat{\xi}$ solves Problem 2. We now consider the possibility of finding $\hat{\xi}$ by the method of steepest descent.

LEMMA. *If A is strictly convex and compact in E^n, then the support mapping for A is continuous and is the gradient of the support function.*

Proof. Assume $\xi_i \to \xi$. By definition, $[\alpha(\xi), \xi] \geqq [\alpha(\xi_i), \xi]$ and $[\alpha(\xi_i), \xi_i] \geqq [\alpha(\xi), \xi_i]$. Thus

$$[\alpha(\xi), \xi] + [\alpha(\xi_i), \xi_i - \xi] \geqq [\alpha(\xi_i), \xi_i] \geqq [\alpha(\xi), \xi_i].$$

Hence

$$\lim [\alpha(\xi_i), \xi_i] = [\alpha(\xi), \xi].$$

Since $\alpha(\xi)$ is the unique maximizer of $[\cdot, \xi]$ every cluster point of $\{\alpha(\xi_i)\}$ is $\alpha(\xi)$, showing that α is continuous at ξ.

We next show that given $\epsilon > 0$, there exists $\delta > 0$ such that for all $\| h \| < \delta$,

$$| \varphi(\xi + h) - \varphi(\xi) - [\alpha(\xi), h] | < \epsilon \| h \|.$$

Verify that

$$\varphi(\xi + h) - \varphi(\xi) - [\alpha(\xi), h] = [\alpha(\xi + h) - \alpha(\xi), \xi] - [\alpha(\xi), h] + [\alpha(\xi + h), h]$$

and that $[\xi + h, \alpha(\xi) - \alpha(\xi + h)] \leqq 0$ and $[-\xi, \alpha(\xi) - \alpha(\xi + h)] \leqq 0$, so that

$$[h, \alpha(\xi) - \alpha(\xi + h)] \leqq [\xi, \alpha(\xi + h) - \alpha(\xi)] \leqq 0$$

and

$$|[h, \alpha(\xi) - \alpha(\xi + h)]| \geqq |[\xi, \alpha(\xi + h) - \alpha(\xi)]|.$$

Finally

$$| \varphi(\xi + h) - \varphi(\xi) - [\alpha(\xi), h]| \leqq 2 \| h \| \cdot \| \alpha(\xi) - \alpha(\xi + h) \|.$$

Thus the continuity of α implies the differentiability of φ. Furthermore we have $\nabla f(\xi) = \alpha(\xi) - a^0$, and $\nabla f(\xi) = 0$ if and only if $a^0 = \alpha(\xi)$.

LEMMA. *The set $\{\xi \in J : F(\xi) \leqq C\}$ is bounded for arbitrary C.*

Proof. $F(\xi) = \max \{[\mathfrak{a}, \xi] - b(\mathfrak{a}) : \mathfrak{a} \in \underline{A} - \underline{a}^0\}$. Let $q = \inf \{\| \mathfrak{a} \| : \mathfrak{a} + \underline{a}^0 \in \delta \underline{A}\}$ and $r = \sup \{b(\mathfrak{a}) : \mathfrak{a} \in \underline{A} - \underline{a}^0\}$. Because $\underline{a}^0 \in \operatorname{int}(\underline{A})$, $q > 0$. Given ξ, for some $\mathfrak{a} + a^0$ in $\delta \underline{A}$, $\mathfrak{a}$ is a positive multiple of ξ. Hence $F(\xi) \geqq \| \xi \| \cdot \| q \| - r$. Therefore $\| \xi \| \leqq (K + r)/q$.

Thus F is convex, has compact convex level sets, a continuous gradient, and a unique minimizer. It follows that the techniques of [11] may be used to generate a sequence $\{\xi^k\}$ converging to $\hat{\xi}$.

REFERENCES

[1] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.

[2] A. A. GOLDSTEIN, *Minimizing functionals on Hilbert space*, Computing Methods in Optimization Problems, A. V. Balakrishnan and Lucien W. Neustadt, eds., Academic Press, New York, 1964, pp. 159–165.

[3] ———, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[4] L. W. NEUSTADT, *Optimization, a moment problem and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[5] L. W. NEUSTADT AND B. PAIEWONSKY, *On synthesizing optimal controls*, Proceedings Second IFAC Congress, Butterworths, London, 1965.

[6] J. S. MEDITCH AND L. W. NEUSTADT, *An Application of Optimal Control to Midcourse Guidance*, Ibid.

[7] GEORGE DANTZIG, *Control processes and mathematical programming*, Operations Research Center, Berkeley, California, ORC 6431.

[8] R. VAN SLYKE, *Linear control processes and mathematical programming*, Thesis, University of California, Berkeley, 1965.

[9] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton's method for convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.

[10] ———, *A finite algorithm for the solution of consistent linear equations and inequalities and for the Tchebycheff approximation of inconsistent linear equations*, Pacific J. Math., 8 (1958), pp. 415–427.

[11] A. A. GOLDSTEIN, *On steepest descent*, this Journal, 3 (1965), pp. 147–151.

[12] H. G. EGGLESTON, *Convexity*, Cambridge University Press, Cambridge, 1963, p. 46.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

ON STEEPEST DESCENT*

A. A. GOLDSTEIN†

This paper continues studies initiated in [1] concerning iterative methods of driving the gradient to 0. In [1] we were concerned with functions which were twice differentiable; in this paper only first derivatives will be assumed. Other results included are a fixed point theorem for "gradient" operators and a simple proof of the classical method of steepest descent.

Let H be a Hilbert space, x_0 an arbitrary point of H and f a functional on H. Let S denote the level set of f at x_0, viz., $S = \{x \in H : f(x) \leqq f(x_0)\}$. Let $f'(x, \cdot)$ denote the Fréchet derivative of f at x, $\nabla f(x)$ its representer in H, and $[x, y]$ the inner product of x and y. Set

$$\Delta(x, \rho) = f(x) - f(x - \rho\varphi(x)),$$

$$g(x, \rho) = \frac{\Delta(x, \rho)}{[\nabla f(x), \varphi(x)]\rho}$$

and fix σ, $0 < \sigma \leqq \frac{1}{2}$. Here φ denotes a bounded map from S to H, satisfying

$$[\nabla f(x), \varphi(x)] \geqq 0,$$

such that given $\epsilon > 0$, there exists $\delta > 0$ for which

$$[\nabla f(x), \varphi(x)] < \delta \quad \text{implies} \quad \| \nabla f(x) \| < \epsilon.$$

For example, see Remark 4, below.

THEOREM 1. *Assume that, on S, f is Fréchet differentiable and bounded below, and that ∇f is uniformly continuous on S. Set $x_{k+1} = x_k$ when $[\nabla f(x_k), \varphi(x_k)] = 0$. Otherwise choose ρ_k so that $\sigma \leqq g(x_k, \rho_k) \leqq 1 - \sigma$ when $g(x_k, 1) < \sigma$, or $\rho_k = 1$ when $g(x_k, 1) \geqq \sigma$, and set $x_{k+1} = x_k - \rho_k\varphi(x_k)$. Then:*

(a) *$\nabla f(x_k)$ converges to 0 while $f(x_k)$ converges downward to a limit L.*

(b) *If the sequence $\{x_k\}$ has cluster points, then every cluster point satisfies $\nabla f(z) = 0$. Assume ∇f has finitely many zeros on S, S is compact, and $\{\varphi(x_k)\}$ converges to 0. Then the sequence $\{x_k\}$ converges.*

(c) *If S and f are convex, and $\varphi(x) = \nabla f(x)$, then $L = \inf \{f(x) : x \in H\}$. If $\{x_k\}$ has weak cluster points, then all such minimize f.*

* Received by the editors April 27, 1965. Presented at the Symposium on the Mathematical Theory of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Mathematics, University of Washington, Seattle, Washington. This research was sponsored in part by the Boeing Scientific Research Laboratories and the Air Force Office of Scientific Research.

Proof. (a) We have that

$$\Delta(\rho, x) = \rho[\nabla f(\xi), \varphi(x)] = \rho[\nabla f(x), \varphi(x)] + \rho[\nabla f(\xi) - \nabla f(x), \varphi(x)],$$

where ξ lies on the open line segment joining x and $x - \rho\varphi(x)$. Thus

$$g(\rho, x) = 1 + \frac{[\nabla f(\xi) - \nabla f(x), \varphi(x)]}{[\nabla f(x), \varphi(x)]}.$$

Since ∇f is continuous at x and $\| \xi - x \| \leqq \rho \| \varphi(x) \|$, $g(0, x) = 1$. If $g(1, x) < \delta$, then g, being continuous, takes on all values between 1 and δ; therefore there exist numbers $\rho > 0$ so that $\delta \leqq g(\rho, x) \leqq 1 - \delta$. Set $\nabla f(x_k) = \nabla f_k$, $\varphi(x_k) = \varphi_k$, and assume $\nabla f_k \neq 0$, and $x_k \in S$. Thus

$$\Delta(x_k, \rho_k) = \rho_k g(x_k, \rho_k)[\nabla f_k, \varphi_k] \geqq \rho_k \delta[\nabla f_k, \varphi_k] > 0.$$

Thus $f(x_{k+1}) < f(x_k)$ and $x_{k+1} \in S$.

Assume $[\nabla f_k, \varphi_k]$ does not converge to 0. This implies that there exist a subsequence $\{x_k\}$ and a number $\epsilon > 0$ such that $[\nabla f_k, \varphi_k] \geqq \epsilon$. It follows, moreover, that $\{\rho_k\}$ is bounded away from 0. For if not, take a thinner subsequence $\{\rho_k\}$, if necessary, such that $\{\rho_k\} \to 0$. Since φ_k is bounded on S, $\{\| x_k - \xi_k \|\} \to 0$. By the uniform continuity of ∇f on S, it follows that $\{[\nabla f(\xi_k) - \nabla f(x_k), \varphi(x_k)]\} \to 0$, and therefore $g(\rho_k, x_k) \to 1$, contradicting that $g(\rho_k, x_k) \leqq 1 - \delta$ for all k. There exists therefore a number $q > 0$ such that $\rho_k \geqq q$. Hence $\Delta(x_k, \rho_k) \geqq q\delta\epsilon$, from which we may contradict the hypothesis that f is bounded below. Thus $[\nabla f_k, \varphi_k] \to 0$, showing (a).

(b) Let z be a cluster point of the sequence $\{x_k\}$. Since $\{\nabla f_k\} \to 0$, we have $\nabla f(z) = 0$. Thus the number of roots of $\nabla f(z)$ is equal to or greater than the number of cluster points of $\{x_k\}$. Therefore if ∇f has a unique root on S, then $\{x_k\}$ converges to it. If the roots of ∇f are finite in number we may suppose the cluster points of $\{x_k\}$ to be finite in number also. Let x_1 be a cluster point and let x_2 be a closest neighboring cluster point to x_1. Set $\epsilon = \| x_1 - x_2 \|$ and let $N(x_i)$ denote the sphere of radius $\epsilon/3$ centered at x_i : $\{x_k\} \sim \cup N(x_i)$ contains a finite number of points at most, say m. Since $\| x_{k+1} - x_k \|$ converges to 0, we can choose k so that $x_k \in N(x_1)$ and $\| x_{k+1} - x_k \| < \epsilon/3m$. But this implies that $\{x_k\} \sim \cup N(x_i)$ contains more than m points, which is a contradiction. This contradiction persists unless we suppose that $\{x_k\}$ has a unique cluster point.

(c) We need the inequality

$$f(y) \geqq f(x) + [\nabla f(x), y - x]$$

which is valid for all x and y in S. By the convexity of f,

$$\frac{[f(x + t(y - x)) - f(x)]}{t} \leqq f(y) - f(x) \qquad \text{for} \quad t \in [0, 1].$$

By the differentiability of f,

$$\lim_{t\to 0+} \frac{[f(x + t(y - x)) - f(x)]}{t} = [\nabla f(x), y - x].$$

The remainder of the proof is as in [1].

Remarks.

1. The hypotheses on φ for the first part of (b) can be stated: ∇f and φ continuous on S, and $[\varphi(x), f(x)] \geqq 0$ for all $x \in S$, $[\varphi(x), f(x)] = 0$ only if $\nabla f(x) = 0$.

2. In (b), if the sequence $\{x_k\}$ has an isolated cluster point, then it converges.

3. The inequality in the proof of (c) may sometimes be used in (b) for terminating computations. We have

$$f(x_k) > f(z) \geqq f(x_k) + [\nabla f_k, z - x_k] \geqq f(x_k) - \| \nabla f_k \| D,$$

where D is the diameter of S.

4. An example for φ is the following. Let $H = E^n$ and let δ_i, $1 \leqq i \leqq n$, denote the rows of the identity matrix. Choose i_0 so that

$$[\nabla f(x), \delta_{i_0}] \geqq [\nabla f(x), \delta_i], \qquad 1 \leqq i \leqq n.$$

Set $\varphi(x) = \| \nabla f(x) \| \delta_{i_0}$; then

$$[\nabla f(x), \varphi(x)] \geqq \frac{\| \nabla f(x) \|^2}{\sqrt{n}}.$$

COROLLARY. *Let Q be a uniformly continuous nonlinear operator on H which is the gradient of a convex "potential" φ. Assume φ satisfies $\varphi \leqq \rho \| x \|^2/2$, $0 < \rho < 1$. Then Q has a fixed point.*

Proof. Define

$$f(x) = \tfrac{1}{2}[x, x] - \varphi(x).$$

Thus

$$f(x) \geqq \tfrac{1}{2} \| x \|^2 (1 - \rho)$$

and f is bounded below. Clearly, for every x_0, the set $S = \{x \in H : f(x) \leqq f(x_0)\}$ is closed, convex and bounded, and ∇f is uniformly continuous. Applying (c) we obtain a point for which $\nabla f(z) = 0$. Thus z is a fixed point of Q.

Observe that if H is finite dimensional we need only assume Q continuous.

Because the method of steepest descent is usually more laborious than the algorithm above, it is not as practical. We shall discuss it below, however, because the proofs require slightly different techniques.

For simplicity we restrict our attention to gradient directions and to sequences with cluster points. In the sequel assume S is bounded.

THEOREM 2 (Steepest descent). *Assume ∇f continuous on S. Choose $\rho > 0$*

to minimize $f(x_k - \rho\nabla f(x_k))$; *set* $x_{k+1} = x_k - \rho_k\nabla f(x_k)$, *and assume that* $\{x_k\}$ *has cluster points. Then* (b) *of Theorem* 1 *may be asserted.*

Proof. Let z be a cluster point of x_n; clearly $f(x_n)$ converges downward to $f(z)$. We shall show that $\nabla f(z) = 0$. For suppose not, and let $\rho_0 > 0$ minimize $f(z - \rho\nabla f(z))$. The number ρ_0 exists because S is bounded. Thus

$$f(z) = f(z - \rho_0\nabla f(z)) + \eta$$

for some $\eta > 0$. But for every x_n,

$$f(x_n - \rho_0\nabla f(x_n)) = f(z - \rho_0\nabla f(z) + (x_n - z) + \rho_0(\nabla f(z) - \nabla f(x_n)))$$
$$= f(z - \rho_0\nabla f(z)) + [\nabla f(\xi_n), x_n - z + \rho_0(\nabla f(z), - \nabla f(x_n))],$$

with ξ_n "between" $z - \rho_0\nabla f(z)$ and $x_n - \rho_0\nabla f(x_n)$. Let $\{x_n\}$ be a subsequence converging to z. Then $\nabla f(\xi_n)$ converges to $\nabla f(z - \rho_0\nabla f(z))$ and $x_n - z + \rho_0[\nabla f(z) - \nabla f(x_n)]$ converges to 0. Hence for n sufficiently large,

$$f(x_n - \rho_0\nabla f(x_n)) \leqq f(z - \rho_0\nabla f(z)) + \eta/2 = f(z) - \eta/2.$$

But

$$f(z) < f(x_n - \rho_n\nabla f(x_n)) \leqq f(x_n - \rho_0\nabla f(x_n)) \leqq f(z) - \eta/2,$$

a contradiction; hence $\nabla f(z) = 0$.

A modified steepest descent due to H. Curry is more difficult to prove. Curry's proof is geometric, intricate, and much abbreviated. We prove his assertion below.

THEOREM 3 (Curry [2]). *Assume* ∇f *continuous on* S *and choose* ρ_k *to be the first zero of* $\dfrac{d}{dp}(f(x - \rho\nabla f(x)))$ *on the halfray* $\{x - \rho f(x) : \rho \geqq 0\}$. *Assume that the sequence* $\{x_k\}$ *has cluster points. Then* (b) *of Theorem* 1 *can be asserted.*

Proof. Similar to the above, set

$$\Delta f(x, \rho) = f(x) - f(x - \rho\nabla f(x)).$$

Then

$$\frac{d\Delta f}{d\rho} = -\left[\nabla f(u), \frac{du}{d\rho}\right] = [\nabla f(x - \rho\nabla f(x), \nabla f(x)],$$

and

$$\rho g(x, \rho) = \frac{f(x, \rho)}{\| \nabla f(x) \|^2}.$$

Let $\hat{\rho}$ be the least positive root of

$$(\rho g(x, \rho))' = \frac{[\nabla f(x - \rho\nabla f(x)), \nabla f(x)]}{\| \nabla f(x) \|^2}.$$

Thus

$$\hat{\rho}g(x, \hat{\rho}) = \int_0^{\hat{\rho}} \frac{[\nabla f(x) - \rho f(x), \nabla f(x)]}{\| \nabla f(x) \|^2} \, d\rho.$$

Assume now $\nabla f(x_k)$ does not converge to 0. Then for some cluster point z of $\{x_k\}$, $\nabla f(z) \neq 0$. For each x, consider the above integrand as a function of ρ. The integrand is 1 at $\rho = 0$ and 0 at $\rho = \hat{\rho}$. We shall choose ρ_0 so that the integrand $I(x, \rho) \geqq \frac{1}{2}$. Take a subsequence $\{x_k\}$ which converges to z. This sequence with its limit point is a compact subset of S, so that ∇f is uniformly continuous on this compactum, which we call K. Therefore there exist positive numbers q and r such that for all K, $q < \| \nabla f(x_k) \| < r$. By the uniform continuity of ∇f on K, given $\epsilon = q/2$, there exists δ such that

$$\| \nabla f(x - \rho \nabla f(x)) - \nabla f(x) \| < \frac{q}{2} < \frac{\| \nabla f(x) \|}{2}$$

whenever $\rho \| \nabla f(x) \| < \delta$, for all $x \in K$. Choose $\rho_0 = \delta/r$. Then if $\rho \leqq \rho_0$, we have $\rho \| \nabla f(x) \| < \delta$. Thus if $\rho \leqq \rho_0$ and $x \in K$, $\| \nabla f(x - \rho \nabla f(x)) - \nabla f(x) \| < \| \nabla f(x) \|/2$. Thus

$$\begin{aligned} \frac{1}{2} &> \frac{1}{\| \nabla f(x) \|} \| \nabla f(x - \rho \nabla f(x)) - \nabla f(x) \| \\ &\geqq \frac{1}{\| \nabla f(x) \|} \left| \left[\nabla f(x - \rho \nabla f(x)) - \nabla f(x), \frac{\nabla f(x)}{\| \nabla f(x) \|} \right] \right| \\ &= \left| \left[\frac{\nabla f(x - \rho \nabla f(x))}{\| \Delta f(x) \|}, \frac{\nabla f(x)}{\| \Delta f(x) \|} \right] - 1 \right|, \end{aligned}$$

which proves that if $x \in K$ and $\rho \leqq \rho_0$, $I(x, \rho) > \frac{1}{2}$ and therefore $\rho_0 < \hat{\rho}$. It follows that

$$\hat{\rho}g(\hat{\rho}) > \int_0^{\rho_0} \tfrac{1}{2} \, d\rho = \tfrac{1}{2}\rho_0$$

and thus $\Delta f(x_k, \rho_k) \geqq q\rho_0/2$, contradicting that f is bounded below on K.

Acknowledgment. I am indebted to Professor Alexander Ostrowski for suggesting to me that $\{x_k\}$ in Theorem 1 converges if ∇f has finitely many zeros.

REFERENCES

[1] A. A. Goldstein, *Minimizing functionals on Hilbert space*, Computer Methods in Optimization Problems, Academic Press, New York, 1964, pp. 159–165.

[2] H. Curry, *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math., 2 (1944), pp. 258–261.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 1
Printed in U.S.A., 1965

OPTIMAL CONTROL PROBLEMS IN BANACH SPACES*

A. V. BALAKRISHNAN†

1. Introduction. This paper is concerned with a systematic study of a class of control problems in which the state and input or control variables are allowed to range in Banach spaces. Specifically, the state equation is of the form

$$\dot{x}(t) = f(x(t), u(t), t), \tag{1.1}$$

where for each t, $u(t)$ and $x(t)$ are Banach space valued. This extension to infinite dimensions is more than of purely mathematical interest. In the first place, control problems involving distributed parameter systems where the state dynamics are described by partial differential equations are conveniently formulated in this way. Secondly, stochastic control problems can also be handled in this way. Thus, the study of control problems in Banach spaces has the merit of providing a measure of unification for a wide range of problems.

The Banach spaces in what follows will always be function spaces. In the case of partial differential equations, the spaces will usually be L_p spaces (with respect to Lebesgue measure) of functions defined on some region R, not necessarily bounded, of a Euclidean space of one or more dimensions. For example, let the zero-input equation be the partial differential equation,

$$\frac{\partial x(t, r)}{\partial t} = a(t, r) \frac{\partial^2 x(t, r)}{\partial r^2}, \qquad r \in R = \text{interval in } E_1 . \tag{1.2}$$

We formulate this as an abstract Cauchy problem in the chosen Banach space in which the functions $x(t, r)$ are required to lie for each t, and write it as

$$\dot{x}(t) = A(t)x(t), \tag{1.3}$$

where $A(t)$ is identified as the differential operator with the given boundary conditions, and is thus unbounded in general. It may be noted that the derivatives in (1.1) and (1.3) are taken in the "strong" sense, that is, in the topology of the Banach space involved. In the case of stochastic prob-

* Received by the editors February 19, 1965, and in revised form March 10, 1965. Presented at the Symposium on the Mathematical Theorem of Optimal Control, held at the University of Michigan, October 5–7, 1964.

† Department of Engineering, University of California, Los Angeles, California. This research was supported in part by the Air Force Office of Scientific Research, Applied Mathematics Division, United States Air Force, under Grant No. AFOSR 700-65.

lems we have a probability measure space (Ω, β, μ) in the usual notation, Ω being the whole space, β the Borel field, and μ the probability measure. The Banach space here will be an L_p space, $L_p(\Omega, \beta, \mu)$. Again we note that "strong" limits will now be in the topology of the chosen L_p space.

Since we are thinking of partial differential equations, the almost exclusive attention to linear problems needs no particular apology. For the linear case we shall take the *state-input* equation as

$$\dot{x}(t) = f(x(t), u(t), t) = A(t)x(t) + B(t)u(t) + z(t), \tag{1.4}$$

where $x(t)$, $z(t)$ are elements of a Banach space X_1 for each t;
$u(t)$, the control, is an element of another Banach space X_2 for each t;
$A(t)$, $B(t)$ are linear operators for each t;
$B(t)$ is a linear bounded transformation mapping X_2 into X_1;
and $A(t)$ is a closed (and not necessarily bounded) operator with domain dense in X_1 and range in X_1. The fact that $A(t)$ is unbounded means in particular that the right side of (1.4) is not continuous in $x(t)$, let alone differentiable.

The first question of course is that of existence and uniqueness of solutions for (1.4). The homogeneous equation (1.3) does not necessarily have unique solutions without additional conditions on $A(t)$. First let us consider the time-invariant case so that we have

$$\dot{x}(t) = Ax(t). \tag{1.5}$$

Then we may invoke the theory of semigroups of linear operators [1]. Indeed, Phillips [2] has shown that a necessary and sufficient condition for (1.5) to have a unique solution in $(0, \infty)$ for each initial value $x(0)$ in the domain A (with nonvacuous resolvent set) such that

$$\| x(t) - x(0) \| \to 0 \quad \text{as} \quad t \to 0+ \tag{1.6}$$

is that A be the infinitesimal generator of a semigroup $T(t)$ of linear bounded transformations over X_1 which is strongly continuous at the origin. The solution itself is then given by

$$x(t) = T(t)x(0), \tag{1.7}$$

where of course

$$T(t + s) = T(t)T(s), \qquad T(0) = I,$$

and

$$\frac{d}{dt} T(t)x = AT(t)x = T(t)Ax,$$

for $x \in D(A)$, where $D(A)$ denotes the domain of A. Necessary and suf-

ficient conditions for a closed operator with dense domain to be an infinitesimal generator of such a semigroup are given in [1]. We state one such result [1] due to Feller-Miyadera-Phillips: a necessary and sufficient condition for a closed linear operator with dense domain to generate a strongly continuous semigroup is that $(\lambda I - A)$ have a bounded inverse $R(\lambda, A)$ for each $\lambda > 0$, with

$$\| R(\lambda, A)^n \| \leqq M(\lambda - \omega)^{-n}, \qquad \lambda > \omega, \tag{1.8}$$

for some M, $\omega > 0$ and all n.

The formal solution of

$$\dot{x}(t) = Ax(t) + v(t), \tag{1.9}$$

by analogy with the finite dimensional case, is

$$x(t) = T(t)x(0) + \int_0^t T(t - s)v(s)\, ds. \tag{1.10}$$

However, we need to postulate some additional conditions before (1.10) is the actual solution to (1.9).

THEOREM 1.1. *Let A be the infinitesimal generator of a strongly continuous semigroup $T(t)$. Let $v(t)$ be strongly measurable and Bochner integrable in every finite interval in $(0, \infty)$. Further let $v(t) \in D(A)$ for almost every t and let $\| Av(t) \|$ be integrable in each finite interval in $(0, \infty)$. Then, for every $t > 0$ and for each initial value $x(0)$ in $D(A)$, (1.10) is the unique solution of (1.9) satisfying (1.6).*

Proof. Since $v(t)$ is strongly measurable, and $v(t) \in D(A)$ almost everywhere, $Av(t)$ is also strongly measurable. From the assumed integrability of $\| Av(t) \|$, it follows that $Av(t)$ is Bochner integrable in each finite interval. Now the integral in (1.10) is a Bochner integral. Moreover,

$$\frac{x(t + \Delta) - x(t)}{\Delta} = \frac{T(t + \Delta) - T(t)}{\Delta} x(0)$$
$$+ \frac{1}{\Delta} \int_t^{t+\Delta} T(t + \Delta - s)v(s)\, ds + \int_0^t T(t - s) \left(\frac{T(\Delta) - I}{\Delta} \right) v(s)\, ds.$$

The first term tends to $AT(t)x(0)$ as $\Delta \to 0$. Because $v(t)$ is Bochner integrable, the second term goes to $v(t)$ almost everywhere in t. In the third term, the integrand is bounded in norm by

$$\text{const.} \| Av(s) \|,$$

since

$$\left\| \left(\frac{T(\Delta) - I}{\Delta} \right) v(s) \right\| \leqq \left\| \frac{1}{\Delta} \int_0^\Delta T(\sigma)Av(s)\, d\sigma \right\| \leqq M \| AV(s) \|;$$

and hence the third term converges to

$$\int_0^t T(t - s)Av(s)\, ds.$$

But since A is closed linear, this implies that the third term tends to

$$A \int_0^t T(t - s)v(s)\, ds.$$

In other words, (1.9) is satisfied for almost all t, and in particular for every point of continuity of $v(t)$.

The uniqueness of the solution is as usual a trivial consequence of the uniqueness of the solution to the homogeneous equation. Whether the condition that $Av(t)$ be Bochner integrable is necessary is not clear. However, it is possible to show that if (1.10) is a solution of (1.9) for every t, then $v(t)$ must be in the domain of A almost everywhere in t. The situation is more complicated when $A(t)$ depends on t. The homogeneous equation

$$\dot{x}(t) = A(t)x(t) \tag{1.11}$$

has been called the "evolution equation". No necessary conditions for existence and uniqueness of solutions are available, in general, although a variety of sufficient conditions have been given [3]. One such condition of interest to us is the following. Let

$$A(t) = A + F(t), \tag{1.12}$$

where A is the infinitesimal generator of a strongly continuous semigroup and $F(t)$ is a linear bounded transformation for each t and is strongly continuous in t on each finite interval in $(0, \infty)$. Then (1.11) has a solution given by

$$x(t) = S(t, \sigma)x(\sigma), \qquad t \geqq \sigma, \qquad x(\sigma) \in D(A), \tag{1.13}$$

where $S(t, \sigma)$ is a two-parameter family of endomorphisms, $t - \sigma \geqq 0$, and such that

$$\begin{aligned} S(t, \sigma)S(\sigma, \tau) &= S(t, \tau), \qquad t \geqq \sigma \geqq \tau, \\ S(0, 0) &= I. \end{aligned} \tag{1.14}$$

From (1.12) it follows that $A(t)$ is a closed linear operator for each t and that

$$D = \bigcap_t D(A(t))$$

is dense in X. This would appear to be a minimal condition on $A(t)$. If we assume that (1.11) has a unique solution with the requisite continuity

properties, then the solution must be given by (1.13) with $S(t, \sigma)$ as in (1.14). The formal solution of the nonhomogeneous equation

$$\dot{x}(t) = A(t)x(t) + v(t)$$

now becomes for $x(\sigma) \in D$,

$$x(t) = S(t, \sigma)x(\sigma) + \int_{\sigma}^{t} S(t, \sigma)v(\sigma)\, d\sigma. \tag{1.15}$$

The analogue of Theorem 1.1 now holds if we postulate that $v(t)$ and

$$A(t)S(t, \sigma)v(t), \qquad t \geqq \sigma, \tag{1.16}$$

are Bochner integrable in each finite interval in (σ, ∞).

The solution to (1.4) can now be obtained from (1.10) and (1.15) under the appropriate sufficiency conditions on $A(t)$ and $B(t)$. In what follows, we shall be interested primarily in the time-invariant case. In particular, we shall suppose that

$$B(t) = B, \tag{1.17}$$

where B is a bounded linear transformation mapping X_2 into D, the domain of A. Then $ABu(t)$ is Bochner integrable since $u(t)$ has this property and AB is linear and bounded.

2. Optimal control problems. Of the endless variety of finite-dimensional control problems that can be extended into the present setting we shall discuss two main classes—the so-called *final-value* problem and the *time-optimal* problem, and some variations thereon.

Final value problem. This is the problem of minimizing some prescribed functional $g(x(T))$ of the final state at a fixed terminal time T, starting from a given initial state $x(0)$ at time zero, with constraints on the control $u(t)$. In this case it is convenient to introduce the space $B_p\{T; X_2\}$ of strongly measurable functions $u(t)$, $0 \leqq t \leqq T$, with range in X_2 such that

$$\int_0^T \| u(t) \|^p \, dt < \infty, \qquad 1 \leqq p \leqq \infty,$$

and constrain the control function $u(t)$ to be in some subset C therein.

Time-optimal problem. Let x_1, x_2 be two given states in X_1. A system is controllable* if it is possible to "transfer" x_1 to x_2 in some time interval; that is, to make

$$x(0) = x_1, \qquad x(T) = x_2,$$

with the control constrained to be in C. Thus controllability depends on

* The problem of determining conditions for controllability is still largely open in this setting.

x_1, x_2, and C. The time optimal problem is to find the control that yields the minimum T.

Before proceeding to a detailed examination of these problems we shall first collect certain general results pertaining to constrained extremum problems for functionals on Banach spaces. All functionals will be real valued.

THEOREM 2.1. *Let X be a reflexive Banach space. Let $g(\cdot)$ be a real-valued continuous functional which is convex on a closed bounded convex subset C of X. Then there exists an element u_0 in C such that*

$$\inf_{u \in C} g(u) = g(u_0). \tag{2.1}$$

Proof. Let $\{x_n\}$ be a sequence in C such that $g(x_n)$ is monotone decreasing (nonincreasing) and

$$\lim_n g(x_n) = \inf_{x \in C} g(x).$$

Since C is bounded, we choose a weakly convergent subsequence, renumbered as $\{x_n\}$ again, whose limit is x_0. Since C is convex and closed, x_0 belongs to C. Now a continuous convex functional is weakly lower semi-continuous,* so that

$$\underline{\lim}\, g(x_n) = \lim g(x_n) \geqq g(x_0),$$

and thus x_0 is the extremal element sought.

If X is uniformly convex, then there exists a unique element of minimal norm at which the infimum is attained. We note that the spaces $B_p(T, X_2)$ are uniformly convex for $p > 1$, provided X_2 is.

In many problems the convex set is actually characterized by convex inequalities, and we shall need a specialization of Theorem 2.1 in terms of the inequality conditions. For finite-dimensional versions, see [4], [5].

THEOREM 2.2. *Let X be a Banach space and let C be a subset of X characterized by the functional inequalities,*

$$f_i(x) \leqq 0, \qquad i = 1, \cdots, n, \tag{2.2}$$

where $f_i(\cdot)$ is continuous and convex on X. Let $f_0(\cdot)$ be a continuous convex functional on X and

$$f_0(x_0) = \inf_{x \in C} f_0(x).$$

Then there exist constants α_i, $i = 0, 1, \cdots, n$, such that

* A real-valued functional $f(\cdot)$ is said to be weakly lower semi-continuous, if whenever x_n converges weakly to x,

$$f(x) \leqq \underline{\lim} f(x_n).$$

$$\alpha_i \geqq 0, \qquad \sum_0^n \alpha_i > 0,$$

and

$$\inf_{x \in X} \sum_{i=0}^n \alpha_i f_i(x) = \sum_{i=0}^n \alpha_i f_i(x_0). \tag{2.3}$$

Proof. Let $T(x)$ be the mapping of X into E_{n+1} defined by

$$T(x) = \{f_0(x) - f_0(x_0), f_1(x) - f_1(x_0), \cdots, f_n(x) - f_n(x_0)\}.$$

Let E be the subset in E_{n+1} defined by

$$E = \{Y \mid Y \geqq T(x) \quad \text{for some} \quad x \quad \text{in} \quad X\}.$$

We note that E is convex. Indeed, if $Y_1, Y_2 \in E$, then we know that for some x_1, x_2 in X,

$$Y_1 \geqq T(x_1), \qquad Y_2 \geqq T(x_2).$$

But since $f_i(\cdot)$ are convex, $T(x)$ is convex so that

$$\alpha Y_1 + (1 - \alpha)Y_2 \geqq \alpha T(x_1) + (1 - \alpha)T(x_2) \geqq T(\alpha x_1 + (1 - \alpha)x_2),$$

showing that E is convex. Moreover E has interior points since the $f_i(\cdot)$ are assumed continuous. Next let us note that the origin in E_{n+1} is a boundary point of E. For let Y be negative (i.e., all the coordinates of Y are negative),

$$Y = \{-|y_i|\}.$$

Then suppose there is a point x in X such that

$$-|y_i| \geqq f_i(x) - f_i(x_0), \qquad i = 0, \cdots, n.$$

This would mean that

$$f_i(x) < f_i(x_0) \leqq 0, \qquad i = 1, \cdots, n,$$
$$f_0(x) < f_0(x_0),$$

which is a contradiction. Clearly any positive Y is in E. Hence we can find a supporting plane for E through the origin or, in other words, constants α_i not all zero such that

$$\sum_0^n \alpha_i(y_i) \geqq 0, \qquad \{y_i\} = Y \in E.$$

Hence also

$$\sum_0^n i[|l_i| + f_i(x) - f_i(x_0)] \geqq 0, \qquad \text{for arbitrary } |l_i|.$$

Therefore $\alpha_i \geqq 0$, $i = 0, \cdots, n$, or,

$$\sum_0^n \alpha_i f_i(x) \geqq \sum_0^n \alpha_i f_i(x_0),$$

as required.

When the functions $f_i(x)$ are not convex, there may not be a global extremum. However, a version of Theorem 2.2 still holds, since this is primarily a local statement. Thus we have (see [4] for finite dimensional versions):

THEOREM 2.3. *Let C be the subset of a Banach space X characterized by n functional inequalities,*

$$C = [x \in X \mid f_i(x) \leqq 0, \qquad i = 1, \cdots, n],$$

where $f_i(\cdot)$ are given continuous functionals. Let x_0 be in C and such that, for the continuous functional $f_0(x)$ on X,

$$\inf_{x \in C} f_0(x) = f_0(x_0).$$

Suppose the functionals are all Gateaux differentiable on a "c-star" about x_0. Then we can find $\{\alpha_i\}$,

$$\alpha_i \geqq 0, \qquad \sum_1^n \alpha_i > 0,$$

such that

$$\sum_1^n \alpha_i \delta f_i(x_0 ; h) \geqq 0 \tag{2.4}$$

for every h such that $x_0 + h$ is in the c-star about x_0, where $\delta f_i(x_0 ; h)$ denotes the Gateaux differential of f_i at x_0.

Proof. The proof is quite similar to that of the previous theorem. For any h such that $x_0 + h$ is in the c-star about x_0, let us define the function $T(h)$ with values in E_{n+1} by

$$T(h) = \{\delta f_i(x_0 ; h)\}, \qquad i = 0, \cdots, n.$$

Let E be the set in E_{n+1} such that

$$E = [Y \in E_{n+1} \mid Y \geqq T(h) \text{ for some } h].$$

Then E is clearly convex; it is actually a cone ("derived cone" in the terminology of Hestenes [5]). Next by taking h to be the zero element of X, we see that the origin of E_{n+1} is in E. Also, let Y be any negative vector,

$$Y = \{-|y_i|\}, \qquad |y_i| > 0.$$

Then Y cannot be in E. Indeed, suppose

$$-|\, y_i \,| \geqq \delta f_i(x_0\,;\, h).$$

Then since, for $0 < \zeta < 1$,

$$f_i(x_0 + \zeta h) = \sum_0^\infty \zeta^n \frac{\delta^n f_i(x_0\,;\, h)}{n!},$$

it follows that for ζ sufficiently small

$$f_i(x_0 + \zeta h) < f_i(x_0) \leqq 0, \qquad i = 1, \cdots, n,$$

and

$$f_0(x_0 + \zeta h) < f_0(x_0),$$

which is impossible. As in Theorem 2.2, we note that E has interior points and the origin is a boundary point. Hence we can find a supporting plane through the origin. The rest of the argument proceeds as in Theorem 2.2.

COROLLARY. *Suppose $f_i(\cdot)$ are all Frechet differentiable in some sphere about x_0. Let $\nabla f_i(x_0)$ denote the gradient so that the Frechet derivative*

$$\delta f(x_0\,,\, h) = \nabla f_i(x_0)(h), \qquad \nabla f_i(x_0) \in X^*.$$

Then we can find constants α_i such that $\alpha_i \geqq 0$, $\sum_0^n \alpha_i > 0$, and

$$\sum_0^n \alpha_i \nabla f_i(x_0) = 0. \tag{2.5}$$

Proof. From the theorem it follows that for the α_i therein,

$$\sum_0^n \alpha_i \nabla f_i(x_0)(h) \geqq 0 \quad \text{for every} \quad h \quad \text{in} \quad X,$$

so that

$$\sum_0^n \alpha_i \nabla f_i(x_0) = 0,$$

as required.

A simple case. The simplest case which is of interest in the applications arises when we set

$$f_0(u) = \| Lu - y \|, \tag{2.6}$$

where L is a linear bounded operator mapping X_2 into X_1 and y is a fixed nonzero element of X_1. Moreover we shall take X_1 and X_2 to be Hilbert spaces and denote them by H_1 and H_2, respectively, to indicate this. We shall specify the constraint to be

$$f_1(u) = \| u \| - M, \qquad M > 0. \tag{2.7}$$

We can apply Theorems 2.1 and 2.2 since $f_0(\cdot)$ and $f_1(\cdot)$ are continuous

and convex. Hence a (global) minimum exists. Let one of these (if there is more than one) be denoted u_0. Then we can find numbers λ_0, λ_1 such that

$$\lambda_0 f_0(u) + \lambda_1 f_1(u) \geqq \lambda_0 f_0(u_0) + \lambda_1 f_1(u_0),$$

$$0 \leqq \lambda_0 \leqq 1, \quad \lambda_1 = 1 - \lambda_0,$$

or,

$$(2.8)\quad \lambda_0 \| Lu_0 - y \| + \lambda_1 \| u_0 \| \leqq \lambda_0 \| Lu - y \| + \lambda_1 \| u \|, \quad u \in H_2, \cdots.$$

Clearly, λ_0 cannot be zero, since u_0 cannot be zero, y being nonzero. Hence $f_0(u_0)$ is zero if and only if λ_1 is zero. Again let us note that $f_0(\cdot)$ and $f_1(\cdot)$ are now actually Frechet differentiable, so that

$$\frac{\lambda_0}{f_0(u_0)} (L^*Lu_0 - L^*y) + \frac{\lambda_1 u_0}{\| u_0 \|} = 0.$$

It then follows that, for some $k_0 \geqq 0$,

$$(2.9)\quad L^*Lu_0 - L^*y + k_0 u_0 = 0.$$

If $k_0 = 0$, the minimum is actually zero. It may be noted that if k_0 is positive, then

$$(2.10)\quad u_0 = (L^*L + k_0 I)^{-1} L^*y,$$

since L^*L is nonnegative and self-adjoint, so that $(L^*L + k_0 I)$ has a bounded inverse. Again, for k_0 positive,

$$u_0 = \frac{1}{k_0} L^*(Lu_0 - y),$$

so that u_0 is in the range of L^*. Thus the u_0 given by this formula is autotically the unique element of minimal norm that minimizes (2.6) subject to (2.7). When k_0 is zero, it is obvious that we can take any sequence of positive numbers k_n, whose limit is zero and let (see (2.9))

$$(2.11)\quad \begin{aligned} u_n &= (L^*L + k_n I)^{-1} L^*y \\ &= (L^*L + k_n I)^{-1} L^*Lu_0 . \end{aligned}$$

Since L^*L is self-adjoint and nonnegative, it follows from this that u_n converges to the projection of u_0 on the orthogonal complement of the null space of L^*L. If we call this projection v_0, we note that v_0 is the unique element of minimal norm that minimizes (2.6) subject to (2.7). Also, of course

$$\lim_n f_0(u_n) = f_0(u_0) = f_0(v_0).$$

This generalizes the results in [6] and [7]. In particular the operator L is not required to be compact.

Retaining (2.6) we may generalize (2.7) so that $f_1(\cdot)$ is now the support function of a closed bounded convex set C (with the origin as an interior point) which is smooth enough so that $f_1(\cdot)$ is Frechet differentiable. Then (2.9) generalizes to

$$L^*Lu_0 + k_0\nabla f_1(u_0) = L^*y, \tag{2.12}$$

which is a nonlinear equation when k_0 is positive. If k_0 is zero, we can, as before, use the sequence u_n determined by

$$L^*Lu_n + k_n\nabla f_1(u_n) = L^*y, \qquad 0 \leqq k_n \to 0.$$

We note parenthetically that $\nabla f_1(\cdot)$ is a "monotone" operator, and so indeed is $L^*L + k_0\nabla f_1(\cdot)$ in (2.12). We have already proved the existence of a unique solution for (2.12), although this could be established directly as in [8], where some related considerations are also to be found. While iteration methods for solving (2.12) are of interest, our concern here is with the more difficult problem of producing a minimizing sequence for (2.6) and an algorithm for this will be given later.

3. Final value problems. Let us now return to a detailed consideration of a class of final value problems. Let the state equation be given by (1.4). Let us also assume that the system is time-invariant since extensions to time-varying systems that have unique solutions can be readily made. Thus we assume that the state equation is

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{3.1}$$

where, for each t, the state $x(t)$ is in X_1 and the control $u(t)$ is in X_2, A is the infinitesimal generator of a strongly continuous semigroup $S(t)$, and B maps X_2 into the domain of A. We shall assume that X_1 and X_2 are reflexive spaces. We shall consider the final value problem of minimizing for fixed T, y:

$$\| x(T) - y \|, \qquad y \in X_1, \tag{3.2}$$

starting with the known initial state $x(0)$ at time zero, with the constraint on the control described as follows: consider the space $B_p[X_2\,;\,(0, T)]$ (shortened to $B_p[X_2\,;\,T]$ later on) of strongly measurable functions $u(t)$ with range in X_2 such that

$$\int_0^T \| u(t) \|^p\, dt < \infty, \qquad \text{for some } p > 1,\ p < \infty.$$

Then the control $u(\cdot)$ is subject to being in a closed bounded convex set

C in $B_p[X_2\ ;\ T]$. For example, C may consist of all $u(\cdot)$ such that

$$\operatorname*{ess\,sup}_{0 \leqq t \leqq T} \| u(t) \| \leqq M < \infty, \tag{3.3}$$

or such that

$$\int_0^T \| u(t) \|^{p'}\, dt \leqq M < \infty, \qquad p' \geqq p. \tag{3.4}$$

Note that if C is defined by (3.3), the origin in $B_p[X_2\ ;\ T]$ is not an interior point of C. To handle this problem we can write, following (1.10),

$$x(T) = \int_0^T S(T - \sigma)Bu(\sigma)\, d\sigma + S(T)x(0). \tag{3.5}$$

If we define the mapping

$$L(T)u = \int_0^T S(T - \sigma)Bu(\sigma)\, d\sigma,$$

from $B_p[X_2\ ;\ T]$ into X_1, then $L(T)$ is clearly linear bounded. Setting

$$f_0(u) = \| L(T)u + S(T)x(0) - y \|, \tag{3.6}$$

we note that $f_0(\cdot)$ is a continuous convex functional on $B_p[X_2\ ;\ T]$. Hence by Theorem 2.1, we are assured of a global minimum subject to $u(\cdot)$ being in C. In fact, we can find a nonnegative constant k_0 such that, for the minimizing element u_0,

$$f_0(u_0) + k_0 f_1(u_0) \leqq f_0(u_0) + k_0 f_1(u) \quad \text{for every } u \text{ in } B_p[X_2\ ;\ T], \tag{3.7}$$

and assuming Frechet differentiability for $f_1(\cdot)$,

$$\nabla f_0(u_0) + k_0 \nabla f_1(u_0) = 0, \tag{3.8}$$

where $\nabla f_0(\cdot)$ and $\nabla f_1(\cdot)$ range over $B_q[X_2\ ;\ T]$, $1/p + 1/q = 1$. Next let us note that the set $L(T)C$ is also convex and bounded whenever C is bounded. Minimization of (3.6) can then be viewed as minimizing the distance (in X_1) from $\bar{y} = y - S(T)x(0)$ to $L(T)C$. Thus we know that there is an element y_0 in the closure of $L(T)C$ in X_1 such that

$$\min \| y - \bar{y} \| = \| y_0 - y \| = m_0.$$

But since we know that u_0 already provides a global minimum, we have

$$y_0 = L(T)u_0.$$

Again suppose k_0 in (3.7) is positive. From (3.7) it follows that if k_0 is positive, then y_0 is a bounding point of $L(T)C$. Then we can find a support-

ing plane through y_0 and corresponding (nonzero) functional x^* in X_1^* such that

$$\text{Re } [x^*(y) - x^*(y_0)] \leqq 0, \qquad y \in L(T)C,$$

and

$$\text{Re } [x^*(\bar{y})] \geqq \text{Re } [x^*(y_0)].$$

We shall, for convenience in what follows, drop the Re in such inequalities. It is clear that we can, by a suitable multiplying factor, arrange matters so that

$$x^*(y) \leqq x^*(\bar{y}) - m_0 = x^*(y_0), \qquad y \in L(T)C.$$

Now, let $y^* = L(T)^* x^*$, so that

$$y^* \in B_q[X_2^*; T], \qquad 1/p + 1/q = 1. \tag{3.9}$$

Then we have

$$y^*(u) \leqq y^*(u_0), \qquad u \in C. \tag{3.10}$$

It follows from this that u_0 is a boundary point of C. To characterize u_0 further, let us assume that X_2 is a Hilbert space. Then

$$y^*(u) = \int_0^T [y^*(\sigma), u(\sigma)]\, d\sigma,$$

where

$$y^*(\sigma) = B^* S(T - \sigma)^* x^*,$$

and (3.10) becomes

$$\int_0^T [y^*(\sigma), u(\sigma)]\, d\sigma \leqq \int_0^T [y^*(\sigma), u_0(\sigma)]\, d\sigma, \qquad u \in C. \tag{3.11}$$

In special cases, (3.11) is enough to determine $u_0(\cdot)$. Thus let us consider the case where C is characterized by (3.3). Let us define

$$v(\sigma) = \frac{My^*(\sigma)}{\| y^*(\sigma) \|}, \qquad y^*(\sigma) \neq 0. \tag{3.12}$$

Then $v(\cdot)$ belongs to C, and, in fact, is on the boundary of C. By the usual limiting arguments, it is readily seen that

$$u_0(\sigma) = v(\sigma), \qquad y^*(\sigma) \neq 0. \tag{3.13}$$

If C is characterized by (3.4), we set

$$w(\sigma) = k \frac{y^*(\sigma)}{\| y^*(\sigma) \|^{2-q'}}, \qquad 1/p' + 1/q' = 1, \tag{3.14}$$

and

$$k = M\left[\int_0^T \| y^*(\sigma) \|^{q'}\right]^{p'/q'}.$$

Then

$$\int_0^T \| w(\sigma) \|^{p'} d\sigma = M.$$

Now for $u(\cdot)$ in C defined by (3.4),

$$y^*(u) \leqq M\left[\int_0^T \| y^*(\sigma) \|^{q'} d\sigma\right]^{1/q'}.$$

But

$$y^*(w) = M\left[\int_0^T \| y^*(\sigma) \|^{q'} d\sigma\right]^{1/q'},$$

and there is at most one element in C satisfying (3.10). Hence,

$$u_0(\sigma) = w(\sigma). \tag{3.15}$$

We further note that since

$$y^*(\sigma) = B^*S(T - \sigma)^*x^*,$$

$y^*(\sigma)$ is actually continuous in σ, and for any x in X_1,

$$y^*(\sigma)(x) = x^*(S(T - \sigma)Bx),$$

and

$$\begin{aligned} \frac{d}{d\sigma} y^*(\sigma)(x) &= -x^*[AS(T - \sigma)Bx] \\ &= -A^*x^*[S(T - \sigma)Bx] \\ &= -A^*y^*(\sigma)(x), \qquad 0 \leqq \sigma \leqq T, \end{aligned} \tag{3.16}$$

or $y^*(\sigma)$ is actually weakly differentiable and indeed satisfies the differential equation (3.1). This is one direction of generalization of the Pontryagin maximum principle. It is unlikely that results similar to (3.14) and (3.15) can be obtained when X_1 is not a Hilbert space or when C is allowed to be an arbitrary closed bounded convex set. It is, however, of interest to consider a case where C is no longer bounded. Let C in $B_p[X_2\,;\,T]$ be characterized by

$$\int_0^T \| u(\sigma) \| d\sigma \leqq M < \infty, \qquad p > 1, \tag{3.17}$$

and let us continue to assume that X_2 is a Hilbert space. We note that the origin is an interior point of C. Let S_n denote the sphere of radius n, and let

$$C_n = S_n \cap C.$$

Then, since C_n is closed, bounded, and convex, there is a unique element, which we shall denote by u_n, that provides the minimum of (3.6) under the constraint that $u \in C_n$. We note next that

$$\lim_n f_0(u_n) = \inf_{u \in C} f_0(u), \tag{3.18}$$

and $f_0(u_n)$ is a monotone decreasing sequence. The sequence $\{u_n\}$ need not converge even weakly, and there may not be an element in C such that the infimum in (3.18) is attained. It is clear, in fact, that a necessary and sufficient condition for the existence of a minimizing element in C is that

$$\| u_n \| \leqq \text{const.} < \infty.$$

The problem of finding u_n can be handled by setting

$$f_1(u) = \int_0^T \| u(\sigma) \| \, d\sigma - M,$$

$$f_2(u) = \| u \| - n,$$

and noting that the minimum must satisfy

$$f_0(u) + \lambda_1 f_1(u) + \lambda_2 f_2(u) \geqq f_0(u_n) + \lambda_1 f_1(u_n) + \lambda_2 f_2(u_n),$$

$$\lambda_1 + \lambda_2 > 0, \qquad 0 \leqq \lambda_1, \lambda_2.$$

It may be noted, however, that $f_1(\cdot)$ is no longer Frechet differentiable so that (2.5) does not hold. However, (3.10) still holds, provided $L(T)u_n$ is a bounding point of $L(T)C_n$.

When C is characterized by (3.17) we can also proceed in a slightly different manner, if we can assume that X_2 is a Hilbert space. For this we note that C is a closed *bounded* convex subset of $B_1[X_2 ; T]$. But this space is no longer reflexive; hence Theorem 2.1 does not apply. However, following (3.18), let

$$\lim f_0(u_k) = \inf_{u \in C} f_0(u).$$

Then since

$$f_0(u_k) = \| L(T)u_k - \bar{y} \|, \tag{3.19}$$

where $\| \quad \|$ now denotes the norm in X_1, it follows that, if X_1 is reflexive, for any x^* in $X_1{}^*$,

$$x^*[L(T)u_k] = \int_0^T [B^*S(T-\sigma)^*x^*, u_k(\sigma)]\, d\sigma,$$

where

$$y^*(\sigma) = B^*S(T-\sigma)^*x^*$$

is actually continuous in σ in the norm of X_2. From the Helly theorem extended to $BV[X_2\,;\,T]$, the space of functions of (strong) bounded variation, we note that, given any subsequence of $\{u_k\}$, we can find a further subsequence $\{u_{nk}\}$ such that

$$\int_0^T [B^*S(T-\sigma)^*x^*, u_{nk}(\sigma)]\, d\sigma \to \int_0^T [B^*S(T-\sigma)^*x^*, du_0(\sigma)], \tag{3.20}$$

where $u_0(\cdot)$ is now an element of $BV[X_2\,;\,T]$. In other words, we consider the mapping

$$\int_0^T S(T-\sigma)B\, d\beta(\sigma) = L(T),$$

where $\beta(\cdot)$ is an element of $BV[X_2\,;\,T]$, and note that $L(T)$ is linear bounded. Moreover C is a convex, closed bounded subset of $BV[X_2\,;\,T]$ and $L(T)C$ is convex and bounded. Hence there is an element y_0 in the closure of $L(T)C$ such that

$$\| y_0 - \bar{y} \| = \inf \| y - \bar{y} \| = m_0\,.$$

But from (3.19),

$$\lim \| L(T)u_k - \bar{y} \| = \| y_0 - \bar{y} \|,$$

and from (3.20) it follows that for the subsequence u_{nk}, for any x^* in ${X_1}^*$,

$$\lim x^*[L(T)u_{nk} - \bar{y}] = x^*[L(T)u_0 - \bar{y}],$$

and hence

$$\| L(T)u_0 - \bar{y} \| \leqq \underline{\lim} \| L(T)u_{nk} - \bar{y} \| = \| y_0 - \bar{y} \|.$$

It follows then that

$$m_0 = \| L(T)u_0 - \bar{y} \|,$$

and this is independent of what subsequence of $\{u_n\}$ we chose. If $\bar{y}$ is not an interior point of $L(T)C$, we also have that for some x^* in ${X_1}^*$,

$$\int_0^T [B^*S(T-\sigma)^*x^*, u(\sigma)]\, d\sigma \leqq \int_0^T [B^*S(T-\sigma)^*x^*, du_0(\sigma)], \tag{3.21}$$

$u \in C$. If in addition X_1 is reflexive and uniformly convex, then y_0 is unique and

$$\lim_n L(T)u_n = L(T)u_0 = y_0 . \tag{3.22}$$

"Feedback" solutions. So far we have been concerned with the determination of the control $u(\sigma)$ as a solution of functional equations or inequalities. In many problems it is desirable to obtain "feedback" control; that is, with reference to (3.1), we want to determine $u(\sigma)$ as a "function" of $x(\sigma)$. The question of whether this is possible in every case is largely unsettled even in the classical finite-dimensional state space problems. Here we shall consider a specific class of problems in which (3.1) is (slightly generalized to)

$$\dot{x}(t) = Ax(t) + Bu(t) + Z(t), \tag{3.23}$$

where $Z(t)$ is in the domain of A for each t and $Z(t)$ and $AZ(t)$ are Bochner integrable on finite intervals. We shall assume that both X_1 and X_2 are Hilbert spaces and that we want to minimize (3.2) subject to the control $u(\cdot)$ being in $B_2[X_2 ; T]$ and such that

$$\int_0^T \| u(t)\|^2 \, dt \leqq M < \infty. \tag{3.24}$$

We have seen that the solution $u_0(\cdot)$ is then given by (2.12) which now reads

$$(L^*(T)L(T) + k_0 I)u_0(\sigma) = L(T)^* \bar{y}, \tag{3.25}$$

where

$$\bar{y} = y - S(T)x(0) - \int_0^T S(T - \sigma)Z(\sigma) \, d\sigma.$$

First let us assume that k_0 is positive. Then we know that

$$u_0(\sigma) = L(T)^* y_0$$

for some y_0 in X_1, where y_0 satisfies

$$L(T)L^*(T)y_0 + k_0 y_0 = \bar{y}. \tag{3.26}$$

For each t let us define a linear bounded operator $R(t)$ by

$$R(t)x = \int_0^t S(T - \sigma)BB^*S(T - \sigma)^* x \, d\sigma, \quad 0 \leqq t \leqq T.$$

Then

$$u_0(\sigma) = B^*S(T - \sigma)^* y_0 = L(T)^*[R(T) + k_0 I]^{-1} y \tag{3.27}$$

We shall now show that we can find a linear bounded operator $\theta(t)$ for each t, $0 \leqq t \leqq T$, such that

$$(3.28)\quad u_0(t) = \theta(t)\left[y - S(T-t)x(t) - \int_t^T S(T-\sigma)Z(\sigma)\,d\sigma\right],$$

where, in fact,

$$(3.29)\qquad \theta(t) = B^*S(T-t)^*[R(T) - R(t) + k_0 I]^{-1},$$

where the inverse indicated is a linear bounded operator. It is clear that (3.28) provides the "feedback" control sought, in that it depends only on $x(t)$ and other given or a priori known data. In fact the term in square brackets in (3.28) is the difference between the "target" y and what the state at T would be if there were no control after t. To prove (3.28), let $x(t)$ be the unique solution of

$$\dot{x}(t) = Ax(t) + Bu_0(t) + Z(t),$$

where $u_0(t)$ is given by (3.27). Then

$$x(t) = S(t)x(0) + \int_0^t S(t-\zeta)BB^*S(T-\zeta)^*[R(T) + k_0 I]^{-1}\bar{y}\,d\zeta + \int_0^t S(t-\zeta)Z(\zeta)\,d\zeta,$$

so that in (3.28), the quantity in square brackets equals

$$y - S(T)x(0) - \int_0^t S(T-\zeta)BB^*S(T-\zeta)^*[R(T) + k_0 I]^{-1}\bar{y}\,d\zeta - \int_0^t S(T-\zeta)Z(\zeta)\,d\zeta - \int_t^T S(T-\sigma)Z(\sigma)\,d\sigma$$

$$= \bar{y} - R(t)[R(T) + k_0 I]^{-1}\bar{y}$$

$$= (R(T) - R(t) + k_0 I)(R(T) + k_0 I)^{-1}\bar{y},$$

so that

$$\theta(t)\left[y - S(T-t)x(t) - \int_t^T S(T-\sigma)Z(\sigma)\,d\sigma\right] = L(t)(R(T) - R(t) + k_0 I)[R(T) + k_0 I]^{-1}\bar{y} = B^*S(T-t)^*[R(T) + k_0 I]^{-1}\bar{y} = u_0(t),$$

as required. It may be noted that

$$(3.30)\qquad [R(T) - R(t)]x = \int_t^T S(T-\sigma)BB^*S(T-\sigma)^*x\,d\sigma,$$

so that $(R(T) - R(t))$ is self-adjoint and nonnegative for $0 \leqq t < T$. When k_0 is zero, we have already noted that $u_0(t)$ is the limit in $B_2[X_2 ; T]$ of elements of the form

$$L(T)^*[R(T) + k_nI]^{-1}\bar{y}, \qquad 0 \leqq k_n \to 0;$$

and hence the feedback solution can be written as the limit of

$$u_n(t) = \theta_n(t)\left[y - S(T - t)x_n(t) - \int_t^T S(T - \sigma)Z(\sigma)\, d\sigma\right],$$

where

$$\theta_n(t)[R(T) - R(t) + k_nI] = B^*S(T - t)^*,$$

$$\dot{x}_n(t) = Ax_n(t) + Bu_n(t) + Z(t).$$

Also, for k_0 positive,

$$\dot{x}(t) = Ax(t) + B\theta(t)\left[y - S(T - t)x(t) - \int_t^T S(T - \sigma)Z(\sigma)\, d\sigma\right]$$

$$+ Z(t) = [A - B\theta(t)S(T - t)]x(t) + B\theta(t)y$$

$$- B\theta(t)\int_t^T S(T - \sigma)Z(\sigma)\, d\sigma + Z(t)$$

has a unique solution for each initial value $x(0)$.

The stochastic version of this problem is of interest since the feedback aspect becomes an essential in the formulation of the problem. For example, if in (3.23) the term $Z(t)$ is identified as "noise", then the control cannot contain the third term in (3.28) and in fact has to be determined as operations on the "observed data" $x(t)$ alone. To illustrate the considerations involved, we shall go into some detail on a specific example of a stochastic control problem where the state equation is given by

$$\dot{\zeta}(t) = A\zeta(t) + Bu(t), \tag{3.32}$$

where $\zeta(t)$ is an $n \times 1$ matrix, A is an $n \times n$ matrix, B is an $n \times m$ matrix and $u(t)$ is an $m \times 1$ matrix, and

$$y(t) = \zeta(t) + n(t) \tag{3.33}$$

is observed, where $n(t)$ is random noise (n-dimensional stochastic process) with zero mean. We may assume that $\zeta(0)$ is a random variable with known mean. Similarly we may assume that the desired state at time T is also a random variable ζ but with known mean y. The optimal control has to be dependent on the observed variable $y(t)$ and hence is also random. The constraints take the form

$$\int_0^T |E(u(t))|^2\, dt \leqq M_1,$$

$$\int_0^T \mathrm{Var}\,[u(t)]\, dt \leqq M_2,$$

and we want to minimize

$$E\,\|\zeta - \zeta(T)\|^2. \tag{3.34}$$

We shall show that the optimal control can be expressed as operations on $y(t)$. First of all, substituting (3.33) into (3.32) we obtain, assuming $n(t)$ is differentiable,

$$\dot{y}(t) = Ay(t) + Bu(t) + (\dot{n}(t) - An(t)), \tag{3.35}$$

and the similarity to (3.23) is obvious. To make it more precise, let us introduce the space X_1 as

$$X_1 = L_2[\Omega, B, \mu],$$

the space of vector random variables of n dimensions with finite second moments in which the random variables $n(t)$, ζ, $\zeta(0)$ are defined. We also introduce

$$X_2 = L_2'[\Omega, B, \mu],$$

the space of vector random variables of m dimensions with finite second moment defined on the same measure space. To avoid confusion we shall use the notation

$$x(t) \sim y(t, \omega),$$

$$V(t) \sim u(t, \omega),$$

$$Z(t) \sim \dot{n}(t, \omega) - An(t, \omega),$$

the "sample point" $\omega \in \Omega$ denoting the fact that these functions are now random variables, so that (3.35) becomes

$$\dot{x}(t) = Ax(t) + BV(t) + Z(t). \tag{3.36}$$

When deterministic variables are involved, we may consider them as functions defined to be constants over Ω, and hence consider them also as elements of X_1 or X_2 as required. For any element x in X_1 or X_2, let $E[x]$ denote the expectation of x. This expectation corresponds to a (finite-dimensional) linear functional on X_1 or X_2. Paraphrasing (3.34), we have to minimize

$$\|\zeta - x(T)\|^2, \tag{3.37}$$

subject to the constraints

$$\int_0^T \| E[V(t)] \|^2 \, dt \leqq M_1 ,$$
$$\int_0^T \| V(t) - E[V(t)] \|^2 \, dt \leqq M_2 . \tag{3.38}$$

We shall now show that the optimal $V(t)$ can be determined as

$$V(t) = L_1(t)[\hat{\xi} - S(T-t)(x(t) - p(t)] + L_2(t)[y - S(T-t)p(t)], \tag{3.39}$$

where $L_1(t)$ and $L_2(t)$ are continuous matrix functions, $\hat{\xi} = \xi - y$, and $p(t)$ is deterministic and is the solution of

$$\dot{p}(t) = Ap(t) + BL_2(t)[y - S(T-t)p(t)], \quad p(0) = E[\xi(0)]. \tag{3.40}$$

First of all, we let

$$m(t) = E[x(t)].$$

Then

$$\dot{m}(t) = Am(t) + BL_1(t)[-S(T-t)(m(t) - p(t))] + BL_2(t)[y - S(T-t)p(t)],$$

so that

$$\dot{m}(t) - \dot{p}(t) = A(m(t) - p(t)) + BL_1(t)[-S(T-t)(m(t) - p(t))]. \tag{3.41}$$

But

$$m(0) = E(x(0)) = E(\xi(0)) = p(0),$$

so that by the uniqueness of the solution of (3.41),

$$m(t) = p(t), \qquad 0 \leqq t \leqq T,$$

and hence

$$\dot{m}(t) = Am(t) + BL_2(t)[y - S(T-t)m(t)], \tag{3.42}$$

$$\dot{\eta}(t) = A\eta(t) + BL_1(t)[\hat{\xi} - S(T-t)\eta(t)], \tag{3.43}$$

where

$$\eta(t) = x(t) - m(t).$$

But the criterion (3.37) can now be written as

$$\| y - m(T) \|^2 + \| \hat{\xi} - \eta(T) \|^2, \tag{3.44}$$

and each term can now be minimized separately, each with the constraint read off from (3.38). From the previous theory, we know that a feedback control of the required form can be found. Of course the constant k_0 will in general be different for $L_1(t)$ and $L_2(t)$.

This example by no means exhausts the class of stochastic problems, but the main feature that the form of the control has to be "feedback" or involve operations on observed data is of course essential in all such problems, and our purpose is merely to indicate how the abstract space set-up gives us a convenient frame-work for such problems.

4. Time-optimal problems. We shall examine the time-optimal problem for the system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{4.1}$$

where A is the infinitesimal generator of a strongly continuous semigroup $S(t)$, X_1 and X_2 are Hilbert spaces, and B, as before, is a linear bounded mapping of X_2 into the domain of A. The constraint is taken as

$$\| u(t)\| \leqq M \quad \text{almost everywhere.} \tag{4.2}$$

Let $x(0)$ be given, and let it be assumed that we can find a control satisfying (4.2) such that, for some finite time T_1,

$$x(T_1) = x_1 .$$

We want, first, to show that there exists a control that takes $x(0)$ to the origin in minimum time, and, secondly, to determine the conditions that this optimal control must satisfy. It is of course assumed that $u(t)$ is strongly measurable. The classical results on this problem, due to Gamkrelidze et al., are documented in [9].

The existence of the optimal control will be proved first. Thus let T be the minimum time and let T_n be a sequence monotone decreasing and converging to T and let $u_n(\cdot)$ be the corresponding controls. Then

$$\| u_n(\sigma)\| \leqq M \quad \text{almost everywhere in} \quad (0, T_n),$$

and

$$\begin{aligned} x_1 &= S(T_n)x(0) + \int_0^{T_n} S(T_n - \sigma)Bu_n(\sigma)\, d\sigma \\ &= S(T_n)x(0) + \int_0^{T} S(T_n - \sigma)Bu_n(\sigma)\, d\sigma \\ &\qquad + \int_T^{T_n} S(T_n - \sigma)Bu_n(\sigma)\, d\sigma. \end{aligned} \tag{4.3}$$

It is clear that

$$S(T_n)x(0) \to S(T)x(0),$$

and that the third term in (4.3) goes to zero, so our main concern is with the second term. Let us now consider $B_2[T; X_2]$, and consider $u_n(\cdot)$ as elements in this space. Let C be the set in $B_2[T; X_2]$ defined by (4.2). Then $u_n(\cdot)$ will be in C and C is closed and convex. Since C is weakly compact, we can find a weakly convergent subsequence (renumber it $\{u_n(\cdot)\}$ for convenience) converging to $u_0(\cdot)$. But since C is closed and convex, $u_0(\cdot)$ must also belong to C, so that $u_0(\cdot)$ satisfies (4.2). For any x in X_1,

$$B^*S(T_n - \sigma)^*x \in B_2[T; X_2],$$

since $S(t)^*$ is also a strongly continuous semigroup because X_1 is a Hilbert space. Now for any x in X_1,

$$\left[\int_0^T S(T_n - \sigma)Bu_n(\sigma)\, d\sigma, x\right] - \left[\int_0^T S(T - \sigma)Bu_0(\sigma)\, d\sigma, x\right]$$

$$= \int_0^T [u_n(\sigma) - u_0(\sigma), B^*S(T - \sigma)^*x]\, d\sigma$$

$$+ \int_0^T [u_n(\sigma), B^*S(T - \sigma)^*(S(T_n - T)x - x)]\, d\sigma.$$

The first term goes to zero by the weak convergence of the $u_n(\cdot)$. The second term, in magnitude, does not exceed

$$\text{const.} \parallel S(T_n - T)x - x \parallel,$$

and tends to 0 as $n \to \infty$ by the strong continuity of the semigroup $S(t)$. Hence, for every x in X_1,

$$[x_1, x] = \left[S(T)x(0) + \int_0^T S(T - \sigma)Bu_0(\sigma)\, d\sigma, x\right],$$

and therefore

$$S(T)x(0) + \int_0^T S(T - \sigma)Bu_0(\sigma)\, d\sigma = x_1,$$

or $u_0(\cdot)$ is the sought optimal control.

The next step is to characterize $u_0(\cdot)$. Here we shall eventually need to make more assumptions on the semigroup. To simplify the notation, let us now set $x(0) = 0$. For each t let us define $\Omega(t)$ to be the set

$$\Omega(T) = \left\{y \;\middle|\; \int_0^T S(T - \sigma)Bu(\sigma)\, d\sigma = y, \parallel u(\sigma) \parallel \leqq M \text{ a.e.}\right\}.$$

Then $\Omega(T)$ is convex and closed. For let

$$\int_0^T S(T - \sigma)Bu_n(\sigma)\, d\sigma = y_n ,$$

and let

$$\| y_n - y_0 \| \to 0.$$

With C defined in $B_2[X_2 ; T]$ as before, we know that we can find an element $V(\cdot)$ in C such that a subsequence of $\{u_n(\cdot)\}$ (renumbered $\{u_n\}$ again) converges weakly in $B_2[X_2 ; T]$ to $V(\cdot)$. But for any x in X_1 ,

$$B^*S(T - \sigma)^*x, \qquad 0 \leqq \sigma \leqq T,$$

is in $B_2[X_2 ; T]$, and hence

$$[y_n , x] = \int_0^T [u_n(\sigma), B^*S(T - \sigma)^*x]\, d\sigma;$$

and taking limits on each side,

$$\begin{aligned}[y_0 , x] &= \int_0^T [V(\sigma), B^*S(T - \sigma)^*x]\, d\sigma \\ &= \int_0^T [S(T - \sigma)BV(\sigma)\, d\sigma, x],\end{aligned}$$

or $\Omega(T)$ is closed. Let T_n be a monotone increasing sequence of positive numbers converging to T. Let $u_0(\cdot)$ be an optimal control corresponding to x_1 , so that

$$x_1 = \int_0^T S(T - \sigma)Bu_0(\sigma)\, d\sigma,$$

and x_1 does not then belong to $\Omega(T_n)$ for any n. But $\Omega(T_n)$ is convex and closed and bounded. Since X_1 is a Hilbert space there is a unique element, say y_n , in $\Omega(T_n)$ that is closest to x_1 ,

$$\| x_1 - x \| \geqq \| x_1 - y_n \| > 0, \qquad x \in \Omega(T_n). \tag{4.4}$$

Now let

$$x_0(t) = \int_0^t S(t - \sigma)Bu_0(\sigma)\, d\sigma.$$

In particular then, we have

$$\| x_1 - x_0(T_n) \| \geqq \| x_1 - y_n \| ; \tag{4.5}$$

but the left side goes to zero, and hence y_n converges (strongly) to x_1 .

Again from (4.4) it follows that (taking real parts, as usual)

$$[x_1 - y_n, x] \leqq [x_1 - y_n, y_n] < [x_1 - y_n, x_1], \qquad x \in \Omega(T_n), \tag{4.6}$$

or, equivalently, for every $u(\cdot)$ in C, setting

$$p_n = x_1 - y_n, \qquad y_n = \int_0^{T_n} S(T_n - \sigma)Bu_n(\sigma)\, d\sigma,$$

we have

$$\int_0^{T_n} [u(\sigma), B^*S(T_n - \sigma)^*p_n]\, d\sigma \leqq \int_0^{T_n} [u_n(\sigma), B^*S(T_n - \sigma)^*p_n]\, d\sigma. \tag{4.7}$$

Since y_n is unique and $u(\cdot)$ in C is arbitrary, it follows that if we set

$$y_n(\sigma) = k_nS(T_n - \sigma)^*p_n,$$

where k_n is an arbitrary positive number, then (except possibly for a set of measure zero)

$$u_n(\sigma) = \frac{MB^*y_n(\sigma)}{\| B^*y_n(\sigma) \|} \text{ if } B^*y_n(\sigma) \neq 0, \qquad 0 \leqq \sigma \leqq T_n. \tag{4.8}$$

It should be noted that (4.8) is independent of the arbitrary positive constant k_n.

Suppose now that the range of the operator $L(t)$ mapping $B_2[X_2; t]$ into X_1 defined by

$$L(t)u = x, \qquad x = \int_0^t S(t - s)Bu(s)\, ds,$$

is finite-dimensional in a neighborhood of T so that the p_n are confined to a fixed finite-dimensional subspace of X_1. In this case, if we set

$$k_n = \frac{1}{\| p_n \|},$$

we can find a subsequence of k_np_n that converges strongly to an element p_0, which must automatically be nonzero, in fact of norm one; and from (4.8) and (4.5) it follows that

$$x_1 = L(T)v_0,$$

where v_0 is an optimal control and

$$v_0(s) = \frac{MB^*y_0(s)}{\| B^*y_0(s) \|}, \qquad B^*y_0(s) \neq 0, \qquad 0 \leqq s \leqq T,$$

where

$$y_0(s) = S(T - s)^*p_0. \tag{4.9}$$

Next let us consider the case where the operator $L(T)$ is compact, and X_1 is not finite-dimensional. Then we note that the convex set $L(T)C$ cannot have interior points. Hence in particular x_1 is automatically a boundary point. Let $\{z_n\}$ be a sequence of points in the complement of $L(T)C$ that converges to x_1. Let y_n be the point in $L(T)C$ closest to z_n. Then

$$y_n = L(T)u_n, \qquad u_n \in C,$$

and

$$\| z_n - y_n \| \leqq \| x_1 - z_n \| \to 0.$$

Moreover for every u in C,

$$[L(T)u, z_n - y_n] \leqq [L(T)u_n, z_n - y_n],$$

and, as before, letting $p_n = k_n(z_n - y_n)$, this implies that

$$u_n(s) = \frac{MB^*h_n(s)}{\| B^*h_n(s) \|}, \qquad B^*h_n(s) \neq 0,$$

where

$$h_n(s) = S(T - s)^*p_n. \tag{4.10}$$

We note that $L(T)u_n$ converges to x_1, and from any subsequence of $\{u_n\}$ we can find a further subsequence which converges weakly to an optimal control. Also we can take p_n of norm one, and converging weakly to p_0; but p_0 may be zero in this case.

Finally let us consider the general case where X_1 is infinite dimensional and $L(T)$ is not necessarily compact. First let Δ_0 be the smallest positive number such that

$$x_1 = S(\Delta)y, \qquad y \in \Omega(T - \Delta).$$

It is clear that Δ_0 has to be less than T. Consider first the case where Δ_0 is actually zero. Let $T_n < T$ be a sequence of positive numbers converging monotonically to T. Then x_1 does not belong to the set $S(T - T_n)\Omega(T_n)$ for any n. But each set is convex, bounded, and closed, so that there is an element, $S(T - T_n)z_n$, say, closest to x_1, and we have

$$S(T - T_n)z_n = L(T)u_n,$$

where

$$u_n(s) = 0, \qquad T_n < s < T; \qquad \| u_n(s) \| \leqq M \text{ a.e. in } [0, T_n].$$

Moreover for each n,

$$[S(T - T_n)x, x_1 - L(T)u_n] \leqq [L(T)u_n, x_1L(T)u_n] \tag{4.11}$$

for $x \in \Omega(T_n)$. Letting

$$p_n = k_n(x_1 - L(T)u_n),$$

where k_n is positive, we have from (4.11) that

$$\int_0^{T_n} [u(s), B^*S(T-s)^*p_n]\, ds \leqq \int_0^{T_n} [u_n(s), B^*S(T-s)^*p_n] \| u(s) \| \leqq M,$$

from which it follows that

$$u_n(s) = \frac{MB^*S(T-s)^*p_n}{\| B^*S(T-s)^*p_n \|}$$

for s in $[0, T_n]$ such that the denominator is not zero. Also, if $x_0(t)$ is the state corresponding to an optimal control $u_0(t)$,

$$\| x_1 - L(T)u_n \| \leqq \| x_1 - S(T - T_n)x_0(T_n) \| \to 0.$$

Now let us define

$$v_n(s) = \begin{cases} u_n(s) \text{ if } 0 < s < T_n, \\ \dfrac{MB^*S(T-s)^*p_n}{\| B^*S(T-s)^*p_n \|} \text{ if } T_n < s < T, \\ 0 \text{ if the denominator is zero and } T_n < s < T. \end{cases}$$

Then it is clear that $L(T)v_n$ converges to x_1, and

$$v_n(s) = \frac{MB^*h_n(s)}{\| B^*h_n(s) \|}, \qquad B^*h_n(s) \not\equiv 0,$$

where

$$h_n(s) = S(T-s)^*p_n. \tag{4.12}$$

As before, p_n can be chosen to converge weakly to p_0, which may be zero.

If Δ_0 is positive, we have

$$x_1 = S(\Delta_0)y.$$

Then since T is the minimal time, it is necessary that

$$y = \int_0^{T-\Delta_0} S(T - \Delta_0 - \sigma)Bu(\sigma)\, d\sigma,$$

and also that y does not belong to $\Omega(t)$ for t less than $T - \Delta_0$. Hence we have the time-optimal problem for y with the minimal time $T - \Delta_0$. Moreover, since y has to be a boundary point of $\Omega(T - \Delta_0)$, the problem is now reduced to the case already considered. In other words, the optimal

control can be approximated as

$$v_n(s) = \begin{cases} 0 \text{ for } s > T - \Delta_0, \\ \dfrac{MB^*S(T - \Delta_0 - s)^*p_n}{\| B^*S(T - \Delta_0 - s)^*p_n \|} \text{ for } s \leqq T - \Delta_0, \end{cases}$$

$$B^*S^*(T - \Delta_0 - s)^*p_n \neq 0, \quad p_n \in X_1.$$

We note that $L(T)$ is compact if the semigroup $S(t)$ is compact for each positive t. Moreover in this case the semigroup is then [1, p. 304] uniformly continuous for $t > 0$ and this implies that for each x in X_1, $S(t)x$ is infinitely differentiable for $t > 0$. Since X_1 is reflexive this implies that $S^*(t)$ also has similar properties.

If the semigroup $S(t)$ is actually analytic, we can characterize (4.9 et seq.) further in terms of the solution of the adjoint equation. Thus consider the equation

$$\frac{d}{d\sigma} y(\sigma) = - A^*y(\sigma), \qquad 0 \leqq \sigma < T, \tag{4.13}$$

with the condition $y(T) = y_0$. Then (4.13) has a unique solution. For, let $Z(\sigma)$ be a null solution corresponding to $Z(T) = 0$. Then we note that

$$\frac{d}{d\sigma} S^*(\sigma)Z(\sigma) = A^*S^*(\sigma)Z(\sigma) - S^*(\sigma)A^*Z(\sigma) = 0$$

for $0 < \sigma < T$. Hence

$$S^*(\sigma)Z(\sigma) = 0, \qquad 0 < \sigma < T.$$

But since the semigroup is analytic, zero cannot be in the point spectrum of $S^*(\sigma)$ and this implies that

$$A(\sigma) = 0, \qquad 0 < \sigma < T,$$

or that (4.17) has a unique solution, namely,

$$y(\sigma) = S(T - \sigma)^*y_0.$$

REFERENCES

[1] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, American Mathematical Society Colloquium Publications, Providence, 1957.

[2] R. S. Phillips, *A note on the abstract Cauchy problem*, Proc. Nat. Acad. Sci. U. S. A., 40 (1954), pp. 244–248.

[3] T. Kato and H. Tanabe, *On the abstract evolution equation*, Osaka Math. J., 14(1962), pp. 107–133.

[4] S. Karlin, *Mathematical Methods and Theory in Games, Programming and Economics*, vol. 1, Addison-Wesley, Reading, Massachusetts, 1959.

[5] M. R. HESTENES, *Variational theory and optimal control theory*, Computing Methods in Optimization Problems, Academic Press, New York, 1964.
[6] R. BELLMAN, I. GLICKSBERG, AND O. GROSS, *Some aspects of the mathematical theory of control processes*, The RAND Corporation, Report R-313, 1958.
[7] A. V. BALAKRISHNAN, *An abstract formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.
[8] M. M. VAINBERG, *Variational Methods for the Study of Non-linear Operators*, Holden-Day, San Francisco, 1964.
[9] L. S. PONTRYAGIN, ET AL., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

THE PROBLEM OF BOUNDED SPACE COORDINATES AS A PROBLEM OF HESTENES*

T. GUINN†

1. Introduction. The problem here considered is that of first order necessary conditions for a problem in the calculus of variations which involves inequality constraints on the space variables independent of the control variables.

In 1961, Gamkrelidze [1] using methods developed by the Pontryagin school [2]–[6] obtained necessary conditions for a somewhat less general problem than is considered here. Berkovitz [7], using the classical variational theory as presented by Hestenes in [8], obtained all but one of Gamkrelidze's results for a similar problem. Both approach the problem by dividing an optimizing arc into subarcs with special properties. This gives rise to difficulty in matching multipliers at points at which the subarcs are joined. To avoid normality assumptions it appears necessary to consider arcs as a whole as is done here. For a discussion of normality, see Bliss [11]. The concepts there must be modified for this problem.

We here show that a general problem of this nature can be reduced to the problem of Hestenes [9], [10]. The development uses a device of Gamkrelidze to obtain one result. Not only are the results of [1], [7] obtained but it is further shown that the multipliers can be chosen to have less discontinuities than theirs.

2. The problem. Consider the problem of finding in a class of arcs

$$x: x^i(t), \qquad u^k(t), \qquad w^\sigma;$$

$$t^0 \leqq t \leqq t^1; \; i = 1, \cdots, n; \; k = 1, \cdots, m; \; \sigma = 1, \cdots, r;$$

satisfying conditions of the form

$$\dot{x}^i = f^i(t, x, u, w), \tag{2.1}$$

$$\psi_\alpha(t, x, w) \geqq 0, \qquad \alpha = 1, \cdots, s, \tag{2.2}$$

$$\phi_\beta(t, x, u, w) \geqq 0, \qquad \beta = 1, \cdots, p, \tag{2.3}$$

$$t^\tau = T^\tau(w), \qquad x^i(t^\tau) = X^{i\tau}(w), \qquad \tau = 0, 1, \tag{2.4}$$

* Received by the editors September 14, 1964, and in revised form November 25, 1964.

† Douglas Aircraft Company, Incorporated, Santa Monica, California. Part of a dissertation submitted in partial satisfaction of the requirements for the Ph.D. degree in mathematics at the University of California, Los Angeles. This research was supported in part by the United States Army Research Office (Durham) and in part by Douglas Aircraft Company.

$$
\begin{aligned}
I_\gamma &\geqq 0, && \gamma = 1, \cdots, q'; \\
I_\gamma &= 0, && \gamma = q' + 1, \cdots, q;
\end{aligned} \tag{2.5}
$$

where

$$I_\gamma = g_\gamma(w) + \int_{t^0}^{t^1} f_\gamma(t, x(t), u(t), w)dt\,;$$

one which minimizes

$$I = g(w) + \int_{t^0}^{t^1} f(t, x(t), u(t), w)\, dt.$$

We assume that all functions of (t, x, u, w) are of class C' in a region R, that $\psi_\alpha(t, x, w)$ is of class C'' and that grad $\psi_\alpha(t, x, w)$ does not vanish. Let R_0 be the set of all elements (t, x, u, w) satisfying (2.2) and (2.3). Let S be the set of elements in (t, x, w)-space which satisfy (2.2) with boundary S^*. Assume that if (t, x, w) is in S^* then the vectors grad $\psi_\alpha(t, x, w)$ are linearly independent for all α such that $\psi_\alpha(t, x, w) = 0$.

Define an *admissible control* $u(t)$ to be a piecewise continuous function such that if $x(t)$ is a solution of (2.1) for $u = u(t)$, then $(t, x(t), u(t), w)$ is in R_0 for $t^0 \leqq t \leqq t^1$. The corresponding arc

$$x\colon x^i(t), u^k(t), w^\sigma, \qquad t^0 \leqq t \leqq t^1,$$

will be called *differentially admissible* and, if (2.4) and (2.5) are also satisfied, *totally admissible.*

Assume an arc x_0 affords a strong relative minimum to I in the class of totally admissible arcs. Appropriate necessary conditions for a special problem will be first derived, then stated in Theorem 4.1. These results will then be extended to a more general problem in Theorem 5.1.

We remark that we can assume that only T^τ, $X^{i\tau}$, g, g_γ depend on w, since we can always let

$$\begin{aligned}
\dot{x}^{n+\sigma} &= 0, \qquad \sigma = 1, \cdots, r, \\
x^{n+\sigma}(t^0) &= w^\sigma,
\end{aligned}$$

and thus treat w^σ as a space variable.

3. Preliminary results. The problem of Hestenes is the same as given in §2 with the deletion of (2.2). We will first state some results from [9], [10] for this reduced problem which will then be used for the case when (2.2) is present. Here R_0 is the set of all elements (t, x, u, w) satisfying (2.3).

As in [9], a *broken field* $\mathfrak{F}$ is a region in (t, x, w)-space and a set of functions $U^k(t, x, w)$ such that (1) the functions $U^k(t, x, w)$, $U^k{}_{xi}(t, x, w)$ are continuous in (t, x, w) except for a finite set of values $t_1, \cdots, t_N$, at which

points they have left and right hand limits which are continuous in x and w; and (2) the elements $(t, x, U(t, x, w), w)$ are in R_0 for all (t, x, w) in $\mathfrak{F}$. By an *arc* in $\mathfrak{F}$ we will mean a solution

$$x: x^i(t),\ u^k(t),\ w^\sigma, \qquad t^0 \leqq t \leqq t^1,$$

of the system

$$\dot{x}^i = f^i(t, x, U(t, x, w), w), \quad u^k = U^k(t, x, w).$$

We now introduce a requirement on conditions (2.3). Given a point $(\bar{t}, \bar{x}, \bar{u}, \bar{w})$ for which (2.3) holds, let $\alpha_1, \cdots, \alpha_h$ be the set of values for which $\phi_\alpha(\bar{t}, \bar{x}, \bar{u}, \bar{w}) = 0$. We assume the matrix

$$\left(\frac{\partial \phi_\alpha}{\partial u^k}\right), \qquad \alpha = \alpha_1, \cdots, \alpha_h,$$

has rank h at $(\bar{t}, \bar{x}, \bar{u}, \bar{w})$. With this assumption we have:

LEMMA 3.1. *A differentially admissible arc*

$$x_0 : x_0{}^i(t),\ u_0{}^k(t),\ w_0^\sigma, \qquad t^0 \leqq t \leqq t^1,$$

is an arc of a broken field $\mathfrak{F}_0$ with control functions $U_0(t, x, w)$. Moreover, if $(\bar{t}, \bar{x}, \bar{u}, \bar{w})$ is in R_0, there is a broken field $\mathfrak{F}_1$ defined over a neighborhood of $(\bar{t}, \bar{x}, \bar{w})$ with control functions $\bar{U}(t, x, w)$ such that $\bar{U}^k(\bar{t}, \bar{x}, \bar{w}) = \bar{u}^k$.

This is a trivial modification to Lemma 16 in [9]. Similar results without introduction of broken fields are given in [10, §4].

THEOREM 3.1. *Let x_0 be an arc of a broken field which affords a strong relative minimum to I in the class of totally admissible arcs. Then there exist multipliers $\lambda_0, \lambda_\gamma, \mu_\beta, p_i(t)$ such that if we set*

$$\begin{aligned} H(t, x, u, p, \lambda, \mu, w) &= p_i f^i + \lambda_\gamma f_\gamma - \lambda_0 f + \mu_\beta \phi_\beta, \\ G(w) &= \lambda_0 g - \lambda_\gamma g_\gamma, \end{aligned} \tag{3.1}$$

then:

(1) *The multipliers $\lambda_0, \lambda_1, \cdots, \lambda_q$, are nonnegative constants. Further, $\lambda_\gamma = 0$ for each $\gamma \geqq 1$ for which $I_\gamma(x_0) > 0$. The multipliers $\lambda_0, \lambda_\gamma, p_i(t)$ do not vanish simultaneously at any point of $t^0 \leqq t \leqq t^1$. The multipliers $\mu_\beta(t)$ are nonnegative functions, continuous except possibly at discontinuities of $u_0(t)$. Also $\mu_\beta(t) = 0$, $\beta = 1, \cdots, p$, when $\phi_\beta(t, x_0(t), u_0(t)) > 0$.*

(2) *Along x_0,*

$$\dot{p}_i = -H_{x^i}, \qquad H_{u^k} = 0. \tag{3.2}$$

(3) *The inequality*

$$H(t, x_0(t), u, p(t), \lambda, 0, w_0) \leqq H(t, x_0(t), u_0(t), p(t), \lambda, 0, w_0) \tag{3.3}$$

holds for all t on $t^0 \leqq t \leqq t^1$ and for all u such that (t, x_0, u, w_0) is in R_0.

(4) *The transversality condition,*

$$(3.4) \qquad dG + [-H\,dT^\tau + p_i(t^\tau)\,dX^{i\tau}]_{\tau=0}^{\tau=1} - \int_{t^0}^{t^1} H_{w^\sigma}\,dw^\sigma\,dt = 0,$$

holds on x_0 for all dw^σ.

This is equivalent to Theorem 3.1 of [10]. Contained in the proof but not stated explicitly is the additional result that the functions $\mu_\beta(t)$ have as many derivatives with respect to t as the lesser of

$$H_{u^k}(t, x_0(t), u_0(t), p(t), \lambda, 0, w_0)$$

and

$$\phi_{u^k}(t, x_0(t), u_0(t), w_0).$$

4. A special problem. With the simplification of a device due to Gamkrelidze we now reduce a special case of the original problem to that considered in §3.

Suppose that for some a, $(1 \leqq a \leqq s)$, $\psi_a(t, x_0(t)) = 0$ for $t_1 \leqq t \leqq t_2$. Then since $\psi_a(t, x)$ is of class C'', we can choose a neighborhood M_a of $(t, x_0(t))$ and a vector $N_a(t, x)$ of class C' on M_a such that the inner product

$$(4.1) \qquad (N_a(t, x), \operatorname{grad} \psi_a(t, x)) \geqq 0, \qquad (t, x) \text{ in } M_a.$$

The magnitude of $N_a(t, x)$ is at our disposal. For convenience we denote by $y = (y^0, y^1, \cdots, y^n)$, the $(n + 1)$-dimensional vector with components $y^0 = t$, $y^i = x^i$, $i = 1, \cdots, n$. Then for $|\nu| \leqq \delta'$ and δ' sufficiently small, if y satisfies $\psi_a(y + \nu N_a(y)) = 0$, then y is in M_a. Further, if $\nu \leqq 0$ the definition of $N_a(y)$ assures that y is in S. Define $f^0 = 1$ and set

$$\phi_{p+\alpha}(t, x, u, \nu) = \psi_{\alpha y^i}(y + \nu N(y)) f^i(t, x, u), \qquad \alpha = 1, \cdots, s.$$

Then for any point (t, x) in S^*, for admissibility of u we must have that

$$(4.2) \qquad \phi_{p+\alpha}(t, x, u, 0) \geqq 0$$

for all α such that $\psi_\alpha(t, x) = 0$.

We first consider the case where $(t, x_0(t))$ is in S^* for $t^0 \leqq t \leqq t_1$ and is interior to S for $t_1 < t \leqq t^1$. By a suitable change in parameter we may take $t^0 = 0$, $t^1 = 1$. We assume that on $0 \leqq t \leqq t_1$, the matrix

$$(4.3) \qquad \left(\frac{\partial \phi_\rho}{\partial u^k}\right).$$

has maximum rank for all subscripts $\rho = p + \alpha_h$ such that $\phi_{p+\alpha}(t, x, u, \nu) = 0$ and all subscripts $\rho = \beta_k$ such that $\phi_\beta(t, x, u) = 0$. Set

$$\phi_{p+\alpha}(t, x_0(t), u_0(t), 0) = \phi_{p+\alpha}(t), \qquad \alpha = 1, \cdots, s.$$

Then the system of equations

$$\phi_{p+\alpha}(t, x, u, \nu) - \phi_{p+\alpha}(t) = 0, \qquad \alpha = \alpha_1, \alpha_2, \cdots,$$
$$\phi_\beta(t, x, u) = 0, \qquad \beta = \beta_1, \beta_2, \cdots,$$

has a solution for each point t on the extended interval $0 \leqq t \leqq t_1 + \epsilon$ for some $\epsilon > 0$ in a neighborhood of $[x_0(t), u_0(t), 0]$. By Lemma 3.1, x_0 is an arc in a broken field $\mathfrak{F}_1$ on the larger interval. For $\nu \leqq 0$, an arc $x(t)$ in $\mathfrak{F}_1$ is in R_0 provided that $\psi_\alpha[0, x(0)] \geqq 0$. Hence, in addition to (4.2) we must require that

$$I_{q+\alpha} = \psi_\alpha(0, X^0(w)) \geqq 0, \qquad I_{q+s+1} = -\nu \geqq 0. \tag{4.4}$$

Next we assume that on $t_1 \leqq t \leqq 1$ the matrix

$$\left(\frac{\partial \phi_\beta}{\partial u^k}\right) \tag{4.5}$$

has maximum rank for all subscripts β_k as above. Then, again by Lemma 3.1, x_0 is an arc in a broken field $\mathfrak{F}_2$ on $t_1 \leqq t \leqq 1$.

We extend the arcs in $\mathfrak{F}_1 \cap R_0$ to the interval $0 \leqq t \leqq 1$ using those in $\mathfrak{F}_2$ for $t_1 + \epsilon \leqq t \leqq t^1$. The resulting arcs lie in R_0 and the corresponding control functions $U(t, x)$ satisfy the hypotheses of Theorem 3.1 since we have at most introduced a discontinuity at $t_1 + \epsilon$. Hence, the conclusions of Theorem 3.1 hold, where because of (4.4) the function of G in the transversality condition corresponding to (3.4) becomes

$$G = \lambda_0 g - \lambda_\gamma g_\gamma - \lambda_{q+\alpha}\psi_\alpha(0, X^0(w)) + \lambda_{q+s+1}\nu, \tag{4.6}$$

and the function H corresponding to (3.1) becomes

$$H(t, x, u, p, \lambda, \mu, \nu) = p_i f^i + \lambda_\gamma f_\gamma - \lambda_0 f + \mu_\rho \phi_\rho, \qquad \rho = 1, \cdots, p + s. \tag{4.7}$$

If we assume that the partial derivatives of $H(t, x, u, p, \lambda, 0, 0)$ and $\phi_{p+\alpha}(t, x, u, 0)$ with respect to u^k are of class C' on x_0 except at discontinuities of $u_0(t)$, then $\dot{\mu}_{p+\alpha}(t)$ has the same discontinuities as $u_0(t)$ by the remark following Theorem 3.1.

Now, since $(t, x_0(t))$ is in S^* for $0 \leqq t \leqq t_1$, for at least one value of α, $(p + 1 \leqq \alpha \leqq p + s)$, $\phi_\alpha(t, x_0(t), u_0(t), 0) = 0$ at each point t. For simplicity assume that $\alpha = a$, a constant. Take $N_\alpha(t, x) = 0$, $\alpha \neq a$, and $N_a(t, x) = 0$ at 0 and t_1. By considering ν to be an additional parameter w, noting that ν appears in H only in ϕ_a and applying the transversality condition (3.4), we have that

$$(4.8)\quad \lambda_{q+s+1} - \int_0^1 H_\nu \, dt = \lambda_{q+s+1} - \int_0^{t_1} \mu_\alpha \phi_{a\nu} \, dt = 0, \qquad \lambda_{q+s+1} \geqq 0,$$

along x_0. To evaluate $\phi_{a\nu}$ along x_0 we define y as before and set $\eta^i = y^i + \nu N_a{}^i(y)$. Then

$$(4.9)\quad \begin{aligned} \phi_a(t, x, y, \nu) &= \psi_{ay^i}(\eta)\dot{y}^i = \psi_{a\eta^i}\eta_{y^i}^j \dot{y}^i \\ &= \psi_{a\eta^i}\dot{y}^i + \nu\psi_{a\eta^i}N_{ay^i}^j \dot{y}^i, \end{aligned}$$

where a is not summed. Hence, along x_0,

$$\begin{aligned} \phi_{a\nu}(t, x_0(t), u_0(t), 0) &= [\psi_{a\eta^i}\eta_\nu{}^i \dot{y}^i]_{\nu=0} + \psi_{ay^i}N_{ay^i}^j \dot{y}^i \\ &= \psi_{ay^iy^i}N_a{}^j\dot{y}^i + \psi_{ay^i}N_{ay^i}^j\dot{y}^i, \end{aligned}$$

where $\dot{y}^0 = 1$, $\dot{y}^i = f^i(t, x_0(t), u_0(t))$, $i = 1, \cdots, n$. But also

$$(4.10)\quad \frac{d}{dt}(\psi_{ay^i} N_a{}^i(y)) = \psi_{ay^iy^i}\dot{y}^j N_a{}^i + \psi_{ay^i}N_{ay^i}^i \dot{y}^j = \phi_{a\nu}$$

along x_0. Integrating (4.8) by parts we have that

$$-\int_0^{t_1} \mu_a \phi_{a\nu} \, dt = -\psi_{ay^i}(y)N_a{}^i(y) \,|_{t=0}^{t=t_1} + \int_0^{t_1} \dot{\mu}_a \psi_{ay^i} N_a{}^i \, dt \leqq 0.$$

The boundary terms vanish by choice of $N_a(t, x)$. Since also $\psi_{ay^i}N_a{}^i \geqq 0$ by (4.1), $N_a(t, x)$ is arbitrary on $0 < t < t_1$, and $\dot{\mu}_a$ is continuous except possibly at discontinuities of $u_0(t)$, we have that $\dot{\mu}_a \leqq 0$ except at these points. With this we have proved the following.

THEOREM 4.1. *Under the foregoing assumptions, let $x_0(t)$ afford a strong relative minimum to I in the class of totally admissible arcs. There exist multipliers λ_0, λ_γ, $\lambda_{q+\alpha}$, $\mu_\rho(t)$, $p_i(t)$ such that if we set*

$$(4.11)\quad H(t, x, p, w, \mu, \nu) = p_i f^i + \lambda_\gamma f_\gamma - \lambda_0 f + \mu_\rho \phi_\rho,$$

$$(4.12)\quad G = \lambda_0 g - \lambda_\gamma g_\gamma - \lambda_{q+\alpha}\psi_\alpha(t^0, X^0(w)),$$

conclusions of Theorem 3.1 *hold, where $\mu_{p+\alpha}(t) = 0$, $\alpha = 1, \cdots, s$, when $\psi_\alpha(t, x_0(t), w_0) > 0$. Further, $\dot{\mu}_{p+\alpha}$ is a nonpositive function with the same continuity properties as $u_0(t)$.*

Next suppose that for $t^0 \leqq t < t_1$ and $t_2 < t \leqq t^1$, x_0 is interior to S while in S^* for $t_1 \leqq t \leqq t_2$. By a suitable change of parameter we can take $t^0 = -1$, $t_1 = 0$, and $t^1 = 1$. Assume that for $0 \leqq t \leqq t_2$, the condition corresponding to (4.3) holds and similarly for (4.5) on $-1 \leqq t \leqq 0$ and $t_2 \leqq t \leqq 1$.

Let

$$\bar{x}(t) = x(-t), \quad \bar{u}(t) = u(-t), \qquad 0 \leqq t \leqq 1.$$

Set

$$F(t, x, u) = \begin{cases} f(t, x, u), & 0 \leqq t \leqq 1, \\ 0, & \text{otherwise}; \end{cases}$$

$$\bar{F}(t, \bar{x}, \bar{u}) = \begin{cases} f(-t, \bar{x}, \bar{u}), & 0 \leqq t \leqq 1, \\ 0, & \text{otherwise}; \end{cases}$$

and similarly for other functions of (t, x, u) and (t, x).

Consider the problem of finding in a class of arcs

$$x: x^i, \bar{x}^i, u^k, \bar{u}^k, w^\sigma, \qquad i = 1, \cdots, n; k = 1, \cdots, m; \sigma = 1, \cdots, r + n;$$

satisfying conditions of the form

$$\dot{x}^i = F^i(t, x, u), \quad \dot{\bar{x}}^i = -\bar{F}^i(t, \bar{x}, \bar{u}); \tag{2.1'}$$

$$\psi_\alpha(t, x) \geqq 0, \quad \psi_\alpha(t, \bar{x}) \geqq 0; \tag{2.2'}$$

$$\phi_\beta(t, x, u) \geqq 0, \quad \phi_\beta(t, \bar{x}, \bar{u}) \geqq 0; \tag{2.3'}$$

$$\begin{gathered} t^0 = 0, \quad t^1 = 1, \quad x^i(1) = X^{i1}(w), \\ x^i(0) = w^{r+i}, \quad \bar{x}^i(1) = X^{i0}(w); \end{gathered} \tag{2.4'}$$

$$\begin{gathered} I_\gamma = g_\gamma(w) + \int_0^1 (F(t, x, u) + \bar{F}(t, \bar{x}, \bar{u}))\, dt \geqq 0, \\ I_{q+\alpha} = \psi_\alpha(0, x(0)) \geqq 0, \qquad I_{q+s+1} = -\nu \geqq 0; \end{gathered} \tag{2.5'}$$

one which minimizes

$$I = g(w) + \int_0^1 (F(t, x, u) + \bar{F}(t, \bar{x}, \bar{u}))\, dt.$$

This problem corresponds to the previous case. Hence, if x_0 minimizes the original problem, for the transformed problem there exist multipliers $\lambda_0, \lambda_1, \cdots, \lambda_{q+s+1}, \mu_1(t), \cdots, \mu_{p+s}(t), \bar{\mu}_1(t), \cdots, \bar{\mu}_{p+s}(t), p_1(t), \cdots, p_n(t), \bar{p}_1(t), \cdots, \bar{p}_n(t)$, and functions

$$\begin{aligned} \bar{H} = {} & p_i F^i - \bar{p}_i \bar{F}^i + \lambda_\gamma(F_\gamma + \bar{F}_\gamma) \\ & - \lambda_0(F + \bar{F}) + \mu_\rho \phi_\rho + \bar{\mu}_\rho \bar{\phi}_\rho, \end{aligned} \tag{4.13}$$

$$G = \lambda_0 g - \lambda_\gamma g_\gamma - \lambda_{q+\alpha} \psi_\alpha(0, x(0)) + \lambda_{q+s+1}, \tag{4.14}$$

such that the conclusions of Theorem 3.1 hold along x_0 given by

$$x_0 : x_0^i(t), \bar{x}_0^i(t), u_0(t), \bar{u}_0(t), w_0^\sigma,$$

$$i = 1, \cdots, n; j = 1, \cdots, k; \sigma = 1, \cdots, r + n.$$

Applying the transversality condition (3.4) gives

$$\lambda_0\, dg - \lambda_\gamma\, dg_\gamma + p_i(1)X^{i1}(w) + \bar{p}_i(1)X^{i0}(w) = 0, \tag{4.15}$$

$$p_i(0) + \bar{p}_i(0) + \lambda_{q+\alpha}\psi_{\alpha x^i}(0, x_0(0)) = 0. \tag{4.16}$$

We note also that the multipliers $\bar{\mu}_{p+\alpha}(t) \equiv 0$, $\alpha = 1, \cdots, s$, since $\bar{\psi}_\alpha(t, x_0(t)) > 0$.

On $0 \leqq t < 1$ set

$$\mu_\beta(t) = \bar{\mu}_\beta(-t), \qquad \beta = 1, \cdots, p,$$

$$p_i(t) = -\bar{p}_i(-t),$$

and observe that $\dot{x}(t) = -\dot{x}(-t)$. Substituting these in (2.1′) through (2.5′), (4.13) and (4.14) we find that the conclusions of Theorem 4.1 hold except that now $\mu_\alpha(t)$, $p_i(t)$ may be discontinuous at $t = 0$ with the discontinuity for $p_i(t)$ of the form (4.16).

Added in proof. An equation was inadvertently omitted following (4.16). The transversality condition also yields that

$$-\bar{H} + \lambda_{q+\alpha}\,\psi_{\alpha t}(0, X_0(0)) = 0, \tag{4.17}$$

where $\bar{H}$ given by (4.13) is evaluated at $t = 0$ on $x_0(t)$. Thus the function H given in Theorem 4.1 may also be discontinuous at $t = 0$. The statement of Theorem 5.1 should be modified accordingly.

5. The general problem. To generalize this result suppose that for $t = 0$, 1, and 2, x_0 is interior to S while in S^* on one subinterval of $0 < t < 1$ and one subinterval of $1 < t < 2$.

Let

$$\bar{x}(t) = x(t-1), \quad \bar{u}(t) = u(t-1), \qquad 1 \leqq t \leqq 2.$$

Set

$$F(t, x, u) = \begin{cases} f(t, x, u), & 0 \leqq t \leqq 1, \\ 0, & \text{otherwise;} \end{cases}$$

$$\bar{F}(t, x, u) = \begin{cases} f(t-1, \bar{x}, \bar{u}), & 0 \leqq t \leqq 1, \\ 0, & \text{otherwise;} \end{cases}$$

and similarly for other functions of (t, x) and (t, x, u). This reduces to the previous case where now

$$I_{q+\alpha} = \psi_\alpha(0, x_0(0)) > 0,$$

so $\lambda_{q+1}, \cdots, \lambda_{q+s}$ are all zero. Hence, the transversality condition gives

$$p^i(0) = \bar{p}^i(0). \tag{5.1}$$

Setting

$$\mu_\rho(t+1) = \bar{\mu}_\rho(t),$$

$$p_i(t+1) = \bar{p}_i(t),$$

we have that, because of (5.1), the functions $p_i(t)$ are continuous, at $t = 0$.

We define a point $x_0(\bar{t})$ to be a *contact point* if for some $\delta > 0$, $x_0(t)$ is interior to S for $\bar{t} - \delta \leqq t < \bar{t}$ while $x_0(\bar{t})$ is in S^*. We call $\bar{t}$ a *contact time*. Then since Theorem 4.1 as proved includes the case where x_0 lies entirely in S^*, the generalization to where x_0 has a finite number of contact points is clear. Hence we have proved:

THEOREM 5.1. *Under the above assumptions, Theorem* 4.1 *holds except that the multipliers* $p_i(t)$, $\mu_\rho(t)$ *may be discontinuous at contact times, where discontinuities in* $p_i(t)$ *are of the form* (4.16).

The conclusion that $\dot{\mu}_\alpha \leqq 0$, $\alpha = p+1, \cdots, p+s$, was first obtained by Gamkrelidze [1] for the problem where (2.1) is independent of w, (2.2) depends on x alone, (2.3) on u alone, and (2.4) and (2.5) are not given. The method for this result used here is an adaptation of his. He does not determine that μ_α itself is nonnegative. Berkowitz [7], using methods based on [8], derives the condition on μ_α but not its derivative and for a problem where (2.5) does not appear.

Gamkrelidze's definition of "jump point" includes both a contact point as here defined and also a point at which x_0 leaves S^*. The methods of [1] and [7] give a possible discontinuity of $p_i(t)$ at jump points rather than at just contact points. It is not clear how either author obtains that λ_0 can be chosen continuous without an additional hypothesis regarding normality.

REFERENCES

[1] R. V. GAMKRELIDZE, *Optimal processes with bounded phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.

[2] V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND L. S. PONTRYAGIN, *On the theory of optimal processes*, Dokl. Akad. Nauk SSSR, 110 (1957), pp. 7–10.

[3] R. V. GAMKRELIDZE, *On the general theory of optimal processes*, Ibid., 123 (1958), pp. 223–226.

[4] V. G. BOLTYANSKII, *The maximum principle in the theory of optimal processes*, Ibid., 119 (1958), pp. 1070–1073.

[5] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[6] V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND L. S. PONTRYAGIN, *The theory of optimal processes I, The maximum principle*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 3–24.

[7] L. D. BERKOVITZ, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.

[8] M. R. HESTENES, *A general problem in the calculus of variations with applications to paths of least time*, RM-100, The RAND Corporation, Santa Monica, California; see also ASTIA Document 112382.

[9] ———, *Variational theory and optimal control theory*, mimeographed lecture notes, University of California, Los Angeles, 1963.

[10] ———, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[11] G. A. BLISS, *Lectures in the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

AN EXTENDED PONTRYAGIN PRINCIPLE FOR CONTROL SYSTEMS WHOSE CONTROL LAWS CONTAIN MEASURES*

RAYMOND W. RISHEL†

Introduction. In this paper, nonlinear control problems where the control laws are Radon measures are considered. This includes the cases where the control laws are δ-functions or ordinary functions or a combination of δ-functions and ordinary functions. A technique is developed for transforming this type of control problem into an equivalent control problem in which the control law is given by an ordinary function. This equivalence is used to show that Pontryagin's principle for ordinary systems implies an analog of Pontryagin's principle for systems whose control laws are Radon measures. The equivalent system is obtained by filling in the jumps of the trajectory of the original system with straight line segments and reparameterizing by a parameter that could be called "time plus control fuel used." In order to obtain both implications of the equivalence, a simple condition is assumed. This condition assures that no advantage would be gained by varying the control during an impulse and suffices to make the two optimization problems equivalent.

The optimal control of rockets in a gravitational field where the control laws were allowed to contain δ-functions was considered by Lawden [4]. In [2], Friedland and Ladd computed the minimum fuel control of a second-order linear system where the control law was allowed to contain δ-functions. In each of these treatments, the necessary conditions for optimality were obtained by a formal limiting procedure from the necessary conditions when the optimal control laws were bounded functions. There is also a large literature on the optimal impulsive transfer of space vehicles between orbits. Typical of these papers are [1] and [9].

Neustadt, in [6], considered the optimal control of linear systems in which the control laws were allowed to contain δ-functions. His results overlap with the results of this paper in the case where the equation of the system is linear. Schmaedeke, in [8], obtained existence theorems for optimal control laws for systems whose control laws were given by measures. The systems considered were nearly of the same generality as those considered in this paper.

Definitions and preliminary considerations. Let x denote an n-dimen-

* Received by the editors October 5, 1964, and in final revised form December 21, 1964.

† Mathematical Analysis Staff, Aero-Space Division, The Boeing Company, Seattle, Washington.

sional vector of state variables, u an m-dimensional vector of control variables, t the time, and μ a positive Radon measure on the time axis. Let $f_i(t, x, u)$ and $g_i(t, u)$ for $i = 0, 1, \cdots, n$ be continuous real valued functions of the variables indicated. The functions $g_i(t, u)$ and $f_i(t, x, u)$ will be assumed to be differentiable with respect to t and the x variables. The function $g_0(t, u)$ will be assumed to be a nonnegative function $g_0(t)$ of only the variable t. Let $f(t, x, u)$ and $g(t, u)$ denote the vector functions

$$f(t, x, u) = (f_1(t, x, u), \cdots, f_n(t, x, u)),$$
$$g(t, u) = (g_1(t, u), \cdots, g_n(t, u)).$$

A positive Radon measure μ and a Borel measurable vector function $u(t)$ defined almost everywhere on an interval $[t_0, \infty)$ with respect to both Lebesgue measure and μ will be called a control law. It is only a convenience to have control laws defined on an infinite interval. Usually, the values of the control law, after some finite time, play no role.

A vector function $x(t)$ will be said to be the trajectory, on an interval $t_0 \leqq t \leqq t_1$, of a control system with equation of motion

$$\frac{dx}{dt} = f(t, x, u) + g(t, u)\mu \tag{1}$$

corresponding to the control law $(u(t), \mu)$ and initial condition x_0, if

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s), u(s))\, ds + \int_{t_0}^{t} g(s, u(s))\mu(ds) \tag{2}$$

holds for every t, $t_0 \leqq t \leqq t_1$.

The symbol $\int_{t_0}^{t}$ will always be used to denote the integral over the closed interval $[t_0, t]$. This implies that $x(t)$ is right continuous on $[t_0, t_1]$. Formula (2) shows

$$x(t_0) = x_0 + g(t_0, u(t_0))\mu(\{t_0\}); \tag{3}$$

thus, $x(t_0)$ may differ from x_0 if the measure of the single point $\{t_0\}$ is different from zero.

Given a positive Radon measure μ and a constant t_0, define functions $F(t)$, $\eta(t)$, $\tau(\eta)$, and $\tilde{F}(t)$ on the interval $[t_0, \infty)$ by

$$F(t) = \int_{t_0}^{t} \mu(ds), \tag{4}$$

$$\eta(t) = t + F(t), \tag{5}$$

$$\tau(\eta) = \inf\{t : \eta(t) \geqq \eta\}, \tag{6}$$

$$\tilde{F}(\eta) = \eta - \tau(\eta). \tag{7}$$

The following are easy consequences of these definitions:

$$\tau(\eta(t)) = t, \qquad \tilde{F}(\eta(t)) = F(t), \tag{8}$$

$$\eta(t) = \sup\{\eta : \tau(\eta) = t\}. \tag{9}$$

If $\eta_1 < \eta_2$, then

$$\eta_2 - \eta_1 \geqq \tau(\eta_2) - \tau(\eta_1) \geqq 0 \tag{10}$$

and

$$\eta_2 - \eta_1 \geqq \tilde{F}(\eta_2) - \tilde{F}(\eta_1) \geqq 0. \tag{11}$$

Equations (10) and (11) show $\tilde{F}(\eta)$ and $\tau(\eta)$ are nondecreasing and Lipschitzian with Lipschitz constant 1. Therefore, $\tilde{F}(\eta)$ has a derivative $\nu(\eta)$ defined almost everywhere such that $0 \leqq \nu(\eta) \leqq 1$ and the formula

$$\tilde{F}(\eta) = \int_{t_0}^{\eta} \nu(\eta)\, d\eta \tag{12}$$

holds. Formula (7) implies that the derivative of $\tau(\eta)$ is given by $1 - \nu(\eta)$ and

$$\tau(\eta) = t_0 + \int_{t_0}^{\eta} [1 - \nu(\eta)]\, d\eta. \tag{13}$$

LEMMA 1. *Let $\tau(\eta)$ be a continuous nondecreasing function of η defined on an interval $[t_0, \infty)$ such that $\tau(t_0) = t_0$. Let $\eta(t)$ be defined on the interval $[t_0, \infty)$ and satisfy (9). Let t_i and t_j be values of t such that $t_i < t_j$. Then*

$$\tau^{-1}\{(t_i, t_j]\} = (\eta(t_i), \eta(t_j)] \tag{14}$$

and

$$\tau^{-1}\{[t_0, t]\} = [t_0, \eta(t)]. \tag{15}$$

Proof. Since $\eta(t)$ satisfies (9), $\tau(\eta(t)) = t$. If $\eta(t_i) < \eta \leqq \eta(t_j)$, $t_i = \tau(\eta(t_i)) \leqq \tau(\eta) \leqq \tau(\eta(t_j)) = t_j$, since $\tau(\eta)$ is nondecreasing. Property (9) implies $t_i < \tau(\eta)$. Hence

$$\tau^{-1}\{(t_i, t_j]\} \supset (\eta(t_i), \eta(t_j)]. \tag{16}$$

Since τ is nondecreasing, (9) implies that if $t_i < \tau(\eta) \leqq t_j$,

$$\eta(t_i) < \eta \leqq \eta(\tau(\eta)) \leqq \eta(t_j). \tag{17}$$

Hence

$$\tau^{-1}\{(t_i, t_j]\} \subset (\eta(t_i), \eta(t_j)]. \tag{18}$$

Since $\tau(\eta)$ is increasing and $\tau(t_0) = t_0$, (9) implies

$$\tau^{-1}\{t_0\} = [t_0, \eta(t_0)]. \tag{19}$$

Hence,

$$\tau^{-1}\{[t_0, t]\} = \tau^{-1}\{t_0\} \cup \tau^{-1}\{(t_0, t]\} = [t_0, \eta(t)]. \tag{20}$$

LEMMA 2. *Let μ be a positive Radon measure and $\tau(\eta)$ and $\nu(\eta)$ the functions defined above. Then*

$$\mu(A) = \int_{\tau^{-1}(A)} \nu(\eta)\, d\eta \tag{21}$$

for every Borel subset A of the interval $[t_0, \infty)$.

Proof. Since both sides of (21) are countably additive with respect to A, it is sufficient to establish (21) when A is a half-open interval $(t_i, t_j]$ or the single point $\{t_0\}$. Using (4), (8), (12), and (14),

$$\begin{aligned} \mu((t_i, t_j]) &= F(t_j) - F(t_i) \\ &= \tilde{F}(\eta(t_j)) - \tilde{F}(\eta(t_i)) = \int_{\tau^{-1}(t_i, t_j]} \nu(\eta)\, d\eta. \end{aligned} \tag{22}$$

Using (4), (8), and (12) again shows

$$\mu(\{t_0\}) = F(t_0) = \tilde{F}(\eta(t_0)) = \int_{t_0}^{\eta(t_0)} \nu(\eta)\, d\eta = \int_{\tau^{-1}(\{t_0\})} \nu(\eta)\, d\eta; \tag{23}$$

hence, (21) is valid.

LEMMA 3. *Let $\nu(\eta)$ be a measurable function such that $0 \leqq \nu(\eta) \leqq 1$ and let t_0 be a constant. Let*

$$\tau(\eta) = t_0 + \int_{t_0}^{\eta} [1 - \nu(\eta)]\, d\eta. \tag{24}$$

Let $\eta(t)$ satisfy

$$\eta(t) = \sup\{\eta : \tau(\eta) = t\}. \tag{25}$$

Then, $\nu(\eta) = 1$ almost everywhere on the set $\{\eta : \eta(\tau(\eta)) \neq \eta\}$.

Proof. The function $\tau(\eta)$ is nondecreasing and $\eta(t)$ is strictly increasing. Hence, if $\eta(\tau(\eta)) \neq \eta$, η is either an interior point or a left end point of an interval on which $\tau(\eta)$ is constant. There are at most countably many such intervals. The function $\nu(\eta)$ must equal one almost everywhere on each since $\tau(\eta)$ is constant. Hence, the lemma follows.

Let $h(t)$ be a Borel measurable function and A be a Borel measurable subset of $[t_0, \infty)$. Theorem 3C of [3, p. 163], Lemma 2 and (13) imply the following two change of variables formulas hold.

$$\int_{\tau^{-1}(A)} h(\tau(\eta))\nu(\eta)\,d\eta = \int_A h(t)\mu(dt). \tag{26}$$

$$\int_{\tau^{-1}(\tau(A))} h(\tau(\eta))[1 - \nu(\eta)]\,d\eta = \int_{\tau(A)} h(t)\,dt. \tag{27}$$

Since $\tau^{-1}(\tau(A)) - A \subset \{\eta : \eta(\tau(\eta)) \neq \eta\}$, Lemma 3 and (27) imply

$$\int_A h(\tau(\eta))[1 - \nu(\eta)]\,d\eta = \int_{\tau(A)} h(t)\,dt. \tag{28}$$

The optimization problem. The purpose of this paper is to study the following optimization problem. Let U be a subset of m-dimensional space, x_0 and x_1 be two states, and t_0 an initial time. Consider the class of control laws $(u(t), \mu)$ such that $u(t) \in U$ and there is a corresponding solution $x(t)$ of the equation of motion

$$\frac{dx}{dt} = f(t, x, u) + g(t, u)\mu \tag{29}$$

on some interval $[t_0, t_1]$ with initial condition x_0 and terminal condition $x(t_1) = x_1$. Find, in this class of control laws, a control law $(u(t), \mu)$ which minimizes the performance index

$$\int_{t_0}^{t_1} f_0(s, x(s), u(s))\,ds + \int_{t_0}^{t_1} g_0(s)\mu(ds). \tag{30}$$

Call this optimization problem "Problem A." The technique of studying Problem A will be to find another optimization problem, called "Problem B," in which the control laws are given in terms of ordinary functions which is equivalent to Problem A as an optimization problem. With this in mind, consider the following optimization problem which shall be called Problem B.

Consider the class of control laws $(u(\eta), \nu(\eta))$ in which $u(\eta)$ and $\nu(\eta)$ are respectively an m-dimensional vector-valued and a real-valued Borel measurable function on an interval $[t_0, \infty)$. The values of $u(\eta)$ belong to the subset U and $0 \leqq \nu(\eta) \leqq 1$. There are corresponding solutions $z(\eta)$ and $\tau(\eta)$ of the equations

$$\frac{dz}{d\eta} = f(\tau, z, u)(1 - \nu) + g(\tau, u)\nu, \tag{31}$$

$$\frac{d\tau}{d\eta} = 1 - \nu, \tag{32}$$

on an interval $[t_0, \eta_1]$ for which $z(t_0) = x_0$, $\tau(t_0) = t_0$, $z(\eta_1) = x_1$.

In this class of control laws $(u(\eta), \nu(\eta))$, find a control law which mini-

mizes the performance index

$$\int_{t_0}^{\eta_1} [f_0(\tau(\eta), z(\eta), u(\eta))[1 - \nu(\eta)] + g_0(\tau(\eta))\nu(\eta)]\, d\eta. \tag{33}$$

THEOREM 1. *Let* $(u(\eta), \mu)$ *be a control law of the type described in Problem* A. *Let* $\tau(\eta)$, $\eta(t)$, *and* $\nu(\eta)$ *be functions defined in terms of the measure* μ *by* (4)–(7) *and* (12). *Then,* $(u(\tau(\eta)), \nu(\eta))$ *is a control law of the type described in Problem* B *and*

$$z(\eta) = x(\tau(\eta)) - \int_{\eta}^{\eta(\tau(\eta))} g(\tau(\eta), u(\tau(\eta)))\nu(\eta)\, d\eta \tag{34}$$

and $\tau(\eta)$ *are corresponding solutions of* (31) *and* (32) *for which* $z(t_0) = x_0$, $\tau(t_0) = t_0$, $z(\eta_1) = x_1$. *Moreover, the performance index of Problem* A *has the same value as that of Problem* B.

Proof. Since $x(t)$ is a solution of (29), (2) implies that

$$x(\tau(\eta)) = x_0 + \int_{t_0}^{\tau(\eta)} f(s, x(s), u(s))\, ds + \int_{t_0}^{\tau(\eta)} g(s, u(s))\mu(ds). \tag{35}$$

Formulas (28), (26), and Lemma 1 imply

$$\begin{aligned} x(\tau(\eta)) = x_0 &+ \int_{t_0}^{\eta} f(\tau(\eta), x(\tau(\eta)), u(\tau(\eta)))[1 - \nu(\eta)]\, d\eta \\ &+ \int_{t_0}^{\eta(\tau(\eta))} g(\tau(\eta), u(\tau(\eta)))\nu(\eta)\, d\eta. \end{aligned} \tag{36}$$

Lemma 3 states that $\nu(\eta) = 1$ almost everywhere that $\eta \neq \eta(\tau(\eta))$; hence, almost everywhere, either $x(\tau(\eta)) = z(\eta)$ or $\nu(\eta) = 1$. Therefore, substituting $z(\eta)$ for $x(\tau(\eta))$ in the first integral of (36) and subtracting $\int_{\eta}^{\eta(\tau(\eta))} g(\tau(\eta), u(\tau(\eta)))\, d\eta$ from both sides of (36) gives:

$$\begin{aligned} z(\eta) = x_0 &+ \int_{t_0}^{\eta} f(\tau(\eta), z(\eta), u(\tau(\eta)))[1 - \nu(\eta)]\, d\eta \\ &+ \int_{t_0}^{\eta} g(\tau(\eta), u(\tau(\eta)))\nu(\eta)\, d\eta. \end{aligned} \tag{37}$$

Equations (37) and (13) show that (31) and (32) are satisfied by $z(\eta)$ and $\tau(\eta)$. Let $\eta_1 = \eta(t_1)$; then (8) and (34) imply that $z(\eta_1) = x(t_1) = x_1$. Formulas (37) and (13) show that $z(t_0) = x_0$ and $\tau(t_0) = t_0$.

From Lemma 1, $\tau^{-1}([t_0, t_1]) = [t_0, \eta_1]$. Lemmas 2 and 3 and the change of variables formulas (26) and (28) imply

$$\int_{t_0}^{t_1} f_0(s, x(s), u(s))\, ds + \int_{t_0}^{t_1} g_0(s)\mu(ds)$$

$$= \int_{t_0}^{\eta_1} [f_0(\tau(\eta), z(\eta), u(\tau(\eta)))[1 - \nu(\eta)] + g_0(\tau(\eta))\nu(\tau(\eta))]\, d\eta, \tag{38}$$

completing the proof of Theorem 1.

The functions $g_i(t, u)$, $i = 0, 1, \cdots, n$, will be said to satisfy *the constancy condition* if, for each fixed t and each pair of states x_0 and x_1 for which there are values η_0 and η_1 and a Borel measurable function $u(\eta)$ such that

$$\int_{\eta_0}^{\eta_1} g(t, u(\eta))\, d\eta = x_1 - x_0, \tag{39}$$

there exist a vector u^* belonging to U and constants $k, \Phi_0, \Phi_1, \cdots, \Phi_n$ such that

$$\Phi_0 < 0, \tag{40}$$

$$x_1 - x_0 = kg(t, u^*), \tag{41}$$

$$\max_{u \in U} \sum_{i=0}^{n} \Phi_i\, g_i(t, u) = \sum_{i=0}^{n} \Phi_i\, g_i(t, u^*) = 0. \tag{42}$$

LEMMA 4. *Let the functions $g_i(t, u)$ satisfy the constancy condition. Let $(u(\eta), \nu(\eta))$ be a control law in the class described in Problem* B. *There is a control law $(u^*(\eta), \nu^*(\eta))$ in the class which has the property*

$$u^*(\eta) = u^*(\eta^*(\tau^*(\eta))) \tag{43}$$

whose performance index is no larger than the performance index for the control $(u(\eta), \nu(\eta))$.

The statement $u^*(\eta) = u^*(\eta^*(\tau^*(\eta)))$ is the condition that $u^*(\eta)$ is constant when $\tau^*(\eta)$ is constant. In this lemma, $\tau^*(\eta)$ is defined by (24) and $\eta^*(t)$ by (25).

Proof. As in the proof of Lemma 3, it is seen that

$$\{\eta \in [t_0, \eta_1] : \eta(\tau(\eta)) \neq \eta\}$$

is a countable union of intervals on which $\tau(\eta)$ is constant and $\nu(\eta) = 1$ almost everywhere. Formulas (24) and (25) imply these intervals will be left-closed and right-open intervals with the exception that if there is an interval which contains η_1, it will be closed. If this is the case, remove the point η_1 from this interval to obtain a collection of half-open intervals. Order this collection and let $I_n = [\eta_n^0, \eta_n^1)$ denote the nth interval which has endpoints η_n^0 and η_n^1. Let t_n denote the constant value of $\tau(\eta)$ on I_n.

Since $\nu(\eta) = 1$ almost everywhere on I_n,

$$\frac{dz(\eta)}{d\eta} = g(t_n, u(\eta)) \tag{44}$$

holds almost everywhere on I_n. Let $z(\eta_n{}^0) = z_n{}^0$ and $z(\eta_n{}^1) = z_n{}^1$. Since $g_i(t, u)$ satisfies the constancy condition and $g_0(t) \geqq 0$, Theorem A of [5, p. 243] implies there is a value $\eta_n{}^2$ and there is $u_n \in U$ such that

$$\int_{\eta_n{}^0}^{\eta_n{}^2} g(t_n, u_n)\, d\eta = z_n{}^1 - z_n{}^0 \tag{45}$$

and

$$\eta_n{}^0 \leqq \eta_n{}^2 \leqq \eta_n{}^1. \tag{46}$$

Let J_n denote the interval $[\eta_n{}^2, \eta_n{}^1)$. Let $\bar{I}_n$ denote the closure of I_n. Define $\gamma(\eta)$, $\beta(\xi)$, $\bar{u}(\eta)$, and $\bar{z}(\eta)$ by

$$\gamma(\eta) = t_0 + \int_{[t_0,\eta)-\bigcup_{n=1}^{\infty} J_n} d\eta, \tag{47}$$

$$\beta(\xi) = \sup_\eta \{\eta : \gamma(\eta) = \xi\}, \tag{48}$$

$$\bar{u}(\eta) = \begin{cases} u_n, & \text{if } \eta \in \bar{I}_n, \\ u(\eta), & \text{otherwise,} \end{cases} \tag{49}$$

$$\bar{z}(\eta) = \begin{cases} z_n{}^0 + g(t_n, u_n)(\eta - \eta_n{}^0), & \text{if } \eta \in I_n, \\ z(\eta), & \text{otherwise.} \end{cases} \tag{50}$$

Define $\tau^*(\xi)$, $z^*(\xi)$, $u^*(\xi)$, $\nu^*(\xi)$ by

$$\begin{aligned} \tau^*(\xi) &= \tau(\beta(\xi)), & z^*(\xi) &= \bar{z}(\beta(\xi)), \\ u^*(\xi) &= \bar{u}(\beta(\xi)), & \nu^*(\xi) &= \nu(\beta(\xi)). \end{aligned} \tag{51}$$

With these definitions, the formula

$$\int_{\beta^{-1}(A)} d\xi = \int_{A-\bigcup_{n=1}^{\infty} J_n} d\eta \tag{52}$$

holds. To establish this formula, it is sufficient to establish it in the case where A is a half-open interval $[\bar{\eta}_0, \bar{\eta}_1)$ since both sides are countably additive with respect to A. From the definitions (47) and (48), it is seen by an argument similar to Lemma 1 that

$$\beta^{-1}\{[\bar{\eta}_0, \bar{\eta}_1)\} \subset [\gamma(\bar{\eta}_0), \gamma(\bar{\eta}_1)] \quad \text{and} \quad \beta^{-1}\{[\bar{\eta}_0, \bar{\eta}_1)\} \supset [\gamma(\bar{\eta}_0), \gamma(\bar{\eta}_1)) \tag{53}$$

Hence,

$$\int_{\beta^{-1}\{[\bar{\eta}_0,\bar{\eta}_1)\}} d\xi = \gamma(\bar{\eta}_1) - \gamma(\bar{\eta}_0) = \int_{[\bar{\eta}_0,\bar{\eta}_1)-\bigcup_{n=1}^{\infty} J_n} d\eta. \tag{54}$$

Since the range of $\beta(\xi)$ is disjoint from $\bigcup_{n=1}^{\infty} J_n$, it follows from (52) and Theorem 3C of [3, p. 163] that if $h(\eta)$ is a Borel measurable function and A a Borel measurable subset of $[t_0, \infty)$, then the change of variables formula

$$\int_{\beta^{-1}(A)} h(\beta(\xi))\, d\xi = \int_{A-\bigcup_{n=1}^{\infty} J_n} h(\eta)\, d\eta \tag{55}$$

holds. Consider a value of η so that η is not a point of any I_n.

$$\begin{aligned} z(\eta) &= x_0 + \sum_{\eta_n^1 \leqq \eta} (z(\eta_n^1) - z(\eta_n^0)) \\ &+ \int_{[t_0,\eta]-\bigcup_{n=1}^{\infty} I_n} [f(\tau(\eta), z(\eta), u(\eta))(1 - \nu(\eta)) + g(\tau(\eta), u(\eta))\nu(\eta)]\, d\eta. \end{aligned} \tag{59}$$

Since $\nu(\eta) = 1$ almost everywhere on I_n, (45) implies

$$\begin{aligned} z(\eta_n^1) - z(\eta_n^0) = \int_{I_n - J_n} &[f(t_n, \bar{z}(\eta), u_n)(1 - \nu(\eta)) \\ &+ g(t_n, u_n)\nu(\eta)]\, d\eta. \end{aligned} \tag{57}$$

If η is a point of I_n, (50) implies

$$\bar{z}(\eta) = z(\eta_n^0) + \int_{\eta_n^0}^{\eta} [f(t_n, \bar{z}(\eta), u_n)(1 - \nu(\eta)) + g(t_n, u_n)\nu(\eta)]\, d\eta. \tag{58}$$

Hence, using the definition (49) of $\bar{u}(\eta)$ and (50) of $\bar{z}(\eta)$, and (56), (57), and (58),

$$\begin{aligned} \bar{z}(\eta) = x_0 + \int_{[t_0,\eta]-\bigcup_{n=1}^{\infty} J_n} &[f(\tau(\eta), \bar{z}(\eta), \bar{u}(\eta))(1 - \nu(\eta)) \\ &+ g(\tau(\eta), \bar{u}(\eta))\nu(\eta)]\, d\eta \end{aligned} \tag{59}$$

if η is not a point of some J_n.

Since $J_n \subset I_n$ and $\nu(\eta) = 1$ almost everywhere on I_n,

$$\tau(\eta) = t_0 + \int_{[t_0,\eta]-\bigcup_{n=1}^{\infty} J_n} [1 - \nu(\eta)]\, d\eta. \tag{60}$$

Since the range of $\beta(\xi)$ is contained in $[t_0, \infty) - \bigcup_{n=1}^{\infty} J_n$, the change of variables formula (55) and (59) and (60) imply

$$z^*(\xi) = x_0 + \int_{t_0}^{\xi} [f(\tau^*(\xi), z^*(\xi), u^*(\xi))(1 - \nu^*(\xi)) + g(\tau^*(\xi), u^*(\xi))\nu^*(\xi)]\, d\xi \tag{61}$$

and

$$\tau^*(\xi) = t_0 + \int_{t_0}^{\xi} [1 - \nu^*(\xi)]\, d\xi.$$

Let $\xi_1 = \gamma(\eta_1)$. Since η_1 is not a point of an interval I_n, $\beta(\gamma(\eta_1)) = \eta_1$; hence

$$z(\xi_1) = \bar{z}(\beta(\gamma(\eta_1))) = z(\eta_1). \tag{62}$$

This shows $(u^*(\xi), \nu^*(\xi))$ belongs to the class of control laws described in Problem B.

Now, since $\nu(\eta) = 1$ almost everywhere on I_n and $g_0(t) \geqq 0$, (46) implies

$$\begin{aligned} &\int_{I_n - J_n} [f_0(t_n, \bar{z}(\eta), u_n)(1 - \nu(\eta)) + g_0(t_n)\nu(\eta)]\, d\eta \\ &\leqq \int_{I_n} [f_0(\tau(\eta), z(\eta), u(\eta))(1 - \nu(\eta)) + g_0(\tau(\eta))\nu(\eta)]\, d\eta. \end{aligned} \tag{63}$$

Formulas (63) and (55) imply that

$$\begin{aligned} &\int_{t_0}^{\xi_1} [f_0(\tau^*(\xi), z^*(\xi), u^*(\xi))(1 - \nu^*(\xi)) + g_0(\tau^*(\xi))\nu^*(\xi)]\, d\xi \\ &\leqq \int_{t_0}^{\eta_1} [f_0(\tau(\eta), z(\eta), u(\eta))(1 - \nu(\eta)) + g_0(\tau(\eta))\nu(\eta)]\, d\eta. \end{aligned} \tag{64}$$

Since $\eta^*(t) = \sup\{\xi : \tau^*(\xi) = t\}$, either ξ and $(\eta^*(\tau^*(\xi)))$ are identical or belong to the same interval on which τ^* is constant. Since $\tau^*(\xi) = \tau(\beta(\xi))$, and $\beta(\xi)$ is strictly increasing, $\beta(\xi)$ and $\beta(\eta^*(\tau^*(\xi)))$ belong to the same interval $\bar{I}_n$ on which τ is constant. Since $u^*(\xi) = \bar{u}(\beta(\xi)) = u_n$ if $\beta(\xi)$ belongs to $\bar{I}_n$, it follows that (43) holds.

THEOREM 2. *Let $(u(\eta), \nu(\eta))$ be a control law of the type described in Problem* B. *Let $\tau(\eta)$, $\eta(t)$ and μ be defined in terms of $\nu(\eta)$ by* (24), (25), *and* (21). *If*

$$u(\eta(\tau(\eta))) = u(\eta) \tag{65}$$

and $\nu(\eta) = 0$ outside of the interval $[t_0, \eta_1]$, then $(u(\eta(t)), \mu)$ is a control law of the type described in Problem A *and $x(t) = z(\eta(t))$ is a solution of* (29) *with initial condition x_0 and terminal condition $x(t_1) = x_1$. Moreover, the performance index of Problem* A *has the same value as that of Problem* B.

Proof. Since $z(\eta)$ is a solution of (31),

$$z(\eta) = x_0 + \int_{t_0}^{\eta} [f(\tau(\eta), z(\eta), u(\eta))(1 - \nu(\eta)) + g(\tau(\eta), u(\eta))\nu(\eta)]\, d\eta. \tag{66}$$

Lemma 3, (65), Lemma 1 and the change of variables formulas (26) and (28) imply

$$\begin{aligned} z(\eta(t)) &= x_0 + \int_{t_0}^{\eta(t)} [f(\tau(\eta), z(\eta(\tau(\eta))), u(\eta(\tau(\eta))))(1 - \nu(\eta)) \\ &\qquad + g(\tau(\eta), u(\eta(\tau(\eta))))\nu(\eta)]\, d\eta \\ &= x_0 + \int_{t_0}^{t} f(s, z(\eta(s)), u(\eta(s)))\, ds \\ &\qquad + \int_{t_0}^{t} g(s, u(\eta(s)))\mu(ds). \end{aligned} \tag{67}$$

Another application of Lemma 3, (65), Lemma 1, (26), and (28) shows the two performance indices have the same value.

THEOREM 3. *Let $g_i(t, u)$ satisfy the constancy condition. Then, optimization Problems* A *and* B *are equivalent in the sense that if either has an optimal control law so does the other. Optimal controls for Problem* B *may be defined in terms of those of Problem* A *by* (4)–(7) *and* (12). *If Problem* B *has an optimal control law, there is always one which satisfies* (65). *An optimal control law for Problem* A *may be defined in terms of this type of optimal control law for Problem* B *by* (24), (25), *and* (21).

Theorem 3 is a consequence of Theorems 1 and 2 and Lemma 4.

THEOREM 4. *Let $g_i(t, u)$ satisfy the constancy condition. Then necessary conditions for $(u(t), \mu)$ to be an optimal control law for Problem* A *are the existence of a nonpositive constant Φ_0, continuous functions $\Phi_1(t), \cdots, \Phi_n(t)$ and a right continuous function $\Phi_{n+1}(t)$, not all identically zero, such that: the functions $\Phi_i(t)$, $i = 1, \cdots, n$, are solutions of* (68); *$\Phi_{n+1}(t)$ is a solution of* (69); (70) *and* (71) *hold almost everywhere on $[t_0, t_1]$ with respect to Lebesgue measure*; (72) *and* (73) *hold almost everywhere on $[t_0, t_1]$ with respect to μ*; (74) *is satisfied at t_1.*

$$\frac{d\Phi_i(t)}{dt} = -\sum_{j=0}^{n} \frac{\partial f_j(t, x(t), u(t))}{\partial x_i} \Phi_j(t), \quad i = 1, 2, \cdots, n. \tag{68}$$

$$\begin{aligned} \frac{d\Phi_{n+1}(t)}{dt} &= -\sum_{j=0}^{n} \frac{\partial f_j(t, x(t), u(t))}{\partial t} \Phi_j(t) \\ &\qquad - \sum_{j=0}^{n} \frac{\partial g_j(t, u(t))}{\partial t} \Phi_j(t)\mu. \end{aligned} \tag{69}$$

$$\max_{u\in U}\sum_{j=0}^{n} f_j(t, x(t), u)\Phi_j(t) = \sum_{j=0}^{n} f_j(t, x(t), u(t))\Phi_j(t) = -\Phi_{n+1}(t). \tag{70}$$

$$\sup_{u\in U}\sum_{j=1}^{n} g_j(t, u)\Phi_j(t) \leqq 0. \tag{71}$$

$$\max_{u\in U}\sum_{j=0}^{n} g_j(t, u)\Phi_j(t) = \sum_{j=0}^{n} g_j(t, u(t))\Phi_j(t) = 0. \tag{72}$$

$$\sup_{u\in U}\sum_{j=0}^{n} f_j(t, x(t), u)\Phi_j(t) \leqq -\Phi_{n+1}(t). \tag{73}$$

$$\Phi_{n+1}(t_1) = 0. \tag{74}$$

Proof. Theorems 1 and 3 imply that for each optimal control law of Problem A there is a corresponding optimal control law of Problem B for which $u(\eta) = u(\eta(t(\eta)))$. Applying Pontryagin's principle to Problem B gives the following necessary conditions for an optimal control law $(u(\eta), \nu(\eta))$. There exist a nonpositive constant ψ_0 and continuous functions $\psi_i(\eta)$, $i = 1, \cdots, n+1$, such that (75)–(78) hold almost everywhere with respect to Lebesgue measures on the interval $[t_0, \eta_1]$.

$$\frac{d\psi_i(\eta)}{d\eta} = -\sum_{j=0}^{n} \frac{\partial f_j(\tau(\eta), z(\eta), u(\eta))}{\partial x_i}\psi_j(\eta)(1-\nu(\eta)), \qquad i = 1, \cdots, n. \tag{75}$$

$$\frac{d\psi_{n+1}(\eta)}{d\eta} = -\sum_{j=0}^{n} \frac{\partial f_j(\tau(\eta), z(\eta), u(\eta))}{\partial t}\psi_j(\eta)[1-\nu(\eta)] - \sum_{j=0}^{n} \frac{\partial g_j(\tau(\eta), u(\eta))}{\partial t}\psi_j(\eta)\nu(\eta). \tag{76}$$

$$\psi_{n+1}(\eta_1) = 0. \tag{77}$$

$$\begin{aligned}\max_{\substack{u\in U\\ 0\leqq\nu\leqq 1}} &\left[\sum_{j=0}^{n} f_j(\tau(\eta), z(\eta), u)\psi_j(\eta)[1-\nu] + \sum_{j=0}^{n} g_j(\tau(\eta), u)\psi_j(\eta)\nu + \psi_{n+1}(\eta)[1-\nu]\right] \\ &= \sum_{j=0}^{n} f_j(\tau(\eta), z(\eta), u(\eta))\psi_j(\eta)[1-\nu(\eta)] + \sum_{j=0}^{n} g_j(\tau(\eta)u(\eta))\psi_j(\eta)\nu(\eta) + \psi_{n+1}(\eta)(1-\nu(\eta)) = 0.\end{aligned} \tag{78}$$

Since the values of $\nu(\eta)$ for $\eta > \eta_1$ play no role in Problem B, it may be assumed that $\nu(\eta) = 0$ if $\eta > \eta_1$.

The expression to be maximized in (78) is of the form

$$a(u)(1 - \nu) + b(u)\nu. \tag{79}$$

In order that

$$\max_{\substack{u \in U \\ 0 \leqq \nu \leqq 1}} [a(u)(1 - \nu) + b(u)\nu] = 0 \tag{80}$$

it must follow that $a(u) \leqq 0$ and $b(u) \leqq 0$. If either of these quantities were positive there would be a value of ν making (80) positive. If (80) is maximized by u^*, ν^*, it must follow that

$$\begin{aligned} a(u^*) &= 0 && \text{if } \nu^* = 0, \\ b(u^*) &= 0 && \text{if } \nu^* = 1, \\ a(u^*) &= b(u^*) = 0 && \text{if } 0 < \nu^* < 1. \end{aligned} \tag{81}$$

Hence,

$$\begin{aligned} &\max_{u \in U} \sum_{j=0}^{n} f_j(\tau(\eta), z(\eta), u)\psi_j(\eta) \\ &\qquad = \sum_{j=0}^{n} f_j(\tau(\eta), z(\eta), u(\eta))\psi_j(\eta) = -\psi_{n+1}(\eta) \end{aligned} \tag{82}$$

and

$$\sup_{u \in U} \sum_{j=0}^{n} g_j(\tau(\eta), u)\psi_j(\eta) \leqq 0, \qquad \text{if } \nu(\eta) = 0; \tag{83}$$

$$\max_{u \in U} \sum_{j=0}^{n} g_j(\tau(\eta), u)\psi_j(\eta) = \sum_{j=0}^{n} g_j(\tau(\eta), u(\eta))\psi_j(\eta) = 0 \tag{84}$$

and

$$\sup_{u \in U} \sum_{j=0}^{n} f_j(\tau(\eta), z(\eta), u)\psi_j(\eta) \leqq -\psi_{n+1}(\eta), \quad \text{if } \nu(\eta) = 1. \tag{85}$$

Both (82) and (84) hold almost everywhere if $0 < \nu(\eta) < 1$.

Let $\{I_k\}$ denote the set of intervals on which $\eta \neq \eta(\tau(\eta))$ defined in Lemma 4. Lemma 3 asserts that $\nu(\eta) = 1$ almost everywhere on I_k. This and (75) imply that ψ_i, for $i = 1, \cdots, n$, and $z(\eta)$ are constant on I_k. The function $\tau(\eta)$ is constant on I_k from the definition of I_k and $u(\eta)$ has been assumed to be constant on I_k. Hence, the continuity of the quantities involved and the assumption on $u(\eta)$ imply (84) and (85) hold everywhere on $\bar{I}_k$.

Let $x(t) = z(\eta(t))$, $u(t) = u(\eta(t))$, and $\Phi_j(t) = \psi_j(\eta(t))$, $j = 0, 1, \cdots, n + 1$. Let C and D denote, respectively, the sets on which (72),

(73) and (84), (85) do not hold. Since τ is one-to-one and $\eta(\tau(\eta)) = \eta$ on the complement of $\bigcup \bar{I}_k$, and (84) and (85) hold on $\bigcup \bar{I}_k$, $\tau^{-1}(C) = D$. Now, D is the union of a set of Lebesgue measure zero and a set on which $\nu(\eta) = 0$. Hence $\mu(C) = 0$.

Let E and F denote, respectively, the sets on which (70), (71) and (82), (83) do not hold. Since $\eta(\tau(\eta)) = \eta$ if $\eta \notin \bigcup I_k$, $E \subset \tau((\bigcup I_k) \cup F)$. Since $(\bigcup I_k) \cup F$ is the union of a set of measure zero and a set on which $\nu(\eta) = 1$, (28) implies

$$\int_E dt \leqq \int_{(\cup I_k)\cup F} (1 - \nu(\eta))\, d\eta = 0. \tag{86}$$

Using Lemma 3, Lemma 1, (65), (26), and (28), formulas (68) and (69) may be shown to follow from (75) and (76) by the same change of variables used in Theorem 2. Let $t_1 = \tau(\eta_1)$. Since $\nu(\eta) = 0$ if $\eta > \eta_1$, $\eta(t_1) = \eta(\tau(\eta_1)) = \eta_1$. Hence, (77) implies (74).

Remarks on the computation of μ. Since μ does not appear explicitly in relations (68)–(74), it might be thought that they do not furnish information on the determination of μ. This is not the case. Relations (71) and (72) imply that

$$\max_{u \in U} \sum_{j=0}^{n} g(t, u)\Phi_j(t) \leqq 0 \tag{87}$$

almost everywhere with respect to both Lebesgue and μ measure, and if A is a subset of

$$\left\{t: \max_{u \in U} \sum_{j=0}^{n} g(t, u)\Phi_j(t) < 0\right\}, \tag{88}$$

that $\mu(A) = 0$. This implies that (87) must vanish at each impulse (atom) of the measure μ.

If (87) as a function of t on $[t_0, t_1]$ assumes a maximum of zero only a finite number of times, the measure μ must consist of a finite number of impulses located at these maximum points. In some cases the magnitudes of the impulses may be determined by the requirement that $x(t)$ be a solution of (29) with the proper initial and terminal conditions for which (70) holds.

If the restriction of μ to an interval is absolutely continuous with respect to Lebesgue measure and has a positive Radon-Nikodym derivative, (87) and (88) imply that (72) must hold almost everywhere with respect to Lebesgue measure on this interval. This furnishes an additional relation. The simultaneous solution of this relation, together with (70) and the differential equations (68), (69), and (29), may determine μ on this interval.

REFERENCES

[1] R. B. BARRAR, *An analytic proof that Hohman-type transfer is the true minimum two impulse transfer*, Astronaut. Acta, 9 (1963), pp. 1–11.

[2] H. O. LADD, JR. AND B. FRIEDLAND, *Minimum fuel control of a second order linear process with a constraint on time to run*, Trans. ASME Ser. D., 86 (1964), pp. 160–168.

[3] P. HALMOS, *Measure Theory*, Van Nostrand, Princeton, New York, 1950.

[4] D. F. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworths, London, 1963.

[5] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[6] L. W. NEUSTADT, *Optimization, a moment problem, and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[7] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[8] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations with measure coefficients*, Minneapolis-Honeywell Report U-RD 63 18, (1963).

[9] L. TING, *Optimal orbital transfer by several impulses*, Astronaut. Acta, 6 (1960), pp. 256–265.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

A COMMON FRAMEWORK FOR AUTOMATA THEORY AND CONTROL THEORY*

MICHAEL A. ARBIB†

In the abstract theory of automata, our preoccupation is with machines which carry out computations, or logical manipulations, on their inputs to produce their outputs. Automata are digital, in that the inputs and outputs are always assumed to come from some finite "alphabet" of symbols; and the operation of the machines proceeds in discrete steps, i.e., at times $t = 0, 1, 2, 3, \cdots$ on some suitable time scale.

In control theory, however, we consider the inputs of our machines as variables which we may alter in such a way as to control the output or "states" of our machine. Our preoccupation is with executing the control in an economical manner, so as to minimize some "cost function." The inputs and outputs are usually assumed to be continuously variable, and in fact, to take values in some Euclidean space.

Despite the disparity in goals and assumptions, much of the basic apparatus is common to the two theories. In this paper we shall make explicit this commonality, and thus provide a basis for later papers to exploit the interaction between the two theories without duplicating basic material. The paper falls into three sections: §1, States and semigroups; §2, Additivity and duality; §3, Controllability and observability.

1. States and semigroups.

1.1. A system is, for us, something into which something (be it matter, energy, or information) may be put at certain times, and which itself puts out something at certain times. For instance we may think of an electric circuit whose input is the setting on a rheostat and whose output is a current reading. We may think of a network of switching elements whose input is an on-off setting of a number of input switches, and whose output is on-off pattern on an array of lights. In the first case we might think of the time-scale as being continuous with the adjustment of the rheostat varying

* Received by the editors September 15, 1964, and in revised form February 9, 1965.

† Stanford University, Stanford, California. This paper is descended from [1] via mutations induced by an (unpublished) paper on *Some basic notions of abstract automata theory*, presented at the International Summer School on Automata Theory held at Ravello, Italy, in 1964, and by some suggestions of R. E. Kalman.

Most of the work was carried out at the Imperial College of Science and Technology, University of London, London, England. The author thanks John Westcott and Jack Cowan for making his stay there possible and pleasant. The remaining work was supported by the United States Air Force under Contract AF 49(638)-1446.

continuously in time, whereas in the second case we may think of the time-scale as being discrete, the input switches being set every 5 seconds, say.

In either case we think of the system S as having an associated time-scale T. At each moment of time $t \in T$, our system S receives some input $i(t)$ and emits some output $o(t)$. We now assume that there is a fixed set I of possible inputs, and that at any time t, the input $i(t)$ may be chosen arbitrarily from I.

In general, the input segments of a system are *not* allowed to be arbitrary functions i: $[t_a , t_b) \to I$, but rather must belong to a restricted system D. The set of outputs, O, is to include all possible values of $o(t)$ for all times $t \in T$.

Now, we may not be able to predict $o(t)$ without knowing more than just what the present input is. The past history of the machine may have altered S in such a way (e.g., by hysteresis in our first example; by setting internal switches in our second example) as to modify the output. In other words, the output of S is (not surprisingly) a function both of the input of S, and of the history of S. We think of the *state* of a system as being some attribute of the machine at the present moment which together with the input determines the output (cf. [8, Chap. 1]). But to quality as the state of S, it must have one more property, namely that the states and inputs together suffice to determine subsequent states.[1]

We thus demand that the set of internal states be sufficiently rich to carry all the information about the past history of the machine needed to predict the effect of the past upon the future.

We can now define a system as follows (where the time-scale T is usually the half-line $[0, \infty)$ or the discrete set $\{0, 1, 2, 3, \cdots\}$).

DEFINITION. A *system* S is a mathematical structure defined by the following axioms:

AS_1 : There is given a *state-space* Q and a set of values of *time* T at which the behavior of the system is defined; T is an ordered subsemigroup (under addition) of the real numbers.

AS_2 : There is given a space I of admissible *inputs* to the system. An input segment $i_{[t_0,t)}$ is a mapping i: $[t_0 , t) \to I$. If $[t_1 , t_2) \subseteq [t_0 , t)$, we set $i_{[t_1,t_2)} = i \mid [t_1 , t_2)$.

There is given a set D of *admissible* input segments subject to the conditions:[2]

(a) D contains all constant functions.

(b) Let E_1 and E_2 be disjoint unions of intervals, whose union is the

[1] We do *not* consider stochastic systems here. See, however, [1, §8].

[2] These are the input conditions for the proof of the Pontryagin maximum principle used in [9].

interval $[t_1, t_2)$. Then if $i^1_{[t_1,t_2)}$ and $i^2_{[t_1,t_2)}$ belong to D, so does $\chi(E_1)i^1 + \chi(E_2)i^2$.

There is also given a space O of admissible *outputs*.

AS_3 : For any initial time t_0 in T, any initial state $q(t_0)$ in Q, any input segment $i_{[t_0,t)}$, $t \geqq t_0$, from D, the state $q(t)$ and the output $o(t)$ of the system S at time t are determined by the functions λ and δ according to the scheme

$$\begin{aligned} q(t) &= \lambda(q(t_0), i_{[t_0,t)}\ ; t_0) \in Q, \\ o(t) &= \delta(q(t_0), i_{[t_0,t)}\ ; t_0) \in O, \end{aligned} \tag{1.1}$$

i.e., $\lambda\colon Q \times D \times T \to Q$ and $\delta\colon Q \times D \times T \to O$. Moreover, for any fixed $t_0 \leqq t_1 \leqq t$ in T, and any q in Q and any fixed input segment $i\colon [t_0, t) \to I$, the following relations hold

$$\begin{aligned} \lambda(q, i_{[t_0,t_0)}\ ; t_0) &= q, \\ \lambda(q, i_{[t_0,t)}\ ; t_0) &= \lambda(\lambda(q, i_{[t_0,t_1)}\ ; t_0), i_{[t_1,t)}\ ; t_1), \\ \delta(q, i_{[t_0,t)}\ ; t_0) &= \delta(\lambda(q, i_{[t_0,t_1)}\ ; t_0), i_{[t_1,t)}\ ; t_1). \end{aligned} \tag{1.2}$$

We shall denote such an S by the quintuple $(D, O, Q, \lambda, \delta)$. This formulation clearly generalizes the definition of dynamical system in [2, p. 154]. It is clear from (1.1) that the system is *non-anticipatory*—the values of i after time τ do not affect its behavior up to time τ. In all that follows we shall consider only time-invariant systems, which we define next.

DEFINITION. A system S is said to be *time-invariant* if $i(t_0 + \tau) = \hat{\imath}(\hat{t}_0 + \tau)$ for $0 \leqq \tau < t_1 - t_0$ implies

$$\begin{aligned} \lambda(q, i_{[t_0,t_1)}\ ; t_0) &= \lambda(q, \hat{\imath}_{[\hat{t}_0,\hat{t}_1)}\ ; \hat{t}_0), \\ \delta(q, i_{[t_0,t_1)}\ ; t_0) &= \delta(q, \hat{\imath}_{[\hat{t}_0,\hat{t}_1)}\ ; \hat{t}_0) \end{aligned}$$

for all $t_0 \leqq t_1$, all $\hat{t}_0$, all $q \in Q$, and all input segments $i\colon [t_0, t_1) \to I$ (subject to $\hat{t}_1 = t_1 - t_0 + \hat{t}_0$; and that the $\hat{t}_0 - t_0$ translate of $[t_0, t_1) \cap T$ is just $[\hat{t}_0, \hat{t}_1) \cap T$).[3]

For the time-invariant systems we drop explicit mention in λ and δ of the time-variable, and usually consider the input segment of the second variable as defined on a suitable time-interval $[0, t)$.

We now list a few basic definitions, and some simple assertions whose easy proofs are left to the reader.

DEFINITION. Two states q and q' belonging to systems S and S' (where S and S' may or may not be identical but have common D and O) are said to be *equivalent* if and only if for all input segments $i_{[t_0,t)}$ from D the response segment of S starting on state q is identical with the response seg-

[3] This is satisfied for our usual choices of T, and positive $\hat{t}_0 - t_0$ in T.

ment of S' starting in state q', i.e.,

$$q \simeq q' \Leftrightarrow \delta(q; i_{[t_0,t)}) = \delta'(q'; i_{[t_0,t)})$$

for all times t, t_0 ($t_0 \leqq t$) and all input segments $i_{[t_0,t)}$ of S and S'.

DEFINITION. A system S is in *reduced form* if there are no distinct states in its state space which are equivalent to each other.

ASSERTION. If q and q' are equivalent, so are the states into which they are taken by any input segment of S and S'.

DEFINITION. A state q' of S is *reachable* from a state q of S if and only if there exists an input segment $i_{[t_0,t)}$ in D such that $q' = \lambda(q, i_{[t_0,t)})$.

DEFINITION. S is said to be *strongly connected* if every state is reachable from every other state.

DEFINITION. Systems S and S' are *equivalent*, $S \equiv S'$, if and only if to every state in the state-space of S there corresponds an equivalent state in the state-space of S', and vice-versa.

1.2. Our systems become the automata of automata theory if we quantize time and study the behavior of our systems at successive moments, $t = 0, 1, 2, 3, \cdots$, on some appropriate discrete time-scale, and further require that the input and output sets be finite.

Note that the conditions of AS_2 for a *discrete* time scale merely assert that *every* input segment is admissible. Hence, in this case, we need merely give I to determine D.

We shall *not* necessarily demand that there be only finitely many states. If Q does have but finitely many members, we shall say that M is a *finite automaton* or *finite-state machine*. It will be a question of interest to ask: "Given an automaton, does there exist an equivalent *finite automaton*?" Our general considerations then yield the following definition.

DEFINITION. An *automaton* is a quintuple $M = (I, O, Q, \lambda, \delta)$, where

I is a finite set: "the set of inputs",
O is a finite set: "the set of outputs",
Q is a set: "the set of states",
$\lambda: Q \times I \to Q$: "the next-state function",
$\delta: Q \times I \to O$: "the next-output function."

We interpret this formal quintuple as being a mathematical description of a machine which, if at time t is in state q and receives input i, will at time $t + 1$ be in state $\lambda(q, i)$ and will emit output $\delta(q, i)$.

Let us contrast this situation with that in control theory. Here time is usually regarded as continuous, i.e., we regard T as an interval of the real line, e.g., $T = [0, \infty)$. I, O, and Q are then regarded as finite-dimensional Euclidean spaces (or even as Banach spaces). What makes a system S a *control* system, however, is not so much these choices of T, I, O and Q but

rather that we associate with its operation some cost-function, and study problems of the kind "choose inputs to bring the system S from state q_0 to state q_1 with minimum cost." The cost may (of course) involve time, energy, money, etc.

Quite apart from the differences in motivation, we should also point out the differences in technique:

Automata theory emphasizes the *algebraic* and *combinatorial* aspects.

Control theory emphasizes the *analytical* techniques.

The mutual development of the two theories within the common framework developed here will thus savor much of the interplay within group theory between the algebraic theory of finite groups and the analytic theory of Lie groups.

1.3. We devote the remainder of this section to developing for systems in general a number of ideas usually met with only in automata theory, as prolegomena to the semigroup[4] theory of machines.

For definiteness, we now assume T to be either $\{0, 1, 2, 3, \cdots\}$ or $[0, \infty)$. We define $\hat{T}$ to be the set of finite initial segments of T, i.e.,

$$\{0, 1, 2, 3, \cdots\}^{\wedge} = \{\{0, 1, 2, \cdots, n\} \mid n = 0, 1, 2, 3, \cdots\},$$

$$[0, \infty)^{\wedge} = \{[0, a) \mid a \geqq 0\}.$$

Adopting the notation $[a, b) = \{t \in T \mid a \leqq t < b\}$, we have that $[0, n) = \{0, 1, \cdots, n-1\}$ if $T = \{0, 1, 2, 3, \cdots\}$, whereas $[0, b)$ is the usual halfopen interval if T is the real halfline. Then $\hat{T} = \{[0, t) \mid t \in T\}$. Given T and a set A, we define $A^{\hat{T}}$ to be simply the set of all functions from $\hat{T}$ to A. If $\alpha\colon [0, a) \to A$, $\beta\colon [0, b) \to A$, then we define

$$\alpha \cdot \beta\colon [0, a + b) \to A$$

by

$$\alpha \cdot \beta(t) = \begin{cases} \alpha(t) & \text{if } 0 \leqq t < a, \\ \beta(t - a) & \text{if } a \leqq t < a + b. \end{cases}$$

$A^{\hat{T}}$ is clearly a semigroup under this operation, and has for identity the null-function $e\colon \varnothing \to A$.[5] If α is defined on $[a, b)$ we set $l(\alpha) = b - a$, the "length" of α.

In the remainder of this section, we assume that our set D of admissible initial input segments is a *subsemigroup* I^* of $I^{\hat{T}}$.

[4] Or, as M. P. Schutzenberger would insist, the *monoid* theory, since all our semigroups have identities, but no topological structure. A semigroup for us is just a set on which is defined a binary associative operation, $(xy)z = x(yz)$.

[5] This definition is appropriate for *time-invariant* systems. In the general case, we would have to consider $\hat{T} = \{[a, b) \mid a < b,\ a, b \in T\}$ and define $\alpha \cdot \beta$ only in case α is defined on $[a, c)$ and β on $[c, b)$ for some a, b and a mutual c.

In case T is the real halfline, we would thus admit *piecewise* continuous inputs here, but *not* continuous input segments.

In case $T = \{0, 1, 2, 3, \cdots\}$, I^* is the familiar "free semigroup on I," consisting of finite sequences of elements of I, composed under concatenation.

Reexamining (1.1), we see that a time-invariant system is really defined by two functions

$$\begin{aligned} &\lambda: Q \times I^* \to Q, \\ &\delta: Q \times I^* \to O. \end{aligned} \tag{1.3}$$

For an automaton, we are given the state-transition and output functions as $\lambda: Q \times I \to Q$, $\delta: Q \times I \to O$, but these extend immediately to the form (1.3), in which we shall feel free to use them.

Returning now to our general time-invariant systems, we see that (1.2) becomes

$$\lambda(\lambda(q, x), y) = \lambda(q, xy) \quad \text{for all} \quad q \in Q, \quad x, y \in I^*.$$

The input-output function of the time-invariant system S when started in state q is the function

$$S_q : I^* \to O$$

defined by $S_q(x) = \delta(q, x)$ for $x \in I^*$. Defining $L_y(x) = yx$ for all $y, x \in I^*$ and noting that $\delta(q, yx) = \delta(\lambda(q, y), x)$, we see that

$$S_{\lambda(q,y)}(x) = \delta(\lambda(q, y), x) = \delta(q, yx) = S_q(yx) = S_q L_y(x).$$

Thus

$$S_{\lambda(q,y)} = S_q L_y .$$

When our interest in a system is in how it transforms input-sequences into output-sequences, all that we wish to know about q is contained in the function S_q. Returning to our notion of system equivalence, we then have the following.

ASSERTION. Given two time-invariant systems S and S' with state-sets Q and Q', respectively, then they are equivalent if and only if

$$\{S_q \mid q \in Q\} = \{S_r' \mid r \in Q'\}.$$

Clearly, we also have the following.

ASSERTION. S is a system in reduced form if and only if $q \to S_q$ is a 1–1 mapping. We say S is a state-output system if there is a function i: $Q \to O$ such that $\delta(q, x) = i(\lambda(q, x))$, i.e., the output depends only on the state at a given time.

ASSERTION. Let $S = (I^*, O, Q, \lambda, \delta)$. Then there exists a reduced state-

output system equivalent to S. One such system is given by S°, termed the state-output reduction of S, where

$$S^\circ = (I^*, O, Q^\circ, \lambda^\circ, i^\circ \cdot \lambda^\circ),$$

where

$$\begin{aligned} Q^\circ &= \{S_q \mid q \in Q\}, \\ \lambda^\circ(q^\circ, x) &= q^\circ L_x \quad \text{for} \quad q^\circ \in Q^\circ, \qquad x \in I^*, \\ i^\circ(q^\circ) &= q^\circ(e), \end{aligned}$$

e.g.,

$$i^\circ(S_q L_x) = \delta(q, x).$$

In the case of a finite automaton, the reduction is again a finite automaton. The reader *au fait* with linear systems will recognize (cf. §2) that the reduction of a linear system with state space $Q = E^n$ will be a linear system with state space E^m $(m \leqq n)$. However, for a nonlinear system with state-space E^n, the reduced system is often not of interest, since the reduction may well destroy the Euclidean topology.

We close this section with the definition of the semigroup of the system S. First we recall that an equivalence $\equiv$ on a semigroup A is called a congruence if

$$x \equiv y \Rightarrow xz \equiv yz, \quad x \equiv y \Rightarrow zx \equiv zy, \qquad \text{for all } x, y, z \in A;$$

in which case we may define the factor semigroup $A/\equiv$ to have elements $[x]_\equiv$, the equivalence classes under $\equiv$, with multiplication defined by $[x]_\equiv[y]_\equiv = [xy]_\equiv$. Given the system $S = (I^*, O, Q, \lambda, \delta)$, we define an equivalence $\equiv_S$ on the semigroup I^* by

$$x \equiv_S y \quad \text{if and only if} \quad S_q(uxv) = S_q(uyv) \quad \text{for all} \quad q \in Q, \qquad u, v \in I^*.$$

This is clearly a congruence, and so we may define the semigroup of the system S to be the factor semigroup[6] $I^*/\equiv_S$.

For a low-level exposition of further basic notions of automata semigroup theory, see [1]. The reader may find it an easy exercise to extend it to general systems, in the fashion of our present treatment. To see these notions "in action," the reader should consult the papers of Krohn and Rhodes [5] and of Schutzenberger [7].

[6] R. E. Kalman comments: "If the semigroup is simple, we have the abstract generalization of an eigenvalue (when calculating in the complex field) or of a first- or second-order (with complex eigenvalues) linear system, when calculating in the real field. Similar ideas concerning the structure of Lie groups are well known. (See, e.g., K. T. Chen, Math. Ann., 146 (1962), pp. 263–278.)"

2. Additivity and duality.

2.1. So much of modern mathematics—from the humble matrix to the most abstract Banach space—is preoccupied with linearity that it is not surprising that linear systems have been much studied by control theorists. In control theory, the spaces I, O and Q are usually Euclidean, and the theory of linear systems is erected on this basis. Our contribution in this section is to develop some of the basic theory using only the group property of the relevant spaces.[7]

Thus, save in our treatment of duality for automata, we shall assume in this section that I, O, *and* Q *are abelian groups,* and use $+$ for the group operations. Since we make no use of scalar multiplication, we shall use "*additivity*" to refer to our various analogues of the classical *linearity.*

DEFINITION. A state θ of the system S is a *zero state* if

$$\delta(\theta, 0^t) = 0 \quad \text{for all} \quad t,$$

where 0^t is the zero-input, $0^t\colon [0, t) \to 0$ defined by $0^t(\tau) \equiv 0$.

DEFINITION. The system S is *zero-state additive* if

$$\delta(\theta, \alpha - \beta) = \delta(\theta, \alpha) - \delta(\theta, \beta)$$

for all $\alpha, \beta \in I^*$ with $l(\alpha) = l(\beta)$, and with θ the zero-state of S.

DEFINITION. The system S is *additive with respect to an initial state* q if and only if

$$\delta(q, u) - \delta(q, v) = \delta(q, u - v)$$

for all $u, v \in I^*$ with $l(u) = l(v)$.

ASSERTION. If S is zero-state linear, then it is also linear with respect to all states which are reachable from the zero-state.

DEFINITION. A system S is *zero-input additive* if and only if

$$\delta(q', 0^t) - \delta(q'', 0^t) = \delta(q' - q'', 0^t)$$

for all $q', q'' \in Q$, $t \geqq 0$.

We now have the crucial definition.

DEFINITION. The system S is *additive* if and only if

(i) S is additive with respect to all states of Q,

(ii) S is zero-input additive.

We immediately have:

THEOREM. *A system* S *is additive if and only if it has the following three properties*:

[7] For a treatment of linear systems, see, e.g., [8]. Linear automata seem to have been studied only when the state-space is a vector-space over a finite field—for a review and extensive bibliography, see Gill's contribution to *System Theory* (ed., L. A. Zadeh).

(A.1) *The decomposition property*: $\delta(q, u) = \delta(q, 0^{l(u)}) + \delta(\theta, u)$,
(A.2) *Zero-state linearity*: $\delta(\theta, x - y) = \delta(\theta, x) - \delta(\theta, y)$, $(l(x) = l(y))$,
(A.3) *Zero-input linearity*: $\delta(q' - q'', 0^t) = \delta(q', 0^t) - \delta(q'', 0^t)$.

ASSERTION. If S has the decomposition property, state q' is equivalent to state q'' if and only if

$$\delta(q', 0^t) = \delta(q'', 0^t) \quad \text{for all} \quad t \geqq 0.$$

Since $\delta(q' - q'', x) = \delta(q' - q'', 0) + \delta(0, x) = \delta(q', 0) - \delta(q'', 0) + \delta(0, x)$, we see further that two states are equivalent if and only if their difference is equivalent to the zero-state.

LEMMA. *The states equivalent to the zero-state form a normal subgroup N of the group Q of states.*

Proof. If $r, s \in N$, then $r - s \in N$, since

$$\delta(r - s, x) = \delta(0, x) \qquad \text{for all } x \in I^*;$$

i.e., N is a subgroup. But N is also normal, since Q is abelian.

ASSERTION. If q is equivalent to q', and if $x \in I^*$, then

$$\lambda(q, x) - \lambda(q', x) \in N.$$

Thus we may set up the factor group Q/N. The elements of Q/N are the equivalence classes of states of S. Hence the reduced system of S is simply

$$S^\circ = (I^*, O, Q/N, \lambda^\circ, \delta^\circ),$$

where

$$\lambda^\circ([q], x) = [\lambda(q, x)], \qquad \delta^\circ([q], x) = \delta(q, x).$$

COROLLARY. *An additive automaton is equivalent to a finite-state machine if and only if Q/N is a finite group.*

DEFINITION. An additive system is said to be *completely additive* if each of the three properties (A.1–3) of an additive system still holds on replacing δ by λ.

A routine proof yields the following.

THEOREM. *A system $M = (I^*, O, Q, \lambda, \delta)$ operating on $T = \{0, 1, 2, \cdots\}$ for abelian groups is completely additive if and only if there exist homomorphisms*

$$A: Q \to Q, \quad B: I \to Q,$$
$$C: Q \to O, \quad D: I \to O,$$

such that, for all $q \in Q$, $x \in I$, we have

$$\begin{aligned} \lambda(q, x) &= Aq + Bx, \\ \delta(q, x) &= Cq + Dx. \end{aligned} \tag{2.1}$$

We then have

$$\begin{aligned}\lambda(q, x_1x_2 \cdots x_n) &= A^n q + \sum_{m=1}^{n} A^{m-1}Bx_{n-m+1}\,,\\ \delta(q, x_1x_2 \cdots x_n) &= CA^{n-1}q + \sum_{m=1}^{n-1} CA^{m-1}Bx_{n-m} + Dx_n\,.\end{aligned} \tag{2.2}$$

(Clearly conditions 1 to 3 for λ and δ are all satisfied.)

Let us set

$$\Phi(n) = CA^{n-1},$$

$$h(m) = \begin{cases} D & \text{if } m = 0,\\ CA^{m-1}B & \text{if } m > 0.\end{cases}$$

We then have

$$\delta(q,\ x_0x_1 \cdots x_{n-1}) = \Phi(n)q + \sum_{m=0}^{n-1} h(m)x_{n-m}\,.$$

2.2. Now the passage from additive systems to the usual linear systems consists in

(i) replacing group homomorphisms by vector space homomorphisms,

(ii) replacing discrete time by continuous time.

And thus we get a linear system S with state-space $Q = E^n$ and time $T = [0,\ \infty)$ described by

$$y(t) = \Phi(t - t_0)x(t_0) + \int_{t_0}^{t} h(t - \xi)u(\xi)\,d\xi, \qquad t \geqq t_0,$$

where the p-vector $y(t)$ is the output at time t, the n-vector $x(t)$ is the state at time t, the r-vector $u(t)$ is the input at time t, $\Phi(t)$ is the state-transition matrix whose ith column is the response of S at time t to zero-input when started in state $(0, \cdots, 1, \cdots, 0)$ (1 in ith place), and $h(t)$ is the impulse response of S, i.e., response of S when started in the zero-state, to the impulse input

$$\delta(t) = \begin{cases}\text{Dirac delta function for } T = [0,\ \infty),\\ \text{Kronecker delta } \delta_{0t} \text{ for } T = \{0, 1, 2, 3, \cdots\}.\end{cases}$$

If Φ is differentiable, and h is unsullied by delta-functions or their derivatives, we may refer to S as a linear differential system. Then the state equations of S in differential form read [8, §3.7]:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \dot{\mathbf{\Phi}}(0)\mathbf{x}(t) + \mathbf{h}(0+)u(t),\\ y(t) &= \langle \boldsymbol{\phi}(0), \mathbf{x}(t)\rangle,\end{aligned}$$

and the state at time t is given by

$$\mathbf{x}(t) = \mathbf{\Phi}(t - t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{h}(t - \xi)u(\xi)\, d\xi, \qquad t \geqq t_0,$$

where $\mathbf{h}(t)$ is the state impulse response given by

$$\mathbf{h}(t) = \begin{bmatrix} h(t) \\ h'(t) \\ \vdots \\ h^{(n-1)}(t) \end{bmatrix}, \qquad t \geqq 0;$$

$$\mathbf{h}(t) = \mathbf{x}(t)]_{\mathbf{x}(0-)=0, u(t)=\delta(t)} \qquad \text{for all} \quad t \geqq 0,$$

and $\mathbf{\Phi}(t)$ is the state transition matrix and satisfies

$$\dot{\mathbf{\Phi}}(t) = \dot{\mathbf{\Phi}}(0)\mathbf{\Phi}(t), \qquad \mathbf{\Phi}(0) = I.$$

The state impulse response $\mathbf{h}(t)$ is related to the state transition matrix $\mathbf{\Phi}(t)$ by the equation

$$\mathbf{h}(t) = 1(t)\mathbf{\Phi}(t)\mathbf{h}(0+).$$

The input-output-state relation

$$y(t) = \langle \phi(t - t_0), \mathbf{x}(t_0) \rangle + \int_{t_0}^{t} h(t - \xi)u(\xi)\, d\xi \tag{2.3}$$

and the state equations

$$\mathbf{x}(t) = \mathbf{\Phi}(t - t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{h}(t - \xi)u(\xi)\, d\xi,$$
$$y(t) = \langle \phi(0), \mathbf{x}(t) \rangle,$$

hold for *all* t and t_0 provided the basis functions $\phi_1, \cdots, \phi_n$, the state transition matrix $\mathbf{\Phi}(t)$, the impulse response $h(t)$, and the state impulse response $\mathbf{h}(t)$ are understood to be extended rather than one-sided.

$$\mathbf{\Phi}(-t) = \mathbf{\Phi}^{-1}(t).$$

The extended state impulse response is given in terms of the extended $\mathbf{\Phi}(t)$ by

$$\mathbf{h}(t) = \mathbf{\Phi}(t)\mathbf{h}(0).$$

Correspondingly, the extended basis functions and the impulse response are given by the elements of the first row of the extended $\mathbf{\Phi}(t)$ and $\mathbf{h}(t)$, respectively.

COROLLARY [8, §3.7]. *If S is characterized by an input-output-state relation of the form* (2.3)*—which implies that S is in reduced form—then S is initial state determinable in the following sense: Given $u_{(t_0,t]}$ and $y_{(t_0,t]}$ one can uniquely determine the initial state $\mathbf{x}(t_0)$.*

Roughly speaking then: *linear differential systems are reversible.*

DEFINITION. The two systems of equations

$$\dot{\mathbf{x}} = A(t)\mathbf{x}$$

and

$$\dot{\mathbf{y}} = -A^*(t)\mathbf{y},$$

where $A^*(t)$ is the conjugate transpose of $A(t)$, are said to be *adjoint* to one another.

THEOREM [8, §6.2]. *Let $\Psi(t, t_0)$ be the state transition matrix of the adjoint system, i.e.,*

$$\frac{d}{dt}\Psi(t, t_0) = -A^*(t)\Psi(t, t_0), \qquad \Psi(t, t_0) = I.$$

Then $\Psi^*(t, t_0)\Phi(t, t_0) = I$ for all t and t_0. Thus

$$\Psi^*(t, t_0) = \Phi(t, t_0)^{-1} = (t_0, t). \tag{2.4}$$

Conversely, if this holds, then the corresponding systems are adjoint.

DEFINITION. Let the linear system S have a real-valued impulse response $h(t, \xi)$. The linear system $S^{(a)}$ is said to be the *adjoint* of S if its impulse response $h^{(a)}(t, \xi)$ satisfies the relation

$$h^{(a)}(t, \xi) = h(\xi, t) \quad \text{for all} \quad t, \xi, \tag{2.5}$$

where $h^{(a)}(t, \xi)$ is the response of $S^{(a)}$ at time t to a unit impulse applied at time ξ. Thus (2.5) implies that $S^{(a)}$ has a response at time t to a unit impulse applied at time ξ equal to the response of S at time ξ to a unit impulse applied at time t.

Note that, in general, if S is non-anticipatory then $S^{(a)}$ has to be anticipatory—i.e., the present output of $S^{(a)}$ depends on *future* inputs!

Consider now the equations

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + B\mathbf{u}, \\ \mathbf{y} &= C\mathbf{x} + D\mathbf{u}. \end{aligned} \tag{2.6}$$

They have the solution

$$\mathbf{x}(t) = \Phi(t - t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \Phi(t - \tau)B(\tau)\mathbf{u}(\tau)\, d\tau,$$

and thus the impulse-response matrix for input $\boldsymbol{\delta}(t - t_0)$ is simply

$$\mathbf{h}(t, t_0) = C(t)\Phi(t - t_0)B(t_0).$$

Similarly the system

$$\begin{aligned} \dot{\boldsymbol{\xi}} &= A^*\boldsymbol{\xi} + C^*\mathbf{v}, \\ \boldsymbol{\xi} &= B^*\boldsymbol{\xi} - D^*\mathbf{v}, \end{aligned} \tag{2.7}$$

has impulse-response matrix to input $\boldsymbol{\delta}(t - t_0)$ as

$$\mathbf{h}^{(a)}(t, t_0) = B^*(t)\Psi(t - t_0)C^*(t_0).$$

Thus

$$\mathbf{h}^{(a)}(t, t_0) = \mathbf{h}(t_0, t)^*. \tag{2.8}$$

We accept (2.8) as the appropriate generalization of (2.5) to a system with multi-dimensional output, and say that the system (2.7) is the adjoint of system (2.6). (A possible alternative has the signs of B^* and C^* changed.)

2.3. In automata theory, duality has been little studied. The only interesting example I know of is given by Rabin and Scott [6]. We shall slightly modify their definition to ease our later discussion.

DEFINITION. Given $M = (I, O, Q, \lambda, \delta)$ with $Q = \{q_1, \cdots, q_n\}$, let $\bar{M} = (I, \bar{O}, \bar{Q}, \lambda^*, \delta^*)$, where $\bar{Q}$ is the set of subsets of Q, $\bar{O}$ is the set of subsets of O,

$$\lambda^*(q^*, x) = \bigcup_{q \in q^*} \{t \in Q \mid \lambda(t, x) = q\},$$

$$\delta^*(q^*, x) = \bigcup_{q \in q^*} \{\delta(t, x) \mid \lambda(t, x) = q\}.$$

M^*, the dual of M, is defined to be $\bar{M}$ restricted to those states reachable from at least one of the states $\{q_1\}, \cdots, \{q_n\}$.

In general, if M has n states, then M^* has of the order of 2^n states—i.e., the state space "blows up" under taking of duals. It is of some interest to know when we can preserve the state-space, as is possible with linear systems. Now: M^* has states

$\{q_1\}, \cdots, \{q_n\} \Leftrightarrow$ for all $q, q' \in Q$, $x \in I^*$; $\lambda(q, x) = \lambda(q', x)$ implies $q = q'$

$\Leftrightarrow$ for all $q, q' \in Q$, $i \in I$; $\lambda(q, i) = \lambda(q', i)$ implies $q = q'$

$\Leftrightarrow$ each $\lambda(\cdot, i): Q \to Q$ is a permutation.

If this is the case, we may identify M^* with

$$M^R = (I, O, Q, \lambda^R, \delta^R),$$

where $\lambda^R(q, x)$ is the unique q' for which $\lambda(q', x) = q$ and then $\delta^R(q, x) = \delta(q', x)$. We shall say that M is *reversible* and that M^R is the *reverse* of M. Clearly $(M^R)^R = M$.

ASSERTION. M^* has the same number of states as M if and only if M is reversible. Then $M^* = M^R$.

ASSERTION. If M is reversible and there is a state q from which all states of M are reachable, then M is strongly connected.

We should contrast our notion of reversible automata with the general ideas of a converse system (cf. [8, §2.10]).

DEFINITION. $\mathfrak{A}$ and $\mathfrak{B}$ are *converse* systems if every input-output pair $(\mathbf{u}, \mathbf{y})$ for $\mathfrak{A}$[8] has the property that $(\mathbf{y}, \mathbf{u})$ (with $\mathbf{y}$ as input and $\mathbf{u}$ as output) is an input-output pair for $\mathfrak{B}$, and vice versa.

THEOREM. *The converse of a finite automaton is not necessarily finite-state.*

Proof. Let M be a finite automaton with m inputs and n outputs, but let $n < m$.

Then M has m^k input sequences of length k and at most n^k output sequences of length k.

$(x_1 \cdots x_k, y_1 \cdots y_k)$ is an input-output pair for M if there exists a state q of M such that

$$y_n = \delta(q, x_1 \cdots x_n) \qquad \text{for } 1 \leqq n \leqq k. \tag{2.9}$$

Let M_c be the converse of M (it may not be completely specified). If M_c has a finite number of states, then (2.9) implies $m^k \leqq rn^k$, so that $r \geqq (m/n)^k$. Letting k increase, we get a contradiction.

In mathematics, if M^{**} is the dual of the dual of some system M, then usually M is isomorphic either to M^{**} or to some subsystem of M^{**}. Thus, in automata theory, we might be tempted to claim that M is equivalent to the smallest submachine of M^{**} whose states include $\{\{q_1\}\}, \cdots, \{\{q_n\}\}$. Call it $\hat{M}$. However a state $\{q_1\}$ is not in general reachable from other states of M^*—the reversing action of a nonreversible machine in general increases the cardinality of a state-set at each transition. This means that in general, $\lambda^{**}(\{\{q_1\}\}, x) = \emptyset$, which is not interesting. To see this, let us follow a computation:

$$\lambda^{**}(R, x) = \bigcup_{T \in R} \{Q' \leqq Q \mid \lambda^*(Q', x) = T\},$$

where

$$\lambda^*(Q', x) = \bigcup_{q' \in Q'} \{q \in Q \mid \lambda(q, x) = q'\}.$$

If $R = \{\{q_1\}\}$, then

$$\lambda^{**}(R, x) = \{Q' \leqq Q \mid \lambda^*(Q', x) = \{q_i\}\}.$$

But

$$\{q_i\} = \bigcup_{q' \in Q'} \{q \in Q \mid \lambda(q, x) = q'\}$$

[8] I.e., there is a $q \in Q\mathfrak{A}$ such that $y(t) = \delta\mathfrak{A}(q, U_{[0,t)})$.

if and only if (a) $Q' = \{\lambda(q_i, x)\}$ and (b) $\lambda(q, x) = \lambda(q_i, x)$ implies $q = q_i$. We thus arrive at the following result.

THEOREM. *M is isomorphic to the submachine $\hat{M}$ of M^{**} if and only if M is reversible.*

3. Controllability and observability.

3.1. A new domain of study in control theory is that of controllability and observability. The basic work here (cf. [2], [3], [4]) has been in terms of linear differential systems S described by such equations as

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + B\mathbf{u}, \\ \mathbf{y} &= C\mathbf{x} + D\mathbf{u}, \end{aligned} \tag{3.1}$$

where A, B, C, D are, respectively, $n \times n$, $n \times r$, $p \times n$ and $p \times r$ matrices. The n-vector $\mathbf{x}$ is the state of the system, the r-vector $\mathbf{u}$ is the input, and the p-vector $\mathbf{y}$ is the output of S.

In our treatment we shall present as much of the theory as possible in a form applicable to general systems (and so, in particular, to automata). Many results will turn out to require only our additivity conditions rather than the linearity conditions used elsewhere.

We say a state q is controllable if we may so choose the input as to bring our system from $\mathbf{q}$ to the zero state. More formally, for any system S with a designated state $\mathbf{0}$, we have the following definition.

DEFINITION. State q of system S is *controllable* if and only if there exists $u \in I^*$ such that

$$\lambda(\mathbf{q}; u) = \mathbf{0}.$$

The system S is said to be controllable if and only if every state of S is controllable.

ASSERTION. If S is a system in which every state is reachable from θ, then: S is controllable $\Leftrightarrow$ S is strongly connected.

Let us recall the additive system M of (2.1). It is reversible (cf. (2.8)) if for each x, $\lambda(\cdot, x)$ is invertible, i.e., if for each $r \in Q$, there is a unique solution of

$$Aq = r - Bx,$$

i.e., if and only if A is an automorphism of Q with inverse A^{-1}, say. Then define

$$\begin{aligned} \lambda^R(r, x) &= A^{-1}r - A^{-1}Bx, \\ \delta^R(r, x) &= \delta(\lambda^R(r, x), x) \\ &= C[A^{-1}(r - Bx)] + Dx \\ &= CA^{-1}r + [D - CA^{-1}B]x. \end{aligned}$$

Then $M^R = (I, O, Q, \lambda^R, \delta^R)$ is also completely linear. Note that we do indeed have $M^{RR} = M$ since

$$(A^{-1})^{-1} = A, CA^{-1}AA^{-1}B - CA^{-1}B + D = D, \text{ etc.}$$

Now, our last assertion tells us that the reversible machine M is controllable if and only if all states can be reached from the zero-state. Consulting (2.2) we immediately have the following result.

THEOREM. *A reversible completely additive system is controllable if and only if each state q can be represented as a linear combination*

$$\sum A^{m-1}Bx, \qquad x \in I.$$

Turning now to the system (3.1) and recalling that the power A^n of an $n \times n$ matrix may be represented as a linear combination of I, A, $A^2, \cdots, A^{n-1}$, we have "essentially" proved:

THEOREM 3.1 [3, p. 201]. *The system S of* (3.1) *is controllable if and only if the column vectors of the matrix*

$$[B, AB, \cdots, A^{n-1}B]$$

span the state-space of S.

Turning to observability, we formulate the general definition:

DEFINITION. S is observable if and only if it is in reduced form.

Now we know that our completely additive system (2.1) is in reduced form only if $N = \{0\}$—i.e., the zero state is only equivalent to itself. But inspecting (2.2) we obtain the next result.

THEOREM. *The completely additive system S is observable if and only if*

$$CA^{k-1}q = 0 \quad \textit{for all} \quad k \implies q = 0.$$

This "immediately" yields the familiar consequence:

COROLLARY 3.2 [8, 11.4]. *The system S of* (3.1) *is observable if and only if the column vectors of the matrix*

$$[C^*, A^*C^*, \cdots, A^{*(n-1)}C^*]$$

span the state-space of S.

Theorem 3.1 and Corollary 3.2 combine to give the following result.

KALMAN DUALITY THEOREM. *Let Σ be the system dual to S of* (3.1); *it is defined by*

$$\xi = -A^*\xi + C^*\mathbf{v},$$
$$\mathbf{n} = B^*\xi - D^*\mathbf{v},$$

where the state ξ is an n-vector, the input $\mathbf{v}$ is a p-vector, and the output $\mathbf{n}$ is an r-vector. Then S is controllable (respectively, observable) if and only if Σ is observable (respectively, controllable).

Recalling our discussion of the reduced form of an additive system, we see the following.

THEOREM. *An additive system with state group Q, and subgroup N of states equivalent to the zero-state, can be partitioned into two subsystems S_1, which is observable, and S_2, which is unobservable, if and only if*

$$Q = N \times Q/N,$$

i.e., the decomposition of Q by N splits.

Note that we have only used conditions (i) and (ii) in the definition of *additive* systems. In the case of *linear* systems, Q becomes a linear vector space, and N becomes a subspace—and such a decomposition always splits.

REFERENCES

[1] M. A. ARBIB, *Automata theory and control theory—A rapprochement*, System Theory, L. A. Zadeh, ed., McGraw-Hill, New York, to appear.

[2] R. E. KALMAN, *A mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[3] R. E. KALMAN, Y. C. HO, AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1962), pp. 189–213.

[4] R. E. KALMAN, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 596–600.

[5] K. B. KROHN AND J. L. RHODES, *Algebraic theory of machines. I. The main decomposition theorem*, Technical Report, Department of Mathematics, University of California, Berkeley, 1963.

[6] M. O. RABIN AND D. SCOTT, *Finite automata and their decision problems*, IBM J. Res. Develop., 3 (1959), pp. 114–125.

[7] M. P. SCHUTZENBERGER, *On the definition of a family of Automata*, Information and Control, 4 (1961), pp. 245–270.

[8] L. A. ZADEH AND C. A. DESOER, *Linear System Theory: The State-Space Approach*, McGraw-Hill, New York, 1963.

[9] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

THE GENERALIZED LIENARD EQUATION*

T. A. BURTON†

1. Introduction. Consider the generalized Lienard equation

$$\ddot{x} + f(x, \dot{x}) + g(x) = 0, \tag{1}$$

or equivalently

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -f(x, y) - g(x), \end{aligned} \tag{2}$$

where

$$\begin{aligned} xg(x) &> 0 \qquad \text{if } x \neq 0, \\ yf(x, y) &> 0 \qquad \text{if } y \neq 0. \end{aligned} \tag{3}$$

Assume that f and g satisfy a Lipschitz condition with respect to x and y throughout the plane.

We give sufficient conditions that the null solution to (2) be globally asymptotically stable and necessary and sufficient conditions that the null solution to

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -p(x)|\, y \,|^{\alpha} y - g(x), \end{aligned} \tag{4}$$

be globally asymptotically stable when $p(x) > 0$ and $0 \leqq \alpha < 1$.

A good summary of results regarding the stability properties of the null solution to (1) may be found in either [1] or [2, pp. 139–156]. Although a great deal of work has been done on (1), the main result on global asymptotic stability of the null solution is embodied in the following theorem by Bushaw [1, pp. 17–18].

THEOREM. *If* $\int_0^x g(u)\, du \to +\infty$ *as* $|\, x \,| \to +\infty$, *then the null solution for* (2) *is globally asymptotically stable.*

Notable in this theorem is the fact that no mention is made of $f(x, y)$ (except for (3)). Bushaw has proposed that a similar theorem might be proved without using this hypothesis, but adds that no one seems to have done it [1, p. 31]. We are able to prove the same result by imposing a condition on f with none on g (except (3)).

* Received by the editors February 18, 1965, and in revised form March 1, 1965.
† Department of Mathematics, University of Alberta, Edmonton, Alberta.

2. Preliminaries. In order to avoid certain difficulties later, we state the following for reference [1, pp. 13, 16].

THEOREM. *If $f(x, y) + g(x)$ is continuous everywhere, and t_0, x_0, and y_0 are any real numbers, then (2) has a solution $(x(t), y(t))$ which:*
(a) *is defined over an interval $t_{-1} \leqq t \leqq t_1$, where $t_{-1} < t_0 < t_1$;*
(b) *satisfies the initial conditions $x(t_0) = x_0$, $y(t_0) = y_0$;*
(c) *(if $f(x, y) + g(x)$ is Lipschitzian) is unique; and*
(d) *(if (3) holds) may be continued over the interval $t_{-1} \leqq t < \infty$.*

LEMMA 1. *Equations (2) have the following properties.*
(i) *The null solution is asymptotically stable.*
(ii) *For any (x_0, y_0), the solution $(x(t), y(t))$ through (x_0, y_0) satisfies $|y(t)| \leqq k(x_0, y_0)$ for all $t \geqq 0$.*
(iii) *Every bounded solution tends to $(0, 0)$ as $t \to \infty$.*

Proof. For (x, y) sufficiently close to $(0, 0)$,

$$V(x, y) = \frac{y^2}{2} + \int_0^x g(u)\, du$$

is a positive definite function. Also

$$\frac{dV}{dt} = -yf(x, y) \leqq 0,$$

so $(0, 0)$ is Lyapunov stable.

The following result will be used to complete the proof.

THEOREM [3, p. 66]. *Let $V(x, y)$ be a scalar function with continuous first partial derivatives for all (x, y). Suppose that $V(x, y) > 0$ for all $(x, y) \neq (0, 0)$ and $dV/dt \leqq 0$. Let E be the locus $dV/dt = 0$ and M be the largest invariant set contained in E. Then all solutions bounded for $t > 0$ tend to M as $t \to \infty$.*

Now there exists a constant $A > 0$ such that the locus

$$V(x, y) = \frac{y^2}{2} + \int_0^x g(u)\, du = A$$

is a simple closed curve L. For any (x', y') inside L we have $V(x', y') < A$ and for any (x', y') outside L we have $V(x', y') > A$. Since $dV/dt \leqq 0$ no solution starting inside L crosses L for $t > 0$. So $\dot{x} = y$ and the locus of $dV/dt = 0$ is the y-axis; thus M is the origin. Hence $(0, 0)$ is asymptotically stable.

Although there may exist $B > A$ such that the locus $V(x, y) = B$ is not a simple closed curve, $V \to \infty$ as $|y| \to \infty$ and $dV/dt \leqq 0$, so there exists $k(x_0, y_0)$ for each (x_0, y_0) such that $|y(t)| \leqq k(x_0, y_0)$ for all $t > 0$.

This theorem explicitly covers part (iii) of the lemma.

3. Analysis of (2).

THEOREM 1. *If there exists a nonnegative continuous function $h(x)$ satisfying*

$$\left|\frac{f(x, y) + g(x)}{y}\right| \geqq h(x)$$

for (x, y) in quadrants I and III with

$$\int_0^x h(u)\, du \to \pm\infty \quad as \quad x \to \pm\infty,$$

then $(0, 0)$ is globally asymptotically stable.

Remark. This assumption may appear somewhat unnatural and can be replaced by the weaker form $|f(x, y)| \geqq h(x)|y|$. If (1) has the form

$$\ddot{x} + h(x)\dot{x} + g(x) = 0,$$

then the statement is perfectly straightforward. Either of these two forms, however, seriously weakens the result.

Proof of Theorem 1. Using Lemma 1 we see that we need only prove boundedness of solutions. Let (x_0, y_0) be any point and $(x(t), y(t))$ the solution through it. By Lemma 1 there is some $k(x_0, y_0)$ bounding $|y(t)|$. We shall bound the solution within a simple closed curve. From (2) we obtain

$$\frac{dy}{dx} = \frac{-[f(x, y) + g(x)]}{y}$$

which defines the slope of the orbits (paths of solutions). By assumption there exists $h(x) \geqq 0$ such that $dy/dx \leqq -h(x)$ in quadrants I and III. Now $\dot{x} = y$, so for $y > 0$ the solution moves from left to right. Consider the curve defined by

$$y = -\int_{|x_0|}^{x} h(t)\, dt + k(x_0, y_0)$$

starting at $(|x_0|, k(x_0, y_0))$. It intersects the x-axis at $(x_1, 0)$ for some $x_1 > 0$ since $\int_0^x h(t)\, dt \to \infty$ as $x \to \infty$. This curve bounds $(x(t), y(t))$ above and to the right since $dy/dx \leqq -h(x)$ and $y(t) \leqq k(x_0, y_0)$. For $y < 0$ the direction of the field is to the left since $\dot{x} = y$, so continue this curve by drawing a vertical line from $(x_1, 0)$ to $(x_1, -k(x_0, y_0))$. Then continue with a horizontal line to $(-|x_0|, -k(x_0, y_0))$. The curve from this last point defined by

$$y = -\int_{-|x_0|}^{x} h(t)\, dt - k(x_0, y_0)$$

bounds the solution on the left and intersects the negative x-axis at $(x_2, 0)$. Continue with a vertical line to $(x_2, k(x_0, y_0))$. The horizontal line from $(x_2, k(x_0, y_0))$ to $(|x_0|, k(x_0, y_0))$ completes the curve bounding the solution. Application of Lemma 1 completes the proof.

THEOREM 2. *Let*

$$\inf_x |f(x, y)| = h(y).$$

If

$$\int_B^0 \frac{y\,dy}{h(y)}$$

is finite for every finite B, then the null solution to (2) is globally asymptotically stable.

Proof. By Lemma 1 we need only show boundedness. Let (x_0, y_0) be any point in the plane. From (2) we obtain the differential equation for the orbits

$$\frac{dy}{dx} = -\frac{f(x, y) + g(x)}{y}.$$

By assumption, in quadrant I we have

$$-[f(x, y) + g(x)]/y < -h(y)/y.$$

Since $\dot{x} = y$, every solution in quadrant I moves from left to right. Also $|y(t)| \leqq k(x_0, y_0)$ for all $t > 0$, so the curve defined by

$$\int_{k(x_0,y_0)}^{y} \frac{u\,du}{h(u)} = -x + |x_0|$$

bounds $(x(t), y(t))$ on the right in quadrant I. Since the integral is bounded, this curve crosses the x-axis at some point $(x_1, 0)$.

Thus if the solution enters quadrant I it also enters quadrant IV. Now in quadrants IV and III every solution moves from right to left since $\dot{x} = y$. Continue the curve vertically from $(x_1, 0)$ to $(x_1, -k(x_0, y_0))$ and from there horizontally to $(-|x_0|, -k(x_0, y_0))$. Then continue with the curve defined by

$$\int_{-k(x_0,y_0)}^{y} \frac{u\,du}{h(u)} = -x + |x_0|,$$

which bounds the solution on the left and intersects the negative x-axis at $(x_2, 0)$ since the integral is bounded. Continue the curve with a vertical line to $(x_2, k(x_0, y_0))$ and then with a horizontal line to $(|x_0|, k(x_0, y_0))$. This simple closed curve bounds $(x(t), y(t))$, and so the null solution of (2) is globally asymptotically stable.

4. Analysis of (4).

THEOREM 3. *Let* $0 \leqq \alpha < 1$. *The null solution to* (4) *is globally asymptotically stable if and only if*

$$\int_0^{\pm\infty} [p(x) + |g(x)|]\, dx = \pm\infty.$$

Proof. Suppose the integrals diverge. Let (x_0, y_0) be any point and $(x(t), y(t))$ the solution through it. By Lemma 1 there is some $k(x_0, y_0)$ bounding $|y(t)|$. Just as before, we shall construct a simple closed curve bounding the solution.

The orbits of (4) are given by

$$\frac{dy}{dx} = \frac{-p(x)\,|y|^{\alpha} y - g(x)}{y}. \tag{5}$$

For (x, y) in quadrants I and III we have

$$\frac{dy}{dx} \leqq -p(x)\,|y|^{\alpha} \tag{6}$$

and

$$\frac{dy}{dx} \leqq -\frac{g(x)}{y}. \tag{7}$$

The proof is exactly the same as that of Theorem 1, so we shall only show that the curve can be constructed in quadrant I.

Assume first that $\int_0^{\infty} p(x)\, dx = \infty$. Consider the curve starting at $(|x_0|, k(x_0, y_0))$ defined by

$$\int_{k(x_0,y_0)}^{y} u^{-\alpha}\, du = -\int_{|x_0|}^{x} p(u)\, du \tag{8}$$

obtained from (6). There are two possibilities. If $\alpha = 0$, then (8) becomes

$$y - k(x_0, y_0) = -\int_{|x_0|}^{x} p(u)\, du.$$

Since the integral diverges this curve intersects the positive x-axis at some point $(x_1, 0)$. By (6) the solution is bounded from above and to the right by this curve. If $0 < \alpha < 1$, then (8) becomes

$$\frac{y^{-\alpha+1}}{-\alpha+1} - \frac{k(x_0, y_0)^{-\alpha+1}}{-\alpha+1} = -\int_{|x_0|}^{x} p(u)\, du.$$

Since $0 < \alpha < 1$, this curve also intersects the positive x-axis.

Now assume that $\int_0^{\infty} g(x)\, dx = \infty$. From (7) we obtain

$$y^2 - k(x_0, y_0)^2 = -2\int_{|x_0|}^{x} g(u)\,du,$$

which defines a curve starting at $(|x_0|, k(x_0, y_0))$. This curve bounds the solution from above and to the right and intersects the positive x-axis.

The arguments for obtaining the curve in quadrant III are just the same. The simple closed curve obtained bounds the solution and proves the sufficiency of the statement.

Assume that the null solution is globally asymptotically stable. We shall show that the integrals must diverge. We prove the statement only for $\int_0^\infty [p(x) + g(x)]\,dx < \infty$, since the other proof is symmetric and is carried out in quadrant III.

Let (x_0, y_0) be a point in quadrant I with $x_0 > 0$ and arbitrary, but y_0 large and to be specified later. We shall bound the solution through (x_0, y_0) below by a curve which never crosses the x-axis. This will show that $x(t) \to \infty$, contradicting the assumption.

Let $r(x) = p(x) + g(x)$. Since $r(x) > p(x)$ and $r(x) > g(x)$ for $x > 0$, from (5) we obtain

$$\frac{dy}{dx} \geqq \frac{-r(x)\,|\,y\,|^\alpha\, y - r(x)}{y} = -\frac{r(x)[|\,y\,|^\alpha\, y + 1]}{y}.$$

Consider the curve through (x_0, y_0) defined by

$$\int_{y_0}^{y} \frac{u\,du}{u^{\alpha+1} + 1} = -\int_{x_0}^{x} r(u)\,du.$$

Now

$$\int_{y_0}^{0} \frac{u\,du}{u^{\alpha+1} + 1}$$

diverges as $y_0 \to \infty$ so if the curve is to intersect the positive x-axis for y_0 as large as we please, then $\int_0^\infty r(u)\,du$ must diverge. This proves the necessity of the condition.

We are unable to obtain such a sharp theorem for $\alpha \geqq 1$, but some results can be given. Define

$$h(x) = \min\{p(x), |\,g(x)\,|\}$$

and

$$q(x) = \max\{p(x), |\,g(x)\,|\}.$$

THEOREM 4. *Let $\alpha \geqq 1$. The null solution of* (4) *is globally asymptotically*

stable if

$$\int_0^{\pm\infty} h(x)\,dx = \pm\infty.$$

Proof. From (5) we obtain

$$\frac{dy}{dx} \leqq -\frac{h(x)[|\,y\,|^{\alpha}\,y + 1]}{y}$$

for (x, y) in quadrant I. The curve through $(|\,x_0\,|, k(x_0, y_0))$ defined by

$$\int_{k(x_0,y_0)}^{y} \frac{u\,du}{u^{\alpha+1} + 1} = -\int_{|x_0|}^{x} h(u)\,du$$

bounds the solution through (x_0, y_0) from above and to the right. Since the integral on the left is bounded as $y \to 0$ for any finite $k(x_0, y_0)$, the curve so defined crosses the positive x-axis. The remainder of the proof proceeds just as in Theorem 1.

THEOREM 5. *If* $\alpha \geqq 1$ *and if either* $\int_0^{\infty} q(x)\,dx$ *or* $\int_0^{-\infty} q(x)\,dx$ *converges, then there exist unbounded solutions.*

Proof. We prove the statement only for $\int_0^{\infty} q(x)\,dx$ convergent.

From (5) we obtain

$$\frac{dy}{dx} \geqq -\frac{q(x)[y^{\alpha+1} + 1]}{y}$$

for (x, y) in quadrant I. There exist $\epsilon > 0$, $x_0 > 0$, and $y_0 > 0$ such that

$$\int_{x_0}^{\infty} q(x)\,dx < \epsilon \quad \text{and} \quad -\int_{y_0}^{0} \frac{y\,dy}{y^{\alpha+1} + 1} > 2\epsilon.$$

Thus the curve through (x_0, y_0) defined by

$$\int_{y_0}^{y} \frac{u\,du}{u^{\alpha+1} + 1} = -\int_{x_0}^{x} g(u)\,du$$

does not intersect the positive x-axis, but it bounds the solution to (4) through (x_0, y_0) from below. Hence the solution becomes unbounded.

Remark. It is clear that the same treatment can be given for the stability properties of

$$\ddot{x} + p(x)m(\dot{x}) + g(x) = 0.$$

5. Acknowledgments. The author gratefully acknowledges the assistance of Professor D. W. Bushaw, who, in 1959, suggested the problem and guided the work through a modified form of Theorem 1. Also, the referee

kindly suggested improvements which shortened the proofs and generalized the results of the last section.

REFERENCES

[1] D. W. Bushaw, *The differential equation* $\ddot{x} + g(x, \dot{x}) + h(x) = e(t)$, Terminal Report on Contract AF 29(600)-1003, Holloman Air Force Base, New Mexico, 1958.

[2] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Springer-Verlag, Berlin, 1963.

[3] J. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

OPTIMAL CONTROL THEORY FOR NONLINEAR VECTOR DIFFERENTIAL EQUATIONS CONTAINING MEASURES*

W. W. SCHMAEDEKE†

1. Introduction. In an ordinary optimum control problem, one is given a real ordinary differential equation system

$$(\mathfrak{S}) \qquad \frac{dx^i}{dt} = f^i(t, x^1, \cdots, x^n, u^1, \cdots, u^m), \qquad i = 1, \cdots, n,$$

which is a mathematical model of some physical process. The problem of control is to select the real functions $u^j(t)$, $j = 1, \cdots, m$, (called control variables) on an interval of time $t_0 \leqq t \leqq t_1$ such that the solution $x(t)$ of $(\mathfrak{S})$ behaves in a prescribed manner on $[t_0, t_1]$. The usual prescribed behavior consists of requiring that the $u^j(t)$ be found such that the solution $x(t)$ moves from some given initial point x_0 to a prescribed moving target $G(t)$ so as to minimize some cost function or performance index. The control problem for $(\mathfrak{S})$ has been studied extensively; cf. [1], [2], [6], and the bibliographies contained therein.

This paper generalizes the results in [1] by considering in place of $(\mathfrak{S})$ the system

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

$$i = 1, \cdots, n, \quad j = 1, \cdots, m,$$

where a summation convention is implied for the last term with repeated indices. Because it is a problem of great interest to consider the system $(\mathfrak{M})$ for instances in which the functions $u(t)$ have discontinuities of the first kind, it is necessary to develop a theory for treating $(\mathfrak{M})$ when impulse control inputs represented by $\frac{du}{dt}$ arise. Accordingly, the functions $u^j(t)$ are restricted to the class of functions of bounded variation and the derivatives in $(\mathfrak{M})$ are taken in the sense of distribution derivatives. Because the distribution derivative of a function of bounded variation can be identified with a Stieltjes measure, $(\mathfrak{M})$ is rewritten in the notation

* Received by the editors November 27, 1964.

† School of Mathematics, Institute of Technology, University of Minnesota, Minneapolis, Minnesota. This work was supported partly by O. N. R. Contract No. Nonr-3776(00). The results presented in this paper are contained in the doctoral dissertation submitted by the author to the University of Minnesota. The author gratefully acknowledges the inspiration provided by Professor L. Markus under whose guidance this work was completed.

$$(\mathfrak{M}) \qquad Dx^i = f^i(t, x^1, \cdots, x^n, u^1, \cdots, u^m) + g_j{}^i(t)\, Du^j,$$

and the equation is called a measure differential equation. The notation Dx means the distribution derivative of $x(t)$. For example, if

$$u(t) = \begin{cases} 1 & \text{for } t \geqq 0, \\ 0 & \text{for } t < 0, \end{cases}$$

then Du is the Dirac measure $\delta(t)$ (note that the ordinary derivative of $u(t)$ is the zero function almost everywhere). Thus the solution of

$$Dx = 1 + \delta(t) \quad \text{with } x(0) = 0$$

is

$$x(t) = \int_0^t dt + \int_0^t du = \begin{cases} t & \text{for } t \geqq 0, \\ t - 1 & \text{for } t < 0. \end{cases}$$

This paper develops the theory of measure differential equations in a manner entirely analogous to the theory of ordinary differential equations as developed, for example, in [3] and [5]. §6 treats the problem of existence of an optimal control for the system $(\mathfrak{M})$, and Theorems 12 and 13 state conditions such that, if there exists even one allowable control which compels $(\mathfrak{M})$ to behave in the prescribed manner, then an optimal control will exist.

In §7, the optimal control problem is treated for the ordinary system (§). The replacement of the hypothesis that the class of allowable controls be measurable by the hypothesis that they be of uniform bounded variation with values in some subset Ω of R^m allows one to relax the usual hypotheses regarding the linearity of $f^i(t, x^1, \cdots, x^n, u^1, \cdots, u^m)$ with respect to u^j; cf. [1].

2. Ordinary differential equations of the first order containing measures. In this section we study the problem of existence and uniqueness of a solution of a measure differential equation. Before proceeding, we here set the notation and recall some standard definitions. Let t denote a single real variable and let x denote a variable vector $(x^1, x^2, \cdots, x^n)$ in the real n-dimensional number space R^n. Define the magnitude $|x|$ of x by $|x| = |x^1| + \cdots + |x^n|$. The class of m-times continuously differentiable complex-valued functions in R^n is denoted by C^m, $0 \leqq m \leqq \infty$, and the subclass consisting of those functions which have compact support is denoted by $C_c{}^m$. The elements of the conjugate space of $C_c{}^\infty$ are called distributions. Thus distributions are continuous linear functionals on $C_c{}^\infty$. The application of a distribution T to $\varphi \in C_c{}^\infty$ is denoted by $T(\varphi)$ or $T \cdot \varphi$.

A measure μ is a totally additive complex-valued function defined on

the bounded Borel sets of R^n, i.e., on all bounded sets which are obtained from the open sets of R^n by taking countable unions of finite intersections. There exists a one-to-one correspondence between measures and a linear subset of the conjugate space of C_c^∞ given by

$$\mu(\varphi) = \int_{R^n} \varphi(x)\, d\mu(x). \tag{2.1}$$

A theorem of F. Riesz asserts that any continuous linear functional of C_c^0 can be represented in the form (2.1) where $\mu(x)$ is a measure. If T is a distribution identified with a measure and if $\alpha \in C^0$, then αT is again a distribution defined by

$$\alpha T(\varphi) = T(\alpha\varphi). \tag{2.2}$$

The derivative $D_i T \left(\text{or } \frac{\partial T}{\partial x^i}\right)$ of a distribution T is defined by

$$D_i T(\varphi) = -T(D_i\varphi). \tag{2.3}$$

If f is a complex-valued function defined on an interval I (interval as used here shall not include the degenerate case of a single point), the total variation of f on I is defined by

$$v(f, I) = \sup \sum_{i=1}^{N} |f(b_i) - f(a_i)|, \tag{2.4}$$

where the supremum is taken over all finite sets of points a_i, b_i in I with $a_1 < b_1 < a_2 < \cdots < a_n < b_n$. If $v(f, I)$ is finite then f is said to be *of bounded variation* on I. The space $BV(I)$ is defined for an interval I and consists of all scalar functions on I which are of bounded variation. If a is the left endpoint of I, then the norm of f is

$$\|f\| = v(f, I) + |f(a+)|. \tag{2.5}$$

With this norm the space $BV(I)$ is a Banach space. The space $NBV(I)$ is defined for an interval I and consists of those functions f in $BV(I)$ which are normalized by the requirement that f is continuous on the right at each interior point of I and that $f(a+) = 0$, where a is the left endpoint of I. The norm of f is given by the equation

$$\|f\| = v(f, I), \tag{2.6}$$

and with this norm, $NBV(I)$ is a Banach space. If the function $f(t)$ belonging to $BV(I)$ is continuous on the right at every point of the interval $I = [a, b]$, then the measure function $\mu([a, d]) = f(d) - f(a)$ and $\mu((c, d]) = f(d) - f(c)$ for $a < c < d \leqq b$ has a regular countably additive extension to the σ-field Σ of all Borel sets in I. This extension is called the Borel-

Stieltjes measure in I determined by the function f. Now let Σ^* consist of all sets of the form $E \cup N$, where E is in Σ and N is a subset of a set M in Σ with $v(\mu, M) = 0$. Then Σ^* is a σ-field and if the domain of μ is extended to Σ^* by defining $\mu(E \cup N) = \mu(E)$, the extended function is countably additive on Σ^* and the measure space (I, Σ^*, μ) is called the *Lebesgue extension* of the measure space (I, Σ, μ). The function μ with domain Σ^* is the Lebesgue-Stieltjes measure on I determined by the function f and the integral $\int_I g(t)\mu(dt)$ is written $\int_a^b g(t)\, df(t)$. A distribution $F(\varphi)$ on an interval I is to be identified with the Stieltjes measure $d\psi(t)$ if for every closed finite interval J contained in I, $\psi(t)$ is of bounded variation on J and

$$F(\varphi) = \int_J \varphi(t)\, d\psi(t)$$

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

$$F(\varphi) = \int_J f(t)\varphi(t)\, dt$$

for all $\varphi \in C_c^\infty(J)$.

Let (S, Σ, μ) be a measure space, f a complex valued μ-integrable function, and

$$\lambda(E) = \int_E f(s)\mu(ds), \qquad E \in \Sigma. \tag{2.7}$$

Then by [4, Corollary 6, p. 180], a function g on S to the Banach space B is λ-integrable if and only if $f \cdot g$ is μ-integrable, and in this case we have

$$\int_E g(s)\lambda(ds) = \int_E f(s)g(s)\mu(ds), \qquad E \in \Sigma. \tag{2.8}$$

The fact that vector and matrix notation will be used throughout the remainder of this paper requires that some of the previous notions be extended to vector valued functions. The space $BV(I)^*$ is defined for an interval I and consists of all vector functions with values in R^n whose individual components belong to $BV(I)$. The norm of f is

$$\|f\|^* = \sum_{i=1}^n \|f^i\| = \sum_{i=1}^n \{v(f^i, I) + |f^i(a+)|\}, \tag{2.9}$$

where a is the left endpoint of I. With this norm the space $BV(I)^*$ is a Banach space. Similarly, the space $NBV(I)^*$ of all vector functions f with values in R^n and whose individual components belong to $NBV(I)$ is a

Banach space with norm given by

$$\|f\|^* = \sum_{i=1}^{n} \|f^i\| = \sum_{i=1}^{n} v(f^i, I). \tag{2.10}$$

Let S be a domain (i.e., an open connected set) in the (t, x) space R^{n+1} and let $f(t, x)$ be a real n-vector function defined on S. Let $u(t)$ be a real m-vector function of bounded variation, continuous from the right on an interval I_1, and let $G(t)$ be a continuous $n \times m$ matrix defined on I_1. Let (t_0, x_0) be a point in S with t_0 also in I_1 and let a differential equation

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du, \quad x(t_0) = x_0,$$

involving f, G, u and x be given where the operations of differentiation are to be understood in the sense of distribution derivatives with respect to the real variable t. We shall call $(\mathfrak{M})$ a measure differential equation of the first order because the distribution derivative Du of a function of bounded variation may always be identified with a measure. Then the central problem of this section consists of finding an interval I contained in I_1 (such that t_0 belongs to I) and a real bounded variation n-vector $x(t)$ defined on the interval I such that $(t, x(t))$ belongs to S for all t in I, the initial point x_0 is equal to $x(t_0)$, and the distribution derivative of $x(t)$ on I is $f(t, x) + G(t)\, Du$. For convenience we summarize the foregoing as a definition.

DEFINITION 1. A solution $x(t)$ of $(\mathfrak{M})$ is a real bounded variation n-vector $x(t)$ together with an interval I containing the given initial time t_0 such that $x(t)$ is continuous from the right on I and

(i) $$(t, x(t)) \in S \quad \text{for} \quad t \in I;$$

(ii) $$x(t_0) = x_0;$$

(iii) the distribution derivative of $x(t)$ on I is $f(t, x) + G(t)\, Du$.

Now consider the integral equation

$$(\mathfrak{g}) \qquad x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds + \int_{t_0}^{t} G(s)\, du(s).$$

DEFINITION 2. A solution $x(t)$ of $(\mathfrak{g})$ is a real bounded variation n-vector $x(t)$ together with an interval I such that

(i) $$(t, x(t)) \in S \quad \text{for} \quad t \in I;$$

(ii) $x(t)$ satisfies the integral equation.

Note. A solution $x(t)$ of $(\mathfrak{g})$ is necessarily continuous from the right. If $x(t)$ is any continuous function satisfying the integral equation, then $x(t)$ is of bounded variation and is a solution of $(\mathfrak{g})$. By examining the last integral in $(\mathfrak{g})$ it is apparent that $x(t)$ only has discontinuities where $u(t)$ does.

THEOREM 1. *A solution $x(t)$ of $(\mathfrak{s})$ is a solution $x(t)$ of $(\mathfrak{M})$ and conversely.*

Proof. Let $x(t)$ be a solution of $(\mathfrak{s})$ and let $F^i(\varphi)$ be the distribution on I to be identified with the ith component $x^i(t)$; we then have for any closed interval J contained in I that

$$F^i(\varphi) = \int_J \left[x_0{}^i + \int_{t_0}^t f^i(s, x(s))\, ds + \int_{t_0}^t [G(s)\, du(s)]^i \right] \varphi(t)\, dt \tag{2.11}$$

for all φ in $C_c^\infty(J)$. The derivative-distribution is then

$$\begin{aligned} DF^i(\varphi) &= -F^i(\varphi') \\ &= -\int_J \left[x_0{}^i + \int_{t_0}^t f^i(s, x(s))\, ds + \int_{t_0}^t [G(s)\, du(s)]^i \right] \varphi'(t)\, dt. \end{aligned} \tag{2.12}$$

Integrate by parts and obtain

$$\begin{aligned} DF^i(\varphi) = \int_J \varphi(t)\, dx_0{}^i + \int_a^b \varphi(t) f^i(t, x(t))\, dt \\ + \int_J \varphi(t)\, d\left\{ \int_{t_0}^t \sum_{j=1}^m g_{ij}(s)\, du^j(s) \right\}, \end{aligned} \tag{2.13}$$

where $g_{ij}(s)$ is the i,jth element of $G(s)$ and $u^j(s)$ is the jth component of $u(s)$.

Consider a typical term from the last integral and apply (2.8) to obtain

$$\int_J \varphi(t)\, d\left\{ \int_{t_0}^t g_{ij}(s)\, du^j(s) \right\} = \int_J \varphi(t) g_{ij}(t)\, du^j(t). \tag{2.14}$$

Hence

$$DF^i(\varphi) = \int_J \varphi(t) f^i(t, x(t))\, dt + \int_J \varphi(t)[G(t)\, du(t)]^i. \tag{2.15}$$

Since $G(t)$ is continuous, the last continuous linear functional in (2.15) is, according to (2.2), identified with the measure $[G(t)\, du(t)]^i$ while the first continuous linear functional there is identified with the function $f^i(t, x(t))$. This holds for each $i = 1, 2, \cdots, n$ and therefore the derivative-distribution $DF(\varphi)$ is identified with $f(t, x(t)) + G(t)\, Du$ and $x(t)$ is also a solution of $(\mathfrak{M})$.

Conversely, suppose $x(t)$ is a solution of $(\mathfrak{M})$. Then we have

$$\int_J \varphi(t) Dx^i(t) = \int_J \varphi(t) f^i(t, x(t))\, dt + \int_J \varphi(t)[G(t)\, du(t)]^i. \tag{2.16}$$

By using (2.8) again we may write

$$\int_J \varphi(t)[G(t)\, du(t)]^i = \int_J \varphi(t)\, d\left(\int_{t_0}^t [G(s)\, du(s)]^i \right) \tag{2.17}$$

and then integrate the right hand side by parts to obtain

$$\int_J \varphi(t)[G(t)\,du(t)]^i = -\int_J \varphi'(t)\left[\int_{t_0}^t [G(s)\,du(s)]^i\right] dt. \tag{2.18}$$

Next, integrate the first two integrals in (2.16) by parts to obtain

$$\begin{aligned}\int_J \varphi'(t)[x^i(t) - x_0{}^i]\,dt \\ = \int_J \varphi'(t)\left[\int_{t_0}^t f^i(s, x(s))\,ds + \int_{t_0}^t [G(s)\,du(s)]^i\right] dt.\end{aligned} \tag{2.19}$$

From (2.19) we have almost everywhere in J that

$$x^i(t) = x_0{}^i + \int_{t_0}^t f^i(s, x(s))\,ds + \int_{t_0}^t [G(s)\,du(s)]^i. \tag{2.20}$$

But, since $x^i(t)$ is continuous from the right, being a solution of $(\mathfrak{M})$, and since the right hand side of (2.20) is a function of t that is continuous from the right, then equality holds everywhere in J for (2.20) and thus $x(t)$ is a solution of $(\mathcal{g})$. This completes the proof.

Remark. If $u(t)$ is absolutely continuous, then all of the preceding definitions reduce to the conventional theory for absolutely continuous solutions $x(t)$ of systems of ordinary differential equations, since the distribution derivative is then the usual derivative.

We also note that there is no distribution solution $x(t)$ more general than a function of bounded variation. This follows since $f(t, x) + G(t)\,Du$ is a measure—and, in fact, it can only be defined for nonlinear f if $x(t)$ is a function.

Before proceeding to the local existence theorem, we make the following normalization. Let $y(t) = x(t) - x_0$; then

$$y(t) = \int_{t_0}^t f(s, y(s) + x_0)\,ds + \int_{t_0}^t G(s)\,du(s), \tag{2.21}$$

and $y(t_0) = 0$. Hence a solution of $(\mathcal{g})$ with initial point $x(t_0) = x_0$ is given by $y(t) + x_0$, where $y(t)$ is a solution of (2.21). Next make a change of variable $s = r + t_0$ in (2.21) to obtain

$$\begin{aligned}y(t) = \int_0^{t-t_0} f(r + t_0, y(r + t_0) + x_0)\,dr \\ + \int_0^{t-t_0} G(r + t_0)\,du(r + t_0).\end{aligned} \tag{2.22}$$

Finally, substitute $t = \tau + t_0$, obtaining

$$\begin{aligned}y(\tau + t_0) = \int_0^{\tau} f(r + t_0, y(r + t_0) + x_0)\,dr \\ + \int_0^{\tau} G(r + t_0)\,du(r + t_0),\end{aligned} \tag{2.23}$$

and change the notation so that

$$\begin{aligned} \hat{y}(\tau) &= y(\tau + t_0), \\ \hat{f}(r, \hat{y}(r)) &= f(r + t_0, y(r + t_0) + x_0), \\ \hat{G}(r) &= G(r + t_0), \\ \hat{u}(r) &= u(r + t_0). \end{aligned}$$

Then (2.23) becomes

$$\hat{y}(\tau) = \int_0^\tau \hat{f}(r, \hat{y}(r))\, dr + \int_0^\tau \hat{G}(r)\, d\hat{u}(r), \tag{2.24}$$

which is of the same form as $(\mathcal{I})$ but with the difference that the initial point is zero at time $\tau = 0$, i.e., $\hat{y}(0) = 0$. Thus a solution of $(\mathcal{I})$ with $x(t_0) = x_0$ is just $x(t) = \hat{y}(t - t_0) + x_0$, where $\hat{y}(t)$ is a solution of (2.24) with $\hat{y}(0) = 0$. It is apparent that we may consider from now on the equation

$$Dx = f(t, x) + G(t)\, Du, \quad x(0) = 0, \tag{2.25}$$

or equivalently, the equation

$$x(t) = \int_0^t f(s, x(s))\, ds + \int_0^t G(s)\, du(s). \tag{2.26}$$

THEOREM 2. (LOCAL EXISTENCE AND UNIQUENESS). *Consider the measure differential equation*

$$Dx = f(t, x) + G(t)\, Du, \quad x(0) = 0, \tag{$\mathfrak{M}$}$$

where x is an n-vector, $G(t)$ is a continuous $n \times m$ matrix on $[-a, a]$ for $a > 0$, and $u(t)$ is a bounded variation m-vector which is continuous from the right on $[-a, a]$. Let $\left\| \int_0^t G(s)\, du(s) \right\|^* < b$, where the variation referred to in the norm is taken over $[-a, a]$, and let $f(t, x)$ be defined on

$$R_{ab}: \quad -a \leqq t \leqq a, \quad |x| \leqq b.$$

The following assumptions, hereafter referred to as Assumptions A, *will be made with regard to $f(t, x)$.*

A1. *$f(t, x)$ is measurable in t for each fixed x with $|x| \leqq b$;*

A2. *$f(t, x)$ satisfies a Lipschitz condition in R_{ab} with respect to x for a constant K;*

A3. *There exists a summable function $r(t)$ on $[-a, a]$ such that $|f(t, x)| \leqq r(t)$ for (t, x) in R_{ab}.*

Conclusion: Then there exists a constant a' such that $0 < a' < a$ and for

which we have

$$\text{(i)} \qquad \int_{-a'}^{a'} r(t)\,dt < \frac{\theta}{2}, \qquad \theta \equiv b - \left\| \int_0^t G(s)\,du(s) \right\|^*,$$

$$\text{(ii)} \qquad 4Ka' < 1,$$

and there exists a unique solution $x(t)$ *of* $(\mathfrak{M})$ *on* $[-a', a']$ *with* $x(0) = 0$.

Proof. Let W be the subspace of the Banach space $BV(I)^*$ (where $I = [-a', a']$) for which $x \in W$ implies that $\| x \|^* \leqq b$. We shall show that the mapping T defined by

$$Ty(t) = \int_0^t f(s, y(s))\,ds + \int_0^t G(s)\,du(s) \tag{2.27}$$

is a contraction mapping of W into W and thus by the principle of contraction mappings there is a unique fixed point.

First we must show that T maps W into W. Now it is evident that $\int_0^t f(s, y(s))\,ds$ and $\int_0^t G(s)\,du(s)$ are functions of bounded variation on $[-a', a']$ and thus Ty is a function of $BV(I)^*$. To show $\| Ty \|^* \leqq b$ we proceed as follows:

$$\| Ty \|^* \leqq \left\| \int_0^t f(s, y(s))\,ds \right\|^* + \left\| \int_0^t G(s)\,du(s) \right\|^*, \tag{2.28}$$

but

$$\begin{aligned} &\left\| \int_0^t f(s, y(s))\,ds \right\|^* \\ &\quad = \sum_{i=1}^n \left\{ v\left(\int_0^t f^i(s, y(s))\,ds, [-a', a'] \right) + \left| \int_0^{-a'+} f^i(s, y(s))\,ds \right| \right\} \\ &\quad \leqq \sum_{i=1}^n \left\{ \int_{-a'}^{a'} |f^i(s, y(s))|\,ds + \int_{-a'}^0 |f^i(s, y(s))|\,ds \right\} \qquad (2.29) \\ &\quad \leqq 2 \int_{-a'}^{a'} \sum_{i=1}^n |f^i(s, y(s))|\,ds \leqq 2 \int_{-a'}^{a'} |f(s, y(s))|\,ds \\ &\quad \leqq 2 \int_{-a'}^{a'} r(s)\,ds. \end{aligned}$$

Since $\left\| \int_0^t G(s)\,du(s) \right\|^* < b$, let $\theta \equiv b - \left\| \int_0^t G(s)\,du(s) \right\|^*$. Then (i) follows immediately, because of the summability of $r(t)$, where a' is chosen so small that (ii) is also satisfied. It then follows from (2.29) that

$$\left\| \int_0^t f(s, y(s))\,ds \right\|^* < \theta, \tag{2.30}$$

and furthermore by (2.28) and (2.30) it follows that

$$\| Ty \|^* < \theta + b - \theta = b. \tag{2.31}$$

Thus T maps W into itself.

Finally, we must show that T is a contraction. To this end, consider $\| Ty - Tz \|^*$ in the same fashion as (2.29) to obtain

$$\begin{aligned} \| Ty - Tz \|^* &= \left\| \int_0^t f(s, y(s))\, ds - \int_0^t f(s, z(s))\, ds \right\|^* \\ &\leqq 2 \int_{-a'}^{a'} | f(s, y(s)) - f(s, z(s)) |\, ds \\ &\leqq 2K \int_{-a'}^{a'} | y(s) - z(s) |\, ds. \end{aligned} \tag{2.32}$$

But it is easily shown that

$$| y(s) - z(s) | \leqq \| y - z \|^* \tag{2.33}$$

for all t such that $-a' \leqq t \leqq a'$. Thus

$$\| Ty - Tz \|^* \leqq 4K \cdot a' \| y - z \|^*, \tag{2.34}$$

but $4Ka' < 1$, and T is therefore a contraction as was asserted.

Remark. By reversing the steps in the normalization procedure it is observed that the local existence and uniqueness theorem holds for arbitrary initial points (t_0, x_0) which are centered in an appropriate rectangle wherein Assumptions A are known to hold.

It is possible to establish the local existence and uniqueness theorem under less restrictive hypotheses on the size of the rectangle domain if one only desires a solution for times t greater than the initial time t_0.

THEOREM 3. *Consider the measure differential equation*

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du, \quad x(0) = 0,$$

where x is an n-vector, $G(t)$ is a continuous $n \times m$ matrix on $[0, a]$ for $a > 0$ and $u(t)$ is a bounded variation m-vector which is continuous from the right on $[0, a]$. Let $f(t, x)$ be defined on a rectangle

$$R_{ab}: \quad 0 \leqq t \leqq a, \quad | x | \leqq b.$$

The following assumptions will be made with regard to $f(t, x)$.

B1. *$f(t, x)$ is measurable in t for each fixed x with $| x | \leqq b$;*

B2. *$f(t, x)$ satisfies a Lipschitz condition in R_{ab} with respect to x for a constant K:*

B3. *There exists a summable function $r(t)$ on $[0, a]$ such that $| f(t, x) | \leqq r(t)$ for (t, x) in R_{ab}.*

Conclusion: *Then there exists a constant* a' *such that* $0 < a' < a$ *and for which we have*

(i) $$\int_0^{a'} r(t)\, dt < \frac{b}{2},$$

(ii) $$Ka' < 1,$$

(iii) $$\left\| \int_0^t G(s)\, du(s) \right\|^* < \frac{b}{2} \qquad (\textit{norm is taken on } [0, a']),$$

and there exists a unique solution $x(t)$ *of* $(\mathfrak{M})$ *on* $[0, a']$ *with* $x(0) = 0$.

Proof. The existence of an a' such that (i) and (ii) are true is obvious. The fact that (iii) is true follows from the right continuity of $u(t)$ on $[0, a]$ which makes the indefinite variation of the function $\int_0^t G(s)\, du(s)$ right continuous also.

Let W be the subspace of the complete metric space $NBV([0, a'])^*$ for which $x \in W$ implies $\| x \|^* \leqq b$. It can be shown, just as in the proof of Theorem 2, that the mapping T defined by

(2.35) $$Ty(t) = \int_0^t f(s, y(s))\, ds + \int_0^t G(s)\, du(s)$$

is a contraction mapping of W into W and thus by the principle of contraction mappings there is a unique fixed point in W. This function, call it $x(t)$, does indeed originate at $x(0) = 0$ because of the normalization requirement on functions in W.

Remark. A solution $x(t)$ of a measure differential equation $(\mathfrak{M})$ in a rectangle R_{ab} can be found for times greater than the initial time t_0 no matter how small the number b regulating the distance of points x from the initial point x_0. Contrast this with the situation in Theorem 2 in which the desire for a solution in an interval containing the initial time in its interior placed a requirement on the size of b in terms of the variation of $\int_{t_0}^t G(s)\, du(s)$ on $[t_0 - a', t_0 + a']$.

3. Global extension of solutions. The fact that a solution $x(t)$ of $(\mathfrak{M})$ may have discontinuities presents some difficulty whenever we try to consider a solution near the boundary of a domain S wherein $f(t, x)$ is defined because the solution may take a jump at some particular time which would carry it out of S. To prevent such an anomaly from complicating the analysis, it is desirable to consider systems of equations in domains which will not exhibit such abnormal behavior. Following the idea in the theory of ordinary differential equations of defining a Carathéodory system, we make the following definition.

DEFINITION 3. Consider the measure differential equation

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du,$$

where x is an n-vector, $G(t)$ is a continuous $n \times m$ matrix on an interval I, and $u(t)$ is a bounded variation m-vector continuous from the right on this same interval. Let $f(t, x)$ be defined in a neighborhood of a domain S of R^{n+1} such that for each point (t_0, x_0) in S there exist a rectangle R_{ab} centered at (t_0, x_0), a constant $K > 0$, and a function $r(t)$ summable on the interval $[t_0 - a, t_0 + a]$, a subinterval of I, such that

(1) $f(t, x)$ is measurable in t for each fixed x such that $(t, x) \in R_{ab}$;

(2) $f(t, x)$ satisfies a Lipschitz condition with constant K with respect to x for all $(t, x) \in R_{ab}$:

$$(3) \qquad |f(t, x)| \leqq r(t) \text{ in } R_{ab} ;$$

$$(4) \qquad \left\| \int_{t_0}^{t} G(s)\, du(s) \right\|^* < b \qquad \text{(norm taken on } [t_0 - a, t_0 + a]).$$

A system $(\mathfrak{M})$ satisfying these conditions will be called a *Carathéodory Measure System* in S and will be denoted for brevity by the symbols CMS.

Remark. The rectangles R_{ab} centered at points of S need not themselves be contained in S. The definition of CMS requires that $f(t, x)$ be defined in a neighborhood of S which contains all the rectangles centered in S. More insight to the nature of this remark is given in the example below.

Remark. A CMS has a unique solution $x(t)$ on an interval $[t_0 - a, t_0 + a]$ (for appropriately chosen a) such that $x(t_0) = x_0$ for every point (t_0, x_0) belonging to S.

Example. Consider the measure differential equation

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du$$

in a domain S in R^{n+1}. Let $G(t)$ be a continuous matrix on an interval I and let $u(t)$ be of bounded variation and continuous from the right on I. Let

$$(3.1) \qquad S^* = \{(t, x) \mid (t, x) \in S,\ t \in I\},$$

and let I^* be the subset of I consisting of those t for which there exists an x such that (t, x) is in S^*. Thus I^* is open and we may assume that it is an interval for, in the contrary case, we can restrict our attention to one of its interval components. Thus S^* is a domain in R^{n+1}. Now for any t_0 in I^*, define

$$(3.2) \qquad J(t) = \int_{t_0}^{t} G(s)\, du(s) - \int_{t_0}^{t-} G(s)\, du(s)$$

for all t in I^*. Notice that $J(t)$ does not depend on t_0 ; in fact we have

$$J(t) = G(t)[u(t) - u(t -)]. \tag{3.3}$$

For each t_1 in I^* define

$$S^*_{t_1} = \{(t_1, x) \mid (t_1, x) \in S^*\} \tag{3.4}$$

and define

$$S^{**}_{t_1} = \{(t_1, x) \mid (t_1, x) \in S^*_{t_1},\ d(x, \partial S^*_{t_1}) > |J(t_1)|\}. \tag{3.5}$$

Finally let

$$S^{**} = \bigcup_{t_1 \in I^*} S^{**}_{t_1}.$$

It is easy to conceive of situations in which S^{**} is not connected (i.e., $|J(t_1)|$ being so large that $S^{**}_{t_1}$ is empty) but S^{**} is open in R^{n+1}. To see the latter, observe that S^{**} is merely S^* with certain constricting cuts taken at times corresponding to discontinuities of $u(t)$. Since $u(t)$ is of bounded variation it can have at most a denumerable number of discontinuities and furthermore the depths of the cuts in S^* cannot all be the same depth—a situation which would cause trouble in proving that S^{**} is open. Thus given any point $P = (t_1, x_1)$ in S^{**} it is easy to find an open set in S^{**} containing it, because any cuts in the vicinity of the point are of necessity isolated, that is, infinitely many cuts do not exist whose depths are greater than the distance of P from the boundary of S^* (measured in the hyperplane $t = t_1$).

Finally, if $f(t, x)$ is continuous in S and $\partial f(t, x)/\partial x$ exists and is continuous in S, then $(\mathfrak{M})$ is a CMS in S^{**}; for given (t_0, x_0) in S^{**}, take R_{ab} centered at (t_0, x_0) such that $b = |J(t_0)| + \epsilon/2$, and choose a so small that

$$\left\| \int_{t_0}^{t} G(s)\, du(s) \right\|^* \leqq |J(t_0)| + \frac{\epsilon}{4},$$

where ϵ is less than the distance from (t_0, x_0) to the boundary of S^{**} measured on the hyperplane $t = t_0$. If t_0 is not a limit point of discontinuities of $u(t)$, it is obvious that such a choice of a can be made. If t_0 is a limit point of discontinuities, then letting t_k be the sequence of times at which the discontinuities occur, it is clear that the depth of the cut at t_k must decrease to zero as k approaches infinity and hence the variation of $\int_{t_0}^{t} G(s)\, du(s)$ on $[t_0 - a, t_0 + a]$ must decrease to the value of the jump at t_0. Thus a choice of a can be made in this case also. Then take

$$r(t) = \max_{(t,x) \in R_{ab}} |f(t, x)| = \text{const.},$$

and for the Lipschitz constant K use

$$K = n \max_{(t,x)\in R_{ab}} \left| \frac{\partial f^i}{\partial x^j} \right|.$$

Then

$$| f^i(t, y) - f^i(t, z) | \leqq \sum_{j=1}^{n} \left| \frac{\partial f^i}{\partial x^j} \right| | y^j - z^j |,$$

and summing on i we obtain

$$| f(t, y) - f(t, z) | \leqq K | y - z |.$$

Hence the verification that $(\mathfrak{M})$ is a CMS is completed.

THEOREM 4. *Consider the CMS*

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du$$

in a domain S. Then there exists a unique solution $\varphi(t, t_0, x_0)$ of $(\mathfrak{M})$ such that $\varphi(t_0, t_0, x_0) = x_0$ for every point $(t_0, x_0) \in S$, where $\varphi(t, t_0, x_0)$ is defined on a maximal open interval (τ_-, τ_+). Every solution of $(\mathfrak{M})$ through the initial point (t_0, x_0) is merely a restriction of $\varphi(t, t_0, x_0)$ to some subinterval containing t_0.

DEFINITION 4. Call $\varphi(t, t_0, x_0)$ on (τ_-, τ_+) the *maximal solution through* (t_0, x_0).

Proof of Theorem 4. Without loss of generality take $t_0 = 0$, $x_0 = 0$. Let Σ be the set of all solutions of $(\mathfrak{M})$ through $(0, 0)$. The set Σ is not empty since the local existence and uniqueness theorem provides at least one such solution. Let $\varphi(t)$ and $\psi(t)$ be any two solutions in Σ. Then they must coincide on their common interval of definition. To see this, let the interval common to the two intervals of definition be denoted by (t_1, t_2). If they do not coincide on this interval, then let τ, where $t_1 < \tau < t_2$ be such a time for which $\varphi(\tau) \neq \psi(\tau)$. There are two cases which must be considered: (i) $\tau < 0$ and (ii) $\tau > 0$.

Case (i), $\tau < 0$: Since the solutions are continuous from the right, there is a first time $t' < 0$ such that $t' > \tau$ and $\varphi(t') = \psi(t')$. If this were not so, then the set of points to the right of τ for which the solutions were equal would have a limit point $t^* < 0$ such that for points sufficiently near t^* but greater than t^* the two solutions would be equal. But continuity from the right implies that the two solutions are equal at t^* also, contradicting the existence of t^*. Now apply the local existence and uniqueness theorem for the initial point $(t', \varphi(t'))$ to deduce that φ and ψ coincide on some interval about t', thus contradicting the definition of t'. Hence the solutions cannot differ on the part of their common interval of definition which lies to the left of the initial time $t = 0$.

Case (ii), $\tau > 0$: Let t' be the least upper bound of all times $\tilde{t}$ for which φ and ψ are equal to the left of $\tilde{t}$ as well as at $\tilde{t}$ itself. It is clear that $t' > 0$. On one hand we have that $\varphi(t') \neq \psi(t')$ because if they were equal at t', an application of the local existence and uniqueness theorem would yield the result that the two solutions were equal on an interval about t', thus contradicting its definition. On the other hand, if t' is assumed to be less than τ, then a glance at the two formulas

$$\varphi(t') = \int_0^{t'} f(s, \varphi(s))\, ds + \int_0^{t'} G(s)\, du(s),$$

$$\psi(t') = \int_0^{t'} f(s, \psi(s))\, ds + \int_0^{t'} G(s)\, du(s),$$

shows that $\varphi(t') = \psi(t')$ in contradiction to the previous result. Thus t' is in fact the right hand endpoint t_2 of the interval common to both of the intervals of definition of φ and ψ and they do indeed coincide on their common interval of definition.

Now consider the set I formed by taking the union of all intervals of definition of solutions in Σ. The set I is clearly an interval and we define $\varphi(t, t_0, x_0)$ on I by taking for its value the value of any solution of the class Σ which is itself defined at t. By the previous calculations it is to be noticed that it will not make any difference which of the solutions (defined at t of course) is taken from Σ because they all coincide on their common intervals of definition. Thus, by construction, every solution Σ is just a restriction of $\varphi(t, t_0, x_0)$ to some subinterval of I containing t_0.

If we consider the function $\varphi(t)$ to be a solution on the interval $a < t < b$ (where (a, b) may be any subinterval of (τ_-, τ_+), including (τ_-, τ_+) itself) we may wonder whether or not the solution $\varphi(t)$ has a limit as t approaches b from the left or a from the right. The following theorem establishes that these limits exist for a CMS.

THEOREM 5. *Consider the CMS*

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du$$

in a domain S of R^{n+1}. Given a solution $\varphi(t)$ of $(\mathfrak{M})$ on an interval (a, b), suppose that $\varphi(t)$ is contained in a subset C of S such that $|f(t, x)| \leqq r(t)$, where $r(t)$ is integrable on (a, b) for all (t, x) in C. Then both limits

$$\varphi(a+) = \lim_{t \to a+} \varphi(t), \quad \varphi(b-) = \lim_{t \to b-} \varphi(t)$$

exist. If furthermore C is compact then the maximal extension of $\varphi(t)$ in S must reach $S - C$ both as t increases and as t decreases.

Proof. From the representation

$$(3.6) \qquad \varphi(t) = x_0 + \int_{t_0}^{t} f(s, \varphi(s))\, ds + \int_{t_0}^{t} G(s)\, du(s)$$

for $a < t < b$, we have for $a < t_1 < t_2 < b$ that

$$\begin{aligned} | \varphi(t_1) - \varphi(t_2) | &\leqq \int_{t_1}^{t_2} | f(s, \varphi(s)) | \, ds + Mv(u, [t_1, t_2]) \\ &\leqq \int_{t_1}^{t_2} r(t) \, dt + Mv(u, [t_1, t_2]), \end{aligned} \tag{3.7}$$

where $| G(s) | \leqq M$ for $a \leqq s \leqq b$. Then since $u(t)$ is of bounded variation on $[a, b]$ and continuous from the right, we can maintain the sense of the inequality in (3.7) by adding the nonnegative term $Mv(u, [a, t_1])$ to obtain

$$| \varphi(t_1) - \varphi(t_2) | \leqq \int_{t_1}^{t_2} r(t) \, dt + Mv(u, [a, t_2]). \tag{3.8}$$

Now choose t_2 so close to a (but $a < t_2$) that

$$v(u, [a, t_2]) < \frac{\epsilon}{2M},$$

and then choose t_1 so close to t_2 that

$$\int_{t_1}^{t_2} r(t) \, dt < \frac{\epsilon}{2}.$$

Then from (3.8) we obtain

$$| \varphi(t_1) - \varphi(t_2) | < \frac{\epsilon}{2} + M \frac{\epsilon}{2M} = \epsilon \tag{3.9}$$

for all $a < t_1 < t_2$ such that t_2 is sufficiently close to a and by Cauchy's criterion, $\lim_{t \to a+} \varphi(t)$ exists. In fact, since $\varphi(t)$ is continuous from the right, then $\varphi(a+) = \varphi(a)$.

For the other limit we define

$$\begin{aligned} \psi(t) &= \int_{t_0}^{t} f(s, \varphi(s)) \, ds, \\ V(t) &= \int_{t_0}^{t} G(s) \, du(s). \end{aligned} \tag{3.10}$$

From (3.6) it is clear that

$$\varphi(b-) = \psi(b-) + V(b-). \tag{3.11}$$

But for $a < t_1 < t_2 < b$ we have

$$| \psi(t_1) - \psi(t_2) | \leqq \int_{t_1}^{t_2} r(t) \, dt, \tag{3.12}$$

so that as t_1 and t_2 approach b from the left with $t_1 < t_2$, $\psi(b-)$ exists by Cauchy's criterion. The function $V(t)$ on the other hand is of bounded

variation on the interval $[t_0, b]$ and hence $V(b-)$ exists. Thus from (3.11) $\varphi(b-) = \lim_{t\to b-} \varphi(t)$ exists.

Finally, we consider what results if C is compact but the maximal extension of $\varphi(t)$ in S does not reach $S - C$. Treating the extension to the left first, we note that $\varphi(\tau_-+) \equiv \varphi(\tau_-)$ exists and belongs to S. By the local existence and uniqueness theorem we can find a solution $\psi(t)$ through the point $(\tau_-, \varphi(\tau_-))$ on an interval $[\tau_- - a, \tau_- + a]$ which coincides with $\varphi(t)$ on $[\tau_-, \tau_- + a]$ and hence provides an extension of φ to the left of τ_-, contradicting the definition of that point. Thus $\varphi(t)$ must have reached $S - C$ as t approached τ_-. In fact, it is easy to see that the limit point $(\tau_-, \varphi(\tau_-))$ cannot belong to S or else an application of the local existence and uniqueness theorem would provide a contradiction of the definition of τ_-. Thus a maximal solution of $(\mathfrak{M})$ in S can be continued up to the boundary of S as t decreases.

Lastly, we treat the case when t approaches τ_+ from the left. We have that $\varphi(\tau_+ -)$ exists and is in C and hence by the definition of a CMS, there is a rectangle $R_{\alpha\beta}$ about $(\tau_+, \varphi(\tau_+ -))$ such that

$$\left\| \int_{\tau_+}^{t} G(s)\, du(s) \right\|^* < \beta,$$

where the norm is taken over $[\tau_+ - \alpha, \tau_+ + \alpha]$. But then clearly

$$| J(\tau_+) | < \left\| \int_{\tau_+}^{t} G(s)\, du(s) \right\|^* < \beta, \tag{3.13}$$

in other words, $\varphi(\tau_+ -) + J(\tau_+)$ belongs to $\bar{R}_{\alpha\beta}$. Now $(\mathfrak{M})$ satisfies the conditions of Theorem 3 for some rectangle $R_{\alpha'\beta'}$ contained in $\bar{R}_{\alpha\beta}$ and centered at $(\tau_+, \varphi(\tau_+ -) + J(\tau_+))$, and hence a solution $\lambda(t)$ of $(\mathfrak{M})$ exists on $[\tau_+, \tau_+ + \hat{\alpha}]$ for sufficiently small $\hat{\alpha}$. We can represent this solution by

$$\lambda(t) = \varphi(\tau_+-) + J(\tau_+) + \int_{\tau_+}^{t} f(s, \lambda(s))\, ds + \int_{\tau_+}^{t} G(s)\, du(s). \tag{3.14}$$

This serves to extend $\varphi(t)$ to the right of τ_+ because we may rewrite (3.14) as

$$\begin{aligned} \lambda(t) &= x_0 + \int_{t_0}^{\tau_+} f(s, \varphi(s))\, ds + \int_{t_0}^{\tau_+-} G(s)\, du(s) \\ &\quad + \int_{t_0}^{\tau_+} G(s)\, du(s) - \int_{t_0}^{\tau_+-} G(s)\, du(s) \\ &\quad + \int_{\tau_+}^{t} f(s, \lambda(s))\, ds + \int_{\tau_+}^{t} G(s)\, du(s) \\ &= x_0 + \int_{t_0}^{\tau_+} f(s, \varphi(s))\, ds + \int_{\tau_+}^{t} f(s, \lambda(s))\, ds \\ &\quad + \int_{t_0}^{t} G(s)\, du(s). \end{aligned} \tag{3.15}$$

Consider the function $\omega(t)$ defined by

$$\omega(t) = \begin{cases} \varphi(t) & \text{if } \tau_- < t < \tau_+, \\ \lambda(t) & \text{if } \tau_+ \leqq t < \tau_+ + \hat{\alpha}. \end{cases} \tag{3.16}$$

Then for $t \geqq \tau_+$ we have from (3.15) that

$$\omega(t) = x_0 + \int_{t_0}^{t} f(s, \omega(s))\, ds + \int_{t_0}^{t} G(s)\, du(s),$$

and for $t < \tau_+$ we have from (3.6) that

$$\omega(t) = x_0 + \int_{t_0}^{t} f(s, \omega(s))\, ds + \int_{t_0}^{t} G(s)\, du(s).$$

Hence for $\tau_- < t < \tau_+ + \hat{\alpha}$ we have

$$\omega(t) = x_0 + \int_{t_0}^{t} f(s, \omega(s))\, ds + \int_{t_0}^{t} G(s)\, du(s), \tag{3.17}$$

which is a solution of $(\mathfrak{M})$ on an interval which contains the maximal extension interval, thus contradicting the definition of τ_+. Again the conclusion is that $\varphi(t)$ must have reached $S - C$ as t approached τ_+ from the left and that, in fact, the limit point $(\tau_+, \varphi(\tau_+-))$ cannot belong to S or else the preceding argument can be repeated.

THEOREM 6. *Consider the CMS*

$$Dx = f(t, x) + G(t)\, Du \tag{$\mathfrak{M}$}$$

in a domain S of R^{n+1}. Assume for each compact set C of S there exists a bound B_C such that $|f(t, x)| \leqq B_C$ for (t, x) in C. Let $\varphi(t)$ be a maximal solution of $(\mathfrak{M})$ on $\tau_- < t < \tau_+$. Then $\varphi(t)$ must lie entirely in $S - C$, for every compact set C in S, both as t approaches τ_- and as t approaches τ_+.

Proof. If $\tau_+ = +\infty$ or if $\tau_- = -\infty$, the assertion is obvious. Consider then the case when $\tau_+ < +\infty$. If $\varphi(t)$ is never in C the assertion is true, but if $\varphi(t)$ lies in C for some time $t < \tau_+$ then the solution must meet $S - C$ by Theorem 5. The solution could not remain in C until the time τ_+ because then the limit point $(\tau_+, \varphi(\tau_+-))$ would belong to C and hence by the definition of a CMS, there is a rectangle R_{ab} centered at $(\tau_+, \varphi(\tau_+-))$ such that $\varphi(\tau_+-) + J(\tau_+)$ belongs to R_{ab}. But by using the method in the last part of the proof of Theorem 5, we can extend $\varphi(t)$ to the right of τ_+, contradicting the definition of τ_+. Thus the solution $\varphi(t)$ must leave C at a time prior to τ_+. If it then remains in $S - C$ we are done, but if it returns to C it must meet $S - C$ again at a time prior to τ_+. We thus have only to show that $\varphi(t)$ cannot meet C infinitely many times as t approaches τ_+.

Suppose $\varphi(t)$ meets C infinitely many times as t approaches τ_+. There

are two cases which can occur: (i) $\varphi(t)$ also leaves, infinitely often, a second compact set C_1 contained in S such that the interior of C_1 contains C or (ii) no compact set C_1 in S can be found such that the solution $\varphi(t)$ also leaves C_1 before returning to C. Consider case (ii) first. This means that a compact set C_1 can be found which ultimately contains $\varphi(t)$ from some time on. But this is a contradiction since φ must meet $S - C_1$ at some time prior to τ_+. Hence only case (i) is left to be considered. This implies that the solution $\varphi(t)$ crosses infinitely often the space common to the exterior of C, denoted by $E(C)$, and the interior of C_1, denoted by $I(C_1)$. Denoting the closure of a set by placing a bar over the symbol for the set, we observe that there exists an $\epsilon > 0$ such that for any two times t_1 and t_2 for which $\varphi(t_1) \in C$ and $\varphi(t_2) \in \overline{E(C_1)}$ we have

$$| \varphi(t_1) - \varphi(t_2) | > \epsilon. \tag{3.18}$$

We shall show that this leads to a contradiction.

Let t_0 be a time at which φ belongs to $C_1 - C$ and choose a sequence of times $t_0 < t_1 < t_2 < \cdots < t_k < t_{k+1} < \cdots < \tau_+$ such that φ belongs to C for odd numbered times and to $\overline{E(C_1)}$ for even numbered times. Starting at each odd numbered time t_{2k-1} we extend the solution to the right and observe that at some time $t'_{2k} \leqq t_{2k}$ the solution first belongs to $\overline{E(C_1)}$. Now each of the points $(t'_{2k}, \varphi(t'_{2k}))$ lies on a hyperplane $t = t'_{2k}$ so that the distance of $\varphi(t'_{2k})$ from ∂C_1 is less than $| J(t'_{2k}) |$, i.e., if we define

$$H_{2k} = \{(t'_{2k}, x) \mid (t'_{2k}, x) \in E(C_1), d(x, \partial C_1) \leqq | J(t'_{2k}) |\}, \tag{3.19}$$

then each of the points $(t'_{2k}, \varphi(t'_{2k}))$ lies in the set H_{2k} and the set

$$C_1' = C_1 \bigcup_{k=1}^{\infty} H_{2k} \tag{3.20}$$

contains all of the points $\varphi(t_{2k-1})$ and $\varphi(t'_{2k})$. It is easy to show C_1 is compact by noting it is bounded and by showing its complement is open in the same manner that set S^{**} was shown to be open in the example following the definition of a CMS.

For convenience, let us drop the primes from the even numbered times obtained above and summarize the results:

$$\begin{gathered} t_1 < t_2 < t_3 < \cdots < t_{2k-1} < t_{2k} < \cdots < \tau_+, \\ \varphi(t_{2k-1}) \in C, \quad \varphi(t_{2k}) \in C_1'. \end{gathered} \tag{3.21}$$

Now compute by virtue of (3.18):

$$\begin{aligned} \epsilon < | \varphi(t_{2k-1}) - \varphi(t_{2k}) | &= \left| \int_{t_{2k-1}}^{t_{2k}} f(s, \varphi(s))\, ds + \int_{t_{2k-1}}^{t_{2k}} G(s)\, du(s) \right| \\ &\leqq \int_{t_{2k-1}}^{t_{2k}} | f(s, \varphi(s)) |\, ds + \left| \int_{t_{2k-1}}^{t_{2k}} G(s)\, du(s) \right|. \end{aligned} \tag{3.22}$$

But $|f(s, \varphi(s))|$ is bounded by some constant $B_{C_1'}$ because the solution $\varphi(t)$ is contained in the compact set C_1' during the interval $[t_{2k-1}, t_{2k}]$. Thus we have from (3.22),

$$\epsilon \leqq B_{C_1'}(t_{2k} - t_{2k-1}) + Gv(u, [t_{2k-1}, t_{2k}]), \tag{3.23}$$

where $G = \max_{\tau_- \leqq s \leqq \tau_+} |G(s)|$.

By summing (3.23) for $k = 1, 2, 3, \cdots$ we obtain

$$\infty \leqq B_{C_1'}(\tau_+ - t_1) + Gv(u, [t_1, \tau_+]), \tag{3.24}$$

which is a contradiction since the right hand side is finite.

A similar method proves the assertion for the case when t approaches τ_-. That case is, in fact, simpler because $\varphi(t)$ is continuous from the right.

COROLLARY 1. *Consider the CMS*

$$Dx = f(t, x) + G(t)\, Du \tag{$\mathfrak{M}$}$$

with $f(t, x)$ continuous in R^{n+1}, $G(t)$ continuous on R^1, and $u(t)$ of bounded variation and continuous from the right on R^1. If a solution $\varphi(t)$ of ($\mathfrak{M}$) is bounded in R^n for $t > t_0$, then $\tau_+ = +\infty$, i.e., the solution can be extended for all times in the future.

COROLLARY 2. *Consider the CMS*

$$Dx = f(x) + G(t)\, Du, \tag{$\mathfrak{M}$}$$

where $f(x)$ is of class C^1 in an open set S_1 in R^n, $G(t)$ is continuous on R^1, and $u(t)$ is of bounded variation and continuous from the right on R^1. Let $\varphi(t)$ on $\tau_- < t < \tau_+$ be a maximal solution in $S = R^1 \times S_1$. If $\varphi(t)$ lies in a compact subset C of S_1 then $\tau_- = -\infty$ and $\tau_+ = +\infty$. If $\tau_+ < +\infty$, then for each compact subset C of S the curve $\varphi(t)$ lies in $S - C$ as $t \to \tau_+$. The same result is true if $\tau_- > -\infty$.

THEOREM 7. *Consider the CMS*

$$Dx = f(t, x) + G(t)\, Du \tag{$\mathfrak{M}$}$$

in $S = I \times R^n$, where I is an open interval of R^1. Then each maximal solution $\varphi(t, t_0, x_0)$ for t_0 in I is defined on the whole interval I in case, for some constant $K > 0$,

(1) $|f(t, x)| < K$ in S, or

(2) $|f(t, x_1) - f(t, x_2)| \leqq K |x_1 - x_2|$ in S.

Proof. Suppose that φ is defined only on a proper subinterval (τ_-, τ_+) of I. Then take the compact set C of R^{n+1} defined by

$$C = [\tau_-, \tau_+] \times S(\rho),$$

where $S(\rho)$ is a closed sphere of radius ρ centered at the origin of R^n. By Theorem 6, $\varphi(t)$ lies entirely outside of $S(\rho)$ as t approaches τ_+. Thus for times $t' < \tau_+$ but arbitrarily close to τ_+ we have $|\varphi(t')|$ unbounded because we can take ρ as large as we please. On the other hand we have

$$
\begin{aligned}
|\varphi(t')| &\leqq |\varphi(t_0)| + \int_{t_0}^{t'} |f(t, \varphi(t))|\, dt + Gv(u, [t_0, \tau_+]) \\
&\leqq |\varphi(t_0)| + K(t' - t_0) + Gv(u, [t_0, \tau_+]) < \infty.
\end{aligned} \tag{3.25}
$$

Hence by choosing ρ sufficiently large we obtain a contradiction and thus $\varphi(t)$ is defined on the whole open interval I.

4. Dependence of solutions on initial conditions. A solution of a measure differential equation can be considered from several points of view. For example, if we write $\varphi(t, t_0, x_0)$ for the solution through the point (t_0, x_0), then we can consider φ as a function of t_0 and x_0 as well as a function of t. Another way to consider solutions is with regard to changes in the coefficients in the equation.

Let us treat the latter named problem first. We suppose that we have two equations with coefficients that differ little from each other over an interval I and we estimate the size of the difference in the two solutions.

THEOREM 8. *Consider the two CMS's,*

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du,$$

$$(\mathfrak{M}') \qquad Dy = h(t, y) + F(t)\, Du,$$

in a domain S of R^{n+1}. The coefficients are related as follows:
(i) $|f(t, x) - h(t, x)| \leqq \epsilon$ *for all* (t, x) *in* S, *and*
(ii) $|G(t) - F(t)| \leqq \eta$ *on an interval* (a, b).
Assume further that f satisfies a Lipschitz condition in S with constant K, that $|f(t, x)| \leqq M$ for all points in S, and that $|G(t)| \leqq G$ on $[a, b]$. Let $x(t)$ and $y(t)$ be solutions in S. Then for any times τ, σ, t in (a, b) we have

$$
\begin{aligned}
|y(t) - x(t)| \leqq \{&|y(\sigma) - x(\tau)| + M|\tau - \sigma| + \epsilon|t - \sigma| \\
&+ \eta v(u, [\sigma, t]) + Gv(u, [\tau, \sigma])\} e^{K|t-\sigma|}.
\end{aligned}
$$

Proof. The solutions $x(t)$ and $y(t)$ are represented by

$$
\begin{aligned}
x(t) &= x(\tau) + \int_\tau^t f(s, x(s))\, ds + \int_\tau^t G(s)\, du(s), \\
y(t) &= y(\sigma) + \int_\sigma^t h(s, y(s))\, ds + \int_\sigma^t F(s)\, du(s).
\end{aligned} \tag{4.1}
$$

Thus

$$
\begin{aligned}
| x(t) - y(t) | &\leqq | x(\tau) - y(\sigma) | + \int_\sigma^t | h(s, y(s)) \\
&\quad - f(s, x(s)) | \cdot | ds | + \int_\tau^\sigma | f(s, x(s)) | \cdot | ds | \\
&\quad + \left| \int_\tau^\sigma G(s)\, du(s) \right| + \left| \int_\sigma^t [G(s) - F(s)]\, du(s) \right| \\
&\leqq | x(\tau) - y(\sigma) | + M | \tau - \sigma | + \eta v(u, [\sigma, t]) \\
&\quad + Gv(u, [\tau, \sigma]) + \int_\sigma^t | h(s, y(s)) - f(s, y(s)) + f(s, y(s)) \\
&\qquad\qquad - f(s, x(s)) | \cdot | ds | .
\end{aligned} \tag{4.2}
$$

Concerning the last term on the right in (4.2) we have

$$
\begin{aligned}
&\int_\sigma^t | h(s, y(s)) - f(s, y(s)) + f(s, y(s)) - f(s, x(s)) | \cdot | ds | \\
&\qquad\qquad \leqq \epsilon | \tau - \sigma | + \int_\sigma^t K | y(s) - x(s)) | \cdot | ds | .
\end{aligned} \tag{4.3}
$$

Hence by applying the well-known inequality of Bellman and Gronwall to (4.2) after strengthening the inequality through use of (4.3), we have the stated result.

Notice in particular that if the two solutions take on nearby initial data at the same instant of time, i.e., $| y(\tau) - x(\tau) | < \delta$, then the result of the theorem reduces to

$$
| x(t) - y(t) | \leqq \{\delta + \eta v(u, [\tau, t]) + \epsilon | t - \tau |\} e^{K|t-\tau|}, \tag{4.4}
$$

and thus the two solutions are near to each other in a sense which is very transparent.

We proceed now to the investigation of the behavior of a solution as a function of the initial conditions. According to the definition of a solution, it does not necessarily depend continuously on the time but is instead a function of bounded variation in t. The following simple example shows that a solution need not depend continuously on the initial time t_0 either.

Example. Consider the equation

$$
Dx = x + Du, \tag{4.5}
$$

where x and u are real and

$$
u(t) = \begin{cases} 1, & t \geqq 0, \\ 0, & t < 0, \end{cases} \tag{4.6}
$$

so Du is the "Dirac δ-function," or more properly, the Dirac measure. Let

$t_1 < 0 < t_2$ and find solutions through $(t_1, 1)$ and $(t_2, 1)$ respectively. We shall prove in the next section that for equations with f linear in x we can find a fundamental solution to the homogeneous equation and use a modified form of the variation of parameters formula so widely used in the conventional theory. Proceeding from this fact we obtain

$$\varphi(t, t_1, 1) = e^{t-t_1} + e^t \int_{t_1}^{t} e^{-s}\, du(s), \tag{4.7}$$

$$\varphi(t, t_2, 1) = e^{t-t_2} + e^t \int_{t_2}^{t} e^{-s}\, du(s). \tag{4.8}$$

Now

$$\int_{t_1}^{t} e^{-s}\, du(s) = 0 \quad \text{if} \quad t < 0,$$

hence

$$\varphi(t, t_1, t) - \varphi(t, t_2, 1) = e^t \left[e^{-t_1} - e^{-t_2} - \int_{t_2}^{t} e^{-s}\, du(s) \right]. \tag{4.9}$$

By using integration by parts on $-\int_{t_2}^{t} e^{-s}\, du(s)$ we obtain

$$\begin{aligned} -\int_{t_2}^{t} e^{-s}\, du(s) &= -\int_{t_2}^{t} u(s) e^{-s}\, ds + e^{-t_2} \\ &= \int_{0}^{t_2} e^{-s}\, ds + e^{-t_2} = 1. \end{aligned} \tag{4.10}$$

Hence, combining (4.9) and (4.10) we have

$$\varphi(t, t_1, 1) - \varphi(t, t_2, 1) = e^t[e^{-t_1} - e^{-t_2} + 1]. \tag{4.11}$$

In particular, for values of t, t_1, and t_2 very near zero but such that $t < 0$ and $t_1 < 0 < t_2$, we have that $| \varphi(t, t_1, 1) - \varphi(t, t_2, 1) |$ is approximately equal to one.

If φ is a solution of a CMS $(\mathfrak{M})$ on some interval I, then it follows from the local existence and uniqueness theorem that $(\mathfrak{M})$ has a unique solution through any point (τ, ξ) close enough to the given solution. That theorem, however, assures the existence of the solution through (τ, ξ) only over some short interval containing τ. Actually it can be shown that the solution through a close neighboring point exists over each compact subinterval of I and hence that $\varphi(t, t_0, x_0)$ is defined on an open set in R^{n+2}. Furthermore, φ is continuous with respect to x_0 and of bounded variation with respect to t_0 (as well as with respect to t) in this open domain. The precise conditions under which this is true form the next theorem.

THEOREM 9. *Consider a CMS*

$$(\mathfrak{M}) \qquad Dx = f(t, x) + G(t)\, Du$$

in a domain S *of* R^{n+1}. *Let* $\psi(t)$ *be a solution of* $(\mathfrak{M})$ *on an interval* I: $a \leqq t \leqq b$. *Then there exists* $\delta > 0$ *such that for any point* (τ, ξ) *in the domain* C *defined by*

$$C: \quad a \leqq t \leqq b, \, | x - \psi(t) | < \delta,$$

there exists a unique solution φ *of* $(\mathfrak{M})$ *on* I *with* $\varphi(\tau, \tau, \xi) = \xi$. *Hence the subset* $\mathfrak{D} \subset R^{n+2}$, *defined by* $(t_0, x_0) \in S$ *and* $\tau_- < t < \tau_+$ (*for* $\varphi(t, t_0, x_0)$), *is open. In* $\mathfrak{D}$, $\varphi(t, t_0, x_0)$ *is continuous with respect to* x_0 *and is of bounded variation in* t *and* t_0.

Proof. Let $\delta_1 > 0$ be chosen so that the (t, x) region U given by

$$U: \quad a \leqq t \leqq b, \, | x - \psi(t) | \leqq \delta_1$$

belongs to S. Let τ be an arbitrary time in $a < \tau < b$ and choose t_1 such that $\tau < t_1 < b$. We wish to show that a solution starting near $\psi(\tau)$ can be extended so that it exists on the interval $[\tau, t_1]$. Then since t_1 is arbitrary, we can extend the solution to the half-open interval $[\tau, b)$.

Consider a covering of the compact set U by the open rectangles centered at each point of U (their existence is given by the definition of a CMS). We can select a finite subcovering of these rectangles and we define K to be the maximum of the finitely many Lipschitz constants associated with these rectangles. Choose $\delta < \delta_1 e^{-K|b-a|}$ and with this δ define C by

$$C: \quad a < t < b, \, | x - \psi(t) | < \delta.$$

Now if (τ, ξ) is in C, there is a solution $\varphi(t, \tau, \xi)$ of $(\mathfrak{M})$ through (τ, ξ) on some interval $[\tau - \alpha, \tau + \alpha]$ and φ can be represented as

$$(4.12) \qquad \varphi(t, \tau, \xi) = \xi + \int_\tau^t f(s, \varphi(s, \tau, \xi))\, ds + \int_\tau^t G(s)\, du(s)$$

for $\tau - \alpha \leqq t \leqq \tau + \alpha$. Also for any t in (a, b) we have

$$(4.13) \qquad \psi(t) = \psi(\tau) + \int_\tau^t f(s, \psi(s))\, ds + \int_\tau^t G(s)\, du(s).$$

Hence if we look at the difference between φ and ψ on $[\tau - \alpha, \tau + \alpha]$ we find

$$(4.14) \qquad | \psi(t) - \varphi(t, \tau, \xi) | \leqq | \psi(\tau) - \xi | \, e^{K|t-\tau|} < \delta_1$$

and thus φ cannot leave the compact set U_1. By Theorem 6 then, φ can be extended over the whole interval $[\tau, b)$. A similar procedure serves to extend φ to the left to a, and thus $\varphi(t, \tau, \xi)$ is defined on an open set V in R^{n+2} given by

$$V: \quad a < t < b, \qquad a < \tau < b, \qquad |x - \psi(t)| < \delta.$$

Next we must show that $\varphi(t, \tau, \xi)$ is a continuous function of ξ on V. Let (τ, ξ) and (τ, x_0) belong to C. Then we know there exist solutions $\varphi(t, \tau, \xi)$ and $\varphi(t, \tau, x_0)$ defined on V such that

$$|\varphi(t, \tau, \xi) - \varphi(t, \tau, x_0)| \leqq |\xi - x_0| e^{K|t-\tau|} \tag{4.15}$$

for all (t, τ) such that $a < t < b$, $a < \tau < b$. Now, given $\epsilon > 0$, choose $|\xi - x_0| < \epsilon e^{-K|b-a|}$ and then from (4.15),

$$|\varphi(t, \tau, \xi) - \varphi(t, \tau, x_0)| < \epsilon \tag{4.16}$$

uniformly with respect to t and τ.

Finally, we show $\varphi(t, \tau, \xi)$ is of bounded variation with respect to τ on (a, b). Let Π be a partition $a_0 < a_1 < \cdots < a_N$ of (a, b). Then for any a_i we have

$$\varphi(t, a_i, \xi) = \xi + \int_{a_i}^{t} f(s, \varphi(s, a_i, \xi))\, ds + \int_{a_i}^{t} G(s)\, du(s). \tag{4.17}$$

Hence from (4.17),

$$\begin{aligned} &|\varphi(t, a_{i+1}, \xi) - \varphi(t, a_i, \xi)| \\ &\leqq \left| \int_{a_{i+1}}^{t} f(s, \varphi(s, a_{i+1}, \xi))\, ds - \int_{a_i}^{t} f(s, \varphi(s, a_i, \xi))\, ds \right| \\ &\qquad + \left| \int_{a_{i+1}}^{t} G(s)\, du(s) - \int_{a_i}^{t} G(s)\, du(s) \right|. \end{aligned} \tag{4.18}$$

Now

$$\begin{aligned} &\left| \int_{a_{i+1}}^{t} f(s, \varphi(s, a_{i+1}, \xi))\, ds - \int_{a_i}^{t} f(s, \varphi(s, a_i, \xi))\, ds \right| \\ &\leqq \int_{a_{i+1}}^{t} K |\varphi(s, a_{i+1}, \xi) - \varphi(s, a_i, \xi)| \cdot |ds| \\ &\qquad + \int_{a_i}^{a_{i+1}} |f(s, \varphi(s, a_i, \xi))|\, ds, \end{aligned} \tag{4.19}$$

and if we define $M = \sup_{\bar{\Omega}} |f(t, x)|$, then

$$\int_{a_i}^{a_{i+1}} |f(s, \varphi(s, a_i, \xi))|\, ds \leqq M(a_{i+1} - a_i). \tag{4.20}$$

Furthermore, defining $G = \max_{[a,b]} |G(s)|$, we have

$$\begin{aligned} \left| \int_{a_{i+1}}^{t} G(s)\, du(s) - \int_{a_i}^{t} G(s)\, du(s) \right| &= \left| \int_{a_i}^{a_{i+1}} G(s)\, du(s) \right| \\ &\leqq G v(u, [a_i, a_{i+1}]). \end{aligned} \tag{4.21}$$

From (4.18) through (4.21) we obtain

$$
\begin{aligned}
&| \varphi(t, a_{i+1}, \xi) - \varphi(t, a_i, \xi) | \leqq M(a_{i+1} - a_i) \\
&+ Gv(u, [a_i, a_{i+1}]) + K \int_{a_{i+1}}^{t} | \varphi(s, a_{i+1}, \xi) - \varphi(s, a_i, \xi) | \cdot | ds | ,
\end{aligned} \tag{4.22}
$$

and by using the Bellman-Gronwall inequality we can write

$$
\begin{aligned}
| \varphi(t, a_{i+1}, \xi) - \varphi(t, a_i, \xi) | \leqq \{ & M(a_{i+1} - a_i) \\
& + Gv(u, [a_i, a_{i+1}]) \} e^{K|a_{i+1} - t|} .
\end{aligned} \tag{4.23}
$$

Let $A = \max \{ | a - t |, | b - t | \}$; then the inequality (4.23) can be maintained by replacing $e^{K|a_{i+1} - t|}$ by e^{AK}. Finally, by summing (4.23) over i there results

$$
\sum_{i=1}^{N} | \varphi(t, a_{i+1}, \xi) - \varphi(t, a_i, \xi) | \leqq \{ M(b - a) + Gv(u, [a, b]) \} e^{AK}, \tag{4.24}
$$

and thus $\varphi(t, \tau, \xi)$ is of bounded variation on $[a, b]$ with respect to τ.

THEOREM 10. *Consider the CMS*

$$
Dx = f(t, x) + G(t)\, Du \tag{$\mathfrak{M}$}
$$

in a domain S of R^{n+1} and suppose that f_x exists and is continuous in S. If $\varphi(t) = \varphi(t, \tau, \xi)$ is a solution, then $\partial\varphi/\partial\xi$ is continuous with respect to ξ for (t, τ, ξ) in $\mathfrak{D} = \{ (t, \tau, \xi) \mid (\tau, \xi) \in S, \tau_- < t < \tau_+ \}$.

Proof. Consider the case of $\partial\varphi/\partial\xi^1$, where (τ, ξ) is in S and τ and ξ are temporarily held fixed and $\xi = (\xi^1, \cdots, \xi^n)$. Let $h = (h^1, 0, \cdots, 0)$ and let $\tilde{\xi} = \xi + h$ for h^1 so chosen that $(\tau, \tilde{\xi})$ is in S. Define χ by the equation

$$
\chi(t, \tau, \xi, h) = \frac{\varphi(t, \tau, \tilde{\xi}) - \varphi(t, \tau, \xi)}{h^1} \tag{4.25}
$$

for (t, τ, ξ) in $(a, b) \times S \subset \mathfrak{D}$. Then we must show that $\lim_{h \to 0} \chi(t, \tau, \xi, h)$ exists. For convenience let

$$
\theta(t, \tau, \xi, h) = \varphi(t, \tau, \tilde{\xi}) - \varphi(t, \tau, \xi) \tag{4.26}
$$

and observe that

$$
\theta(t, \tau, \xi, h) = \tilde{\xi} - \xi + \int_{\tau}^{t} [f(s, \varphi(s, \tau, \tilde{\xi})) - f(s, \varphi(s, \tau, \xi))]\, ds. \tag{4.27}
$$

Let us shorten the notation since (τ, ξ) is being held fixed and combine (4.25) with (4.27) to obtain

$$
\chi(t, h) = \frac{\theta(t, h)}{h^1} = \frac{h}{h^1} + \int_{\tau}^{t} \frac{[f(s, \varphi(s, \tau, \tilde{\xi})) - f(s, \varphi(s, \tau, \xi))]}{h^1}\, ds. \tag{4.28}
$$

Applying the mean value theorem to the integrand in (4.28) there results

$$\chi(t, h) = \frac{h}{h^1} + \int_\tau^t \frac{\partial f}{\partial x}(s, \hat{\varphi}) \left[\frac{\varphi(s, \tau, \tilde{\xi}) - \varphi(s, \tau, \xi)}{h^1}\right] ds, \tag{4.29}$$

or

$$\chi(t, h) = \frac{h}{h^1} + \int_\tau^t \frac{\partial f}{\partial x}(s, \varphi(s, \tau, \xi) + \epsilon(s, h))\chi(s, h)\, ds, \tag{4.30}$$

where for each fixed s, $\epsilon(s, h)$ approaches zero as h approaches zero. Define

$$\epsilon(s, 0) = 0; \tag{4.31}$$

then (4.30) is equivalent to a linear differential equation with a parameter:

$$\frac{dy}{dt} = A(t, h)y(t, h), \quad y(\tau) = \frac{h}{h^1}, \tag{4.32}$$

in which the coefficient matrix is continuous at the parameter value $h = 0$. Such an equation has a unique continuous solution $\chi(t, h)$ for which the limit as h approaches zero exists uniformly for t in (a, b), i.e.,

$$\lim_{h \to 0} \chi(t, h) = \chi(t, 0) \tag{4.33}$$

exists uniformly for t in (a, b).

5. Linear measure differential equations. If one considers an ordinary linear differential equation

$$\frac{dx}{dt} = A(t)x + f(t), \quad x(\tau) = \xi, \tag{$\mathcal{O}$}$$

for t in some interval I where $A(t)$ is an $n \times n$ summable matrix on a real t interval I and $f(t)$ is a summable n-vector on I, then it is well-known (cf. [3, p. 74] that a solution of ($\mathcal{O}$) on I may be represented as

$$\varphi(t) = \Phi(t)\xi + \Phi(t) \int_\tau^t \Phi^{-1}(s)f(s)\, ds, \tag{5.1}$$

where $\Phi(t)$ is a fundamental solution matrix for the homogeneous equation corresponding to ($\mathcal{O}$) with $\Phi(\tau) = E$, the $n \times n$ identity matrix.

The next theorem shows that there is a corresponding variation of parameters formula for linear measure differential equations.

THEOREM 11. *Consider the linear measure differential equation*

$$Dx = A(t)x + f(t) + G(t)\, Du, \qquad x(t_0) = x_0, \tag{$\mathcal{L}$}$$

where $A(t)$ is a summable $n \times n$ matrix on an interval I, $f(t)$ is a summable n-vector on I, $G(t)$ is a continuous $n \times m$ matrix on I, and $u(t)$ is a bounded

variation m-vector continuous from the right on I. If $\Phi(t)$ is a fundamental solution matrix of the homogeneous ordinary differential equation $\dot{x} = A(t)x$ with $\Phi(t_0) = E$, the $n \times n$ identity matrix, then the variation of parameters formula

$$x(t) = \Phi(t)x_0 + \Phi(t)\int_{t_0}^{t} \Phi^{-1}(s)f(s)\,ds + \Phi(t)\int_{t_0}^{t} \Phi^{-1}(s)G(s)\,du(s) \tag{5.2}$$

is a solution of ($\mathfrak{L}$) *on I such that $x(t_0) = x_0$.*

Proof. The fact that a solution of ($\mathfrak{L}$) exists follows easily from Theorem 4 and Theorem 7. We must therefore show that (5.2) is a solution of the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} [A(s)x(s) + f(s)]\,ds + \int_{t_0}^{t} G(s)\,du(s). \tag{5.3}$$

We proceed by substituting the right hand side of (5.2) into (5.3) and showing that (5.3) reduces to an identity. Let the right hand side of (5.2) be temporarily denoted by $\psi(t)$ and define

$$\begin{aligned}
\beta(t) &= x_0 + \int_{t_0}^{t} [A(s)\psi(s) + f(s)]\,ds + \int_{t_0}^{t} G(s)\,du(s) \\
&= x_0 + \int_{t_0}^{t} A(s)\Big\{\Phi(s)x_0 + \int_{t_0}^{s} \Phi(s)\Phi^{-1}(\sigma)\,d\sigma \\
&\quad + \int_{t_0}^{s} \Phi(s)\Phi^{-1}(\sigma)G(\sigma)\,du(\sigma)\Big\}\,ds + \int_{t_0}^{t} f(s)\,ds + \int_{t_0}^{t} G(s)\,du(s) \\
&= x_0 + \int_{t_0}^{t} A(s)\Phi(s)x_0\,ds \\
&\qquad + \int_{t_0}^{t} A(s)\Phi(s)\int_{t_0}^{s} \Phi^{-1}(\sigma)f(\sigma)\,d\sigma\,ds \\
&\qquad + \int_{t_0}^{t} A(s)\Phi(s)\int_{t_0}^{s} \Phi^{-1}(\sigma)G(\sigma)\,du(\sigma)\,ds \\
&\qquad + \int_{t_0}^{t} f(s)\,ds + \int_{t_0}^{t} G(s)\,du(s).
\end{aligned} \tag{5.4}$$

For convenience, we break $\beta(t)$ up into a sum of five integrals as follows:

$$\beta(t) = x_0 + I_1 + I_2 + I_3 + \int_{t_0}^{t} f(s)\,ds + \int_{t_0}^{t} G(s)\,du(s), \tag{5.5}$$

where

$$
\begin{aligned}
I_1 &= \int_{t_0}^{t} A(s)\Phi(s)x_0\, ds, \\
I_2 &= \int_{t_0}^{t} A(s)\Phi(s) \int_{t_0}^{s} \Phi^{-1}(\sigma)f(\sigma)\, d\sigma\, ds, \\
I_3 &= \int_{t_0}^{t} A(s)\Phi(s) \int_{t_0}^{s} \Phi^{-1}(\sigma)G(\sigma)\, du(\sigma)\, ds.
\end{aligned} \tag{5.6}
$$

We investigate the integrals I_1, I_2, and I_3 separately making use of integration by parts formulas and properties of Φ.

By definition of $\Phi(t)$ we have

$$
A(s)\Phi(s) = \frac{d\Phi}{ds}(s), \tag{5.7}
$$

and thus

$$
I_1 = \int_{t_0}^{t} A(s)\Phi(s)x_0\, ds = \int_{t_0}^{t} \frac{d\Phi(s)}{ds} x_0\, ds = \Phi(t)x_0 - x_0. \tag{5.8}
$$

Define

$$
\Lambda(t) = \int_{t_0}^{t} \Phi^{-1}(\sigma)f(\sigma)\, d\sigma; \tag{5.9}
$$

then

$$
I_2 = \int_{t_0}^{t} A(s)\Phi(s)\Lambda(s)\, ds = \int_{t_0}^{t} \frac{d\Phi(s)}{ds} \Lambda(s)\, ds = \int_{t_0}^{t} d\Phi(s)\Lambda(s). \tag{5.10}
$$

Using integration by parts on the last integral in (5.10) there results

$$
I_2 = -\int_{t_0}^{t} \Phi(s)\, d\Lambda(s) + \Phi(t)\Lambda(t) - \Phi(t_0)\Lambda(t_0). \tag{5.11}
$$

But $\Lambda(t_0) = 0$ and $d\Lambda(s) = \Phi^{-1}(s)f(s)\, ds$; thus

$$
\begin{aligned}
I_2 &= -\int_{t_0}^{t} \Phi(s)\Phi^{-1}(s)f(s)\, ds + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)f(s)\, ds \\
&= -\int_{t_0}^{t} f(s)\, ds + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)f(s)\, ds.
\end{aligned} \tag{5.12}
$$

In a similar manner it follows that if we define

$$
\Omega(t) = \int_{t_0}^{t} \Phi^{-1}(\sigma)G(\sigma)\, du(\sigma), \tag{5.13}
$$

we have

$$
\begin{aligned}
I_3 &= \int_{t_0}^{t} A(s)\Phi(s)\Omega(s)\, ds = \int_{t_0}^{t} \frac{d\Phi(s)}{ds}\, \Omega(s)\, ds \\
&= \int_{t_0}^{t} d\Phi(s)\Omega(s) = -\int_{t_0}^{t} \Phi(s)\, d\Omega(s) + \Phi(t)\Omega(t).
\end{aligned} \tag{5.14}
$$

By (2.8) we can write

$$
\int_{t_0}^{t} \Phi(s)\, d\Omega(s) = \int_{t_0}^{t} \Phi(s)\Phi^{-1}(s)G(s)\, du(s), \tag{5.15}
$$

and thus from (5.14) and (5.15) there results

$$
I_3 = -\int_{t_0}^{t} G(s)\, du(s) + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)G(s)\, du(s). \tag{5.16}
$$

By substituting now (5.16), (5.12), and (5.8) into (5.5) we obtain

$$
\begin{aligned}
\beta(t) = \Phi(t)x_0 &+ \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)f(s)\, ds \\
&+ \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)G(s)\, du(s),
\end{aligned} \tag{5.17}
$$

but the right hand side is precisely $\psi(t)$.

6. Existence of an optimal control. In this section, the control problem for the optimal control of a nonlinear dynamical system is defined. We assume that the system under consideration is described by a canonical system of measure differential equations for the state variables. These measure differential equations depend on certain parameters called control variables. The problem of control is to select the control variables so that the vector solution (called the response) of the measure differential equations will satisfy given initial and terminal conditions and further so that the control variables, together with the response vector, minimize a given functional.

Consider the measure differential equation

$$
(\mathfrak{M}'') \qquad Dx = f(t, x, u) + G(t)\, Du,
$$

where $f^i(t, x, u)$ together with $\partial f^i(t, x, u)/\partial x^k$, $i, k = 1, 2, \cdots, n$, are real continuous functions in $R^1 \times R^m \times \Omega$, where R^n is the real n-dimensional number space ($|x| = \sum_{i=1}^{n} |x^i|$), and Ω is a nonempty compact subset of R^m. We further suppose that the elements $g_j^{\,i}(t)$ of the $n \times m$ matrix $G(t)$ are continuous functions on R^1 for $i = 1, \cdots, n, j = 1, \cdots, m$, and that the functions $u^j(t)$ are of bounded variation and continuous from the right on appropriate time intervals such that the graph of $u(t)$ lies in Ω.

For each choice of a function $u(t)$ on $-\infty < t_0 \leqq t \leqq t_1 < \infty$ as a vector

valued function whose value is in Ω and whose components $u^j(t)$, $j = 1, 2, \cdots, m$, are of bounded variation and continuous from the right on $[t_0, t_1]$, the measure differential equation

$$(\mathfrak{M}'') \qquad Dx = f(t, x, u) + G(t)\, Du$$

has a unique bounded variation solution $x(t)$ (called a response) on $t_0 \leqq t \leqq t_1$ (or a subinterval) through a prescribed initial point (t_0, x_0). This is the result of the existence and uniqueness theorems of the previous section for a CMS. The representation of the response is, of course, the unique bounded variation solution of the integral equation

$$(g) \qquad x(t) = x_0 + \int_{t_0}^{t} f(s, x(s), u(s))\, ds + \int_{t_0}^{t} G(s)\, du(s).$$

DEFINITION 5. A control for the equation $(\mathfrak{M}'')$, where a nonempty compact set Ω contained in R^m and an initial point x_0 in R^n have been prescribed, is a vector valued function $u(t)$ of bounded variation and continuous from the right on a finite interval $t_0 \leqq t \leqq t_1$ with its values in Ω such that its response $x(t)$ with $x(t_0) = x_0$ is also defined in R^n on $t_0 \leqq t \leqq t_1$.

The main considerations herein will be directed towards the problem in which a control $u(t)$ is sought whose response $x(t)$ initiates at the given point x_0 at time $t = t_0$ and terminates at time $t = t_1$ in some given moving target $T(t)$, i.e., $x(t_1) \in T(t_1)$. More precisely:

DEFINITION 6. For each t on a given finite interval $\tau_0 \leqq t \leqq \tau_1$, we specify a nonempty compact set $T(t)$ contained in R^n and called the *target*. This set is to vary continuously with t in the sense that the distance d between two nonempty compact subsets is taken to be the Hausdorff metric (i.e., $d = d(X, Y)$ is the smallest real number such that X lies in the d-neighborhood of Y and Y lies in the d-neighborhood of X, cf. [8]). If $T(t)$ is a point for each t, then the target is a continuous curve.

The criterion by which the control is to be selected is incorporated as follows.

DEFINITION 7. For a given real valued continuous function $f^0(t, x, u)$ defined on $R^1 \times R^n \times \Omega$ we define the *cost functional* $C(u)$ of a control $u(t)$ on $t_0 \leqq t \leqq t_1$ with response $x(t)$ by

$$(6.1) \qquad C(u) = \int_{t_0}^{t_1} f^0(t, x(t), u(t))\, dt.$$

If $f^0(t, x, u) \equiv 1$, then $C(u) = t_1 - t_0$, the time duration over which the control is exerted. Such a cost functional is then called *time optimal*.

DEFINITION 8. Given the data:

(a) $Dx = f(t, x, u) + G(t)\, Du$, the equation;

(b) a nonempty compact set $\Omega \subset R^m$, the restraint set (which contains the graphs of the controls);
(c) $x_0 \in R^n$, the initial point;
(d) $T(t) \subset R^n$ on $\tau_0 \leqq t \leqq \tau_1$, the compact target;
(e) the real number $E > 0$.
Define $\Delta = \Delta(f(t, x, u), G(t), \Omega, x_0, T(t), E)$ as the set of all controls $u(t)$ contained in Ω with $u(t)$ of bounded variation and continuous from the right on various subintervals $t_0 \leqq t \leqq t_1$ (but $\tau_0 \leqq t_0 < t_1 \leqq \tau_1$) such that $v(u, [t_0, t_1]) \leqq E$ and such that the responses $x(t)$ satisfy $x(t_0) = x_0$ and $x(t_1) \in T(t_1)$. This set Δ is called the *set of admissible controls.*

DEFINITION 9. A control $u^*(t)$ in Δ is called *optimal* in case

$$C(u^*) \leqq C(u) \tag{6.2}$$

for every $u(t)$ in Δ ($C(u)$ is given by Definition 7).

The search for an optimal control associated with the data in Definitions 8 and 9 is simply termed hereafter as "the control problem for the given data."

Remark. The hypothesis concerning the uniform bounded total variation of the admissible controls is concerned with the fact that in a large class of problems the total variation is a mathematical manifestation of the motion of some process. It is those processes which contain devices capable of sustaining only a bounded amount of movement, regardless of the control that is applied, to which the following theory pertains.

THEOREM 12. *Given the control problem for the data:*

(a) $Dx = f(t, x, u) + G(t)\, Du$, *with* $f^i(t, x, u)$ *and* $\partial f^i(t, x, u)/\partial x^k$, $i, k = 1, \cdots, n$, *continuous on* $R^1 \times R^n \times R^m$ *and* $g_j{}^i(t)$ *continuous on an interval* $\tau_0 \leqq t \leqq \tau_1$ *for* $i = 1, \cdots, n, j = 1, \cdots, m$;

(b) *a nonempty compact restraint set* $\Omega \subset R^m$;

(c) *the initial point* $x_0 \in R^n$;

(d) *the continuously moving nonempty target set* $T(t) \subset R^m$ *defined on* $\tau_0 \leqq t \leqq \tau_1$;

(e) *the cost functional*

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

where $f^0(t, x, u)$ *is continuous on* $R^1 \times R^n \times R^m$;

(f) *the set* $\Delta = \Delta(f(t, x, u), G(t), \Omega, x_0, T(t), E, h(t))$ *of admissible controls defined on the fixed subinterval* $[\tau_0, \tau_1]$, *where* $h(t)$ *is a nondecreasing function, continuous from the right, such that all* $u(t)$ *in* Δ *satisfy the inequalities*

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

on every subinterval of the interval $[t_0, t_0 + \delta]$ *for some appropriate* $\delta > 0$, *however small.*

Assume that set Δ *is such that*

(A) Δ *is nonempty,*

(B) *there exists a real bound* $B < \infty$ *such that for all responses* $x(t)$ *corresponding to controls in* Δ *we have* $| x(t) | \leqq B$.

Conclusion: *Then there exists an optimal control in* Δ.

Remark. We assume that x_0 is not in the target $T(t_0)$, and then $| \Delta u | \leqq \Delta h$ on $[t_0, t_0 + \delta]$ guarantees that $x(t)$ lies outside $T(t)$ for all responses and all t sufficiently near t_0.

If the functions $u(t)$ in Δ satisfy a uniform Lipschitz condition

$$| u(t') - u(t) | \leqq K | t' - t |$$

for all t, t' near t_0, then the function $h(t)$ may be taken to be Kt.

Proof. First let us show that the responses $x(t)$ to controls in Δ are of uniform bounded total variation. We may write

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s), u(s))\, ds + \int_{t_0}^{t} G(s)\, du(s). \tag{6.3}$$

We compute for the variation of $x(t)$ on $[t_0, t_1]$:

$$v(x, [t_0, t_1]) \leqq \int_{t_0}^{t_1} | f(s, x(s), u(s)) |\, ds + Gv(u, [t_0, t_1]), \tag{6.4}$$

where $G = \max_{t \in [t_0, t_1]} | G(t) |$. But $| f(s, x(s), u(s)) |$ is uniformly bounded because the responses $x(t)$ are uniformly bounded and hence $v(x, [t_0, t_1])$ is uniformly bounded for all x that are responses to controls in Δ.

Now, since Δ is nonempty and the corresponding responses are uniformly bounded, $\inf C(u) = \tilde{m} > -\infty$, where the infimum is taken over all u in Δ. Either Δ is a finite set, in which case the theorem is trivially true, or we can select from Δ a sequence of controls $u^{(k)}(t)$ on $t_0 \leqq t \leqq t_1$ for which $C(u^{(k)})$ decreases monotonically to $\tilde{m}$. By [5, Theorem 33, Chap. XII] there exist a subsequence (which we shall allow to retain the same notation $u^{(k)}$) and a function of bounded variation $u^*(t)$ such that

$$\lim_{k \to \infty} u^{(k)}(t) = u^*(t) \tag{6.5}$$

everywhere on $[t_0, t_1]$ and moreover

$$v(u^*, [t_0, t_1]) \leqq \lim_{k \to \infty} v(u^{(k)}, [t_0, t_1]). \tag{6.6}$$

Now $u^*(t)$ is not necessarily continuous from the right in $[t_0, t_1)$; therefore we shall redefine it where necessary so that it is, and call the result $\tilde{u}(t)$. Since this would require that the value of $u^*(t)$ be redefined on at most a

denumerable set of points we notice that $u^*(t) = \tilde{u}(t)$ a.e. in $[t_0, t_1]$ and in particular, as we shall see, at t_0, t_1, and the points of continuity of $\tilde{u}(t)$. We must show that $\tilde{u}(t) \in \Delta$. Since $\tilde{u}(t)$ is continuous from the right by construction, this necessitates showing (i) that $\tilde{u}(t) \in \Omega$ for $t_0 \leqq t \leqq t_1$, (ii) that $v(\tilde{u}, [t_0, t_1]) \leqq E$, (iii) that $\tilde{u}(t)$ has a response $\tilde{x}(t)$ defined on $[t_0, t_1]$ such that $\tilde{x}(t_0) = x_0$ and $\tilde{x}(t_1) \in T(t_1)$, and (iv) that $|\Delta\tilde{u}(t)| \leqq \Delta h$ on $[t_0, t_0 + \delta]$.

(i) Show that $\tilde{u}(t) \in \Omega$. It follows from (6.5) that $u^*(t)$ is in Ω since Ω is compact. Now suppose t' is any point in $[t_0, t_1]$. We know

$$\tilde{u}(t') = \lim_{\eta \to 0} u^*(t' + |\eta|) \tag{6.7}$$

and we may restrict the points $t' + |\eta|$ to be in the set of points on which the value of u^* was not altered. Again we would have $\tilde{u}(t')$ belonging to Ω. Thus $\tilde{u}(t)$ belongs to Ω for all t in $[t_0, t_1]$.

(ii) By (6.6) we have

$$v(u^*, [t_0, t_1]) \leqq E. \tag{6.8}$$

Since we redefine u^* to be continuous from the right it is easily shown that $\tilde{u}(t)$ will have the same (or less) variation as compared with $u^*(t)$. Thus $v(\tilde{u}, [t_0, t_1]) \leqq E$.

(iii) Define $\tilde{x}(t)$ to be the solution of the integral equation

$$\tilde{x}(t) = x_0 + \int_{t_0}^{t} f(s, \tilde{x}(s), \tilde{u}(s))\, ds + \int_{t_0}^{t} G(s)\, d\tilde{u}(s). \tag{6.9}$$

We must show $\tilde{x}(t)$ is a response to $\tilde{u}(t)$, i.e., that $\tilde{x}(t_0) = x_0$ and $\tilde{x}(t_1) \in T(t_1)$. Now the former is obvious since $\tilde{u}(t)$ is continuous from the right. To prove the latter we denote the response to $u^{(k)}(t)$ by $x^{(k)}(t)$; then

$$x^{(k)}(t) = x_0 + \int_{t_0}^{t} f(s, x^{(k)}(s), u^{(k)}(s))\, ds + \int_{t_0}^{t} G(s)\, du^{(k)}(s). \tag{6.10}$$

Since the total variations of the $x^{(k)}(t)$ are uniformly bounded then there exist a subsequence (still labeled $x^{(k)}(t)$) and a function $x^*(t)$ such that

$$\lim_{k \to \infty} x^{(k)}(t) = x^*(t) \tag{6.11}$$

everywhere on $[t_0, t_1]$. We have by Lebesgue's theorem on dominated convergence that for all $t \in [t_0, t_1]$,

$$\lim_{k \to \infty} \int_{t_0}^{t} f(s, x^{(k)}(s), u^{(k)}(s))\, ds = \int_{t_0}^{t} f(s, x^*(s), \tilde{u}(s))\, ds, \tag{6.12}$$

because $u^{(k)}(t) \to \tilde{u}(t)$ a.e. in $[t_0, t_1]$, $x^{(k)}(t) \to x^*(t)$ everywhere in $[t_0, t_1]$, and $|f(s, x^{(k)}(s), u^{(k)}(s))|$ is uniformly bounded by a constant on $[t_0, t_1]$.

We want to show next that $u^*(t)$ was not altered at $t = t_0$, i.e., that $u^*(t)$ is continuous from the right at t_0. Since $|\Delta u^{(k)}| \leqq \Delta h$ on every subinterval of $[t_0, t_0 + \delta]$ and h has a right hand limit at each point of this interval then it follows that

$$\lim_{s \to 0+} u^{(k)}(t + s) = u^{(k)}(t + 0) \tag{6.13}$$

uniformly with respect to k (cf. [5, Theorem 1, Chap. VII]). Since the one-sided limits $u^{(k)}(t + 0)$, $u^*(t + 0)$ exist, and since $\lim_{k\to\infty} u^{(k)}(t + s) = u^*(t + s)$ for $t + s$ in $[t_0, t_0 + \delta]$ we may apply the Moore theorem on interchange of order of repeated limits to obtain

$$\lim_{k \to \infty} u^{(k)}(t + 0) = u^*(t + 0) \tag{6.14}$$

for t in $[t_0, t_0 + \delta]$. Thus $u^*(t)$ is continuous from the right at t_0 and $u^{(k)}(t_0) \to u^*(t_0)$ and in particular $u^{(k)}(t_0) \to \tilde{u}(t_0)$.

(iv) Show $|\Delta u| \leqq \Delta h$ on every subinterval of $[t_0, t_0 + \delta]$. It follows from (6.14) that $u^*(t)$ is continuous from the right in $[t_0, t_0 + \delta]$ and hence $u(t) = u^*(t)$ for t in $[t_0, t_0 + \delta]$. Now $\Delta\tilde{u} = \lim_{k\to\infty} \Delta u^{(k)}$. Therefore, given $\epsilon > 0$ there exists a $K > 0$ such that $|\Delta\tilde{u}| < |\Delta u^{(k)}| + \epsilon$ or $|\Delta u| < \Delta h + \epsilon$. Since ϵ was arbitrary, then $|\Delta u| \leqq \Delta h$.

Since $u^{(k)}(t)$ are of uniform bounded total variation on $[t_0, t_1]$ and $u^{(k)}(t) \to \tilde{u}(t)$ at the points t_0, t_1, and at the points of continuity of $\tilde{u}(t)$, we have by the theorem of Helly and Bray (cf. [11, p. 54]) that

$$\lim_{k \to \infty} \int_{t_0}^{t} G(s)\, du^{(k)}(s) = \int_{t_0}^{t} G(s)\, d\tilde{u}(s) \tag{6.15}$$

for at least all t in $[t_0, t_1]$ which are points of continuity and in particular for $t = t_1$. Thus from (6.12) and (6.13),

$$\lim_{k \to \infty} x^{(k)}(t) = x_0 + \int_{t_0}^{t} f(s, x^*(s), \tilde{u}(s))\, ds + \int_{t_0}^{t} G(s)\, d\tilde{u}(s) \tag{6.16}$$

exists for almost all $t \in [t_0, t_1]$. By (6.11) this limit is equal to $x^*(t)$ almost everywhere in $[t_0, t_1]$ but in particular it exists for $t = t_1$. Now compare $\tilde{x}(t)$ with $x^*(t)$:

$$\begin{aligned} |\tilde{x}(t) - x^*(t)| &= \left| \int_{t_0}^{t} f(s, \tilde{x}(s), \tilde{u}(s))\, ds - \int_{t_0}^{t} f(s, x^*(s), \tilde{u}(s))\, ds \right| \\ &\leqq \int_{t_0}^{t} K\, |\tilde{x}(s) - x^*(s)|\, ds. \end{aligned} \tag{6.17}$$

Hence $|\tilde{x}(t) - x^*(t)|$ is zero for almost all t and in particular it is zero for $t = t_1$, i.e., $\tilde{x}(t_1) = x^*(t_1)$. But we know from (6.11) that $x^{(k)}(t_1)$

$\to x^*(t_1) = \tilde{x}(t_1)$, and since $x^{(k)}(t_1) \in T(t_1)$ and $T(t_1)$ is compact then $\tilde{x}(t_1) \in T(t_1)$. Thus $\tilde{u}(t)$ belongs to Δ.

Now compute the cost of $\tilde{u}(t)$. We have by the definition of $C(u)$ that

$$C(u^{(k)}) = \int_{t_0}^{t_1} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds.$$

But $x^{(k)}(t) \to \tilde{x}(t)$ a.e. in $[t_0, t_1]$ and $u^{(k)}(t) \to \tilde{u}(t)$ a.e. in $[t_0, t_1]$. Thus by Lebesgue's theorem on dominated convergence we have

$$\lim_{k\to\infty} C(u^{(k)}) = \int_{t_0}^{t_1} f^0(s, \tilde{x}(s), \tilde{u}(s))\, ds = C(\tilde{u}) \tag{6.18}$$

and by uniqueness of the limit of $C(u^{(k)})$ for a subsequence we have

$$\lim_{k\to\infty} C(u^{(k)}) = C(\tilde{u}) = \tilde{m}. \tag{6.19}$$

Therefore $\tilde{u}(t)$ on $[t_0, t_1]$ is an optimal control.

The assumption of a fixed terminal time t_1 in the previous theorem eliminates the application of that theorem to time optimal problems. We can treat this problem, and, in fact, a more general problem in which the terminal time is not fixed by requiring that $G(t)$ is of class $C^1[\tau_0, \tau_1]$ and that the function $h(t)$ in the definition of Δ hypothesis satisfies $|\Delta u| \leqq \Delta h$ for every subinterval of the interval $[t_0, \tau_1]$.

THEOREM 13. *Given the control problem for the data* (a) *through* (e) *of Theorem* 12 *with the further requirements that in* (a) *the elements of* $G(t)$ *are of class* $C^1[\tau_0, \tau_1]$ *and that* (f) *be replaced by*:

(f) *the set* $\Delta = \Delta(f(t, x, u), G(t), \Omega, x_0, T(t), E, h(t))$ *of admissible controls* $u(t)$ *defined on subintervals* $[t_0, t_1]$ *contained in* $[\tau_0, \tau_1]$ *with the same left endpoint* t_0 (*and perhaps different right endpoints* $t_1 > t_0$) *is such that* $|\Delta u| \leqq \Delta h$ *on each subinterval of* $[t_0, t_1]$ *for the given nondecreasing right continuous function* $h(t)$ *defined on* $[t_0, \tau_1]$.

Assume further that

(A) Δ *is nonempty,*

(B) *there exists a real bound* $B < \infty$ *such that for all responses* $x(t)$ *corresponding to controls in* Δ, $|x(t)| \leqq B$.

Conclusion: *Then there exists an optimal control in* Δ.

Proof. The computation at the beginning of the proof of the previous theorem was designed to show that responses $x(t)$ to controls in Δ are of uniform bounded total variation whether they be defined on the same interval or whether the interval depends on each separate control.

Proceeding as before, we let $\inf C(u) = \tilde{m} > -\infty$. If Δ is finite the theorem is trivially true; hence we assume Δ has infinitely many controls and we select a sequence $u^{(k)}(t)$ from Δ such that the $u^{(k)}$ are defined on intervals $[t_0, t_1^{(k)}]$ for which $C(u^{(k)})$ decreases monotonically to $\tilde{m}$. Select

a subsequence (which we shall still label $u^{(k)}(t)$) such that $t_1^{(k)} \to t_1^*$ in a monotonically decreasing fashion. We will denote this by $t_1^{(k)} \downarrow t_1^*$ (the case $t_1^{(k)} \uparrow t_1^*$ will be considered later). Next choose $\hat{t}_1$ such that $\hat{t}_1^* < t_1^{(k_0)} \leqq \hat{t}_1 < t_1^{(k_0+1)} < \tau_1$ for some k_0. (From now on all reference to the index k tacitly assumes $k > k_0$.) Then extend the controls $u^{(k)}(t)$ to the interval $[t_0, \hat{t}_1]$ by defining

$$\hat{u}^{(k)}(t) = \begin{cases} u^{(k)}(t) & \text{if } t_0 \leqq t \leqq t_1^{(k)}, \\ u^{(k)}(t_1^{(k)}) & \text{if } t_1^{(k)} < t \leqq \hat{t}_1. \end{cases} \tag{6.20}$$

Since $u^{(k)}(t) \in \Omega$ for all $t_0 \leqq t \leqq t_1^{(k)}$, then it is evident that $\hat{u}^{(k)}(t) \in \Omega$ for $t_0 \leqq t \leqq \hat{t}_1$. Also we have $\hat{u}^{(k)}(t)$ continuous from the right for every $t \in [t_0, \hat{t}_1)$ and when $\hat{u}^{(k)}$ is restricted to the interval $[t_0, t_1^{(k)}]$ it is simply $u^{(k)}(t)$ and thus $\hat{u}^{(k)}(t) \in \Delta$.

Now $|\Delta u| \leqq \Delta h$ on each subinterval of $[t_0, t_1]$ implies that the u in Δ are of uniform bounded variation, in fact, $v(u, [t_0, t_1]) \leqq h(\tau_1) - h(t_0)$. Thus $\hat{u}^{(k)}$ are of uniform bounded variation and, as before, there exist a subsequence of $\hat{u}^{(k)}$ (which we still label $\hat{u}^{(k)}$) and a function $u^*(t)$ defined on $[t_0, \hat{t}_1]$ such that

$$\lim_{k \to \infty} \hat{u}^{(k)}(t) = u^*(t) \tag{6.21}$$

everywhere on $[t_0, \hat{t}_1]$ and moreover

$$v(u^*, [t_0, \hat{t}_1]) \leqq \lim_{k \to \infty} v(\hat{u}^{(k)}, [t_0, \hat{t}_1]). \tag{6.22}$$

We know $u^*(t)$ is continuous from the right on $[t_0, \hat{t}_1)$ by the same calculation that verified $u^*(t)$ was continuous from the right in $[t_0, t_0 + \delta)$ in Theorem 12.

Observe that for $t = t_1^{(k)}$ we have

$$\hat{u}^{(k)}(t_1^{(k)}) \to u^*(t_1^*), \tag{6.23}$$

because for any $\tau > t_1^*$ we know $\hat{u}^{(k)}(\tau) = \hat{u}^{(k)}(t_1^{(k)})$ for k sufficiently large (i.e., such that $t_1^{(k)} < \tau$) and $\hat{u}^{(k)}(\tau) \to u^*(\tau)$ implies $\hat{u}^{(k)}(t_1^{(k)}) \to u^*(\tau)$. Thus $u^*(t)$ is constant on $(t_1^*, \hat{t}_1]$ and since $u^*(t)$ is right continuous we have

$$\lim_{t \to t_1^{*+}} u^*(t) = u^*(t_1^*), \tag{6.24}$$

i.e., $u^*(t)$ has the constant value $u^*(t_1^*)$ on the interval $[t_1^*, \hat{t}_1]$. In particular,

$$u^*(t_1^{(k)}) = u^*(t_1^*) \tag{6.25}$$

for all k.

We must show that $u^*(t) \in \Delta$.

(i) Show that $u^*(t) \in \Omega$. This follows from (6.21) and the compactness of Ω.

(ii) The fact that $|\Delta u^*| \leqq \Delta h$ on each subinterval of $[t_0, \hat{t}_1]$ follows trivially.

(iii) Show that u^* has a response $x^*(t)$ on $[t_0, \hat{t}_1]$ such that its restriction to $[t_0, t_1^*]$ satisfies $x^*(t_0) = x_0$ and $x^*(t_1^*) \in T(t_1^*)$. Denote the response to $\hat{u}^{(k)}(t)$ by

$$\hat{x}^{(k)}(t) = x_0 + \int_{t_0}^{t} f(s, \hat{x}^{(k)}(s), \hat{u}^{(k)}(s))\, ds + \int_{t_0}^{t} G(s)\, d\hat{u}^{(k)}(s). \tag{6.26}$$

We must show that each response $x^{(k)}(t)$ on $[t_0, t_1^{(k)}]$ can be extended to the interval $[t_0, \hat{t}_1]$ using the extended controls in such a way that the sequence of extended responses $\hat{x}^{(k)}(t)$ is uniformly bounded on $[t_0, \hat{t}_1]$. Then if the $\hat{x}^{(k)}(t)$ are uniformly bounded they will also have uniform bounded total variation as seen from the fact that calculating the variation in (6.26) yields

$$v(\hat{x}^{(k)}, [t_0, \hat{t}_1]) \leqq \int_{t_0}^{\hat{t}_1} |f(s, \hat{x}^{(k)}(s), \hat{u}^{(k)}(s))|\, ds + Gv(\hat{u}^{(k)}, [t_0, \hat{t}_1]). \tag{6.27}$$

But the last term is less than or equal to $G[h(\tau_1) - h(t_0)]$ because $v(\hat{u}^{(k)}, [t_0, \hat{t}_1]) = v(u^{(k)}, [t_0, t_1^{(k)}])$. Also the integrand in the integral term is uniformly bounded if the sequence $\hat{x}^{(k)}(t)$ is uniformly bounded.

To show $\hat{x}^{(k)}(t)$ is uniformly bounded we note that every point $x^{(k)}(t_1^{(k)})$ is contained in the sphere $|x| \leqq B$. Let $\hat{x}$ be an arbitrary point in this B-sphere and consider the equation

$$\hat{x}(t) = \hat{x} + \int_{t_1}^{t} f(s, \hat{x}(s), u_0)\, ds \tag{6.28}$$

for some constant value $u_0 \in \Omega$. Because $f_x \in C^1$ with respect to x and f is continuous in t and $u \equiv u_0$ is held fixed we have by the Carathéodory existence theorem for ordinary differential equations [3] that (6.28) has a unique absolutely continuous response $\hat{x}(t)$ on the interval $[t_1, t_1 + a]$ for which $\hat{x}(t_1) = \hat{x}$. We have only to consider the closed sphere $|x - \hat{x}| \leqq B$ and a time t_1 sufficiently near to t_1^* (but greater than t_1^*); then every solution of (6.28) is in the sphere $|x| \leqq 2B$ if t is in the interval $[t_1^*, t_1^* + a]$ for an appropriate value of a which is nonzero and depends on f and f_x. Then from (6.26) we can write

$$\begin{aligned} \hat{x}^{(k)}(t) &= x_0 + \int_{t_0}^{t_1^{(k)}} f(s, x^{(k)}(s), u^{(k)}(s))\, ds + \int_{t_0}^{t_1^{(k)}} G(s)\, du^k(s) \\ &\qquad + \int_{t_1^{(k)}}^{t} f(s, \hat{x}^{(k)}(s), u^{(k)}(t_1^{(k)}))\, ds \\ &= x^{(k)}(t_1^{(k)}) + \int_{t_1^{(k)}}^{t} f(s, \hat{x}^{(k)}(s), u^{(k)}(t_1^{(k)}))\, ds, \end{aligned} \tag{6.29}$$

which is of the form (6.28), and hence it follows that $\hat{x}^{(k)}(t)$ are uniformly bounded on $[t_0, \hat{t}_1]$.

Since the total variation of $\hat{x}^{(k)}(t)$ is uniformly bounded for all k there exist a subsequence of $\hat{x}^{(k)}(t)$ on $[t_0, \hat{t}_1]$ (still retaining the notation $\hat{x}^{(k)}$) and a function $x^*(t)$ such that everywhere on $[t_0, \hat{t}_1]$ we have

$$\lim_{k\to\infty} \hat{x}^{(k)}(t) = x^*(t). \tag{6.30}$$

By selecting the corresponding subsequence from $\hat{u}^{(k)}$ we do not change any of the preceding limiting operations satisfied by $\hat{u}^{(k)}(t)$.

By (6.21) and (6.30) we have

$$\lim_{k\to\infty} f(t, \hat{x}^{(k)}(t), \hat{u}^{(k)}(t)) = f(t, x^*(t), u^*(t)) \tag{6.31}$$

everywhere on $[t_0, \hat{t}_1]$ and since $| f(t, \hat{x}^{(k)}(t), \hat{u}^{(k)}(t)) |$ is uniformly bounded on $[t_0, \hat{t}_1]$ we have by Lebesgue's theorem on dominated convergence that

$$\lim_{k\to\infty} \int_{t_0}^{t} f(s, \hat{x}^{(k)}(s), \hat{u}^{(k)}(s))\, ds = \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds \tag{6.32}$$

for each fixed $t \in [t_0, \hat{t}_1]$. Also by (6.21) and the fact that $\hat{u}^{(k)}(t)$ are of uniform bounded total variation we obtain by an application of the theorem of Helly-Bray that for all continuous $G(s)$,

$$\lim_{k\to\infty} \int_{t_0}^{t} G(s)\, d\hat{u}^{(k)}(s) = \int_{t_0}^{t} G(s)\, du^*(s) \tag{6.33}$$

for each fixed $t \in [t_0, \hat{t}_1]$. Now consider

$$\begin{aligned} \lim_{k\to\infty} \hat{x}^{(k)}(t) &= x_0 + \lim_{k\to\infty} \int_{t_0}^{t} f(s, \hat{x}^{(k)}(s), \hat{u}^{(k)}(s))\, ds \\ &\qquad + \lim_{k\to\infty} \int_{t_0}^{t} G(s)\, d\hat{u}^{(k)}(s) \\ &= x_0 + \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds + \int_{t_0}^{t} G(s)\, du^*(s) \end{aligned} \tag{6.34}$$

by virtue of (6.32) and (6.33). Thus $\lim_{k\to\infty} \hat{x}^{(k)}(t)$ exists for each $t \in [t_0, \hat{t}_1]$ and by uniqueness of the limit we have from (6.30) that

$$x^*(t) = x_0 + \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds + \int_{t_0}^{t} G(s)\, du^*(s) \tag{6.35}$$

is the response to $u^*(t)$. It is obvious that $x^*(t_0) = x_0$ and so we must finally show that $x^*(t_1^*) \in T(t_1^*)$. To this end consider

$$\begin{aligned} | x^{(k)}(t_1^{(k)}) - x^*(t_1^*) | &\leqq | x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)}) | \\ &\qquad + | x^*(t_1^{(k)}) - x^*(t_1^*) |. \end{aligned} \tag{6.36}$$

Since $x^*(t)$ is continuous from the right (because u^* is) then $x^*(t_1^{(k)}) \to x^*(t_1^*)$, and hence the last term on the right offers no trouble. To handle the first term we proceed as follows:

$$
\begin{aligned}
x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)}) &= \int_{t_0}^{t_1^{(k)}} f(s, x^{(k)}(s), \hat{u}^{(k)}(s))\, ds \\
&\quad + \int_{t_0}^{t_1^{(k)}} G(s)\, d\hat{u}^{(k)}(s) \\
&\quad - \int_{t_0}^{t_1^{(k)}} f(s, x^*(s), u^*(s))\, ds - \int_{t_0}^{t_1^{(k)}} G(s)\, du^*(s) \\
&= \int_{t_0}^{t_1^{(k)}} [f(s, x^{(k)}(s), u^{(k)}(s)) - f(s, x^*(s), u^*(s))]\, ds \\
&\quad + \int_{t_0}^{t_1^{(k)}} G(s)\, d[\hat{u}^{(k)}(s) - u^*(s)] \\
&\equiv I_1 + I_2.
\end{aligned} \tag{6.37}
$$

Now

$$
\begin{aligned}
I_1 &= \int_{t_0}^{t_1^*} [f(s, x^{(k)}(s), \hat{u}^{(k)}(s)) - f(s, x^*(s), u^*(s))]\, ds \\
&\quad + \int_{t_1^*}^{t_1^{(k)}} [f(s, x^{(k)}(s), \hat{u}^{(k)}(s)) - f(s, x^*(s), u^*(s)]\, ds,
\end{aligned} \tag{6.38}
$$

and since the integrand of the first term is uniformly bounded on $[t_0, t_1^*]$ and approaches zero everywhere in that interval by (6.31), the first term on the right approaches zero. The second term on the right has a bounded integrand over a path of integration which goes to zero. Hence $\lim_{k\to\infty} I_1 = 0$. Next we consider I_2 and use integration by parts:

$$
\begin{aligned}
I_2 &= -\int_{t_0}^{t_1^{(k)}} \dot{G}(s)[\hat{u}^{(k)}(s) - u^*(s)]\, ds \\
&\quad + G(t_1^{(k)})[\hat{u}^{(k)}(t_1^{(k)}) - u^*(t_1^{(k)})] - G(t_0)[\hat{u}^{(k)}(t_0) - u^*(t_0)].
\end{aligned} \tag{6.39}
$$

The first term on the right approaches zero for exactly the same reasons as those in establishing (6.38). The second term approaches zero by virtue of (6.23) and (6.24) while the third term approaches zero by virtue of (6.21). We have thus established that

$$
\lim_{k\to\infty} x^{(k)}(t_1^{(k)}) = x^*(t_1^*). \tag{6.40}
$$

Now $x^{(k)}(t_1^{(k)}) \in T(t_1^{(k)})$ for $k = 1, 2, 3, \cdots$ and $x^*(t_1^*) = \lim_{k\to\infty} x^{(k)}(t_1^{(k)})$. If $x^*(t_1^*)$ were not in $T(t_1^*)$ then there would exist a neighborhood N of the compact set $T(t_1^*)$ so that $x^*(t_1^*)$ is not in the closure $\bar{N}$ of N. But

$T(t) \subset N$ for t sufficiently near t_1^* and thus $x^{(k)}(t_1^{(k)}) \in N$ for large k. But $x^*(t_1^*) \notin \bar{N}$ and this is a contradiction; therefore $x^*(t_1^*) \in T(t_1^*)$ and the control $u^*(t)$ on $t_0 \leqq t \leqq t_1^*$ belongs to Δ.

It is an easy matter to compute the cost of $u^*(t)$, for

$$
\begin{aligned}
C(\hat{u}^{(k)}) &= \int_{t_0}^{t_1^{(k)}} f^0(s, x^{(k)}(s), \hat{u}^{(k)}(s))\, ds \\
&= \int_{t_0}^{t_1^*} f^0(s, x^{(k)}(s), \hat{u}^{(k)}(s))\, ds \\
&\qquad + \int_{t_1^*}^{t_1^{(k)}} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds.
\end{aligned} \tag{6.41}
$$

Now $| f^0(s, x^{(k)}(s), u^{(k)}(s)) |$ is uniformly bounded on $[t_0, t_1^{(k)}]$ and thus the last term approaches zero because the path of integration approaches zero. Then by applying Lebesgue's theorem on dominated convergence to the first term we obtain

$$
\lim_{k\to\infty} C(\hat{u}^{(k)}) = \int_{t_0}^{t_1^*} f^0(s, x^*(s), u^*(s))\, ds = C(u^*) = \tilde{m} \tag{6.42}
$$

and thus $u^*(t)$ on $t_0 \leqq t \leqq t_1^*$ is an optimal control.

Return now to the assumption that $t_1^{(k)} \downarrow t_1^*$. Suppose instead that $t_1^{(k)} \uparrow t_1^*$. Then extend each control $u^{(k)}(t)$, for sufficiently large k, to the interval $[t_0, \hat{t}_1]$ by setting $\hat{t}_1 = t_1^* + \delta$ for appropriately small $\delta > 0$ and defining

$$
\hat{u}^{(k)}(t) = \begin{cases} u^{(k)}(t) & \text{if } t_0 \leqq t \leqq t_1^{(k)}, \\ u^{(k)}(t_1^{(k)}) & \text{if } t_1^{(k)} < t \leqq \hat{t}_1. \end{cases}
$$

As before there exist a subsequence (still to be labelled $\tilde{u}^{(k)}(t)$) and a function $u^*(t)$, necessarily of bounded variation and continuous from the right, such that

$$
\lim_{k\to\infty} \tilde{u}^{(k)}(t) = u^*(t) \tag{6.43}
$$

everywhere on $[t_0, \hat{t}_1]$. Furthermore it is clear that the extended controls did not receive an increase in variation and hence

$$
v(u^*, [t_0, t_1^*]) \leqq \varliminf_{k\to\infty} v(\tilde{u}^{(k)}, [t_0, t_1^*]) \leqq [h(\tau_1) - h(t_0)].
$$

We know each extended response $\tilde{x}^{(k)}(t)$ on $[t_0, \hat{t}_1]$ is again uniformly bounded by the same calculations as those in (6.27)–(6.29). Thus the total variation of $\tilde{x}^{(k)}(t)$ is uniformly bounded and there exist a subsequence (still to be labelled $\tilde{x}^{(k)}(t)$) and a function of bounded variation $x^*(t)$ such that everywhere on $[t_0, \hat{t}_1]$ we have

$$\lim_{k\to\infty} \tilde{x}^{(k)}(t) = x^*(t). \tag{6.44}$$

By (6.43) and (6.44) we have

$$\lim_{k\to\infty} f(t, \tilde{x}^{(k)}(t), \tilde{u}^{(k)}(t)) = f(t, x^*(t), u^*(t)) \tag{6.45}$$

everywhere on $[t_0, \hat{t}_1]$; and since $|f(t, x, u)|$ is uniformly bounded on $[t_0, \hat{t}_1]$ for u in Δ (and x therefore bounded by hypothesis B) we have by Lebesgue's theorem on dominated convergence that

$$\lim_{k\to\infty} \int_{t_0}^{t} f(s, \tilde{x}^{(k)}(s), \tilde{u}^{(k)}(s))\, ds = \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds \tag{6.46}$$

for each fixed t in $[t_0, \hat{t}_1]$. Also by (6.43) and the fact that $\tilde{u}^{(k)}(t)$ are of uniform bounded total variation we have by an application of the Helly-Bray theorem that for all continuous $G(s)$,

$$\lim_{k\to\infty} \int_{t_0}^{t} G(s)\, d\tilde{u}^{(k)}(s) = \int_{t_0}^{t} G(s)\, du^*(s) \tag{6.47}$$

for each fixed t in $[t_0, \hat{t}_1]$. Therefore

$$\lim_{k\to\infty} \tilde{x}^{(k)}(t) = x_0 + \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds + \int_{t_0}^{t} G(s)\, du^*(s) \tag{6.48}$$

and by uniqueness of the limit it follows from (6.44) that

$$x^*(t) = x_0 + \int_{t_0}^{t} f(s, x^*(s), u^*(s))\, ds + \int_{t_0}^{t} G(s)\, du^*(s) \tag{6.49}$$

is the response to $u^*(t)$. It is evident that $x^*(t_0) = x_0$; so all that remains is the verification that $x^*(t_1^*) \in T(t_1^*)$.

Consider

$$\begin{aligned} |x^{(k)}(t_1^{(k)}) - x^*(t_1^*)| \leqq\ & |x^{(k)}(t_1^{(k)}) - \tilde{x}^{(k)}(t_1^*)| \\ & + |\tilde{x}^{(k)}(t_1^*) - x^*(t_1^*)|. \end{aligned} \tag{6.50}$$

Since $x^{(k)}(t_1^*) \to x^*(t_1^*)$, the last term on the right offers no trouble. To handle the first term on the right we proceed as follows:

$$\begin{aligned} x^{(k)}(t_1^{(k)}) - \tilde{x}^{(k)}(t_1^*) &= \int_{t_0}^{t_1^{(k)}} f(s, x^{(k)}(s), \tilde{u}^{(k)}(s))\, ds \\ &\quad + \int_{t_0}^{t_1^{(k)}} G(s)\, d\tilde{u}^{(k)}(s) - \int_{t_0}^{t_1^*} f(s, \tilde{x}^{(k)}(s), \tilde{u}^{(k)}(s))\, ds \\ &\quad - \int_{t_0}^{t_1^*} G(s)\, du^{(k)}(s) \\ &= -\int_{t_1^{(k)}}^{t_1^*} f(s, \tilde{x}^{(k)}(s), \tilde{u}^{(k)}(s))\, ds - \int_{t_1^{(k)}}^{t_1^*} G(s)\, d\tilde{u}^{(k)}(s) \equiv I_1 + I_2. \end{aligned} \tag{6.51}$$

The integral I_1 obviously goes to zero as $k \to \infty$. As for I_2 we have

$$\int_{t_1^{(k)}}^{t_1^*} G(s)\, d\tilde{u}^{(k)}(s) = -\int_{t_1^{(k)}}^{t_1^*} \dot{G}(s)\tilde{u}^{(k)}(s)\, ds + G(t_1^*)\tilde{u}^{(k)}(t_1^*) - G(t_1^{(k)})\tilde{u}^{(k)}(t_1^{(k)}). \tag{6.52}$$

But $\tilde{u}^{(k)}(t_1^*) \equiv \tilde{u}^{(k)}(t_1^{(k)})$. Hence

$$\int_{t_1^{(k)}}^{t_1^*} G(s)\, d\tilde{u}^{(k)}(s) = -\int_{t_1^{(k)}}^{t_1^*} \dot{G}(s)\tilde{u}^{(k)}(s)\, ds + \tilde{u}^{(k)}(t_1^*)[G(t_1^*) - G(t_1^{(k)})]. \tag{6.53}$$

Both terms on the right approach zero as k approaches ∞ and we have from (6.50) that

$$\lim_{k\to\infty} x^{(k)}(t_1^{(k)}) = x^*(t_1^*). \tag{6.54}$$

Thus, exactly as before, $x^*(t_1^*) \in T(t_1^*)$ and therefore $u^*(t)$ on $[t_0, t_1^*]$ lies in Δ.

The cost of the control $u^*(t)$ is computed exactly as in (6.41) and (6.42) and we have that $u^*(t)$ on $[t_0, t_1^*]$ is an optimal control.

The following example illustrates a situation where the optimal control fails to exist. (A prime on a vector denotes the transpose of that vector.)

Consider the equations

$$Dx^1 = x^2,$$
$$Dx^2 = -x^2 + u(t) + Du.$$

The initial point is $x_0 = (0, 0)'$ and the target is the fixed point $(0, 1)'$. The restraint set Ω is $-1 \leqq u \leqq 1$. The set Δ is the set of all controls of uniform bounded variation with fixed initial time $t = 0$. The actual bound E on the total variation is only assumed to be greater than 1 and no restriction regarding the existence of a nondecreasing function $h(t)$ such that $|\Delta u| \leqq \Delta h$ for all u in Δ is imposed.

Consider the controls

$$u^{(k)}(t) = \begin{cases} 0 & \text{if } 0 \leqq t < \frac{1}{k}, \\ 1 & \text{if } \frac{1}{k} \leqq t, \end{cases}$$

defined on [0, 1]. Then

$$x^2(t) = x^2(0) + \int_0^t -x^2(s)\, ds + \int_0^t u(s)\, ds + u(t) - u(0)$$

and

$$x^2_{(k)}(t) = \begin{cases} 0 & \text{if} \quad 0 \leqq t < \dfrac{1}{k}, \\ 1 & \text{if} \quad t = \dfrac{1}{k}. \end{cases}$$

Hence

$$\int_0^{1/k} x^2(s)\, ds = 0 = x^1\left(\frac{1}{k}\right),$$

and the sequence $u^{(k)}(t)$ is a sequence of controls in Δ. But the infimum of the cost functional is

$$\lim_{k\to\infty} C(u^{(k)}) = \lim_{k\to\infty} \frac{1}{k} = 0.$$

The only control $u^*(t)$ in Δ which will produce a minimal time of 0 is the "multiple valued control" $u(0) = 0$, $u(0) = 1$; hence an optimal control does not exist. Notice that there does not exist a nondecreasing function $h(t)$ such that

$$| \Delta u^{(k)} | \leqq \Delta h,$$

but that all other hypotheses of the theorem are satisfied.

7. Existence of optimal control for ordinary differential system. The foregoing methods may be applied to the ordinary control problem, i.e., no measure appearing in the right side, to extend the results of that problem to the case where the control enters nonlinearly (see [1], [2], [6]). Because the responses are absolutely continuous in this case, it is not necessary to make the controls continuous from one side or the other, but rather that is left open to be decided in a particular application. All that is required is that the class of admissible controls have uniform bounded total variation (rather than being merely measurable as in [1], for example).

THEOREM 14. *Consider the control problem for the data:*

(a) $\dot{x}^i = f^i(t, x^1, \cdots, x^n, u^1, \cdots, u^m)$ *for* $i = 1, \cdots, n$, *with* $f^i(t, x, u)$ *and* $\partial f^i(t, x, u)/\partial x^k$, $k = 1, \cdots, n$, *continuous* on $R^1 \times R^n \times R^m$;

(b) *a nonempty, compact restraint set* $\Omega \subset R^m$;

(c) *the initial point* $x_0 \in R^n$;

(d) *the continuously moving nonempty compact target set* $T(t) \subset R^n$ *on* $[\tau_0, \tau_1]$;

(e) *the cost functional*

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

where $f^0(t, x, u)$ *is continuous on* $R^1 \times R^n \times R^m$.

Assume the set Δ *consists of controls with values in* Ω *and with* $v(u, [t_0, t_1]) \leqq E$ *uniformly for all* u *in* Δ, *and with responses traveling from* x_0 *to* T. *Assume*

(A) Δ *is nonempty; and*

(B) *there exists a real bound* $B < \infty$ *such that for all responses* $x(t)$ *corresponding to* Δ, $|x(t)| \leqq B$ *uniformly for all responses.*

Conclusion: *Then there exists an optimal control in* Δ.

Proof. Since Δ is nonempty and the corresponding responses are uniformly bounded, $\inf C(u) = \tilde{m} > -\infty$, where the infimum is taken over all $u \in \Delta$. Either Δ is a finite set in which case the theorem is trivially true, or we can select from Δ a sequence of controls $u^{(k)}(t)$, defined on various intervals $t_0^{(k)} \leqq t \leqq t_1^{(k)}$, for which $C(u^{(k)})$ decreases monotonically to $\tilde{m}$. Select a subsequence of $t_0^{(k)}$ such that $t_0^{(k)} \to t_0^*$ monotonically. Similarly select a subsequence of $t_1^{(k)}$ such that $t_1^{(k)} \to t_1^*$ monotonically. First treat the case where $t_1^{(k)} \to t_1^*$ from above and where $t_0^{(k)} \to t_0^*$ from below (other cases will be treated later). Now we may select a subsequence of $u^{(k)}$ (still to be labelled $u^{(k)}$) such that there is a function $u^*(t)$ defined on $[t_0^*, t_1^*]$ with $u^*(t)$ necessarily of bounded variation for which

$$\lim_{k\to\infty} u^{(k)}(t) = u^*(t) \tag{7.1}$$

everywhere on $[t_0^*, t_1^*]$. We show that $u^*(t)$ on $[t_0^*, t_1^*]$ belongs to Δ. Note that

$$v(u^*, [t_0^*, t_1^*]) \leqq E \tag{7.2}$$

and that $u^*(t) \in \Omega$ on $[t_0^*, t_1^*]$.

Let $x^{(k)}(t)$ denote the response to $u^{(k)}(t)$, i.e.,

$$x^{(k)}(t) = x_0 + \int_{t_0^{(k)}}^{t} f(s, x^{(k)}(s), u^{(k)}(s))\, ds. \tag{7.3}$$

Since the responses $x^{(k)}(t)$ are uniformly bounded they also have uniform bounded total variation on their intervals and in particular on $[t_0^*, t_1^*]$. As before, there are a subsequence of $x^{(k)}(t)$ and a function $x^*(t)$ of bounded variation on $[t_0^*, t_1^*]$ such that (retaining same notation)

$$\lim_{k\to\infty} x^{(k)}(t) = x^*(t) \tag{7.4}$$

everywhere on $[t_0^*, t_1^*]$. Now we write

$$\begin{aligned} x^{(k)}(t) = x_0 &+ \int_{t_0^*}^{t} f(s, x^{(k)}(s), u^{(k)}(s))\, ds \\ &+ \int_{t_0^{(k)}}^{t_0^*} f(s, x^{(k)}(s), u^{(k)}(s))\, ds. \end{aligned} \tag{7.5}$$

The last term approaches zero as k approaches infinity and the first integral, by Lebesgue's theorem on dominated convergence, has the limit

$$\lim_{k\to\infty} \int_{t_0^*}^{t} f(s, x^{(k)}(s), u^{(k)}(s))\, ds = \int_{t_0^*}^{t} f(s, x^*(s), u^*(s))\, ds \tag{7.6}$$

for all t in $[t_0^*, t_1^*]$. Thus

$$\lim_{k\to\infty} x^{(k)}(t) = x_0 + \int_{t_0^*}^{t} f(s, x^*(s), u^*(s))\, ds \tag{7.7}$$

exists and because of the uniqueness of the limit it follows from (7.4) that

$$x^*(t) = x_0 + \int_{t_0^*}^{t} f(s, x^*(s), u^*(s))\, ds \tag{7.8}$$

is the response to the control $u^*(t)$ satisfying $x^*(t_0^*) = x_0$.

Now consider

$$\begin{aligned} |\, x^{(k)}(t_1^{(k)}) - x^*(t_1^*)\,| \leqq {} & |\, x^{(k)}(t_1^{(k)}) - x^{(k)}(t_1^*)\,| \\ & + |\, x^{(k)}(t_1^*) - x^*(t_1^*)\,|. \end{aligned} \tag{7.9}$$

We have for an appropriate constant $B > 0$ that

$$|\, x^{(k)}(t_1^{(k)}) - x^{(k)}(t_1^*)\,| \leqq B\,|\, t_1^{(k)} - t_1^*\,|, \tag{7.10}$$

since the derivative $f(t, x, u)$ (i.e., $\dot{x}$) is uniformly bounded for responses in Δ. Thus (7.9) and (7.10) imply

$$\lim_{k\to\infty} x^{(k)}(t_1^{(k)}) = x^*(t_1^*). \tag{7.11}$$

Now $x^{(k)}(t_1^{(k)}) \in T(t_1^{(k)})$ for each k; therefore if $x^*(t_1^*) \notin T(t_1^*)$ there would exist a neighborhood N of the compact set $T(t_1^*)$ such that $x^*(t_1^*)$ is not in the closure of N. But $T(t) \subset N$ for t sufficiently near t_1^* and thus $x^{(k)}(t_1^{(k)}) \in N$ for large k and yet $x^*(t_1^*)$ is not in $\bar{N}$. This is a contradiction and therefore $x^*(t_1^*) \in T(t_1^*)$. Hence the control $u^*(t)$ on $[t_0^*, t_1^*]$ belongs to Δ.

To compute the cost of $u^*(t)$ we consider

$$C(u^{(k)}) = \int_{t_0^{(k)}}^{t_1^{(k)}} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds. \tag{7.12}$$

By arguments similar to those leading to (7.8) we obtain

$$\lim_{k\to\infty} C(u^{(k)}) = \int_{t_0^*}^{t_1^*} f^0(s, x^*(s), u^*(s))\, ds \equiv C(u^*), \tag{7.13}$$

and thus by uniqueness of limit for a subsequence we have

$$C(u^*) = \lim_{k\to\infty} C(u^{(k)}) = \tilde{m}. \tag{7.14}$$

Therefore $u^*(t)$ on $[t_0^*, t_1^*]$ is an optimal control.

Returning to the assumption

$$t_0^{(k)} \uparrow t_0^*, \qquad t_1^{(k)} \downarrow t_1^*,$$

suppose instead we have

$$t_0^{(k)} \leqq t_0^*, \qquad t_1^{(k)} \leqq t_1^*$$

(the other cases can be treated similarly). Extend each control $u^{(k)}(t)$ to the interval $[t_0, t_1^*]$ by defining $u^{(k)}(t) = u^{(k)}(t_0^{(k)})$, a constant on $t_1^{(k)} \leqq t \leqq t_1^*$. Again there are a subsequence of $u^{(k)}(t)$ and a function $u^*(t)$ of bounded variation defined on $[t_0^*, t_1^*]$ such that $v(u^*, [t_0^*, t_1^*]) \leqq E$ and

$$\lim_{k\to\infty} u^{(k)}(t) = u^*(t) \tag{7.15}$$

everywhere in $[t_0^*, t_1^*]$. It is clear that $u^*(t) \in \Omega$.

The extended responses are all uniformly bounded on $[t_0^*, t_1^*]$ because all the unmodified responses lie in some sphere centered at the origin of radius ρ, say, and a calculation similar to that in Theorem 13 shows that for $t_1^* - t_1^{(k)}$ sufficiently small, the extended responses lie in a sphere of at most radius 2ρ centered at the origin. By choosing a subsequence of $x^{(k)}(t)$ there is a corresponding function $x^*(t)$ defined on $[t_0^*, t_1^*]$ and of bounded variation there such that

$$\lim_{k\to\infty} x^{(k)}(t) = x^*(t) \tag{7.16}$$

everywhere on $[t_0^*, t_1^*]$.

Exactly as was done in leading up to (7.8) we obtain that $x^*(t)$ is the response to the control $u^*(t)$ and that $x^*(t_0) = x_0$.

Now $x^{(k)}(t_1^{(k)}) \in T(t_1^{(k)})$ and

$$\lim_{k\to\infty} x^*(t_1^{(k)}) = x^*(t_1^*) \tag{7.17}$$

because $x^*(t)$ is absolutely continuous. Thus

$$x^*(t_1^*) = \lim_{k\to\infty} [x^*(t_1^{(k)}) - x^{(k)}(t_1^{(k)}) + x^{(k)}(t_1^{(k)})], \tag{7.18}$$

but

$$\begin{aligned} &| x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)}) | \\ &\leqq \left| \int_{t_0^{(k)}}^{t_1^{(k)}} [f(s, x^{(k)}(s), u^{(k)}(s)) - f(s, x^*(s), u^*(s))]\, ds \right. \\ &\qquad \left. + \int_{t_1^{(k)}}^{t_1^*} f(s, x^*(s), u^*(s))\, ds \right|, \end{aligned} \tag{7.19}$$

and since $f(s, x, u)$ is continuous in x and u while $|f(s, x^*, u^*)|$ is bounded, we have by Lebesgue's theorem on dominated convergence that

$$|x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)})| \to 0. \tag{7.20}$$

Then from (7.18) it follows that

$$\lim_{k\to\infty} x^{(k)}(t_1^{(k)}) = x^*(t_1^*) \tag{7.21}$$

and, as before, $x^*(t_1^*) \in T(t_1^*)$ as required. Therefore $u^*(t)$ on $[t_0^*, t_1^*]$ lies in Δ.

Finally, we consider

$$C(u^{(k)}) = \int_{t_0^{(k)}}^{t_1^{(k)}} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds. \tag{7.22}$$

Using the extended controls and responses and Lebesgue's theorem, this is seen to approach $C(u^*)$ as $k \to \infty$ and by uniqueness of the limit it also approaches $\tilde{m}$. Thus $u^*(t)$ on $[t_0^*, t_1^*]$ is an optimal control.

Remark 1. The velocity $f(t, x, u)$ need only be defined and satisfy the hypotheses of the theorem for $\tau_0 \leqq t \leqq \tau_1$, $x \in \mathcal{O} \subset R^n$, $u \in \Omega \subset R^m$, where $\mathcal{O}$ is an open set in R^n which contains the initial point x_0, the moving target $T(t)$, and all the responses of Δ in a compact subset.

Remark 2. Note hypothesis (B) is satisfied if for some real α,

$$|f^i(t, x, u)| < \alpha, \qquad i = 1, \cdots, n$$

or if

$$\left|\frac{\partial f^i}{\partial x^k}(t, x, u)\right| < \alpha, \qquad i, k = 1, 2, \cdots, n,$$

in $[\tau_0, \tau_1] \times R^n \times \Omega$. Thus (B) is always satisfied if $f^i(t, x, u)$ are linear in x.

Remark 3. This theorem is an extension of Theorem 1 of [1]. It also includes Theorem 2 of [1] as a special case because the class of controls which satisfy a uniform Lipschitz condition

$$|u(t) - u(t')| \leqq A\,|t - t'|$$

for all pairs t, t' on $[\tau_0, \tau_1]$ is clearly also of uniform bounded total variation.

Remark 4. Consider the set $\Delta(t_0)$ defined to be that subset of Δ for which the control $u(t)$ and the response $x(t)$ initiate at a fixed t_0. If $\Delta(t_0)$ is non-empty and if the responses $x(t)$ for $u(t) \in \Delta(t_0)$ are uniformly bounded then there exists a control $u^*(t) \in \Delta(t_0)$ which is optimal relative to $\Delta(t_0)$. The same applies to the set $\Delta(t_0, t_1) \subset \Delta(t_0)$ where the time interval $[t_0, t_1]$ is fixed.

The following example illustrates a situation where the optimal control fails to exist.

Let the equations be

$$\dot{x} = \sin 2\pi u, \quad \dot{y} = \cos 2\pi u, \quad \dot{z} = -1 \quad \text{in } R^3.$$

The initial point is (0, 0, 1) and the target is the fixed point (0, 0, 0) on the time interval $0 \leqq t \leqq t_1 \leqq 2$. The restraint set Ω is $-1 \leqq u \leqq 1$ and the cost functional is

$$C(u) = \int_0^{t_1} (x^2 + y^2)\, dt.$$

Consider the class Δ to be the set of controls $\Delta(0)$, but without the restriction of uniform bounded total variation. Consider the piecewise linear controls $u^{(k)}(t)$ such that

$$\sin 2\pi u^{(k)}(t) = \sin 2\pi kt,$$

$$\cos 2\pi u^{(k)}(t) = \cos 2\pi kt,$$

for $k = 1, 2, 3, \cdots$. The corresponding responses are

$$x^{(k)}(t) = \frac{1 - \cos 2\pi kt}{2\pi k},$$

$$y^{(k)}(t) = \frac{\sin 2\pi kt}{2\pi k},$$

$$z^{(k)}(t) = 1 - t.$$

Thus $x^{(k)}(1) = 0$, $y^{(k)}(1) = 0$, $z^{(k)}(1) = 0$. The cost functional for $t_1 = 1$ is

$$C(u^{(k)}) = \int_0^1 \frac{1 - \cos 2\pi kt}{2\pi^2 k^2}\, dt = \frac{1}{2\pi^2 k^2}.$$

Thus

$$\lim_{k \to \infty} C(u^{(k)}) = 0,$$

and $\tilde{m} = 0$ is the infimum for all $C(u)$ with u in $\Delta(0)$. Yet there is no optimal control $u^*(t)$ on $0 \leqq t \leqq 1$ for which the cost is

$$C(u^*) = \int_0^1 (x^{*2} + y^{*2})\, dt = 0,$$

for such an optimal control requires the response $x^*(t) = 0$, $y^*(t) = 0$ which in turn implies

$$\sin 2\pi u^*(t) = 0, \quad \cos 2\pi u^*(t) = 0,$$

for almost all t. But this is impossible and hence there does not exist an optimal control for this problem. Note that the control functions $u^{(k)}(t)$

are not of uniform bounded total variation on the interval [0, 1] but that all other hypotheses of the theorem are satisfied.

REFERENCES

[1] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[2] V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, *The theory of optimal processes. I-The maximum principle*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 3–42.

[3] E. A. Coddington and N. L. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[4] N. Dunford and J. Schwartz, *Linear Operators*, Interscience, New York, 1958.

[5] L. M. Graves, *The Theory of Functions of Real Variables*, McGraw-Hill, New York, 1946.

[6] A. F. Filippov, *On certain questions in the theory of optimal regulations*, Vestnik Moskov. Univ., 2 (1959), pp. 25–32.

[7] I. Halperin, *Introduction to the Theory of Distributions*, University of Toronto Press, Toronto, 1952.

[8] P. Alexandroff and H. Hopf, *Topologie*, Springer, Berlin, 1935.

[9] A. Friedman, *Generalized Functions and Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[10] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis. vol. 1, Metric and Normed Spaces*, Graylock Press, Rochester, New York, 1957.

[11] T. H. Hildebrandt, *Convergence of sequences of linear operations*, Bull. Amer. Math. Soc., 28 (1922), pp. 53–58.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

PSEUDO-CONVEX FUNCTIONS*

O. L. MANGASARIAN†

Abstract. The purpose of this work is to introduce pseudo-convex functions and to describe some of their properties and applications. The class of all pseudo-convex functions over a convex set C includes the class of all differentiable convex functions on C and is included in the class of all differentiable quasi-convex functions on C. An interesting property of pseudo-convex functions is that a local condition, such as the vanishing of the gradient, is a global optimality condition. One of the main results of this work consists of showing that the Kuhn-Tucker differential conditions are sufficient for optimality when the objective function is pseudo-convex and the constraints are quasi-convex. Other results of this work are a strict converse duality theorem for mathematical programming and a stability criterion for ordinary differential equations.

1. Introduction. Throughout this work, we shall be concerned with the real, scalar, single-valued, differentiable function $\theta(x)$ defined on the non-empty open set D in the m-dimensional Euclidean space E^m. We let C be a subset of D and let ∇_x denote the $m \times 1$ partial differential operator

$$\nabla_x = \left[\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_m}\right]',$$

where the prime denotes the transpose. We say that $\theta(x)$ is *pseudo-convex* on C if for every x^1 and x^2 in C,

$$(1.1) \qquad (x^2 - x^1)'\nabla_x\theta(x^1) \geqq 0 \quad \text{implies} \quad \theta(x^2) \geqq \theta(x^1).$$

We say that $\theta(x)$ is *pseudo-concave* on C if for every x^1 and x^2 in C,

$$(1.2) \qquad (x^2 - x^1)'\nabla_x\theta(x^1) \leqq 0 \quad \text{implies} \quad \theta(x^2) \leqq \theta(x^1).$$

Thus $\theta(x)$ is pseudo-concave if and only if $-\theta(x)$ is pseudo-convex. In the subsequent paragraphs we shall confine our remarks to pseudo-convex functions. Analogous results hold for pseudo-concave functions by the appropriate multiplication by -1.

We shall relate the pseudo-convexity concept to the previously established notions of convexity, quasi-convexity [1], [2] and strict quasi-convexity [3], [5].

The function $\theta(x)$ is said to be *convex* on C, [2], if C is convex and if for every x^1 and x^2 in C,

$$(1.3) \qquad \theta(\lambda x^1 + (1 - \lambda)x^2) \leqq \lambda\theta(x^1) + (1 - \lambda)\theta(x^2)$$

* Received by the editors March 4, 1965.
† Shell Development Company, Emeryville, California.

for every λ such that $0 \leqq \lambda \leqq 1$. Equivalently, $\theta(x)$ is convex on C if

$$\theta(x^2) - \theta(x^1) \geqq (x^2 - x^1)'\nabla_x\theta(x^1) \tag{1.4}$$

for every x^1 and x^2 in C.

The function $\theta(x)$ is said to be *quasi-convex* on C, [1], [2], if C is convex and if for every x^1 and x^2 in C,

$$\theta(x^2) \leqq \theta(x^1) \quad \text{implies} \quad \theta(\lambda x^1 + (1 - \lambda)x^2) \leqq \theta(x^1) \tag{1.5}$$

for every λ such that $0 \leqq \lambda \leqq 1$. Equivalently, $\theta(x)$ is quasi-convex on C if

$$\theta(x^2) \leqq \theta(x^1) \quad \text{implies} \quad (x^2 - x^1)'\nabla_x\theta(x^1) \leqq 0. \tag{1.6}$$

The function $\theta(x)$ is said to be *strictly quasi-convex* on C, [3], [5], if C is convex and if for every x^1 and x^2 in C, $x^1 \neq x^2$,

$$\theta(x^2) < \theta(x^1) \quad \text{implies} \quad \theta(\lambda x^1 + (1 - \lambda)x^2) < \theta(x^1) \tag{1.7}$$

for every λ such that $0 < \lambda < 1$. It has been shown [5] that every lower semicontinuous strictly quasi-convex function is quasi-convex but not conversely.

In the next section we shall give some properties of pseudo-convex functions and show how these properties can be used to generalize some previous results of mathematical programming, duality theory and stability theory of ordinary differential equations. Theorem 1 generalizes the Arrow-Enthoven version [1, Theorem 1] of the Kuhn-Tucker differential sufficient optimality conditions for a mathematical programming problem. Theorem 2 gives a generalization of Huard's converse duality theorem of mathematical programming [4, Theorem 2] and Theorem 3 generalizes a stability criterion for equilibrium points of nonlinear ordinary differential equations [8, Theorem 1].

2. Properties of pseudo-convex functions and applications. In this section we shall give some properties of pseudo-convex functions and some extensions of the results of mathematical programming and ordinary differential equations.

PROPERTY 0. *Let $\theta(x)$ be pseudo-convex on C. If $\nabla_x\theta(x^0) = 0$, then x^0 is a global minimum over C.*

Proof. For any x in C,

$$(x - x^0)'\nabla_x\theta(x^0) = 0,$$

and hence by (1.1),

$$\theta(x) \geqq \theta(x^0),$$

which establishes the property.

PROPERTY 1. *Let C be convex. If $\theta(x)$ is convex on C, then $\theta(x)$ is pseudo-convex in C, but not conversely.*

Proof. If $\theta(x)$ is convex on C, then by (1.4),

$$(x^2 - x^1)'\nabla_x\theta(x^1) \geqq 0 \qquad \text{implies} \qquad \theta(x^2) \geqq \theta(x^1),$$

which is precisely (1.1). That the converse is not necessarily true can be seen from the example

$$\theta(x) \equiv x + x^3, \qquad x \in E^1,$$

which is pseudo-convex on E^1 but not convex.[1]

PROPERTY 2. *Let C be convex. If $\theta(x)$ is pseudo-convex on C, then $\theta(x)$ is strictly quasi-convex (and hence quasi-convex) on C, but not conversely.*

Proof. Let $\theta(x)$ be pseudo-convex on C. We shall assume that $\theta(x)$ is not strictly quasi-convex on C and show that this leads to a contradiction. If $\theta(x)$ is not strictly quasi-convex on C, then it follows from (1.7) that there exist $x^1 \neq x^2$ in C such that

$$\theta(x^2) < \theta(x^1), \tag{2.1}$$

and

$$\theta(x) \geqq \theta(x^1), \tag{2.2}$$

for some $x \in L$, where

$$L = \{x \mid x = \lambda x^1 + (1 - \lambda)x^2, 0 < \lambda < 1\}. \tag{2.3}$$

Hence there exists an $\bar{x} \in L$ such that

$$\theta(\bar{x}) = \max_{x \in L} \theta(x), \tag{2.4}$$

where

$$\bar{L} = L \cup \{x^1, x^2\}. \tag{2.5}$$

Now define

$$f(\lambda) = \theta((1 - \lambda)x^1 + \lambda x^2), \qquad 0 \leqq \lambda \leqq 1. \tag{2.6}$$

Hence

$$\theta(\bar{x}) = f(\bar{\lambda}), \tag{2.7}$$

where

$$\bar{x} = (1 - \bar{\lambda})x^1 + \bar{\lambda}x^2, \qquad 0 < \bar{\lambda} < 1. \tag{2.8}$$

[1] To see that $x + x^3$ is pseudo-convex, note that $\nabla_x\theta(x) = 1 + 3x^2 > 0$. Hence $(x - x^0)'\nabla_x\theta(x^0) \geqq 0$ implies that $x \geqq x^0$ and $x^3 \geqq (x^0)^3$, and thus

$$\theta(x) - \theta(x^0) = (x + x^3) - (x^0 + (x^0)^3) \geqq 0.$$

We have from (2.4) through (2.7) that $f(\lambda)$ achieves its maximum at $\bar{\lambda}$. Hence it follows by the differentiability of $\theta(x)$ and the chain rule that

$$(x^2 - x^1)'\nabla_x \theta(\bar{x}) = \frac{df(\bar{\lambda})}{d\lambda} = 0. \tag{2.9}$$

Since

$$x^2 - \bar{x} = x^2 - (1 - \bar{\lambda})x^1 - \bar{\lambda}x^2 = (1 - \bar{\lambda})(x^2 - x^1), \tag{2.10}$$

it follows from (2.9) and (2.10) and the fact that $\bar{\lambda} < 1$, that

$$(x^2 - \bar{x})'\nabla_x\theta(\bar{x}) = 0. \tag{2.11}$$

But by the pseudo-convexity of $\theta(x)$, (2.11) implies that

$$\theta(x^2) \geqq \theta(\bar{x}). \tag{2.12}$$

Hence from (2.1) and (2.12),

$$\theta(x^1) > \theta(\bar{x}),$$

which contradicts (2.4). Hence $\theta(x)$ must be strictly quasi-convex on C.

That the converse is not necessarily true can be seen from the example

$$\theta(x) \equiv x^3, \qquad x \in E^1,$$

which is strictly quasi-convex on E^1, but not pseudo-convex.

PROPERTY 3. *Let C be convex. If $\theta(x)$ is pseudo-convex on C, then every local minimum*[2] *is a global minimum.*

Proof. By Property 2, $\theta(x)$ is strictly quasi-convex on C. Now if $\bar{x}$ is a local minimum, then

$$\theta(\bar{x}) \leqq \theta(x) \qquad \text{for every} \quad x \in N(\bar{x}) \cap C, \tag{2.13}$$

where $N(\bar{x})$ is some neighborhood of $\bar{x}$. Let x be any point in C, but not in $N(\bar{x}) \cap C$. Then there exists a $\bar{\lambda}$, $0 < \bar{\lambda} < 1$, such that

$$\tilde{x} = ((1 - \bar{\lambda})\bar{x} + \bar{\lambda}x) \in N(\bar{x}) \cap C.$$

Now if $\theta(x) < \theta(\bar{x})$, then by the strict quasi-convexity of $\theta(x)$,

$$\theta(\tilde{x}) > \theta(\bar{x}),$$

which contradicts (2.13). Hence $\theta(x) \geqq \theta(\bar{x})$, which proves Property 3.

THEOREM 1. *Let $\theta(x)$, $g_1(x)$, $\cdots$, $g_n(x)$ be differentiable functions on E^m. Let C be a convex set in E^m and $\theta(x)$ be pseudo-convex on C and $g_1(x)$, $\cdots$, $g_n(x)$ be quasi-convex on C. If there exist an $x^0 \in C$ and $y^0 \in E^n$ satisfy-*

[2] A local minimum is an $\bar{x} \in C$ such that $\theta(\bar{x}) \leqq \theta(x)$ for all $x \in N(\bar{x}) \cap C$, where $N(\bar{x})$ is some neighborhood of $\bar{x}$.

ing the Kuhn-Tucker differential conditions [7], *namely,*

$$\nabla_x \theta(x^0) + \nabla_x \sum_{i=1}^{n} y_i^0 g_i(x^0) = 0, \tag{2.14}$$

$$\sum_{i=1}^{n} y_i^0 g_i(x^0) = 0, \tag{2.15}$$

$$g_i(x^0) \leqq 0, \qquad i = 1, \cdots, n, \tag{2.16}$$

$$y_i^0 \geqq 0, \qquad i = 1, \cdots, n, \tag{2.17}$$

then

$$\theta(x^0) = \min_{x \in C} \{\theta(x) \mid g_i(x) \leqq 0, i = 1, \cdots, n\}. \tag{2.18}$$

Proof. The proof is similar to part of the proof of [1, Theorem 1]. Let

$$I = \{i \mid g_i(x^0) < 0\}.$$

Hence $g_i(x^0) = 0$ for $i \notin I$. From (2.15), (2.16) and (2.17) it follows that

$$y_i^0 = 0 \quad \text{for} \quad i \in I. \tag{2.19}$$

Let

$$R = \{x \mid g_i(x) \leqq 0, i = 1, 2, \cdots, n, x \in C\}.$$

Then $g_i(x) \leqq g_i(x^0)$ for $i \notin I$, $x \in R$. Hence by the quasi-convexity of the g_i's on R it follows from (1.6) that

$$(x - x^0)' \nabla_x g_i(x^0) \leqq 0 \quad \text{for} \quad i \notin I, x \in R. \tag{2.20}$$

Hence by (2.20) and (2.17) we have that

$$(x - x^0)' \nabla_x \sum_{i \notin I} y_i^0 g_i(x^0) \leqq 0 \quad \text{for} \quad x \in R, \tag{2.21}$$

and from (2.19) we have

$$(x - x^0)' \nabla_x \sum_{i \in I} y_i^0 g_i(x^0) = 0 \quad \text{for} \quad x \in R. \tag{2.22}$$

Hence (2.21) and (2.22) imply

$$(x - x^0)' \nabla_x \sum_{i=1}^{n} y_i^0 g_i(x^0) \leqq 0 \quad \text{for} \quad x \in R,$$

which in turn implies, by (2.14), that

$$(x - x^0)' \nabla_x \theta(x^0) \geqq 0 \quad \text{for} \quad x \in R. \tag{2.23}$$

But by the pseudo-convexity of $\theta(x)$ on R, (2.23) implies that

$$\theta(x) \geqq \theta(x^0) \quad \text{for} \quad x \in R.$$

For the case when the set I is empty, the above proof is modified by deleting (2.19), (2.22) and references thereto. For the case when $I = \{1, 2, \cdots, n\}$ the above proof is modified by deleting that part of the proof *between* (2.19) and (2.22) and references thereto.

It should be noted here that the above theorem is indeed a generalization of Arrow and Enthoven's result [1, Theorem 1]. Every case covered there is covered by the above theorem, but not conversely. An example of a case not covered by Arrow and Enthoven is the following one:

$$\min_{x \in E^1} \{-e^{-x^2} \mid -x \leqq 0\}.$$

Another application of pseudo-convex functions may be found in duality theory. Consider the primal problem

$$\text{(PP)} \qquad \min_{x \in E^m} \{\theta(x) \mid g(x) \leqq 0\},$$

where $\theta(x)$ is a scalar function on E^m and $g(x)$ is an $n \times 1$ vector function on E^m. For the above problem Wolfe [10] has defined the dual problem as

$$\text{(DP)} \qquad \max_{x \in E^m, y \in E^n} \{\psi(x, y) \mid \nabla_x \psi(x, y) = 0, y \geqq 0\},$$

where

$$\psi(x, y) \equiv \theta(x) + y'g(x).$$

Under appropriate conditions Wolfe has shown [10, Theorem 2] that if x^0 solves (PP), then x^0 and some y^0 solve (DP). Conversely, under somewhat stronger conditions, Huard [4, Theorem 2] showed that if (x^0, y^0) solves (DP), then x^0 solves (PP). Both Wolfe and Huard required, among other things, that $\theta(x)$ and the components of $g(x)$ be convex. We will now show that Huard's theorem can be extended to the case where $\theta(x)$ is pseudo-convex and the components of $g(x)$ are quasi-convex, and that Wolfe's theorem is not amenable to such an extension.

THEOREM 2. (*Strict converse duality theorem*)[3]. *Let $\theta(x)$ be a pseudo-convex function on E^m and let the components of $g(x)$ be differentiable quasi-convex functions on E^m.*

(a) *If (x^0, y^0) solves* (DP) *and $\psi(x, y^0)$ is twice continuously differentiable with respect to x in a neighborhood of x^0, and if the Hessian of $\psi(x, y^0)$ with respect to x is nonzero at x^0, then x^0 solves* PP.

(b) *Let x^0 solve* (PP) *and let $g(x) \leqq 0$ satisfy the Kuhn-Tucker constraint qualification* [7]. *It does not necessarily follow that x^0 and some y^0 solve* (DP).

Proof. (a) The assumption that the Hessian of $\psi(x, y^0)$ with respect to x is nonzero at x^0 insures the validity of the following Kuhn-Tucker neces-

[3] For the difference between "duality" and "strict duality," the reader is referred to [9].

sary conditions for some $v^0 \in E^m$:

$$\nabla_x\psi(x^0, y^0) + \nabla_x v^{0\prime}\nabla_x\psi(x^0, y^0) = 0,$$
$$\nabla_y\psi(x^0, y^0) + \nabla_y v^{0\prime}\nabla_x\psi(x^0, y^0) \leqq 0,$$
$$y^{0\prime}\nabla_y\psi(x^0, y^0) + y^{0\prime}\nabla_y v^{0\prime}\nabla_x\psi(x^0, y^0) = 0,$$
$$y^0 \geqq 0,$$

$$\nabla_x\psi(x^0, y^0) = 0.$$

The first and last equations above, together with the assumption that the Hessian of $\psi(x, y^0)$ is nonzero at x^0, imply that $v^0 = 0$. Hence the above necessary conditions become:

$$\nabla_x\psi(x^0, y^0) = 0,$$
$$\nabla_y\psi(x^0, y^0) = g(x^0) \leqq 0,$$
$$y^{0\prime}\nabla_y\psi(x^0, y^0) = y^{0\prime}g(x^0) = 0,$$
$$y^0 \geqq 0.$$

But from Theorem 1, with $C = E^m$, these conditions are sufficient for x^0 to be a solution of (PP).

(b) This part of the theorem will be established by means of the following counter-example:

$$\text{(PP1)} \qquad \min_{x \in E^1} \{-e^{-x^2} \mid -x + 1 \leqq 0\},$$

$$\text{(DP1)} \qquad \max_{x \in E^1, y \in E^1} \{-e^{-x^2} - yx + y \mid 2xe^{-x^2} - y = 0, y \geqq 0\}.$$

The solution of (PP1) is obviously $x^0 = 1$, whereas (DP1) has no maximum solution but has a zero supremum.

Finally, we give an application of pseudo-concavity outside the realm of of mathematical programming. In particular, we extend a stability criterion for equilibrium points of ordinary differential equations [8, Theorem 1].

THEOREM 3. (*Stability criterion*). *Let*

$$\dot{x} = f(t, x)$$

be a system of ordinary differential equations, where x and f are m-dimensional vectors and $0 \leqq t < \infty$. Let $f(t, x)$ be continuous in the (x, t) space and let $f(t, 0) = 0$ for $0 \leqq t < \infty$, so that $x = 0$ is an equilibrium point. If $x'f(t, x)$ is a pseudo-concave function of x on E^m for $0 \leqq t < \infty$, then $x = 0$ is a stable equilibrium point.

Proof. Consider the Lyapunov function

$$V(x, t) = \tfrac{1}{2}x'x,$$

which is obviously positive definite. It follows that for $0 \leqq t < \infty$,

$$\dot{V} = x'\dot{x} = x'f(t, x) \leqq 0,$$

where the last inequality follows from the pseudo-concavity of $x'f(t, x)$ in x and the fact that $f(t, 0) = 0$. Hence by Lyapunov's stability theorem [6], $x = 0$ is a stable equilibrium point.

It should be noted that the above proof would not go through had we merely required that $x'f(t, x)$ be quasi-concave instead of pseudo-concave.

3. Remarks on pseudo-convex functions. Properties 1 and 2 and the fact that every differentiable strictly quasi-convex function is also quasi-convex [5] establish a hierarchy among differentiable functions that is depicted in Fig. 1. In other words, if we let S_1, S_2, S_3, and S_4 represent the sets of all differentiable functions defined on a *convex* set C in E^m that are, respectively, convex, pseudo-convex, strictly quasi-convex, and quasi-convex, then

$$S_1 \subset S_2 \subset S_3 \subset S_4 .$$

Functions belonging to S_1, S_2, or S_3 share the property that a local minimum is a global minimum. Functions belonging to S_4 do not necessarily have this property. The Kuhn-Tucker differential conditions are sufficient for optimality, (see (2.18)), provided that $g_i(x)$, $i = 1, \cdots, n$ belong to S_4 and $\theta(x)$ belongs to S_1 or S_2, but not if $\theta(x)$ belongs to S_3 or S_4. It seems that the pseudo-convexity of $\theta(x)$ and the quasi-convexity of $g_i(x)$ are the weakest conditions that can be imposed so that relations (2.14) to (2.17) are sufficient for optimality.

There does not seem to be a simple extension of the concept of pseudo-convexity to nondifferentiable functions. This may be due to the fact that pseudo-convexity eliminates inflection points, and such points are easily described by derivatives, but not otherwise.

Finally, it should be remarked that the convexity of the set C is inherent in the definition of quasi-convexity. In contrast, the convexity of C is not needed in the definition of pseudo-convexity. Thus, without the convexity of C, we may have a pseudo-convex function that is not quasi-convex. For example, over the nonconvex set

$$C = \{x \mid x \in E^1, x \neq 0\},$$

the function

$$\theta(x) = \begin{cases} x & \text{for } x < 0, \\ x + 1 & \text{for } x > 0, \end{cases}$$

is pseudo-convex but obviously not quasi-convex, since C is nonconvex.

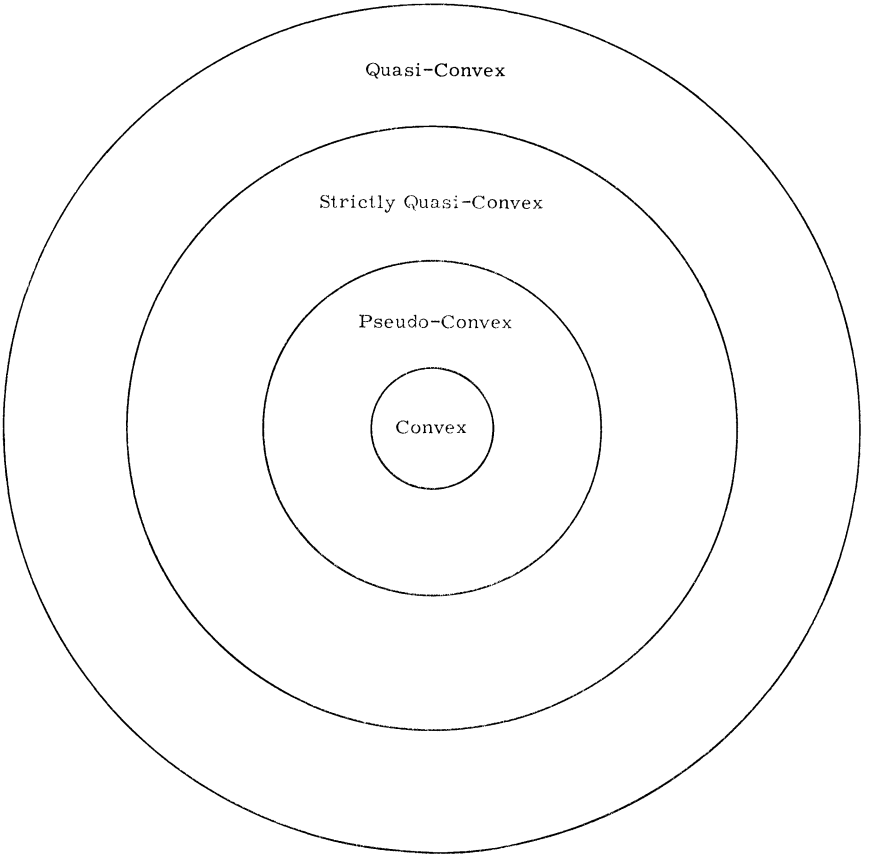

FIG. 1

4. Acknowledgement. I am indebted to my colleagues, S. Karamardian and J. Ponstein, for stimulating discussions on this paper.

REFERENCES

[1] K. J. ARROW AND A. C. ENTHOVEN, *Quasi-concave programming*, Econometrica, 29 (1961), pp. 779–800.

[2] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[3] M. A. HANSON, *Bounds for functionally convex optimal control problems*, J. Math. Anal. Appl., 8 (1964), pp. 84–89.

[4] P. HUARD, *Dual programs*, IBM J. Res. Develop., 6 (1962), pp. 137–139.

[5] S. KARAMARDIAN, *Duality in mathematical programming*, forthcoming.

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

[7] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of the

Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[8] O. L. MANGASARIAN, *Stability criteria for nonlinear ordinary differential equations*, this Journal, 1 (1963), pp. 311–318.

[9] O. L. MANGASARIAN AND J. PONSTEIN, *Minmax and duality in nonlinear programming*, International Symposium on Mathematical Programming, London, 1964; J. Math. Anal. Appl., to appear.

[10] P. WOLFE, *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19 (1961), pp. 239–244.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

OPTIMAL CONTROL, INEQUALITY STATE CONSTRAINTS, AND THE GENERALIZED NEWTON-RAPHSON ALGORITHM*

ROBERT MCGILL†

Introduction. The generalized Newton-Raphson algorithm [1] has been developed as a tool for efficiently extracting solutions to nonlinear problems of optimal control by means of the modern high speed digital computer. This paper presents the extension of the algorithm to the important class of problems characterized by inequality constraints on the state space. The efficacy of the procedure is demonstrated by a numerical example. The problem solved is nonlinear, but admits of a closed form solution. This makes possible a direct comparison between the solution obtained by the algorithm and the analytic solution.

Problem statement. The general problem we are interested in here consists of finding, among all admissible controls $u = u(t)$ which transfer the state point from the position x_0 to the position x_f, the one for which

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

takes on the least possible value. The vector $x = (x^1, \cdots, x^n)$ is required to satisfy the vector differential equation

$$\dot{x} = f(x, u, t), \qquad \dot{x} = \frac{dx}{dt},$$

where $f = (f^1, \cdots, f^n)$, and f^0 are assumed to have the smoothness properties required for the application of the maximum principle [2]. The class of admissible controls will be taken as the collection of functions $u(t)$ differentiable on $[t_0, t_f]$. In addition the trajectory $x(t)$ corresponding to the control $u(t)$ is constrained to satisfy p inequalities of the following form:

$$G^1(x) \leqq 0, \cdots, G^p(x) \leqq 0,$$

where $G^i(x)$, $i = 1, \cdots, p$, are continuously differentiable on the state space X. In the following we shall assume the existence of a solution to the constrained optimal control problem.

* Received by the editors March 4, 1965, and in revised form April 20, 1965.

† Grumman Aircraft Engineering Corporation, Bethpage, New York. This work was partially supported by the Air Force Office of Scientific Research of the Office of Aerospace Research under Contract No. AF49(638)-1207.

Extension of the generalized Newton-Raphson algorithm. The solution of the constrained problem is reduced, by the introduction of an additional state variable x^{n+1}, to the solution of a sequence of *unconstrained* problems. Each subproblem is then solved by the generalized Newton-Raphson algorithm and under appropriate conditions the sequence of solutions converges to the solution of the constrained problem.

We introduce the new state variable x^{n+1} by means of the differential equation and boundary conditions

$$\dot{x}^{n+1} = f^{n+1} = [G^1(x)]^2 H^1(G^1) + \cdots + [G^p(x)]^2 H^p(G^p),$$

$$x^{n+1}(t_0) = 0, \quad x^{n+1}(t_f) = \epsilon_s > 0,$$

where the functions $H^i(G^i)$ are given by

$$H^i(G^i) = \begin{cases} 0 \text{ for } G^i \leqq 0, \\ K \text{ for } G^i > 0, \end{cases} \quad i = 1, \cdots, p, \; K = \text{const.} > 0.$$

We note that $f^{n+1}(x)$ is continuously differentiable with respect to x. The maximum principle is therefore applicable and implies

$$\psi^{n+1} = 0 \Rightarrow \psi^{n+1} = \text{const.},$$

as x^{n+1} is not contained in any of the f^i, $i = 0, \cdots, n+1$. The vector $\psi = (\psi^0, \cdots, \psi^{n+1})$ is the co-state or adjoint vector whose existence as a nonzero, absolutely continuous function is asserted by the maximum principle (see [2, p. 19]).

Suppose $\{\epsilon_s\}$ to be a sequence of positive numbers that converges monotonically to zero. The quantity

$$\epsilon_s = \int_{t_0}^{t_f} \{[G^1(x)]^2 H^1(G^1) + \cdots + [G^p(x)]^2 H^p(G^p)\}\, dt = x^{n+1}(t_f)$$

will be regarded as a measure of penetration of the constraints. To each ϵ_s there corresponds a constant, $\psi_s{}^{n+1}$, through the solution of the corresponding unconstrained optimal control problem. As $s \to \infty$, $\epsilon_s \to 0$ and we may expect the corresponding sequence of solutions to approach the solution to the original *constrained* problem.

This approach is directly related to the "penalty function" technique used by Kelley et al. [3], [4], [5] with the gradient method, and is a generalization of a result due to Courant [6], [7]. The question of convergence and existence of solutions is discussed by Butler and Martin [8].

It is convenient to regard the constant $\psi_s{}^{n+1}$ as a parameter of the boundary value problem associated with the optimal control problem through the application of the necessary conditions, i.e., the conditions implied by

the maximum principle. That is, ψ_s^{n+1} is considered as fixed and the boundary value problem is solved by the generalized Newton-Raphson procedure [1] with $x^{n+1}(t_f) = \epsilon_s$ unspecified. This yields a value for $x^{n+1}(t_f) = \epsilon_s$ corresponding to the constant ψ_s^{n+1}. The parameter ψ_s^{n+1} is now changed automatically, according to the distance between ϵ_s and zero, and the new boundary value problem solved using the previous solution as starting functions for the new iteration.

Numerical example. The example is a modification of the classical brachistochrone problem. We consider a particle, falling for a specified time $t_f - t_0$, under the influence of a constant gravitational acceleration g. The particle has a given initial speed x_0^3 and we wish to find the path that maximizes the final value of the horizontal coordinate, $x^1(t_f)$, with the final values of the vertical coordinate, $x^2(t_f)$, and the speed, $x^3(t_f)$, unspecified. The path is constrained by a given fixed line in the x^1, x^2-plane, the line being chosen such that the unconstrained solution intersects it (Fig. 1).

This particular example was chosen because it has been solved by direct methods [9] (steepest ascent) and has been considered analytically, with regard to necessary conditions, in an informative paper by Bryson et al. [10]. Additionally, it has a known closed form solution [9], [10] making possible a comparison with the analytic solution as well as between the two numerical techniques.

The equations of state including the added state coordinate x^4, which embodies the constraint, are:

$$\dot{x}^1 = f^1 = x^3 \cos u, \quad \dot{x}^2 = f^2 = x^3 \sin u,$$

$$\dot{x}^3 = f^3 = g \sin u, \quad \dot{x}^4 = f^4 = [G(x^1, x^2)]^2 H(G),$$

$$G(x^1, x^2) = x^2 - (ax^1 + b), \quad H(G) = \begin{cases} 0 \text{ for } G \leqq 0, \\ K \text{ for } G > 0. \end{cases}$$

The coordinate to be minimized at the final time is

$$x^0(t_f) = \int_{t_0}^{t_f} f^0 \, dt = \int_{t_0}^{t_f} - x^3 \cos u \, dt = -x^1(t_f).$$

The control variable u is the slope of the path, and boundary conditions are:

$$x^1(t_0) = x_0^1, \quad x^2(t_0) = x_0^2, \quad x^3(t_0) = x_0^3, \quad x^4(t_0) = 0,$$

$$x^1(t_f) \sim \text{payoff}, \quad x^2(t_f) \sim \text{free}, \quad x^3(t_f) \sim \text{free}, \quad x^4(t_f) \sim \epsilon_s .$$

The data for the problem were taken from Dreyfus (see [9, p. 81]) as

follows:

$$x^1(t_0) = 0 \text{ ft}, \quad x^2(t_0) = 0 \text{ ft}, \quad x^3(t_0) = 1 \text{ ft/sec},$$
$$t_0 = 0 \text{ sec}, \quad t_f = .7425 \text{ sec}, \quad g = 32.20 \text{ ft/sec}^2,$$
$$a = .5, \quad b = 1.$$

These data were normalized to obtain:

$$x^1(t_0) = 0, \quad x^2(t_0) = 0, \quad x^3(t_0) = .07195,$$
$$t_0 = 0, \quad t_f = 1.720, \quad g = 1,$$
$$L = 6 \text{ ft} \sim \text{unit of length}, \quad \tau = .4317 \text{ sec} \sim \text{unit of time}, \quad K = 10,$$
$$a = .5, \quad b = .1667.$$

A straightforward application of the maximum principle yields the following nonlinear boundary value problem:

$$\begin{aligned}
\dot{x}^1 &= f^1 = (x^3)^2\psi^1[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2},\\
\dot{x}^2 &= f^2 = (x^3)(\psi^2x^3 + \psi^3g)[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2},\\
\dot{x}^3 &= f^3 = g(\psi^2x^3 + \psi^3g)[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2},\\
\dot{x}^4 &= f^4 = [G(x^1, x^2)]^2 \cdot H(G),\\
(1)\quad \dot{\psi}^1 &= g^1 = \psi^4(2aGH),\\
\dot{\psi}^2 &= g^2 = -\psi^4(2GH),\\
\dot{\psi}^3 &= g^3 = -(\psi^1)^2x^3[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2}\\
&\qquad - \psi^2(\psi^2x^3 + \psi^3g)[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2},\\
\dot{\psi}^4 &= g^4 = 0,
\end{aligned}$$

with boundary conditions:

$$\begin{aligned}
(2)\quad & x^1(t_0) = 0, \quad x^2(t_0) = 0, \quad x^3(t_0) = .07195, \quad x^4(t_0) = 0,\\
& \psi^2(t_f) = 0, \quad \psi^3(t_f) = 0, \quad \psi^1(t_0) = 1, \quad \psi^4(t_0) = \psi_s^4.
\end{aligned}$$

The adjoint variable $\psi^1(t_0)$ has been put equal to one, which scales the adjoint vector (see [2, p. 22]), and $\psi^2(t_f)$ and $\psi^3(t_f)$ are zero from the transversality condition (see [2, p. 50]). The control variable $u(t)$ has been eliminated by the $\max_u H(\psi, x, u, t)$ condition that implies the following two relations:

$$\sin u = (\psi^2x^3 + \psi^3g)[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2},$$
$$\cos u = (\psi^1x^3)[(\psi^1x^3)^2 + (\psi^2x^3 + \psi^3g)^2]^{-1/2}.$$

System (1) may be written as

$$\dot{y} = h(y, t),$$

where $y = (y^1, \cdots, y^8)$, $h = (h^1, \cdots, h^8)$, and $h^i = h^i(y^1, \cdots, y^8, t)$, $i = 1, \cdots, 8$.

The generalized Newton-Raphson algorithm proceeds by solving the following sequence of *linear* boundary value problems [1]:

$$\dot{y}_{k+1} = J(y_k, t)[y_{k+1} - y_k] + h(y_k, t), \quad k = 1, 2, \cdots, \tag{3}$$

where $J(y, t)$ is the Jacobian matrix of partial derivatives of the h^i with respect to the y^j, $i, j = 1, \cdots, 8$.

The adjoint variable ψ_s^4 is fixed at a nominal value ψ_0^4, e.g., $\psi_0^4 = 0$, a starting vector $y_0(t)$ is chosen, and the sequence of linear boundary value problems (3) is solved numerically by the procedure described in detail in [1], with boundary conditions (2).

The iteration continues until a prescribed metric $\bar{\rho}$ becomes $\leqq \beta$, where

$$\bar{\rho} = \sum_{i=1}^{8} \max_{t \in [t_0, t_f]} | y_{k+1}^i(t) - y_k^i(t) |$$

and β is a small positive constant. This results in a nonnegative value for $x^4(t_f) = \epsilon_s$ which is a measure of the penetration of the constraint.

The parameter ψ_s^4 is then adjusted automatically, by a scalar application of the Newton-Raphson algorithm, as follows:

$$\psi_{s+1}^4 = \psi_s^4 + [x_s^4(t_f) - x_{s-1}^4(t_f)]^{-1}[\psi_s^4 - \psi_{s-1}^4]x_s^4(t_f),$$

where the rate of change of ψ^4 with respect to $x^4(t_f)$ has been obtained by a finite difference approximation as shown. Since this recursion formula contains three indices $(s + 1, s, s - 1)$, the value ψ_1^4 was arbitrarily determined by $\psi_1^4 = \psi_0^4 + Kx_0^4(t_f)$, $K = 10$, after which the recursion formula was used. The iteration on y_k is now continued until $\bar{\rho}$ is again $\leqq \beta$. The overall process continues until $\rho \leqq \beta$, where

$$\rho = \bar{\rho} + x_k^4(t_f) = \bar{\rho} + \epsilon_s .$$

The corresponding iterate is accepted as the solution and a final check is made by integrating the *nonlinear* system (1) with a complete set of initial conditions obtained from the final iterate.

The following simple linear starting functions $y_0(t)$ were chosen:

$$y_0^1(t) = x_0^1(t) \equiv 0, \qquad y_0^5(t) = \psi_0^1(t) \equiv 1,$$

$$y_0^2(t) = x_0^2(t) \equiv 0, \qquad y_0^6(t) = \psi_0^2(t) \equiv 0,$$

$$y_0^3(t) = x_0^3(t) = x_0^3 + (x_f^3 - x_0^3)(t_f)^{-1}t,$$

$$y_0^7(t) = \psi_0^3(t) = (\tan 56°)y_0^3(t),$$

$$y_0^4(t) = x_0^4(t) \equiv 0, \qquad y_0^8(t) = \psi_0^4(t) \equiv 0,$$

where the number x_f^3, representing an estimate of the speed at the final time, was put equal to 1, and the starting function $\psi_0^3(t)$ corresponds to a starting control function $u_0(t)$, which is constant at an arbitrary value of 56°. After four iterations the sequence $\{y_k\}$ converged ($\bar{\rho} < 10^{-4}$) to the unconstrained solution (see Fig. 1). This solution corresponds to a value of the parameter $\psi_0^4 = 0$. A new value ψ_1^4 was obtained automatically and the iteration on y_k continued until the sequence $\{y_k\}$ again converged in the function space metric $\bar{\rho}(y_{k+1}, y_k)$. Following a total of 47 iterations, with 13 shifts of the parameter ψ_s^4, the sequence converged to the solution of the *constrained* problem with $\rho < 10^{-4}$. The results are exhibited in Fig. 1 where the analytic solution for the contrained problem is represented by the solid curve. The unconstrained solution and some intermediate curves, corresponding to intermediate values of ψ_s^4, are also indicated in Fig. 1.

We observe that, within the accuracy of our plot, the solution obtained by the extension of the Newton-Raphson algorithm, and the analytic solution, are identical. The program is entirely automatic and required less than 2 minutes of machine time (IBM 7094). The behavior of the measure of penetration ϵ_s with the "penalty" parameter ψ_s^4 is shown in Fig. 2. The optimal trajectory yielded a value for the maximum lateral range $x^1(t_f)$ of 1, both analytically and computationally.

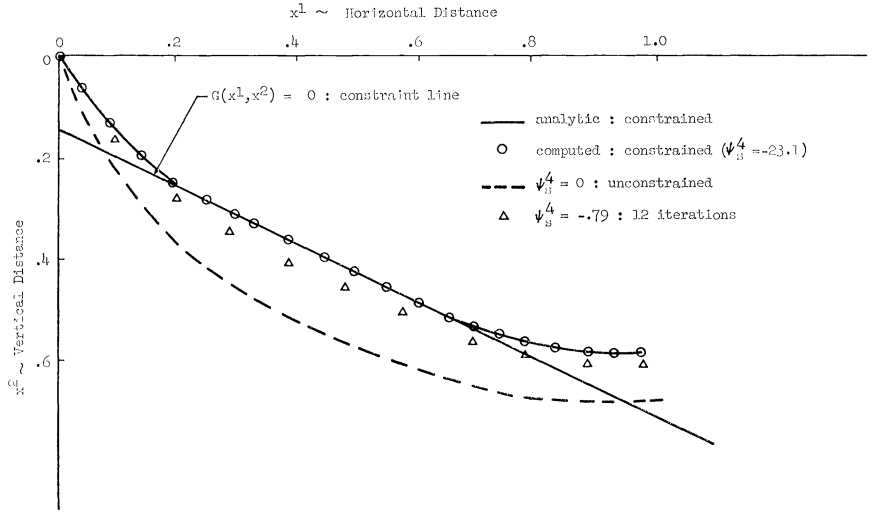

FIG. 1. State space trajectories

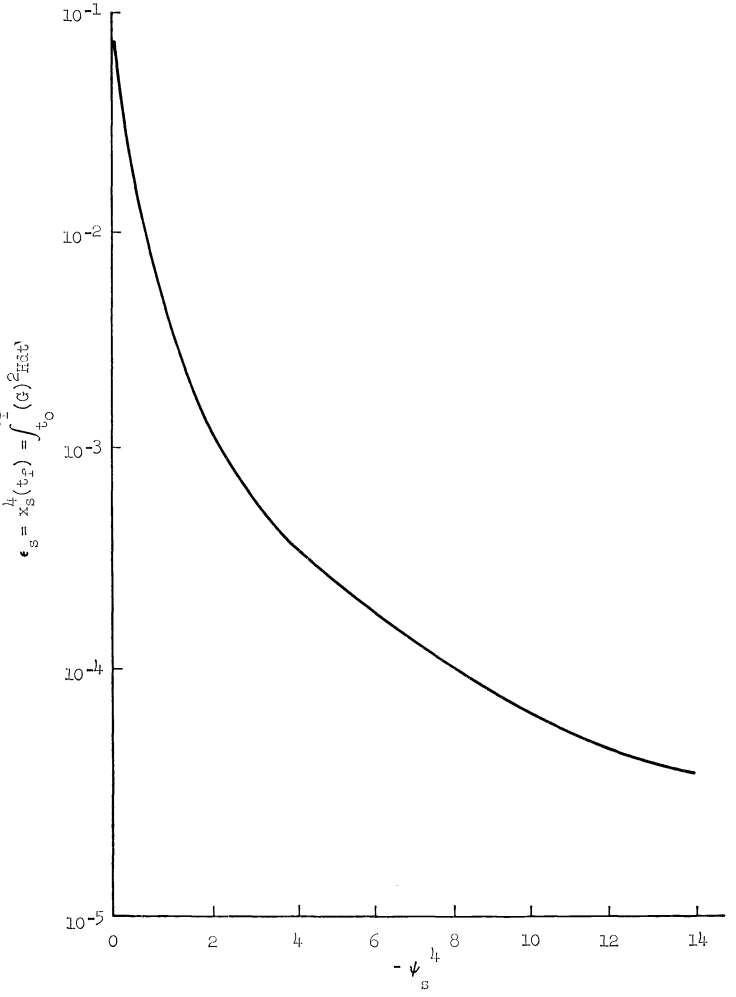

FIG. 2. Constraint penetration ϵ_s versus constraint parameter ψ_s^4

Conclusions. The example presented in this paper suggests that the extension of the generalized Newton-Raphson algorithm, developed herein, may be useful as an efficient method for obtaining solutions to the class of nonlinear optimal control problems with inequalities on the state space. The approach is basically simple, automatic, and appears to be fairly general since it does not require assumptions as to the number or location of junction points, i.e., points of contact between the trajectory and the constraint boundary.

The method does, however, require starting functions. For the example

considered above, simple uninspired functions sufficed: linear functions joining the boundary conditions, or constants. The general question concerning the size of the region of convergence remains open and will likely require further computational experience to clarify.

Acknowledgment. The writer is indebted to Henry J. Kelley, Michael Falco, and Edward J. Beltrami for helpful discussions and to Gerald E. Taylor for his skill in adapting the algorithm to the computer.

REFERENCES

[1] R. McGill and P. Kenneth, *Solution of variational problems by means of a generalized Newton-Raphson operator*, AIAA J., 2 (1964), pp. 1761–1766.

[2] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962, Chap. 1, pp. 9–19.

[3] H. J. Kelley, *Method of gradients*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, Chap. 6, p. 230.

[4] H. J. Kelley, M. Falco, and D. J. Ball, *Air vehicle trajectory optimization*, presented at the Symposium on Multivariable System Theory, Fall Meeting of SIAM, Cambridge, Massachusetts, 1962.

[5] M. Falco, *Supersonic transport climb path optimization including a constraint on sonic boom intensity*, AIAA J., 1 (1963), pp. 2859–2862.

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

[7] ———, *Calculus of variations and supplementary notes and exercises*, mimeographed notes, supplementary notes by M. Kruskal and H. Rubin, revised and amended by J. Moser, New York University, 1956–1957.

[8] T. Butler and A. V. Martin, *On a method of Courant for minimizing functionals*, J. Mathematical Phys., 41 (1962), pp. 291–299.

[9] S. Dreyfus, *Variational problems with state variable inequality constraints*, RAND Corporation Paper P-2605, 1962, pp. 72–85.

[10] A. E. Bryson, W. F. Denham, and S. Dreyfus, *Optimal programming problems with inequality constraints I: Necessary conditions for extremal solutions*, AIAA J., 1 (1963), pp. 2544–2550.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

MINIMIZATION OF FUNCTIONALS WITH EQUALITY CONSTRAINTS*

E. K. BLUM†

1. Introduction. A very general class of problems in the calculus of variations can be formulated as problems of Mayer—or equivalently, as problems of Bolza. As formulated by Bliss [4], the problem of Mayer may be stated as follows:

Let $y = y(t) = (y_1(t), \cdots, y_m(t))$ be a vector function of t in the real interval $t_1 \leqq t \leqq t_2$. y is assumed to be a member of some set of "admissible" functions, S. For example, Bliss takes S to consist of all functions which are piecewise continuously differentiable in $[t_1, t_2]$ and have values $y(t)$ in some open region of R^m, m-dimensional Euclidean space. Further, let

$$\Phi_i(t, y, y') = 0, \qquad i = 1, \cdots, n < m, \tag{1}$$

be a set of differential equations, where the Φ_i are assumed to have continuous third-order derivatives in some suitable $(2m + 1)$-dimensional open region R_1, and the matrix $(\partial\Phi_i/\partial y_j')$ is to have rank n in R_1. Let

$$\psi_j(t_1, y(t_1), t_2, y(t_2)) = 0, \qquad j = 1, \cdots, p \leqq 2m + 2, \tag{2}$$

be endpoint constraints, where the ψ_j have continuous third-order partial derivatives in some $(2m + 2)$-dimensional open region R_2 in which the matrix

$$\left(\frac{\partial\psi_j}{\partial t_1}, \frac{\partial\psi_j}{\partial y_k(t_1)}, \frac{\partial\psi_j}{\partial t_2}, \frac{\partial\psi_j}{\partial y_k(t_2)}\right)$$

has rank p. Finally, let

$$J = J(t_1, y(t_1), t_2, y(t_2)) \tag{3}$$

have continuous third-order partials in the region R_2. The problem of Mayer is to find $y \in S$ which minimizes J while satisfying (1) and (2). (If J is to be maximized, we minimize $-J$.)

A wide class of problems in the theory of optimal control may be formulated as variational problems of the following type.

Let $x = x(t) = (x_1(t), \cdots, x_n(t))$ and $u = u(t) = (u_1(t), \cdots, u_q(t))$ be functions of t on the interval $t_1 \leqq t \leqq t_2$, where x is the "state" function and u is the "control" function. The control is assumed to be in some

* Received by the editors October 20, 1964, and in revised form March 16, 1965.

† Department of Mathematics, Wesleyan University, Middletown, Connecticut, and United Aircraft Research.

specified set of admissible controls. For example, u is usually assumed to be piecewise continuous with values in some region U of R^q. We shall take U to be an open set. Let

$$\frac{dx}{dt} = f(x, u) \tag{1'}$$

be a system of n differential equations, where $f = (f_1, \cdots, f_n)$ and the f_i, $1 \leqq i \leqq n$, are assumed to have continuous partials $\partial f_i/\partial x_j$ and $\partial f_i/\partial u_k$ in some prescribed open set, $\Gamma \subset R^n \times R^q$. We assume that the control region U is contained in the projection of Γ on R^q. Let

$$\psi_j(t_1, x(t_1), t_2, x(t_2)) = 0, \qquad 1 \leqq j \leqq p \leqq 2n + 2, \tag{2'}$$

be endpoint constraints on the state function, where the ψ_j have continuous first-order partial derivatives in some open set R_2, and the matrix

$$\left(\frac{\partial \psi_j}{\partial t_1}, \frac{\partial \psi_j}{\partial x(t_1)}, \frac{\partial \psi_j}{\partial t_2}, \frac{\partial \psi_j}{\partial x(t_2)}\right)$$

has rank p. Finally, let

$$J = J(t_1, x(t_1), t_2, x(t_2)) \tag{3'}$$

be a function of end-values. We assume J has continuous partials in R_2. For each admissible control u, a state function x which satisfies (1') is called a "trajectory" corresponding to u. An "optimal" control is one which has a corresponding trajectory which satisfies the constraints (2') and minimizes J. The "optimal control problem" is to find such an optimal control u and its corresponding trajectory x. The pair (x, u) is an "optimal solution" of the problem.

The optimal control problem becomes a problem of Mayer if we let $y_i = x_i$, $i = 1, \cdots, n$, and $y_{n+j} = u_j$, $1 \leqq j \leqq q$. Conversely, the problem of Mayer is transformed into the optimal control problem by solving (1) for n of the y_i', say $y_1', \cdots, y_n'$, and setting $x_i = y_i$, $1 \leqq i \leqq m$, and $u_j = y_{n+j}'$, $1 \leqq j \leqq m - n$. Further, we adjoin to the n differential equations obtained by solving (1) for y_i' the additional equations $x_{j+n}' = u_j$, $1 \leqq j \leqq m - n$. This yields a system of m differential equations for the m state variables as in (1'). Similarly, (2) and (3) become (2') and (3'), respectively.

The equivalence of the problem of Mayer and the optimal control problem is well-known. Therefore, a necessary condition that (x, u) be a solution of the optimal control problem as formulated above is given by the multiplier rule [4]. The derivation of the multiplier rule in [4] and other sources is based on classical methods of the calculus of variations. It is our purpose in this paper to present an alternate derivation using the methods of functional analysis, insofar as possible. Although the main ideas have

been known for some time (compare, for example, Bolza's results in [7] with our Theorem 4), they do not seem to have been applied as they are here. We believe that the techniques of our proof, being geometrical, are more readily grasped. Furthermore, they are based on natural generalizations of the ideas used in establishing the Lagrange multiplier rule of ordinary calculus. In fact, Theorem 4 below and its proof apply to minimization problems with equality constraints in both the ordinary calculus and in the calculus of variations. It is also the basis for a new derivation of the method of steepest descent in problems with equality constraints. In short, we offer a unifying viewpoint for a large class of problems of minimization (or maximization) with equality constraints.

2. Analysis in abstract spaces. In this section, we assemble some general results of abstract analysis based on ideas going back to Fréchet [10] and Gâteaux [11]. Essentially, we consider the generalization of the differential calculus to normed linear spaces and, conversely, the application of the methods of functional analysis to classical analysis. Similar approaches are taken in [2], [3], [9], [12], [13], [14], [15], [16], [17] and [18], to name but a few recent works. However, the existing literature does not appear to contain results in a form which can be applied directly to the optimal control problem formulated in §1. Results which are close to our Theorems 1–4 below are to be found in [17]. In one respect, the work in [17] is more general, since it deals with functionals on a normed linear space, whereas we deal (ultimately) with a prehilbert space. However, in another respect, important for the intended application, our result is more general, since we require differentials of functionals to exist only in finitely open sets rather than in open sets. Furthermore, our proofs are simpler. For example, they are based on the classical implicit function theorem for real functions of several real variables rather than on the implicit function theorem in abstract spaces. Finally, Theorems 1–4 appear to have just the right amount of generality for the application to variational and control theory.

Throughout this section, E and E_1 will denote normed spaces (usually over the reals) and f will denote an arbitrary mapping from a subset of E to E_1. We denote this, as usual, by "$f: E \to E_1$", which leaves the domain of f as some unspecified subset of E.

DEFINITION 1. Let $f: E \to E_1$ be an arbitrary function defined in a neighborhood of $u \in E$. If there exists a linear continuous operator, $f'(u): E \to E_1$, such that for all h in some neighborhood of 0,

$$f(u + h) - f(u) = f'(u)h + \epsilon(u, h),$$

and

$$\frac{\| \epsilon(u, h) \|}{\| h \|} \to 0 \quad \text{as} \quad \| h \| \to 0,$$

then $f'(u)h$ is the *strong differential of f at u with increment h*, and $f'(u)$ is the *strong derivative* of f at u.

Remark 1. $f'(u)h$ is an element of E_1, the result of the linear operator $f'(u)$ applied to h. $f'(u)h$ is defined for all $h \in E$. For our applications, we must consider a more general notion of differential. First, we introduce the concept of a set "finitely open at u". (Compare [16].)

DEFINITION 2. A subset $D \subset E$ is *finitely open at u* if for any $h_1, \cdots, h_n \in E$, $n \geqq 1$, there exists an open set $T \subset R^n$ such that T contains the origin and $u + \sum_{i=1}^n t_i h_i \in D$ for all $(t_1, \cdots, t_n) \in T$. (Equivalently, there exists $\delta = \delta(h_1, \cdots, h_n) > 0$ such that $u + \sum_{i=1}^n t_i h_i \in D$ whenever $\sum_{i=1}^n |t_i| < \delta$.)

DEFINITION 3. Let $f: E \to E_1$ be defined on some finitely open set at u. If

$$\lim_{|t| \to 0} \frac{f(u + th) - f(u)}{t} \equiv \delta f(u; h)$$

exists for all $h \in E$, it is called the *weak differential* of f at u with increment h.

Remark 2. We also have, by the usual definition of the derivative of a vector function of a scalar t,

$$\frac{d}{dt} f(u + th) \,|_{t=0} = \delta f(u; h).$$

Remark 3. $\delta f(u; h)$ is homogeneous in h. This follows easily from the definition, since

$$\delta f(u; sh) = \lim_{t \to 0} \frac{f(u + tsh) - f(u)}{t} = s \lim_{t \to 0} \frac{f(u + tsh) - f(u)}{ts} = s\delta f(u; h).$$

Remark 4. In the case when f is a real functional such that $\delta f(y; h)$ exists and is continuous in y for y in a set finitely open at u, we obtain additivity of $\delta f(u; h)$ as follows. For any $h_1, h_2 \in E$, choose scalars t_1 and t_2 and define

$$g(t_1, t_2) = f(u + t_1 h_1 + t_2 h_2).$$

Observing that

$$\frac{\partial g}{\partial t_1} = \delta f(u + t_1 h_1 + t_2 h_2 ; h_1)$$

and

$$\frac{\partial g}{\partial t_2} = \delta f(u + t_1 h_1 + t_2 h_2 ; h_2),$$

we obtain

$$f(u + t_1 h_1 + t_2 h_2) - f(u) = t_1 \delta f(u; h_1) + t_2 \delta f(u; h_2) + \epsilon,$$

where $\epsilon/(t_1^2 + t_2^2)^{1/2} \to 0$ as $(t_1, t_2) \to (0, 0)$ along a ray through the origin. Taking $t_1 = t_2$, this yields

$$\lim_{t \to 0} \frac{f(u + t(h_1 + h_2)) - f(u)}{t} = \delta f(u; h_1) + \delta f(u; h_2),$$

since $\epsilon/t \to 0$ as $t \to 0$. The left member is $\delta f(u; h_1 + h_2)$, which establishes the additivity.

DEFINITION 5. Let f be a functional on $D \subset E$, where D is finitely open at u. Let $y \in D$, where D is finitely open at y also. If $f(y + \sum_1^n t_i h_i)$ is continuous in $(t_1, \cdots, t_n)$ at the origin in R^n for all $h_1, \cdots, h_n \in E$, then f is *finitely continuous* at y.

DEFINITION 6. Let J and g be functionals on a domain $D \subset E$. The set $C(g) = \{y \mid g(y) = 0\}$ is called a *constraint*. If there is a neighborhood N_u of $u \in C(g)$ such that $J(y) \geqq J(u)$ for all $y \in C(g) \cap N_u \cap D$, then u is a *relative minimum* of J on the constraint $C(g)$. More generally, if S is any subset of E, u is a relative minimum of J on S if $J(y) \geqq J(u)$ for all $y \in S \cap N_u \cap D$.

A *prehilbert space* is a linear space E with an inner product (u, v) defined for all $u, v \in E$. The space is normed by taking $\| u \|^2 = (u, u)$. Henceforth, we shall assume that E is a prehilbert space. A Hilbert space is a complete prehilbert space. It is a well-known result that if g is a bounded linear functional on a Hilbert space E, then there is a $y \in E$ such that $g(h) = (y, h)$ for all $h \in E$. Although this result may not hold in general in a prehilbert space, we shall see that it holds for the cases which interest us. In particular, since $\delta f(u; h)$ is linear in h, there may exist an element $\nabla f(u) \in E$ such that $\delta f(u; h) = (\nabla f(u), h)$ for all $h \in E$. If such an element exists, we call it the *gradient* of f at u. In this case, $\delta f(u; h)$ is a bounded linear functional of h. Note that this holds only at u and nothing can be said about $\delta f(x; h)$ for x in a neighborhood of u.

THEOREM 1. *Let E be a real prehilbert space. Let J and g be real functionals on a set $D \subset E$ and let u be a relative minimum of J on the constraint $C(g)$. Further, let D be finitely open at u, let $\nabla J(y)$ and $\nabla g(y)$ exist as finitely continuous functions of y for all y in a finitely open set at u. Also let $\nabla g(u) \neq 0$. If $h_t \in E$ is such that $(\nabla g(u), h_t) = 0$, then $(\nabla J(u), h_t) = 0$.*

Proof. For any two scalars λ, μ define

$$h_{\lambda u} = \mu h_t + \lambda \nabla g(u).$$

The real function $F(\lambda, \mu) = g(u + h_{\lambda\mu})$ has the properties:

1. $F(0, 0) = 0$; $F(\lambda, \mu)$ continuous in a neighborhood of $(0, 0)$;

2. $F_\lambda \equiv \partial F/\partial\lambda$ and $F_\mu \equiv \partial F/\partial\mu$ exist as continuous functions of (λ, μ) in some neighborhood of $(0, 0)$;

3. $F_\lambda(\lambda, \mu) \neq 0$ for all (λ, μ) in some neighborhood of $(0, 0)$. In fact,

$$F_\lambda(\lambda, \mu) = \lim_{\Delta\lambda \to 0} \frac{g(u + h_{\lambda\mu} + \Delta\lambda \nabla g(u)) - g(u + h_{\lambda\mu})}{\Delta\lambda} \tag{4}$$

$$= \delta g(u + h_{\lambda\mu}; \nabla g(u)) = (\nabla g(u + h_{\lambda\mu}), \nabla g(u))$$

for all (λ, μ) in some sufficiently small neighborhood of $(0, 0)$. Similarly,

$$F_\mu(\lambda, \mu) = (\nabla g(u + h_{\lambda\mu}), h_t).$$

Since $\nabla g(u + h_{\lambda\mu})$ is continuous in (λ, μ) in a neighborhood of $(0, 0)$, so are $F_\lambda(\lambda, \mu)$ and $F_\mu(\lambda, \mu)$. Also.

$$F_\lambda(0, 0) = (\nabla g(u), \nabla g(u)) = \| \nabla g(u) \|^2 \neq 0.$$

Hence, there is a neighborhood of $(0, 0)$ in which $F_\lambda(\lambda, \mu) \neq 0$, establishing property 3 above. Note also that

$$F_\mu(0, 0) = (\nabla g(u), h_t) = 0.$$

Properties 1 and 3 allow us to invoke the classical implicit function theorem to obtain a function $G(\mu)$ such that $F(G(\mu), \mu) = 0$ for all μ in some neighborhood of $\mu = 0$. Furthermore, by property 2, G is continuously differentiable in a neighborhood of $\mu = 0$ and $G'(0) = -F_\mu(0, 0)/F_\lambda(0, 0) = 0$. Since $G(\mu) = \mu G'(\theta_\mu \mu)$ for μ sufficiently small $(0 < \theta_\mu < 1)$, it follows that

$$\lim_{\mu \to 0} \frac{G(\mu)}{\mu} = G'(0) = 0. \tag{5}$$

Consider the one-parameter family of vectors

$$y_\mu = u + \mu h_t + G(\mu) \nabla g(u).$$

For all μ in some neighborhood of $\mu = 0$,

$$g(y_\mu) = F(G(\mu), \mu) = 0.$$

Since $\lim_{\mu \to 0} G(\mu) = 0$, we have $\lim_{\mu \to 0} y_\mu = u$. Now, let

$$H(\lambda, \mu) = J(u + h_{\lambda\mu})$$

and

$$\Phi(\mu) = H(G(\mu), \mu) = J(y_\mu).$$

We have as before, $H_\lambda(0, 0) = (\nabla J(u), \nabla g(u))$ and $H_\mu(0, 0) = (\nabla J(u), h_t)$.

$$J(y_\mu) - J(u) = \Phi(\mu) - \Phi(0) = \mu \Phi'(0) + \mu\epsilon,$$

where $\epsilon \to 0$ as $\mu \to 0$. Observing that

$$\Phi'(0) = H_\lambda(0, 0) G'(0) + H_\mu(0, 0)$$

and recalling that $G(0) = G'(0) = 0$, we obtain

$$\Phi'(0) = H_\mu(0, 0) = (\nabla J(u), h_t).$$

Hence,

$$J(y_\mu) - J(u) = \mu[(\nabla J(u), h_t) + \epsilon].$$

If $(\nabla J(u), h_t) = a^2 \neq 0$, then for all sufficiently small negative μ, we would have $J(y_\mu) < J(u)$, contradicting the hypothesis that u is a relative minimum of J on $C(g)$. Likewise, if $(\nabla J(u), h_t) = -a^2 \neq 0$, then for $\mu > 0$ and $|\mu|$ sufficiently small, we obtain the same contradiction. Therefore,

$$(\nabla J(u), h_t) = 0,$$

as was to be proven.

As an immediate consequence of Theorem 1, we have the following statement of the multiplier rule for the case of one constraint.

THEOREM 2. *Let J and g be real functionals on a set D in a real prehilbert space. Let D be finitely open at u, a relative minimum of J on the constraint $C(g)$. Let $\nabla J(y)$ and $\nabla g(y)$ be finitely continuous functions of y in a finitely open set at u. Also let $\nabla g(u) \neq 0$. Then*

$$\nabla J(u) = \frac{(\nabla J(u), \nabla g(u))}{\| \nabla g(u) \|^2} \nabla g(u). \tag{6}$$

Proof. Let

$$M = \{h \mid (\nabla g(u), h) = 0\}.$$

Let

$$v = \nabla J(u) - \frac{(\nabla J(u), \nabla g(u))}{\| \nabla g(u) \|^2} \nabla g(u). \tag{7}$$

For any $h \in M$, we have $(v, h) = (\nabla J(u), h) = 0$ by Theorem 1. Also $(v, \nabla g(u)) = 0$. Hence, $v \in M$ and $(v, v) = 0$; that is, $v = 0$ and the result follows.

Theorems 1 and 2 may be generalized to any finite number of constraints.

THEOREM 3. *Let $J, g_1, \cdots, g_p$ be real functionals on a set D in a real prehilbert space E. Let D be finitely open at u and let u be a relative minimum of J on the intersection $\bigcap_{i=1}^{p} C(g_i)$ of the constraints. Let $\nabla J(y)$ and $\nabla g_i(y)$, $i = 1, \cdots, p$, exist as finitely continuous functions of y in a finitely open set at u. Finally, suppose the set of gradients $\{\nabla g_1(u), \cdots, \nabla g_p(u)\}$ is a linearly independent set in E. If $h_t \in E$ is such that*

$$(\nabla g_i(u), h_t) = 0$$

for all $i = 1, \cdots, p$, *then*

$$(\nabla J(u), h_t) = 0.$$

Proof. Let $\lambda = (\lambda_1, \cdots, \lambda_p)$ be a p-dimensional real vector and define

$$y_{\lambda\mu} = u + \mu h_t + \sum_{j=1}^{p} \lambda_j \nabla g_j(u).$$

Then define

$$F_i(\lambda, \mu) = F_i(\lambda_1, \cdots, \lambda_p, \mu) = g_i(y_{\lambda\mu}), \quad i = 1, \cdots, p.$$

The functions F_i have the properties:

$$F_i(0, 0) = g_i(u) = 0,$$

$$F_{ij}(\lambda, \mu) \equiv \frac{\partial F_i}{\partial \lambda_j} = \lim_{\Delta\lambda_j \to 0} \frac{g_i(y_{\lambda\mu} + \Delta\lambda_j \nabla g_j(u)) - g_i(y_{\lambda\mu})}{\Delta\lambda_j}$$

$$= \delta g_i(y_{\lambda\mu}; \nabla g_j(u)) = (\nabla g_i(y_{\lambda\mu}), \nabla g_j(u)).$$

Thus, $F_{ij}(\lambda, \mu)$ exists and is continuous in (λ, μ) in a neighborhood of $(0, 0)$. Similarly,

$$F_{i\mu} = \frac{\partial F_i}{\partial \mu} = (\nabla g_i(y_{\lambda\mu}), h_t)$$

exists and is continuous in such a neighborhood.

$$F_{ij}(0, 0) = (\nabla g_i(u), \nabla g_j(u)),$$

and the Jacobian matrix $(F_{ij}(0, 0))$ has rank p since the $\nabla g_i(u)$ are linearly independent. By continuity, the matrix $(F_{ij}(\lambda, \mu))$ has rank p for (λ, μ) in some neighborhood of $(0, 0)$. Also,

$$F_{i\mu}(0, 0) = (\nabla g_i(u), h_t) = 0$$

for all $i = 1, \cdots, p$.

As in Theorem 1, we invoke the implicit function theorem to obtain p functions $G_i(\mu)$ such that $F_i(G_1(\mu), \cdots, G_p(\mu), \mu) = 0$ in a neighborhood of $\mu = 0$, and $G_i(0)$ and $G_i'(0) = 0$ for $i = 1, \cdots, p$. As in Theorem 1, we find that $\lim_{\mu \to 0} G_i(\mu)/\mu = G_i'(0) = 0$. Consider the vectors

$$y_\mu = u + \mu h_t + \sum_{j=1}^{p} G_j(\mu) \nabla g_j(\mu).$$

For all μ in some neighborhood of $\mu = 0$,

$$g_i(y_\mu) = F_i(G_1(\mu), \cdots, G_p(\mu), \mu) = 0.$$

Since $\lim_{\mu \to 0} G_i(\mu) = 0$, we have $\lim_{\mu \to 0} y_\mu = u$.

Proceeding as in the proof of Theorem 1, we define $H(\lambda, \mu) = J(y_{\lambda\mu})$. We obtain $H_{\lambda_j}(0, 0) = (\nabla J(u), \nabla g_j(u))$ and $H_\mu(0, 0) = (\nabla J(u), h_t)$, which yields

$$J(y_\mu) - J(u) = \mu\left[\sum_{j=1}^{p} (\nabla J(u), \nabla g_j(u))G_j'(0) + (\nabla J(u), h_t) + \epsilon\right]$$

$$= \mu[(\nabla J(u), h_t) + \epsilon].$$

The remainder of the proof follows the proof of Theorem 1.

THEOREM 4. *Let J, $g_1, \cdots, g_p$ be real functionals on a set D in a real prehilbert space E. Let D be finitely open at u and let u be a relative minimum of f on the intersection $\bigcap_{i=1}^{p} C(g_i)$. Let $\nabla J(y)$ and $\nabla g_i(y), i = 1, \cdots, p$, exist as finitely continuous functions of y in a finitely open set at u. Finally, suppose $\{\nabla g_i(u) \mid i = 1, \cdots, p\}$ is a linearly independent set in E. If $\nabla J(u) \neq 0$, then there exist unique scalars $\lambda_1, \cdots, \lambda_p$ not all zero such that*

$$\nabla J(u) = \textstyle\sum_{j=1}^{p} \lambda_j \nabla g_j(u). \tag{8}$$

Proof. Let

$$M = \{h \mid (\nabla g_i(u), h) = 0 \text{ for all } i = 1, \cdots, p\}.$$

Let $(\lambda_1, \cdots, \lambda_p)$ be the unique solution of the linear system

$$\sum_{j=1}^{p} (\nabla g_i(u), \nabla g_j(u))\lambda_j = (\nabla g_i(u), \nabla J(u)), \qquad i = 1, \cdots, p. \tag{9}$$

Define

$$v = \nabla J(u) - \sum_{j=1}^{p} \lambda_j \nabla g_j(u).$$

For any $h \in M$,

$$(v, h) = (\nabla J(u), h) = 0$$

by Theorem 3. We also have for $i = 1, \cdots, p$,

$$(v, \nabla g_i(u)) = (\nabla J(u), \nabla g_i(u)) - \sum_{j=1}^{p} \lambda_j(\nabla g_i(u), \nabla g_j(u)) = 0,$$

by (9) above. Thus $v \in M$ and $(v, v) = 0$, which yields the result.

Remark. The vectors h_t such that $(\nabla g_i(u), h_t) = 0$ form a linear subspace called the *tangent subspace* of the constraint C at the point u. The geometric interpretation of the preceding theorems should be evident.

3. The multiplier rule. We shall now apply the results of §2 to obtain necessary conditions for a solution of the optimal control problem specified

by (1′), (2′), and (3′) in §1. As pointed out in §1, this will also yield a new derivation of the multiplier rule for the problem of Mayer.

Let U be an open set in R^q, Euclidean q-space. A control,

$$u^T = u^T(t) = (u_1(t), \cdots, u_q(t)),$$

is "admissible" on $[t_1, t_2]$ if it is a piecewise continuous function of t in the interval $t_1 \leqq t \leqq t_2$ and $u(t) \in U$ for all t in this interval. (The superscript T denotes the transpose of a matrix.)

The set $\bar{C}_q$ of all piecewise continuous functions $u(t)$ defined on an interval $[a, b]$ to R^q is a linear space. It becomes a prehilbert space if we define the inner product, $\langle u, v \rangle$, of any two functions $u, v \in \bar{C}_q$, where $u^T = u^T(t) = (u_1(t), \cdots, u_q(t))$ and $v^T = v^T(t) = (v_1(t), \cdots, v_q(t))$ as

$$\langle u, v \rangle \equiv \int_a^b (u_1 v_1 + \cdots + u_q v_q)\, dt = \int_a^b u^T v\, dt.$$

For t_1 and t_2 in $[a, b]$, the set of admissible controls on $[t_1, t_2]$ is a subset of $\bar{C}_q$, if we take $u(t) = 0$ for t outside of $[t_1, t_2]$. We shall take $\bar{C}_q$ as the underlying prehilbert space in the ensuing discussion.

Suppose that the initial values $x(t_1)$ and the initial time t_1 are fixed. Suppose further that there exists an admissible control $u = u(t)$ such that $(1')$ has a solution $x(t)$ in the interval $t_1 \leqq t \leqq t_2$, with initial values $x(t_1)$. The points $(x(t), u(t))$ lie in the open region Γ in which $f(x, u)$ has continuous partials. Consider the equations

$$(1'_s) \qquad \frac{dx}{dt} = f(x, u + sh),$$

where $h = h(t)$ is an arbitrary admissible control on $[t_1, t_2]$ and s is a scalar. From the theory of ordinary differential equations (see [8, p. 29], for example), it is known that $(1'_s)$ has a solution in $[t_1, t_2]$ for all sufficiently small $|s|$. In fact, for k arbitrary admissible controls, $h_1, \cdots, h_k$, the equation

$$\frac{dx}{dt} = f(x, u + \sum_{i=1}^{k} s_i h_i),$$

has a solution in $[t_1, t_2]$ for $\sum_{i=1}^{k} |s_i|$ sufficiently small. In all cases, the initial value is taken to be $x(t_1)$.

We shall designate the solution of $(1'_s)$ by $x(t, s)$ when h is being held fixed in the discussion. The final values $x(t_2, s)$ are functionals depending on the control, $y = u + sh$. In the general case, we have solutions $x(t_1, s_1, \cdots, s_n)$ with final values depending on $y = u + \sum_{i=1}^{k} s_i h_i$. Thus, the final values of the solutions of $(1'_s)$ with initial values $x(t_1)$ are func-

tions defined on a domain, $D \subset \bar{C}_q$, which is finitely open at u for any $u \in D$. We express this by writing $x(t_2) = x_2(y)$ for $y \in D$. It follows that the functions ψ_j in (2′) of §1 can also be regarded as functionals on D. Let us define

$$g_j(y) = \psi_j(t_1, x(t_1), t_2, x_2(y)), \qquad j = 1, \cdots, p.$$

Similarly,

$$J(y) \equiv J(t_1, x(t_1), t_2, x_2(y)).$$

If u minimizes J on the intersection $I = \bigcap_{j=1}^{p} C(g_j)$ of the constraints (see Definition 5), then

$$J(y) \geqq J(u)$$

for all $y \in D \cap I \cap N_u$, where N_u is some neighborhood of u. If $\nabla J(y)$ and $\nabla g_j(y)$ exist and satisfy the hypotheses of Theorem 4, then we may apply (8) above to derive the multiplier rule. Thus, the derivation from this point on consists of a calculation of the gradients of J and g_j. We now carry out this calculation using the well-known technique of the adjoint equation.

In the following, let $h = h(t)$ be an arbitrary but fixed function in $\bar{C}_q$. Let $u = u(t)$ be an optimal control on $[t_1, t_2]$. For all $|s|$ sufficiently small, $u + sh$ is an admissible control and, as explained above, has a corresponding trajectory $x(t, s)$ in $[t_1, t_2]$. For $s = 0$, the corresponding trajectory is the optimal trajectory $x = x(t, 0)$. Using the notation explained above, we have $x_2(u + sh) = x(t_2, s)$. Now, let $(\partial J/\partial x_2)_0$ denote the $n \times 1$ matrix of partial derivatives, $\partial J/\partial x_{2j}$, of $J(t_1, x(t_1), t_2, x(t_2))$ with respect to the variables $x_j(t_2)$, $j = 1, \cdots, n$, and evaluated for $x(t_2) = x(t_2, 0) = x_2(u)$, i.e., at the final value of the optimal trajectory. It is assumed that the point $(t_1, x(t_1), t_2, x(t_2, 0))$ lies in the region R_2 in which J and ψ_j have continuous first-order partials. Since R_2 is open, and since $x(t_2, s)$ is continuous in the parameter s (see [8] again), it follows that the points $Q_s : (t_1, x(t_1), t_2, x(t_2, s))$ are also in R_2 for sufficiently small s. Hence, $J(Q_s)$ and $\psi_j(Q_s)$ are defined for s sufficiently small and J and ψ_j have continuous first-order partials at such Q_s. In this discussion, t_1, $x(t_1)$, and t_2 are not being varied. Therefore, we assume that $p \leqq n$ and that the matrix $(\partial\psi_j/\partial x(t_2))$ has rank p.

Again using a superscript T to denote the transpose of a matrix, we have (see Remark 2, Definition 3),

$$\delta J(u; h) = \frac{dJ(u + sh)}{ds}\bigg|_{s=0} = \left(\frac{\partial J}{\partial x_2}\right)_0^T \delta x_2(u; h). \tag{10}$$

Now, let $\partial f/\partial u$ denote the $n \times q$ matrix $(\partial f_i/\partial u_j)$ and $\partial f/\partial x$ the $n \times n$

matrix $(\partial f_i/\partial x_j)$, both evaluated at a point $P_t : (x(t, s), u + sh)$, where $x(t, s)$ is the trajectory corresponding to $u + sh$. Note that $P_t \in \Gamma$ for $t_1 \leqq t \leqq t_2$, so that f has continuous first-order partials at P_t. Writing $x(t, s) = (x_1(t, s), \cdots, x_n(t, s))$, we let $\partial x/\partial s$ be the $n \times 1$ matrix of partials $(\partial x_i/\partial s)$. Once again, by appeal to results in the theory of ordinary differential equations (e.g., see [19, p. 72]), we can assert the existence of the $\partial x_i/\partial s$ for s sufficiently small, and for any such s one obtains the variational equations

$$\frac{d}{dt}\left(\frac{\partial x}{\partial s}\right) = \frac{\partial f}{\partial u} h + \frac{\partial f}{\partial x}\frac{\partial x}{\partial s}.$$

Since

$$\delta x_2(u; h) = \left.\frac{dx_2(u + sh)}{ds}\right|_{s=0} = \left.\frac{\partial x(t_2, s)}{\partial s}\right|_{s=0},$$

it follows that $\delta x_2(u; h)$ exists and is the final value of a solution of the nth order system,

$$\frac{dv}{dt} = \left(\frac{\partial f}{\partial u}\right)_0 h + \left(\frac{\partial f}{\partial x}\right)_0 v, \tag{11}$$

where the zero subscripts indicate that the partials are to be evaluated at the points $(x(t, 0), u(t))$ of the optimal solution. The initial values $v(t_1)$ are to be taken as zero when $x(t_1)$ is not to be varied. Otherwise, $v(t_1)$ will be arbitrary.

The adjoint equations of (11) are given by

$$\frac{dy}{dt} = -\left(\frac{\partial f}{\partial x}\right)_0^T y. \tag{12}$$

Hence, $d(y^T v)/dt = y^T(\partial f/\partial u)_0 h$ for any solutions y of (12) and v of (11). Integration yields

$$y^T(t_2)v(t_2) = \int_{t_1}^{t_2} y^T \left(\frac{\partial f}{\partial u}\right)_0 h dt + y^T(t_1)v(t_1). \tag{13}$$

If we take for y the solution $J_x(t)$ of (12) having final values $J_x(t_2) = (\partial J/\partial x_2)_0$, then since $v(t_2) = \partial x_2(u; h)$, it follows from (10) and (13) that

$$\delta J(u; h) = \int_{t_1}^{t_2} J_x{}^T \left(\frac{\partial f}{\partial u}\right)_0 h dt + J_x{}^T(t_1)v(t_1). \tag{14}$$

The function $J_x{}^T(\partial f/\partial u)_0$ is piecewise continuous on $[t_1, t_2]$, since $J_x{}^T$ is a solution of (12). Thus, $J_x{}^T(\partial f/\partial u)_0 \in \bar{C}_q$. Since t_1, $x(t_1)$, and t_2 are not being varied, we must take $v(t_1) = 0$. Using the inner product notation,

(14) can be written as $\delta J(u;\, h) = \langle J_x{}^T(\partial f/\partial u)_0,\, h\rangle$. This shows that

$$\nabla J(u) = J_x{}^T\left(\frac{\partial f}{\partial u}\right)_0, \tag{15}$$

that is, the gradient of J exists at u and may actually be computed by solving (12) for $J_x{}^T(t)$, integrating backward from t_2 to t_1 with starting values $(\partial J/\partial x_2)_0$. (Since u is piecewise continuous in $[t_1, t_2]$, so is $(\partial f/\partial x)_0$, and a piecewise differentiable solution of (12) exists in $[t_1, t_2]$ for all starting values.)

The gradient of each g_j is obtained similarly. Thus,

$$\nabla g_j(u) = \psi_{jx}^T\left(\frac{\partial f}{\partial u}\right)_0, \qquad j = 1, \cdots, p, \tag{16}$$

where $\psi_{jx}(t)$ is the solution of (12) having final values $\psi_{jx}(t_2) = (\partial\psi_j/\partial x_2)_0$. Finally, as noted earlier, for any admissible controls $h_1, \cdots, h_k$ and all $(s_1, \cdots, s_k)$ with $\sum_1^k |s_i| < \delta$, the control $u + \sum_1^k s_i h_i$ is admissible and has a corresponding trajectory $x(t, s_1, \cdots, s_k)$ which is continuous in $(s_1, \cdots, s_k)$. Hence, $(\partial f/\partial x)$ and $(\partial f/\partial u)$ evaluated at the points $(x(t, s_1, \cdots, s_k),\ u + \sum_1^k s_i h_i)$ are continuous functions of $(s_1, \cdots, s_k)$ in some neighborhood of the origin in R^k. Consequently, any solution $\bar{\psi}_{jx}^T$ of $dy/dt = -(\partial f/\partial x)^T y$ is continuous in the parameters $s_1, \cdots, s_k$ in this neighborhood. Thus,

$$\nabla g_j\left(u + \sum_1^k s_i h_i\right) = \bar{\psi}_{jx}^T\left(\frac{\partial f}{\partial u}\right) \tag{17}$$

exists and is continuous in the s_i, which implies that ∇g_j exists and is finitely continuous in a finitely open set at u, i.e., the set of all controls of the form $u + \sum_1^k s_i h_i$ for arbitrary $h_1, \cdots, h_k \in \bar{C}q$ and $\sum_1^k |s_i| < d$, where $d > 0$ depends on the h_i. This applies to ∇J as well.

Now, consider the gradients $\{\nabla g_j(u)\}$. If they are not linearly independent, then there are scalar multipliers $\lambda_1, \cdots, \lambda_p$ not all zero and such that $\sum_{j=1}^p \lambda_j \nabla g_j(u) = 0$. If the gradients are linearly independent, then by Theorem 4, there are multipliers $\lambda_1, \cdots, \lambda_p$ not all zero such that $\nabla J(u) = \sum_{j=1}^p \lambda_j \nabla g_j(u)$. Both cases may be subsumed under one general principle by asserting the existence of scalars $\lambda_1, \cdots, \lambda_p$ not all zero and a scalar l_0 such that

$$l_0 \nabla J(u) + \sum_{j=1}^p \lambda_j \nabla g_j(u) = 0, \tag{18}$$

where $l_0 = 1$ if the $\{\nabla g_j(u)\}$ are linearly independent and $l_0 = 0$ if they are not.

Applying this to the control problem at hand, we obtain the necessary

conditions

$$-l_0 J_x{}^T \left(\frac{\partial f}{\partial u}\right)_0 = \sum_{j=1}^{p} \lambda_j \psi_{jx}^T \left(\frac{\partial f}{\partial u}\right)_0, \tag{19a}$$

$$\frac{dJ_x}{dt} = -\left(\frac{\partial f}{\partial x}\right)_0^T J_x, \tag{19b}$$

$$J_x(t_2) = \left(\frac{\partial J}{\partial x_2}\right)_0, \tag{19c}$$

$$\frac{d\psi_{jx}}{dt} = -\left(\frac{\partial f}{\partial x}\right)_0^T \psi_{jx}, \qquad j = 1, \cdots, p, \tag{19d}$$

$$\psi_{jx}(t_2) = \left(\frac{\partial \psi_j}{\partial x_2}\right)_0. \tag{19e}$$

Equations (19b) and (19d) can be combined into one set of differential equations as follows. Let

$$l_x(t) = -l_0 J_x(t) - \sum_{j=1}^{p} \lambda_j \psi_{jx}(t). \tag{20}$$

From (19b)–(19e) it follows that l_x is the solution of

$$\frac{dl_x}{dt} = -\left(\frac{\partial f}{\partial x}\right)_0^T l_x, \tag{21}$$

with final values

$$l_x(t_2) = -l_0 \left(\frac{\partial J}{\partial x_2}\right)_0 - \sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial x_2}\right)_0. \tag{22}$$

Equation (19a) becomes

$$l_x{}^T \left(\frac{\partial f}{\partial u}\right)_0 = 0. \tag{23}$$

The components $l_{xi}(t)$, $i = 1, \cdots, n$, of l_x are the multipliers of Bliss' formulation of the multiplier rule (see [4, p. 202]) and (21) and (23) are the Euler-Lagrange equations. To show this, we refer again to [4, p. 203] and the statement of the problem of Mayer in §1 of this paper.

Following Bliss, we introduce the function

$$F(t, y, y') = \sum_{j=1}^{n} l_j(t) \Phi_j(t, y, y'),$$

where the Φ_j are the functions in (1). Let F_y be the $n \times 1$ matrix of partials $(\partial F/\partial y_i)$, $i = 1, \cdots, n$. Similarly, $F_{y'}$ is the $n \times 1$ matrix $(\partial F/\partial y_i')$. The Euler-Lagrange equations may be written in vector form as $dF_{y'}/dt = F_y$.

In the control problem, the Φ_j have the form

$$\Phi_j \equiv \dot{x}_j - f_j(x_1, \cdots, x_n, u_1, \cdots, u_q), \quad j = 1, \cdots, n.$$

To transform this into a Mayer problem, we set $y_i = x_i$, $i = 1, \cdots, n$, and $y_{n+k} = u_k$, $k = 1, \cdots, q$. Hence, for $i = 1, \cdots, n$.

$$\frac{\partial F}{\partial y_i} = -\sum_{j=1}^{n} l_j \frac{\partial f_j}{\partial x_i}, \quad \frac{\partial F}{\partial y_i'} = \frac{\partial F}{\partial \dot{x}_i} = l_i,$$

and the Euler-Lagrange equations are

$$\frac{dl_i}{dt} = -\sum_{j=1}^{n} l_j \frac{\partial f_j}{\partial x_i}. \tag{24}$$

For $i = n + k$, $k = 1, \cdots, q$,

$$\frac{\partial F}{\partial y_{n+k}} = -\sum_{j=1}^{n} l_j \frac{\partial f_j}{\partial u_k}, \quad \frac{\partial F}{\partial y'_{n+k}} = 0,$$

and the Euler-Lagrange equations are

$$\sum_{j=1}^{n} l_j \frac{\partial f_j}{\partial u_k} = 0. \tag{25}$$

Comparing (24), (25), with (21), (23), we see that they are one and the same system of equations. Furthermore, (22) for the final values are the transversality conditions obtained by Bliss [4, (74.9), p. 202] by setting the coefficient of dy_{i2} equal to zero, $i = 1, \cdots, n$. The constants e_μ of Bliss correspond to our λ_j and the multipliers $l_0, l_1, \cdots, l_n$ of Bliss are our l_0, l_{xi}, $i = 1, \cdots, n$. As in Bliss, it is clear that $l_0, l_{xi}(t)$ do not vanish simultaneously at any point $\bar{t}$ in $[t_1, t_2]$. If $l_0 = 1$, this is immediate. If $l_0 = 0$, then $l_{xi}(\bar{t}) = 0$ for all $1 \leqq i \leqq n$ implies that $l_x(t) \equiv 0$ for all t in $[t_1, t_2]$, since l_x is the unique solution of the homogeneous linear equation (21). But then

$$0 = l_x(t_2) = -\sum_{j=1}^{p} \lambda_j \psi_{jx}(t_2) = -\sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial x_2}\right)_0$$

by (20) and (19e). This contradicts the assumption that the matrix $(\partial \psi_j / \partial x_2)_0$ has rank p.

Remark. Let us introduce the (Hamiltonian) function

$$H(l_x, x, u) = \sum_{j=1}^{n} l_{xj} f_j(x, u) = l_x^T f,$$

where x, u, and l_x are regarded as independent vector variables of dimension n, q, and n, respectively. We have

$$\frac{\partial H}{\partial u_i} = \sum_{j=1}^{n} l_{xj} \frac{\partial f_j}{\partial u_i}, \qquad i = 1, \cdots, q.$$

If $(x(t), u(t))$ is an optimal solution of the optimal control problem and $l_x(t)$ is the corresponding solution of (21) given by (20), then (19a), or rather its vector form (23), becomes

$$\frac{\partial H}{\partial u}(l_x(t), x(t), u(t)) = 0, \qquad t_1 \leqq t \leqq t_2.$$

This is a necessary condition for $H(l_x(t), x(t), u)$ to attain a relative maximum with respect to u at the point $u = u(t)$. Hence we obtain in this case (i.e., when the control region U is open) conditions which are a consequence of the maximum principle of Pontryagin, without appeal to that principle.

4. Variation of endpoints and initial conditions. The results of the previous section are readily extended to the case where t_1, $x(t_1)$, t_2, and $x(t_2)$ are varied simultaneously. To do this, the underlying prehilbert space must be chosen to be $E = \bar{C}_q \times R^n \times R^2$ in which an arbitrary element is of the form $e = (u(t), v_1, t_{12})$ with $u = u(t) \in \bar{C}_q$, $v_1 \in R^n$, and $t_{12} = (t_1, t_2) \in R^2$. The inner product of two elements $e, e' \in E$ is defined as

$$\langle e, e' \rangle \equiv \int_a^b u^T u' \, dt + v_1^T v_1' + t_{12}^T t_{12}'.$$

Thus, E is a direct sum of $\bar{C}_q$, the "control space", R^n, the "initial-value space", R^2, the "endpoints space". If, for an arbitrary initial point t_1, an arbitrary final point t_2, an arbitrary set of initial values $x(t_1) \in R^n$, and an arbitrary admissible control u, (1′) has a solution in $[t_1, t_2]$, then the final values $x(t_2)$ can be regarded as a function on E to R^n. Writing $e = (u, x(t_1), t_{12})$, we can denote this function by $x_2(e)$, as in the previous section. Clearly, J and the ψ_j are functionals on E and the results of §3 can be extended in a very natural way to apply to this more general space. For example, (10) becomes

$$\delta J(e; \bar{h}) = \left(\frac{\partial J}{\partial x_2}\right)_0^T \delta x_2(e; \bar{h}) + \left(\frac{\partial J}{\partial x_1}\right)_0^T dx_1 + \left[\left(\frac{\partial J}{\partial t_1}\right)_0 + \left(\frac{\partial J}{\partial x_1}\right)_0^T \left(\frac{dx_1}{dt}\right)_0\right] dt_1 + \left[\left(\frac{\partial J}{\partial t_2}\right)_0 + \left(\frac{\partial J}{\partial x_2}\right)_0^T \left(\frac{dx_2}{dt}\right)_0\right] dt_2,$$

where now we are considering $J(e + s\bar{h})$ with $\bar{h} = (h(t), dx_1, dt_{12}) \in E$. Similarly, $x(t, s)$ is the trajectory corresponding to $e + s\bar{h}$ and $(x(t, 0), e)$ is the optimal solution. The remarks on finitely open sets apply to e and $x(t, s)$. As before, we obtain (13) except that now

$$v(t_1) = \left.\frac{\partial x(t_1, s)}{\partial s}\right|_{s=0} = \frac{d(v_1 + s\, dx_1)}{ds} = dx_1.$$

Hence,

$$\delta J(e; \bar{h}) = \int_{\tau_1}^{\tau_2} J_x{}^T \left(\frac{\partial f}{\partial u}\right)_0 h\, dt + \left[J_x{}^T(\tau_1) + \left(\frac{\partial J}{\partial x_1}\right)_0^T\right] dx_1 + \left(\frac{\partial J}{\partial t_{12}}\right)_0^T dt_{12}.$$

The zero subscript again means that all partials are evaluated at the optimum point $e = (u(t), v_1, \tau_{12})$, where v_1 are the optimum initial values and $\tau_{12} = (\tau_1, \tau_2)$ are the optimum end points. The gradient of J is given by

$$(15') \qquad \nabla J(u) = \left(J_x{}^T \left(\frac{\partial f}{\partial u}\right)_0, \quad J_x(\tau_1) + \left(\frac{\partial J}{\partial x_1}\right)_0, \quad \left(\frac{\partial J}{\partial t_{12}}\right)_0\right),$$

where $J_x(t)$ is again the solution of (12) with values $J_x(\tau_2) = (\partial J/\partial x_2)_0$. An analogous formula holds for $\nabla g_j(u)$ with ψ_j replacing J. The remainder of §3 carries over mutatis mutandis to the present case. In particular, we obtain the additional transversality conditions

$$(26) \qquad -l_x(t_1) + l_0 \left(\frac{\partial J}{\partial x_1}\right)_0 - \sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial x_1}\right)_0 = 0,$$

$$(27) \quad l_0 \left(\frac{\partial J}{\partial t_1}\right)_0 + \sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial t_1}\right)_0 + \left[\left(\frac{\partial J}{\partial x_1}\right)_0^T l_0 + \sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial x_1}\right)_0^T\right] \left(\frac{dx}{dt}\right)_{t_1} = 0,$$

$$(28) \quad l_0 \left(\frac{\partial J}{\partial t_2}\right)_0 + \sum_{j=1}^{p} \lambda_j \left(\frac{\partial \psi_j}{\partial t_2}\right)_0 + \left[\left(\frac{\partial J}{\partial x_2}\right)_0^T l_0 + \sum \lambda_j \left(\frac{\partial \psi_j}{\partial x_2}\right)_0^T\right] \left(\frac{dx}{dt}\right)_{t_2} = 0.$$

Equation (26) arises from the variation of $x(t_1)$. Equations (27) and (28) arise from the variation of t_1 and t_2, respectively.

5. Conclusions. A different approach for obtaining the fundamental general necessary conditions for a solution of the problem of Mayer has been presented in the context of an optimal control problem. The multiplier rule, which sums up these necessary conditions, has been derived using some general theorems from functional analysis, which serves to unify the treatment of minimization problems with constraints. In a sequel to this paper, it will be shown how these theorems and techniques can be applied to steepest descent methods.

REFERENCES

[1] N. I. Akhiezer, *The Calculus of Variations*, Blaisdell, New York, 1962.

[2] H. A. Antosiewicz and W. Rheinboldt, *Numerical analysis and functional analysis*, Survey of Numerical Analysis, J. Todd, ed., McGraw-Hill, New York, 1962.

[3] A. V. Balakrishnan, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1(1963), pp. 109–127.

[4] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[5] E. K. BLUM, *Minimization of functionals with equality constraints*, Abstract 64T-381, Amer. Math. Soc. Notices, 11 (1964), p. 589.

[6] ———, *Minimization of functionals with equality constraints*, United Aircraft Res. Rep. C-110058-14, 1964.

[7] O. BOLZA, *An application of the notions of "general analysis" to a problem of the calculus of variations*, Bull. Amer. Math. Soc., 16 (1910). pp. 402–407.

[8] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

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

[10] M. FRÉCHET, *La notion de differentielle dans l'analyse generale*, Ann. Ecol. Norm Sup., 42 (1925), pp. 293–323.

[11] R. GÂTEAUX, *Sur les fonctionelles continues et les fonctionelles analytiques*, Bull. Soc. Math. France, 50 (1922), pp. 1–21.

[12] A. A. GOLDSTEIN, *Minimizing functionals on Hilbert space*, Proceedings of the Symposium on Computing Methods in Optimization Problems, University of California at Los Angeles, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1964.

[13] H. H. GOLDSTINE, *A multiplier rule in abstract spaces*, Bull. Amer. Math. Soc., 44 (1938), pp. 388–394.

[14] ———, *The calculus of variations in abstract spaces*, Duke Math J., 9 (1942), pp. 811–822.

[15] M. R. HESTENES, *Hilbert space methods in variational theory and numerical analysis*, Proc. of the International Congress of Mathematicians, vol. III, Amsterdam, 1954, pp. 229–236.

[16] E. HILLE, *Functional Analysis and Semi-Groups*, American Mathematical Society Colloquium Publications, vol. 31, 1948.

[17] L. LIUSTERNIK AND V. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.

[18] P. C. ROSENBLOOM, *The method of steepest descent*, A.M.S. Proceedings of Symposia in Applied Mathematics, vol. VI, McGraw-Hill, New York, 1956.

[19] R. A. STRUBLE, *Nonlinear Differential Equations*, McGraw-Hill, New York, 1962.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 2
Printed in U.S.A., 1965

A GENERAL THEORY OF MINIMUM-FUEL SPACE TRAJECTORIES*

LUCIEN W. NEUSTADT†

1. Introduction. This paper is concerned with the trajectories of vehicles moving in free space, i.e., of vehicles that are subject only to gravitational and propulsive forces. The following problem is fundamental in the control of such trajectories: given the vehicle position, velocity, and mass at a specified initial time, find a propulsion program that brings the vehicle to a prescribed terminal state (in a terminal time which may be free or fixed) with a minimum expenditure of fuel. Such a program will be called optimal.

The mathematical treatment of this problem depends very strongly on the model used for the fuel expenditure. In the case of a rocket engine, an excellent approximation is that the rate of fuel consumption is proportional to the magnitude of the thrust vector, and this article will deal exclusively with this representation. For low thrust engines, the rate of fuel consumption is measured by the square of the thrust vector magnitude. Such a model permits a much simpler analysis, and, for the case of linear equations of motion, this problem has been widely studied (e.g., by Billik [13], Meditch [14], and the author [15]).

We shall assume throughout that no constraints are imposed on the vehicle position and velocity. If this assumption results in a trajectory for which the assumed model for the forces acting on the vehicle is incorrect (e.g., if the trajectory intersects a planetary atmosphere), or if the trajectory violates obvious physical constraints (e.g., if the vehicle must pass through the interior of the sun or the earth), the analysis developed in this article is clearly inapplicable. Instead, it will then be necessary to take the additional forces into account, and/or consider a problem with "restricted phase coordinates". However, there is good reason to expect that in many problems arising in current applications, the optimal trajectories will not be physically inconsistent with the model we use.

Further, except for a brief discussion in §9, we shall always suppose that there is no constraint on the allowed value of the thrust vector. Minimum-fuel thrust programs in the absence of any such constraints generally consist of a finite number of impulses. Although impulsive corrections can

* Received by the editors December 3, 1964, and in revised form April 9, 1965.

† Instrumentation Engineering Program, The University of Michigan, Ann Arbor, Michigan. Now at the Department of Electrical Engineering, University of Southern California, University Park, Los Angeles, California. This research was begun while the author was with the Aerospace Corporation, El Segundo, California, but was supported primarily by the United States Air Force through the Air Force Office of Scientific Research under Contract No. AF 49(638)-1318.

never be realized by an actual rocket engine, a knowledge of the optimum impulses will often make it possible to compute the optimum, or near optimum, thrust program in the presence of the thrust amplitude limits which must exist in actual engines.

The problem described above is clearly a variational one. In order to permit impulses, and yet have a precise mathematical formulation, it is necessary to place the problem in a somewhat unorthodox framework, and thereby arrive at a nonclassical variational problem. This development is carried out in §2. In §8 we show that this framework is a reasonable one by proving both an existence theorem for solutions of the resultant variational problem and an approximation theorem which states that solutions of the unorthodox variational problem can be approximated by conventional thrust programs to any desired degree of accuracy.

In §§3–6, necessary conditions that an optimum thrust program and associated trajectory must satisfy are derived. Many of these conditions have been previously obtained by examining the necessary conditions in the presence of a thrust amplitude constraint, and then passing to the limit formally as the maximum allowed amplitude tends to ∞ (see, e.g., Lawden [1]). In §9 we show that this limiting argument is, in a sense, justified, and also prove an existence theorem for optimum trajectories in the presence of thrust amplitude constraints.

The necessary conditions derived in the sequel give rise to a formidable two-point boundary value problem, with some additional unknown quantities to be determined (see the remarks at the end of §5), if it is desired to actually obtain an optimal trajectory. Satisfactory computational methods for handling such problems are only now beginning to be developed.

Some specific examples of contemporary interest are discussed in §7.

Ewing [2] adopted a viewpoint very similar to the one taken in this paper in his investigation of the same problem for the particular case where the gravitational field in which the vehicle moves is uniform. The case where the gravitational field is linear in the space coordinates has been previously treated by the author [3]. While preparing this manuscript, it has come to the author's attention that the problem discussed in this paper has also recently been studied, but from a slightly different viewpoint (basically a change of independent variable to allow "impulses"), by Rishel [4] and Warga [16].

2. Problem formulation. The motion of a vehicle that is subject only to gravitational and propulsive forces can be described by the following differential equations:

$$\ddot{r}_i = G_i(r_1, r_2, r_3, t) + \frac{F_i(t)}{M(t)}, \qquad i = 1, 2, 3,$$

where r_1, r_2, and r_3 are the coordinates of the vehicle center of gravity in some inertial, Cartesian coordinate system; G_1, G_2, and G_3 are the components of the vehicle acceleration due to the action of the gravitational force field; F_1, F_2, and F_3 are the components of the vehicle's thrust vector; and M is the total vehicle mass, including fuel. Denoting the vectors (r_1, r_2, r_3), (G_1, G_2, G_3) and (F_1, F_2, F_3) by $\mathbf{r}$, $\mathbf{G}$, and $\mathbf{F}$, respectively, we may write down the single vector differential equation

$$\ddot{\mathbf{r}} = \mathbf{G}(\mathbf{r}, t) + \frac{\mathbf{F}(t)}{M(t)}. \tag{2.1}$$

The rate of change of mass $\dot{M}$ is the negative of the fuel expenditure rate and, for a single rocket engine, is given by

$$\dot{M} = -\frac{\| \mathbf{F} \|}{gI_{sp}}, \tag{2.2}$$

where $\| \ \|$ denotes the Euclidean norm, and g and I_{sp} (the nominal acceleration due to gravity at the earth's surface and the specific impulse of the fuel, respectively) will be assumed to be known positive constants. We denote (gI_{sp}) by A.

We shall also suppose that $\mathbf{G}(\cdot\ ,\ \cdot)$ is a continuous, bounded function from $E_3 \times E_1$ to E_3 (E_m denotes Euclidean m-space) possessing continuous bounded first partial derivatives with respect to all of its arguments. This assumption is consistent with the conventional models of gravitational fields.

Throughout this paper we shall assume that an initial time t_0 (without loss of generality, and for ease of notation, we shall set $t_0 = 0$) and initial values for (2.1) and (2.2) have been given: $M(0) = M_0 > 0$, $\mathbf{r}(0) = \mathbf{r}_0$, $\dot{\mathbf{r}}(0) = \mathbf{v}_0$. If $\mathbf{F}(\cdot)$ is a summable function from $[0, \infty)$ to E_3 satisfying the inequality

$$\int_0^\infty \| \mathbf{F}(t) \| \, dt < AM_0,$$

then it follows from standard existence theorems that (2.1) and (2.2) have a unique solution[1] for $0 \leqq t < \infty$ that satisfies the above initial conditions.

[1] By a solution of (2.2) we here mean an absolutely continuous function $M(\cdot)$ that satisfies (2.2) for almost all $t > 0$. The inequality $\int_0^\infty \| F \| \, dt < AM_0$ implies that $M(t) > \bar{M}_0$ for some positive constant $\bar{M}_0$ and all $t > 0$. Physically, the first inequality signifies that the rocket cannot provide thrust once the fuel has been consumed. By a solution of (2.1) we mean a function $\mathbf{r}(\cdot)$, whose time derivative $\dot{\mathbf{r}}(\cdot)$ exists for all $t > 0$, with $\dot{\mathbf{r}}(\cdot)$ absolutely continuous, that satisfies (2.1) for almost all $t > 0$.

This solution will be denoted by $\mathbf{r}(t; \mathbf{F})M(t; \mathbf{F})$; $\dot{\mathbf{r}}(t; \mathbf{F})$ denotes the time derivative of $\mathbf{r}(t; \mathbf{F})$.

Finally, we shall suppose that there are given functions $h_i(\cdot\,, \cdot\,, \cdot\,)$ from $E_3 \times E_3 \times [0, \infty)$ to E_1, where $i = 1, \cdots, \nu$ and $\nu \leqq 6$, with the following two properties: (1) The h_i are continuous and have continuous first partial derivatives with respect to all of their arguments. (2) If

$$H(t) = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in E_3\,,\ \mathbf{y} \in E_3\,,\ h_i(\mathbf{x}, \mathbf{y}, t) = 0 \quad \text{for} \quad i = 1, \cdots, \nu\},$$

then $H(t)$ is a smooth manifold in E_6 for each $t \geqq 0$. For each $\bar{t} > 0$, let $\mathfrak{F}(\bar{t})$ denote the class of all summable functions $\mathfrak{F}(\cdot)$ from $[0, \bar{t}]$ to E_3 that satisfy the relations

$$\int_0^{\bar{t}} \| \mathbf{F}(t) \| \, dt < AM_0$$

and $h_i(\mathbf{r}(\bar{t}; \mathbf{F}), \dot{\mathbf{r}}(\bar{t}; \mathbf{F}), \bar{t}) = 0$ for $i = 1, \cdots, \nu$. Physically speaking, $\mathfrak{F}(\bar{t})$ consists of all thrust programs that "transfer" the vehicle from the position $\mathbf{r}_0$ and velocity $\mathbf{v}_0$ at $t = 0$ to a new state (at the time $\bar{t}$) that satisfies the boundary conditions $h_i(\mathbf{r}, \dot{\mathbf{r}}, \bar{t}) = 0$, $i = 1, \cdots, \nu$.

In the sequel, we shall consider two variational problems. The first, the *fixed terminal time problem*, consists in finding, for a given $t_1 > 0$, an element $\hat{\mathbf{F}}(\cdot) \in \mathfrak{F}(t_1)$ such that $M(t_1\,; \hat{\mathbf{F}}) \geqq M(t_1\,; \mathbf{F})$ for every $\mathbf{F} \in \mathfrak{F}(t_1)$. The second, the *variable terminal time problem*, consists in finding a time $t_1 \geqq 0$ and an element $\hat{\mathbf{F}}(\cdot) \in \mathfrak{F}(t_1)$ such that $M(t_1\,; \hat{\mathbf{F}}) \geqq M(t; \mathbf{F})$ for every pair $(t, \mathbf{F})$ with $t > 0$ and $\mathbf{F} \in \mathfrak{F}(t)$. In concrete terms, the basic problem is to find a thrust program that, for the given initial values, achieves prescribed boundary conditions, and that, in so doing, maximizes the terminal mass.

If $\nu = 6$ and $H(t)$ consists of a single point for each t, we shall say that the variational problem is a *fixed endpoint problem*; if $\nu < 6$, the problem will be called a *variable endpoint problem*.

Let us now consider variational problems that are derived from, and equivalent to, the above problems. Namely, replace (2.1) and (2.2) by the equations

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{G}(\boldsymbol{\varrho}, t), \\ \dot{\boldsymbol{\varrho}} &= \mathbf{z} + \mathbf{u}(t), \end{aligned} \tag{2.3}$$

where $\mathbf{z}$, $\boldsymbol{\varrho}$, $\mathbf{G}$, and $\mathbf{u}$ are 3-vectors, $\mathbf{G}(\cdot\,, \cdot\,)$ is the same function that appears in (2.1), and $\mathbf{u}(\cdot)$ is assumed to be an absolutely continuous, bounded function from $[0, \infty)$ to E_3. We shall consider solutions $\mathbf{z}(\cdot)$, $\boldsymbol{\varrho}(\cdot)$ of (2.3) that satisfy the initial conditions $\mathbf{z}(0) = \mathbf{v}_0$ and $\boldsymbol{\varrho}(0) = \mathbf{r}_0$. For a given bounded, absolutely continuous function $\mathbf{w}(\cdot)$ from $[0, \infty)$ to E_3,

we shall denote the solution of (2.3) with $\mathbf{u}(\cdot) = \mathbf{w}(\cdot)$ that satisfies the given initial conditions (it is easily seen that this solution exists and is unique for $0 \leqq t < \infty$) by $\mathbf{z}(t; \mathbf{w})$, $\varrho(t; \mathbf{w})$. We shall also say that $\varrho(t; \mathbf{w})$ is the trajectory that corresponds to $\mathbf{w}$.

For every $\bar{t} > 0$, we denote by $\mathcal{G}(\bar{t})$ the class of absolutely continuous functions $\mathbf{w}(\cdot)$ from $[0, \bar{t}]$ to E_3 for which the relations $\mathbf{w}(0) = 0$ and $h_i(\varrho(\bar{t}; \mathbf{w}), \mathbf{z}(\bar{t}; \mathbf{w}) + \mathbf{w}(\bar{t}), \bar{t}) = 0$ for $i = 1, \cdots, \nu$ are satisfied.

Now the *derived* fixed terminal time problem consists in finding, for a given $t_1 > 0$, an element $\tilde{\mathbf{u}}(\cdot) \in \mathcal{G}(t_1)$ such that

$$\int_0^{t_1} \left\| \frac{d\tilde{\mathbf{u}}(t)}{dt} \right\| dt \leqq \int_0^{t_1} \left\| \frac{d\mathbf{u}(t)}{dt} \right\| dt$$

for every $\mathbf{u} \in \mathcal{G}(t_1)$; the derived variable terminal time problem consists in finding a time $t_1 > 0$ and an element $\tilde{\mathbf{u}}(\cdot) \in \mathcal{G}(t_1)$ such that

$$\int_0^{t_1} \left\| \frac{d\tilde{\mathbf{u}}(t)}{dt} \right\| dt \leqq \int_0^{\tau} \left\| \frac{d\mathbf{u}(t)}{dt} \right\| dt$$

for every pair $(\tau, \mathbf{u})$ with $\tau > 0$ and $\mathbf{u} \in \mathcal{G}(\tau)$. The reason for introducing the derived problem will become clear in what follows.

We shall show that the original and derived variational problems are equivalent. Namely, we shall exhibit a mapping Φ that, for each $t_1 > 0$, is one-to-one from $\mathcal{F}(t_1)$ onto $\mathcal{G}(t_1)$ (if we identify elements in $\mathcal{F}(t_1)$ that differ only on a set of measure zero), and shall prove that $\mathbf{F} \in \mathcal{F}(t_1)$ is a solution of the original problem if and only if $\Phi(\mathbf{F})$ is a solution of the derived problem (whether the problem is fixed or variable terminal time).

Define the mapping Φ as follows. If $\mathbf{F}(\cdot) \in \mathcal{F}(t_1)$, let $\mathbf{u}(\cdot) = \Phi(\mathbf{F}(\cdot))$ be the absolutely continuous function from $[0, t_1]$ to E_3 that is given by

$$\mathbf{u}(t) = \int_0^t \frac{\mathbf{F}(s)}{M(s; \mathbf{F})}\, ds, \qquad 0 \leqq t \leqq t_1. \tag{2.4}$$

We shall show that Φ is one-to-one from $\mathcal{F}(t_1)$ onto $\mathcal{G}(t_1)$, and that $\Phi^{-1} = \Psi$, where $\mathbf{F}(\cdot) = \Psi(\mathbf{u}(\cdot))$ for $\mathbf{u} \in \mathcal{G}(t_1)$, is defined by

$$\mathbf{F}(t) = \exp\{\mu(t; \mathbf{u})\} \frac{d\mathbf{u}(t)}{dt}, \qquad 0 \leqq t \leqq t_1, \tag{2.5}$$

$$\mu(t; \mathbf{u}) = -\frac{1}{A} \int_0^t \left\| \frac{d\mathbf{u}(s)}{ds} \right\| ds + \ln M_0, \qquad 0 \leqq t \leqq t_1. \tag{2.6}$$

Note that $\mathbf{F}(t)$ is defined by (2.5) for almost all $t \in [0, t_1]$, since an absolutely continuous function has a derivative almost everywhere. At points where $d\mathbf{u}/dt$ does not exist, $\mathbf{F}(t)$ may be defined arbitrarily. Since $d\mathbf{u}/dt$ is summable, the integral in (2.6) is finite.

Consider (2.3) with $\mathbf{u} = \Phi(\mathbf{F})$, where $\mathbf{F} \in \mathfrak{F}(t_1)$. It is clear that $\dot{\boldsymbol{\varrho}}(\cdot\,; \mathbf{u})$ is absolutely continuous in $[0, t_1]$, and that, a.e. in $[0, t_1]$,

$$\ddot{\boldsymbol{\varrho}}(t; \mathbf{u}) = \mathbf{G}(\boldsymbol{\varrho}(t; \mathbf{u}), t) + \frac{\mathbf{F}(t)}{M(t; \mathbf{F})}. \tag{2.7}$$

Since

$$\boldsymbol{\varrho}(0; \mathbf{u}) = \mathbf{r}_0, \qquad \dot{\boldsymbol{\varrho}}(0; \mathbf{u}) = \mathbf{z}(0; \mathbf{u}) + \mathbf{u}(0) = \mathbf{v}_0, \tag{2.8}$$

it follows that (replacing $\mathbf{u}$ by $\Phi(\mathbf{F})$), for all $t \in [0, t_1]$,

$$\boldsymbol{\varrho}(t; \Phi(\mathbf{F})) = \mathbf{r}(t; \mathbf{F}), \qquad \mathbf{z}(t; \Phi(\mathbf{F})) + \Phi(\mathbf{F})(t) = \dot{\mathbf{r}}(t; \mathbf{F}). \tag{2.9}$$

Hence, $\Phi(\mathbf{F}) \in \mathcal{G}(t_1)$ by definition of $\mathfrak{F}(t_1)$ and $\mathcal{G}(t_1)$, or $\Phi(\mathfrak{F}(t_1)) \subset \mathcal{G}(t_1)$. Also (by (2.2), (2.4) and (2.6)), for all $t \in [0, t_1]$, we have that

$$M(t; \mathbf{F}) = \exp\{\mu(t; \Phi(\mathbf{F}))\}. \tag{2.10}$$

Let us show that $\Psi(\mathcal{G}(t_1)) \subset \mathfrak{F}(t_1)$. Thus, let $\mathbf{F} = \Psi(\mathbf{u})$, where $\mathbf{u} \in \mathcal{G}(t_1)$. It follows from (2.5) and (2.6) that $\exp\{\mu(\cdot\,; \mathbf{u})\}$ is absolutely continuous in $[0, t_1]$, that $\exp\{\mu(0; \mathbf{u})\} = M_0$ and that, a.e. in $[0, t_1)$,

$$\frac{d}{dt}[\exp\{\mu(t; \mathbf{u})\}] = -A^{-1} \| \mathbf{F}(t) \|, \tag{2.11}$$

so that $\int_0^{t_1} \| \mathbf{F}(t) \| \, dt < AM_0$, and (see (2.2)), for all $t \in [0, t_1]$,

$$\exp\{\mu(t; \mathbf{u})\} = M(t; \Psi(\mathbf{u})). \tag{2.12}$$

It is clear from (2.3) that $\dot{\boldsymbol{\varrho}}(t; \mathbf{u})$ is absolutely continuous in $[0, t_1]$. Also (see (2.3), (2.5) and (2.12)), $\boldsymbol{\varrho}(t; \mathbf{u})$ satisfies (2.7) a.e. in $[0, t_1]$. By the definition of $\mathcal{G}(t_1)$, $\mathbf{u}(0) = 0$, so that relations (2.8) are satisfied. Hence, for all $t \in [0, t_1]$,

$$\boldsymbol{\varrho}(t; \mathbf{u}) = \mathbf{r}(t; \Psi(\mathbf{u})), \qquad \mathbf{z}(t; \mathbf{u}) + \mathbf{u}(t) = \dot{\mathbf{r}}(t; \Psi(\mathbf{u})), \tag{2.13}$$

from which it follows that $\mathbf{F} \in \mathfrak{F}(t_1)$. Now the relation $\Phi^{-1} = \Psi$ is a consequence of (2.2), (2.4)–(2.6), and (2.12), and it only remains to show that Φ maps all of the solutions of the original variational problem onto all of the solutions of the derived problem. But it is a consequence of (2.6) and (2.12) that if $\mathbf{u}_i \in \mathcal{G}(t_i)$, $i = 1, 2$, then $M(t_1\,; \Psi(\mathbf{u}_1)) > M(t_2\,; \Psi(\mathbf{u}_2))$ if and only if

$$\int_0^{t_1} \| \dot{\mathbf{u}}_1(s) \| \, ds < \int_0^{t_2} \| \dot{\mathbf{u}}_2(s) \| \, ds,$$

and this immediately implies the desired result.

Note that (2.9), (2.12), and (2.13) describe the correspondence be-

tween solutions of (2.1) and (2.2) and solutions of (2.3) when $\mathbf{u}$ and $\mathbf{T} = \mathbf{F}$ correspond under the mappings Φ and Ψ.

If $\mathbf{u}(\cdot)$ is an absolutely continuous function from $[0, t_1]$ to E_3, then

$$\int_0^{t_1} \left\| \frac{d\mathbf{u}(t)}{dt} \right\| dt = \text{STV } \mathbf{u}, \tag{2.14}$$

where STV $\mathbf{u}$, the *strong total variation* of $\mathbf{u}$, is defined (see [5, p. 59]) as follows:

$$\text{STV } \mathbf{u} = \sup \sum_{i=1}^{m} \| \mathbf{u}(\tau_i) - \mathbf{u}(\tau_{i-1}) \|,$$

with the supremum taken over all finite partitions $0 = \tau_0 < \tau_1 < \cdots < \tau_m = t_1$ of $[0, t_1]$. For scalar-valued functions (where STV reduces to the total variation), relation (2.14) is well-known. The proof of (2.14) (see, e.g., [6, p. 209]) carries over from the scalar-valued to the vector-valued case with only minor modification.

Thus, the original fixed terminal time problem is equivalent to the problem of finding, for a given $t_1 > 0$, an element $\tilde{\mathbf{u}} \in \mathcal{G}(t_1)$ such that

$$\text{STV } \tilde{\mathbf{u}} = \inf_{\mathbf{u} \in \mathcal{G}(t_1)} STV\ \mathbf{u}; \tag{2.15}$$

and the variable terminal time problem is equivalent to that of finding a number $t_1 > 0$ and an element $\tilde{\mathbf{u}} \in \mathcal{G}(t_1)$ such that

$$\text{STV}_{[0,t_1]}\, \tilde{\mathbf{u}} = \inf_{\substack{\mathbf{u} \in \mathcal{G}(t) \\ t > 0}} \text{STV}_{[0,t]}\, \mathbf{u}, \tag{2.16}$$

where $\text{STV}_{[0,t]}$ denotes the strong total variation over the interval $[0, t]$.

Unfortunately, there is, in general, no element $\tilde{\mathbf{u}} \in \mathcal{G}(t_1)$ that achieves the infimum in the right-hand side of (2.15) or of (2.16). To circumvent this difficulty, we shall embed the sets $\mathcal{G}(t)$ in larger sets $\mathcal{K}(t)$ possessing the following two properties: (1) If $\mathbf{u}(\cdot)$ is any element of $\mathcal{K}(t)$, then there exist functions $\mathbf{u}_n(\cdot) \in \mathcal{G}(t)$, $n = 1, 2, \cdots$, such that $\mathbf{u}_n(s) \to \mathbf{u}(s)$ as $n \to \infty$ for all $s \in [0, t]$, and STV $\mathbf{u}_n \to$ STV $\mathbf{u}$ as $n \to \infty$ (see Theorem 4 in §8). (2) There is an element $\tilde{\mathbf{u}} \in \mathcal{K}(t)$ such that

$$\text{STV}_{[0,t]}\, \tilde{\mathbf{u}} = \inf_{\mathbf{u} \in \mathcal{K}(t)} \text{STV}_{[0,t]} \mathbf{u}$$

(see Theorem 3 in §8). Consequently,

$$\inf_{\mathbf{u} \in \mathcal{G}(t)} \text{STV}_{[0,t]} \mathbf{u} = \inf_{\mathbf{u} \in \mathcal{K}(t)} \text{STV}_{[0,t]} \mathbf{u}.$$

For each $\bar{t}$, $0 < \bar{t} < \infty$, we define $\mathcal{K}(\bar{t})$ as follows. Let

$$\mathcal{B}(\bar{t}) = \{\mathbf{u}(\cdot) : \mathbf{u} \text{ from } [0, \bar{t}] \text{ to } E_3 \text{ and continuous from the right in } (0, \bar{t}),\ \mathbf{u}(0) = 0,\ \text{STV}_{[0,\bar{t}]}\, \mathbf{u} < \infty\}.$$

For every $\mathbf{w} \in \mathcal{B}(\bar{t})$, (2.3) with $\mathbf{u} = \mathbf{w}$ has a unique solution[2] in $[0, \bar{t}]$ that satisfies the initial conditions $\mathbf{z}(0) = \mathbf{v}_0$, $\varrho(0) = \mathbf{r}_0$. We shall also denote this solution by $\mathbf{z}(t; \mathbf{w})$, $\varrho(t; \mathbf{w})$. Then, for each $\bar{t} > 0$, let

$$\mathcal{C}(\bar{t}) = \{\mathbf{w}(\cdot) : \mathbf{w} \in \mathcal{B}(\bar{t}),\ h_i(\varrho(\bar{t}; \mathbf{w}), \mathbf{z}(\bar{t}; \mathbf{w}) + \mathbf{w}(\bar{t}), \bar{t}) = 0 \quad \text{for } i = 1, \cdots, \nu\}. \tag{2.17}$$

It is obvious that $\mathcal{G}(\bar{t}) \subset \mathcal{C}(\bar{t}) \subset \mathcal{B}(\bar{t})$.

We shall denote by $\mathcal{B}(\infty)$ the set of all functions from $[0, \infty)$ to E_3 whose restrictions on $[0, \bar{t}]$, for every $\bar{t} > 0$, belong to $\mathcal{B}(\bar{t})$.

We shall henceforth be concerned with the extended variational problems defined as follows.

The extended variable terminal time problem consists in finding a number t_1, $0 < t_1 < \infty$, *and an element* $\tilde{\mathbf{u}} \in \mathcal{C}(t_1)$ *such that*

$$\mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}} = \inf_{\substack{\mathbf{u} \in \mathcal{C}(t) \\ t > 0}} \mathrm{STV}_{[0,t]}\mathbf{u}.$$

The extended fixed terminal time problem is analogously defined.

§§3–6 are devoted to the derivation of necessary conditions that solutions of the extended variational problems must satisfy. In §5, we consider the variable terminal time problem, and in §6, the fixed terminal time problem.

3. Variational equations. In this section, t_1 is an arbitrary fixed positive number and $\tilde{\mathbf{u}}(\cdot)$ is an arbitrary fixed element of $\mathcal{B}(t_1)$. Denote $\varrho(t; \tilde{\mathbf{u}})$ and $\mathbf{z}(t; \tilde{\mathbf{u}})$, for $0 \leqq t \leqq t_1$, by $\tilde{\varrho}(t)$ and $\tilde{\mathbf{z}}(t)$, respectively. Let $\Lambda(\cdot)$ denote the continuous matrix-valued function on $[0, t_1]$ whose i,jth element Λ_{ij} is given by

$$\Lambda_{ij}(t) = \frac{\partial G_i(\tilde{\varrho}(t), t)}{\partial r_j}, \qquad i, j = 1, 2, 3; \qquad 0 \leqq t \leqq t_1.$$

We shall also use the notation

$$\Lambda(t) = \frac{\partial \mathbf{G}(\tilde{\varrho}(t), t)}{\partial \mathbf{r}}, \qquad 0 \leqq t \leqq t_1. \tag{3.1}$$

For every function $\mathbf{u}(\cdot) \in \mathcal{B}(t_1)$, let $\delta\mathbf{z}(\cdot\ ; \mathbf{u})$ and $\delta\varrho(\cdot\ ; \mathbf{u})$ denote the

[2] If $\mathbf{w}(\cdot) \in \mathcal{B}(\bar{t})$, $\mathbf{w}$ has at most a denumerable number of points of discontinuity, and the discontinuities of $\mathbf{w}$ are of the first kind. By a solution of (2.3), with $\mathbf{u}(t) = \mathbf{w}(t)$, we here mean a continuously differentiable function $\mathbf{z}(\cdot)$ that satisfies the first equation everywhere, and an absolutely continuous function $\varrho(\cdot)$ that satisfies the second equation at all points of continuity of $\mathbf{w}(\cdot)$.

absolutely continuous functions from $[0, t_1]$ to E_3 that satisfy the equations

$$\begin{aligned} \frac{d}{dt}[\delta \mathbf{z}(t; \mathbf{u})] &= \Lambda(t)\delta\varrho(t; \mathbf{u}), \\ \frac{d}{dt}[\delta\varrho(t; \mathbf{u})] &= \delta \mathbf{z}(t; \mathbf{u}) + \mathbf{u}(t) - \tilde{\mathbf{u}}(t), \end{aligned} \tag{3.2}$$

almost everywhere in $[0, t_1]$, and assume the initial values

$$\delta \mathbf{z}(0; \mathbf{u}) = \delta\varrho(0; \mathbf{u}) = 0. \tag{3.3}$$

We shall refer to (3.2) as the variational equations associated with $\tilde{\mathbf{u}}(t)$. Since these equations are linear, their solution is given by the well-known variations of parameters formula, which here takes the form

$$\begin{aligned} \delta \mathbf{z}(t; \mathbf{u}) &= -\int_0^t [\dot{X}_1(t)\dot{Y}_1(s) + \dot{X}_2(t)\dot{Y}_2(s)][\mathbf{u}(s) - \tilde{\mathbf{u}}(s)]\, ds, \\ \delta\varrho(t; \mathbf{u}) &= -\int_0^t [X_1(t)\dot{Y}_1(s) + X_2(t)\dot{Y}_2(s)][\mathbf{u}(s) - \tilde{\mathbf{u}}(s)]\, ds, \end{aligned} \tag{3.4}$$

where the 3×3 matrices $X_i(t)$ and $Y_i(t)$, $i = 1$ or 2, satisfy the differential equations

$$\begin{aligned} \ddot{X}_i(t) = \Lambda(t)X_i(t), \quad \ddot{Y}_i(t) = Y_i(t)\Lambda(t), \\ 0 \leqq t \leqq t_1, \qquad i = 1 \quad \text{or} \quad 2, \end{aligned} \tag{3.5}$$

and the initial conditions

$$\begin{aligned} X_1(0) = \dot{X}_2(0) = -\dot{Y}_1(0) = Y_2(0) = I, \\ \dot{X}_1(0) = X_2(0) = Y_1(0) = \dot{Y}_2(0) = 0, \end{aligned} \tag{3.6}$$

I being the identity matrix. The matrices X_i, Y_i also satisfy the following identity:

$$\begin{pmatrix} X_1(t) & X_2(t) \\ \dot{X}_1(t) & \dot{X}_2(t) \end{pmatrix} \begin{pmatrix} -\dot{Y}_1(t) & Y_1(t) \\ -\dot{Y}_2(t) & Y_2(t) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \qquad 0 \leqq t \leqq t_1. \tag{3.7}$$

In conventional physical models of gravitational fields, the function $\mathbf{G}(\cdot, t)$, for every fixed t, is the gradient of a twice continuously differentiable scalar-valued function on E_3. Under this hypothesis, $\Lambda(t)$ is symmetric for every t, $0 \leqq t \leqq t_1$, in which case (3.5) and (3.6) imply that $Y_1(t) = -X_2^T(t)$ and $Y_2(t) = X_1^T(t)$ for $0 \leqq t \leqq t_1$. This computationally useful result, which is known as Schmidt's theorem, but which is apparently originally due to Siegel [7, p. 14], was brought to my attention by O. K. Smith.

Integrating (3.4) by parts, using (3.7) and the fact that $\mathbf{u}(0) = \tilde{\mathbf{u}}(0) = 0$, we obtain

$$\delta\mathbf{z}(t; \mathbf{u}) = \int_0^t [\dot{X}_1(t)Y_1(s) + \dot{X}_2(t)Y_2(s)]\, d[\mathbf{u}(s) - \tilde{\mathbf{u}}(s)] - \mathbf{u}(t) + \tilde{\mathbf{u}}(t),$$

$$\delta\varphi(t; \mathbf{u}) = \int_0^t [X_1(t)Y_1(s) + X_2(t)Y_2(s)]\, d[\mathbf{u}(s) - \tilde{\mathbf{u}}(s)], \tag{3.8}$$

the integrals in (3.8) being in the sense of Stieltjes.

For every $\mathbf{u}(\cdot) \in \mathcal{B}(t_1)$ and real number α, let $\delta\mathbf{x}(\mathbf{u}, \alpha)$ be the element in $E_3 \times E_3 \times E_1$ given by

$$\delta\mathbf{x}(\mathbf{u}, \alpha) = (\delta\varphi(t_1 ; \mathbf{u}), \delta\mathbf{z}(t_1 ; \mathbf{u}) + \mathbf{u}(t_1) - \tilde{\mathbf{u}}(t_1), \mathrm{STV}_{[0,t_1]}\, \mathbf{u} + \alpha) \tag{3.9}$$

and let

$$\omega_0 = \delta\mathbf{x}(\tilde{\mathbf{u}}, 0) = (0, 0, \mathrm{STV}\ \tilde{\mathbf{u}}). \tag{3.10}$$

Now define the set W in $E_3 \times E_3 \times E_1$ as follows

$$W = \{\delta\mathbf{x}(\mathbf{u}, \alpha) : \mathbf{u} \in \mathcal{B}(t_1),\ \alpha \geqq 0\}. \tag{3.11}$$

Clearly, $\omega_0 \in W$.

Since, for every $\mathbf{u}$ and $\mathbf{w}$ in $\mathcal{B}(t_1)$ and real number β, we have

$$\mathrm{STV}(\mathbf{u} + \mathbf{w}) \leqq \mathrm{STV}\ \mathbf{u} + \mathrm{STV}\ \mathbf{w}, \qquad \mathrm{STV}(\beta\mathbf{u}) = |\beta|\mathrm{STV}\ \mathbf{u}, \tag{3.12}$$

it follows at once that W is convex.

The set W is analogous to the cone of attainability described in [8, Chap. 2], and is also patterned closely after the convex set of variations introduced by Warga in [9, §III].

Let $\mathbf{n}$ be an arbitrary nonzero row vector in E_7. If $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2, \eta_7)$, where $\mathbf{n}_1 \in E_3$, $\mathbf{n}_2 \in E_3$, and $\eta_7 \in E_1$, let $\mathbf{p}(t; \mathbf{n})$ for $0 \leqq t \leqq t_1$, be the row vector defined by

$$\mathbf{p}(t; \mathbf{n}) = \sum_{i=1}^{2} \zeta_i(\mathbf{n})Y_i(t), \quad \zeta_i(\mathbf{n}) = \mathbf{n}_1 X_i(t_1) + \mathbf{n}_2 \dot{X}_i(t_1), \quad i = 1 \text{ or } 2. \tag{3.13}$$

It follows at once from (3.7) and (3.13) that

$$\mathbf{p}(t_1 ; \mathbf{n}) = \mathbf{n}_2, \quad \dot{\mathbf{p}}(t_1 ; \mathbf{n}) = -\mathbf{n}_1. \tag{3.14}$$

If we consider $\mathbf{p}(\cdot\ ; \mathbf{n})$ to be a function from $[0, t_1]$ to E_3 (for $\mathbf{n}$ fixed), we conclude, by virtue of (3.5) and (3.13), that $\mathbf{p}(\cdot\ ; \mathbf{n})$ is twice continuously

differentiable and that

$$\ddot{\mathbf{p}}(t;\,\mathbf{n}) = \mathbf{p}(t;\,\mathbf{n})\Lambda(t), \qquad 0 \leqq t \leqq t_1\,. \tag{3.15}$$

LEMMA 1. *If there is a nonzero vector* $\bar{\mathbf{n}} = (\bar{\eta}_1, \cdots, \bar{\eta}_7) \in E_7$ *such that* $\bar{\mathbf{n}}\cdot\delta\mathbf{x} \leqq \bar{\mathbf{n}}\cdot\boldsymbol{\omega}_0$ *for all* $\delta\mathbf{x} \in W$, *then* $\bar{\eta}_7 < 0$.

Proof. The hypothesis of the lemma, together with the definitions of W, $\boldsymbol{\omega}_0$, and $\mathbf{p}(t;\,\bar{\mathbf{n}})$ (see (3.8)–(3.11) and (3.13)) imply that, for every $\mathbf{u} \in \mathcal{B}(t_1)$,

$$\int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\mathbf{u}(t) + \bar{\eta}_7\,\mathrm{STV}\,\mathbf{u} \leqq \int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\tilde{\mathbf{u}}(t) + \bar{\eta}_7\,\mathrm{STV}\,\tilde{\mathbf{u}}. \tag{3.16}$$

We first show that $\bar{\eta}_7 \neq 0$. Suppose the contrary. Then, (3.16) takes the form

$$\int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\mathbf{u} \leqq \int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\tilde{\mathbf{u}}. \tag{3.17}$$

Since (3.17) must hold for every $\mathbf{u} \in \mathcal{B}(t_1)$, it follows that $\mathbf{p}(t;\,\bar{\mathbf{n}}) \equiv 0$ in $[0,\,t_1]$. Hence, $\dot{\mathbf{p}}(t;\,\bar{\mathbf{n}}) \equiv 0$. In particular, $\mathbf{p}(0;\,\bar{\mathbf{n}}) = \dot{\mathbf{p}}(0;\,\bar{\mathbf{n}}) = 0$, i.e., (see (3.6) and (3.13)) $\zeta_1(\bar{\mathbf{n}}) = \zeta_2(\bar{\mathbf{n}}) = 0$. But this implies that (see (3.7) and (3.13)) $(\bar{\eta}_1, \cdots, \bar{\eta}_6) = 0$, i.e., $\bar{\mathbf{n}} = 0$, and this contradiction shows that $\bar{\eta}_7 \neq 0$.

By hypothesis, $\bar{\mathbf{n}}\cdot\delta\mathbf{x}(\tilde{\mathbf{u}},\,1) \leqq \bar{\mathbf{n}}\cdot\boldsymbol{\omega}_0$, so that (see (3.9)–(3.11)) $\bar{\eta}_7 \leqq 0$. Since $\bar{\eta}_7 \neq 0$, $\bar{\eta}_7 < 0$.

LEMMA 2. *Suppose that* $\tilde{\mathbf{u}}(\cdot) \not\equiv 0$. *If there is a vector* $\bar{\mathbf{n}} = (\bar{\eta}_1, \cdots, \bar{\eta}_7) \in E_7$ *with* $\bar{\eta}_7 = -1$ *such that* $\bar{\mathbf{n}}\cdot\delta\mathbf{x} \leqq \bar{\mathbf{n}}\cdot\boldsymbol{\omega}_0$ *for all* $\delta\mathbf{x} \in W$, *then*

$$\max_{0\leqq t\leqq t_1} \| \mathbf{p}(t;\,\bar{\mathbf{n}}) \| = 1, \tag{3.18}$$

and

$$\int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\tilde{\mathbf{u}}(t) = \mathrm{STV}_{[0,t_1]}\,\tilde{\mathbf{u}}. \tag{3.19}$$

Proof. Let $\Omega = \max_{0\leqq t\leqq t_1} \| \mathbf{p}(t;\,\bar{\mathbf{n}}) \|$, and let $\tau \in [0,\,t_1]$ be such that $\| \mathbf{p}(\tau;\,\bar{\mathbf{n}}) \| = \Omega$. Define $\tilde{\mathbf{w}}(\cdot) \in \mathcal{B}(t_1)$ as follows:

$$\tilde{\mathbf{w}}(t) = \begin{cases} 0 & \text{for } 0 \leqq t < \tau, \\ K[\mathbf{p}(\tau;\,\bar{\mathbf{n}})]^T & \text{for } \tau \leqq t \leqq t_1\,, \end{cases}$$

(an obvious modification must be made if $\tau = 0$), where $K > 0$ is arbitrary. Then

$$\int_0^{t_1} \mathbf{p}(t;\,\bar{\mathbf{n}})\,d\tilde{\mathbf{w}}(t) = K\,\Omega^2. \tag{3.20}$$

As in the proof of Lemma 1, we can show that (3.16) is satisfied for every

$\mathbf{u} \in \mathcal{B}(t_1)$ and, in particular, for $\mathbf{u} = \tilde{\mathbf{w}}$. Since STV $\tilde{\mathbf{w}} = K\Omega$, we conclude that

$$(3.21)\quad \int_0^{t_1} \mathbf{p}(t;\, \bar{\mathbf{n}})\, d\tilde{\mathbf{w}} - \text{STV } \tilde{\mathbf{w}} = K\Omega(\Omega - 1) \leqq \int_0^{t_1} \mathbf{p}(t;\, \bar{\mathbf{n}})\, d\tilde{\mathbf{u}} - \text{STV } \tilde{\mathbf{u}}.$$

But (3.21) must hold for every $K > 0$, which implies that $\Omega \leqq 1$.

It is easily verified that

$$\int_0^{t_1} \mathbf{p}(t;\, \bar{\mathbf{n}})\, d\tilde{\mathbf{u}}(t) \leqq \left[\max_{0 \leqq t \leqq t_1} \| \mathbf{p}(t;\, \bar{\mathbf{n}}) \|\right] \text{STV } \tilde{\mathbf{u}} = \Omega \text{ STV } \tilde{\mathbf{u}},$$

so that

$$(3.22)\quad \int_0^{t_1} \mathbf{p}(t;\, \bar{\mathbf{n}})\, d\tilde{\mathbf{u}} - \text{STV } \tilde{\mathbf{u}} \leqq (\Omega - 1) \text{ STV } \tilde{\mathbf{u}} \leqq 0.$$

But (3.16), with $\mathbf{u} \equiv 0$, gives rise to the inequality

$$(3.23)\quad \int_0^{t_1} \mathbf{p}(t;\, \bar{\mathbf{n}})\, d\tilde{\mathbf{u}} - \text{STV } \tilde{\mathbf{u}} \geqq 0.$$

Combining (3.22) and (3.23), we obtain (3.19).

If $\Omega < 1$, it follows from (3.19) and (3.22) that STV $\tilde{\mathbf{u}} = 0$, i.e., $\tilde{\mathbf{u}} \equiv 0$. This contradiction shows that $\Omega = 1$, i.e., (3.18) holds. This completes the proof of Lemma 2.

LEMMA 3. *The interior of W is not empty.*

Proof. Since W is convex, it is sufficient to show that W does not belong to any flat in E_7 of dimension less than seven. Suppose the contrary. Then there is a nonzero vector $\bar{\mathbf{n}} = (\bar{\eta}_1, \cdots, \bar{\eta}_7)$ in E_7 such that (because $\boldsymbol{\omega}_0 \in W$) $\bar{\mathbf{n}} \cdot \boldsymbol{\omega}_0 = \bar{\mathbf{n}} \cdot \delta\mathbf{x}$ for every $\delta\mathbf{x} \in W$. In particular, $\bar{\mathbf{n}} \cdot \boldsymbol{\omega}_0 = \bar{\mathbf{n}} \cdot \delta\mathbf{x}(\tilde{\mathbf{u}}, 1)$, so that (see (3.9)–(3.11)) $\bar{\eta}_7 = 0$. But this contradicts Lemma 1, and thereby proves Lemma 3.

If $\mathbf{u}_1(\cdot), \cdots, \mathbf{u}_7(\cdot)$ are arbitrary fixed functions in $\mathcal{B}(t_1)$, we define the function $\mathbf{u}(\cdot\,, \cdot)$ from $[0, \infty) \times E_7$ to E_3 as follows:

$$(3.24)\quad \mathbf{u}(t, \delta_1, \cdots, \delta_7) = \mathbf{u}(t, \boldsymbol{\delta}) = \begin{cases} \left(1 - \sum_{j=1}^{7} \delta_j\right) \tilde{\mathbf{u}}(t) + \sum_{j=1}^{7} \delta_j \mathbf{u}_j(t) \\ \qquad \text{for } 0 \leqq t < t_1, \\ \left(1 - \sum_{j=1}^{7} \delta_j\right) \tilde{\mathbf{u}}(t_1^-) + \sum_{j=1}^{7} \delta_j \mathbf{u}_j(t_1^-) \\ \qquad \text{for } t_1 \leqq t < \infty. \end{cases}$$

Note that, for every fixed $\boldsymbol{\delta} \in E_7$, $\mathbf{u}(\cdot\,, \boldsymbol{\delta}) \in \mathcal{B}(\infty)$. For ease of notation, let

$$(3.25)\quad \boldsymbol{\varrho}(t, \boldsymbol{\delta}) = \boldsymbol{\varrho}(t; \mathbf{u}(\cdot\,, \boldsymbol{\delta})), \quad \mathbf{z}(t, \boldsymbol{\delta}) = \mathbf{z}(t; \mathbf{u}(\cdot\,, \boldsymbol{\delta})), \quad 0 \leqq t < \infty.$$

We shall also consider $\varrho(\cdot\,,\cdot)$ and $\mathbf{z}(\cdot\,,\cdot)$ to be functions from $[0, \infty) \times E_7$ to E_3. It is easily verified that, for $0 \leqq t \leqq t_1$,

$$\varrho(t, 0) = \tilde{\varrho}(t), \qquad \mathbf{z}(t, 0) = \tilde{\mathbf{z}}(t). \tag{3.26}$$

It follows from well-known theorems on the dependence of solutions of differential equations on parameters that $\partial \mathbf{z}(t, \boldsymbol{\delta})/\partial \delta_i$ and $\partial \varrho(t, \boldsymbol{\delta})/\partial \delta_i$ exist and are continuous functions of t and $\boldsymbol{\delta}$ in $[0, \infty) \times E_7$. In addition, for fixed $\boldsymbol{\delta}$, these derivatives are absolutely continuous functions of t which, for almost all $t \in [0, t_1]$, satisfy the equations

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial \mathbf{z}(t, \boldsymbol{\delta})}{\partial \delta_i}\right) &= \frac{\partial \mathbf{G}(\varrho(t, \boldsymbol{\delta}), t)}{\partial \mathbf{r}} \cdot \frac{\partial \varrho(t, \boldsymbol{\delta})}{\partial \delta_i}, \\ \frac{d}{dt}\left(\frac{\partial \varrho(t, \boldsymbol{\delta})}{\partial \delta_i}\right) &= \frac{\partial \mathbf{z}(t, \boldsymbol{\delta})}{\partial \delta_i} + \mathbf{u}_i(t) - \tilde{\mathbf{u}}(t), \end{aligned} \tag{3.27}$$

together with the initial conditions

$$\frac{\partial \mathbf{z}(0, \boldsymbol{\delta})}{\partial \delta_i} = \frac{\partial \varrho(0, \boldsymbol{\delta})}{\partial \delta_i} = 0.$$

In particular, it follows from (3.26) and (3.1)–(3.3) that, for $0 \leqq t \leqq t_1$,

$$\left(\frac{\partial \mathbf{z}(t, \boldsymbol{\delta})}{\partial \delta_i}\right)_{\boldsymbol{\delta}=0} = \delta \mathbf{z}(t; \mathbf{u}_i), \qquad \left(\frac{\partial \varrho(t, \boldsymbol{\delta})}{\partial \delta_i}\right)_{\boldsymbol{\delta}=0} = \delta \varrho(t; \mathbf{u}_i), \quad i = 1, \cdots, 7. \tag{3.28}$$

4. A fundamental lemma. In this and the next section we shall suppose that $\tilde{\mathbf{u}}(\cdot)$ is a solution of the extended variable terminal time problem, and that $t_1 > 0$ is the corresponding terminal time. We shall keep the notation introduced in §3.

Let

$$\tilde{\varrho}(t_1) = \bar{\mathbf{r}}, \qquad \tilde{\mathbf{z}}(t_1) + \tilde{\mathbf{u}}(t_1) = \bar{\mathbf{v}}. \tag{4.1}$$

By hypothesis, $(\bar{\mathbf{r}}, \bar{\mathbf{v}}) \in H(t_1)$, or $h_i(\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1) = 0$ for $i = 1, \cdots, \nu$, and

$$\mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}} \leqq \mathrm{STV}_{[0,t]}\,\mathbf{u} \quad \text{for every} \quad \mathbf{u} \in \mathfrak{U}(t) \quad \text{and every} \quad t > 0. \tag{4.2}$$

For $(\mathbf{x}, \mathbf{y}, t)$ in a neighborhood of $(\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1)$, there is a parametric representation of the manifolds

$$H(t) = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in E_3,\ \mathbf{y} \in E_3,\ h_i(\mathbf{x}, \mathbf{y}, t) = 0 \text{ for } i = 1, \cdots, \nu\}$$

of the following form:

$$\begin{aligned} \mathbf{x} &= \lambda_1(\boldsymbol{\sigma}, t), \\ \mathbf{y} &= \lambda_2(\boldsymbol{\sigma}, t), \end{aligned} \tag{4.3}$$

where $\lambda_1(\cdot\,,\cdot)$ and $\lambda_2(\cdot\,,\cdot)$ are functions from a neighborhood of $(0, t_1)$

in $E_{6-\nu} \times E_1$ (let this neighborhood be of the form $N_1 \times N_2$, where $N_1 \subset E_{6-\nu}$ and $N_2 \subset E_1$) to E_3, possessing continuous first partial derivatives with respect to t and the coordinates of $\mathfrak{d}$, and

$$\begin{aligned} \bar{\mathbf{r}} &= \lambda_1(0, t_1), \\ \bar{\mathbf{v}} &= \lambda_2(0, t_1). \end{aligned} \tag{4.4}$$

If we have a fixed endpoint problem, so that $\nu = 6$, λ_1 and λ_2 are continuously differentiable functions from a neighborhood of t_1 to E_3 with

$$\lambda_1(t_1) = \bar{\mathbf{r}}, \qquad \lambda_2(t_1) = \bar{\mathbf{v}}. \tag{4.5}$$

Let $\nabla h_i(\mathbf{x}, \mathbf{y}, t)$ denote the vector in E_6 defined by

$$\nabla h_i = \left(\frac{\partial h_i}{\partial x_1}, \frac{\partial h_i}{\partial x_2}, \frac{\partial h_i}{\partial x_3}, \frac{\partial h_i}{\partial y_1}, \frac{\partial h_i}{\partial y_2}, \frac{\partial h_i}{\partial y_3}\right), \qquad i = 1, \cdots, \nu.$$

By the definition of a smooth manifold, the vectors $\nabla h_i(\mathbf{x}, \mathbf{y}, t)$, $i = 1, \cdots, \nu$, are linearly independent whenever $(\mathbf{x}, \mathbf{y}) \in H(t)$. Let T denote the hyperplane of dimension $6-\nu$ that is tangent to $H(t_1)$ at $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$, i.e.,

$$\begin{aligned} T = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in E_3, \mathbf{y} \in E_3, [\nabla h_i(\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1)] \cdot (\mathbf{x} - \bar{\mathbf{r}}, \mathbf{y} - \bar{\mathbf{v}}) = 0, \\ i = 1, \cdots, \nu\}. \end{aligned} \tag{4.6}$$

If $\nu = 6$, T consists of the single point $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$. Now define the set Q in E_7 as follows:

$$Q = \{(\mathbf{x} - \bar{\mathbf{r}}, \mathbf{y} - \bar{\mathbf{v}}, \text{STV } \tilde{\mathbf{u}} + \alpha) : (\mathbf{x}, \mathbf{y}) \in T, \alpha \leqq 0\}. \tag{4.7}$$

It is easily seen that Q is convex. If $\nu = 6$, Q is the ray $L = \{\delta\mathbf{x}(\tilde{\mathbf{u}}, \alpha) : \alpha \leqq 0\}$.

Let

$$\tilde{\zeta} = \left(\tilde{\mathbf{z}}(t_1) + \tilde{\mathbf{u}}(t_1^-) - \frac{\partial \lambda_1(0, t_1)}{\partial t}, \mathbf{G}(\bar{\mathbf{r}}, t_1) - \frac{\partial \lambda_2(0, t_1)}{\partial t}, 0\right), \tilde{\zeta} \in E_7, \tag{4.8}$$

and, for every $\mathbf{u}(\cdot) \in \mathfrak{B}(t_1)$, $\alpha \in E_1$ and $\Delta \in E_1$, let

$$\delta\xi(\mathbf{u}, \alpha, \Delta) = \delta\mathbf{x}(\mathbf{u}, \alpha) + \Delta\tilde{\zeta}, \tag{4.9}$$

and let

$$V = \{\delta\xi(\mathbf{u}, \alpha, \Delta) : \mathbf{u} \in \mathfrak{B}(t_1), \alpha \geqq 0, -\infty < \Delta < \infty\}. \tag{4.10}$$

It is clear that $W \subset V$, that V is convex, and that $\omega_0 \in V \cap Q$ (see (3.10), (4.6), (4.7), (4.9), and (4.10)).

We now prove a fundamental lemma.

LEMMA 4. *If there is a number $\lambda > 0$ such that $\tilde{\mathbf{u}}(\cdot)$ is continuous in $(t_1 - \lambda, t_1)$, then Q does not meet the interior of V.*

This lemma is similar to [8, Lemma 11, p. 112]. The proof we shall give below is based on the proof given by Warga of an analogous result [9, Lemma 3.1].

Proof. Let us assume that the lemma is false, so that there is a point $\mathbf{q} \in Q$ that belongs to the interior of V. Let

$$\mathbf{q} = (\bar{\mathbf{x}} - \bar{\mathbf{r}}, \bar{\mathbf{y}} - \bar{\mathbf{v}}, \text{STV } \bar{\mathbf{u}} + \bar{\alpha}), \tag{4.11}$$

where $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in T$ and $\bar{\alpha} \leqq 0$. If $\bar{\alpha} = 0$, we replace $\bar{\alpha}$ by $-\epsilon$, where $\epsilon > 0$ is sufficiently small, and thereby obtain a point that also belongs to both Q and the interior of V. Thus, without loss of generality, we shall suppose that $\bar{\alpha} < 0$ in (4.11).

Since $\mathbf{q}$ is an interior point of V, there are seven points $\chi_1, \cdots, \chi_7$, in V, such that $\omega_0, \chi_1, \cdots, \chi_7$ are vertices of a 7-simplex which contains $\mathbf{q}$ in its interior. Let $\chi_j = \delta\xi(\mathbf{w}_j, \alpha_j, \Delta_j)$, where $\mathbf{w}_j(\cdot) \in \mathcal{B}(t_1)$ and $\alpha_j \geqq 0$. For every ϵ, $0 < \epsilon < t_1$, and $j = 1, \cdots, 7$, define $\mathbf{u}_j(\cdot\,; \epsilon) \in \mathcal{B}(t_1)$ as follows:

$$\mathbf{u}_j(t; \epsilon) = \begin{cases} \mathbf{w}_j(t) & \text{if } 0 \leqq t < t_1 - \epsilon \text{ or } t = t_1, \\ \mathbf{w}_j(t_1^-) & \text{if } t_1 - \epsilon \leqq t < t_1. \end{cases}$$

It is easily seen that $\delta\xi(\mathbf{u}_j(\cdot\,; \epsilon), \alpha_j, \Delta_j) \to \chi_j$ as $\epsilon \to 0$ for each j. Hence, for any fixed $\epsilon > 0$ sufficiently small, the points ω_0 and $\delta\xi(\mathbf{u}_j(\cdot\,; \epsilon), \alpha_j, \Delta_j)$, for $j = 1, \cdots, 7$, are vertices of a 7-simplex which contains $\mathbf{q}$ in its interior. Let ϵ_0 be one such ϵ, denote $\mathbf{u}_j(\cdot\,; \epsilon_0)$ simply by $\mathbf{u}_j(\cdot)$, and let

$$\omega_j = \delta\xi(\mathbf{u}_j(\cdot), \alpha_j, \Delta_j), \qquad j = 1, \cdots, 7. \tag{4.12}$$

Then the vectors $(\omega_j - \omega_0)$, $j = 1, \cdots, 7$, are linearly independent, and there are positive numbers $\gamma_0, \gamma_1, \cdots, \gamma_7$ such that

$$\mathbf{q} = \sum_{j=0}^{7} \gamma_j \omega_j, \qquad \sum_{j=0}^{7} \gamma_j = 1. \tag{4.13}$$

Note that the functions $\mathbf{u}_j(\cdot)$, $j = 1, \cdots, 7$, are constant in $[t_1 - \epsilon_0, t_1)$.

Let $\tau(\cdot)$ be the function from E_7 to E_1 defined by

$$\tau(\delta_1, \cdots, \delta_7) = \tau(\boldsymbol{\delta}) = t_1 + \sum_{j=1}^{7} \delta_j \Delta_j, \tag{4.14}$$

let $\lambda_0 = \min\{\lambda, \epsilon_0\}$, and let N be a neighborhood of 0 in E_7 such that $\tau(\boldsymbol{\delta}) \in N_2$ and $\tau(\boldsymbol{\delta}) > t_1 - \lambda_0$ whenever $\boldsymbol{\delta} \in N$. Let $\mathbf{u}(\cdot, \boldsymbol{\delta})$, $\varrho(\cdot, \boldsymbol{\delta})$, and $\mathbf{z}(\cdot, \boldsymbol{\delta})$ be defined as in §3 (see (3.24) and (3.25)). Note that $\mathbf{u}(\cdot, \cdot)$ is continuous in t and $\boldsymbol{\delta}$ for $t \in (t_1 - \lambda_0, \infty)$ and all $\boldsymbol{\delta} \in E_7$.

For each $\boldsymbol{\delta} \in N$, let $\bar{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}) \in \mathcal{B}(t_1)$ and $\hat{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}) \in \mathcal{B}(\tau(\boldsymbol{\delta}))$ be defined as follows:

$$\bar{\mathbf{u}}(t; \boldsymbol{\delta}) = \tilde{\mathbf{u}}(t) + \sum_{j=1}^{7} \delta_j[\mathbf{u}_j(t) - \tilde{\mathbf{u}}(t)], \qquad 0 \leqq t \leqq t_1, \tag{4.15}$$

$$\hat{\mathbf{u}}(t; \boldsymbol{\delta}) = \mathbf{u}(t, \boldsymbol{\delta}) \text{ for } 0 \leqq t < \tau(\boldsymbol{\delta}), \quad \hat{\mathbf{u}}(\tau(\boldsymbol{\delta}); \boldsymbol{\delta}) = \bar{\mathbf{u}}(t_1; \boldsymbol{\delta}). \tag{4.16}$$

It is clear that $\hat{\mathbf{u}}(t; 0) = \tilde{\mathbf{u}}(t)$ for $0 \leqq t \leqq t_1$. If $\tau(\boldsymbol{\delta}) = t_1$, then $\hat{\mathbf{u}}(t; \boldsymbol{\delta}) = \tilde{\mathbf{u}}(t; \boldsymbol{\delta})$ for $0 \leqq t \leqq t_1$, and obviously

$$\mathrm{STV}_{[0,\tau(\boldsymbol{\delta})]}\hat{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}) = \mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}).$$

It follows easily that the same equality holds if $\tau(\boldsymbol{\delta}) > t_1$, and it is not difficult to show that, if $\tau(\boldsymbol{\delta}) < t_1$, then

$$\mathrm{STV}_{[0,\tau(\boldsymbol{\delta})]}\hat{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}) \leqq \mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}}(\cdot\,; \boldsymbol{\delta}), \tag{4.17}$$

i.e., (4.17) is satisfied for all $\boldsymbol{\delta} \in N$.

Note that (see (4.16) and (3.25)), for $0 \leqq t \leqq \tau(\boldsymbol{\delta})$,

$$\boldsymbol{\varrho}(t; \hat{\mathbf{u}}(\cdot\,; \boldsymbol{\delta})) = \boldsymbol{\varrho}(t, \boldsymbol{\delta}), \quad \mathbf{z}(t; \hat{\mathbf{u}}(\cdot\,; \boldsymbol{\delta})) = \mathbf{z}(t, \boldsymbol{\delta}). \tag{4.18}$$

Since $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$ both belong to T, the entire line segment joining the two points (assuming that they are distinct) belongs to T. Denote this segment by $\hat{l}$; $\hat{l}$ is tangent to $H(t_1)$ at $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$. Hence, $\hat{l}$ is tangent at $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$ to a smooth curve Γ on $H(t_1)$. Let Γ be represented parametrically as $\{(\lambda_1(\mathfrak{d}(s), t_1), \lambda_2(\mathfrak{d}(s), t_1)) : -1 < s < 1\}$, where $\mathfrak{d}(\cdot)$ is a continuously differentiable function from $(-1, 1)$ to N_1, $\mathfrak{d}(0) = 0$, and

$$\left(\sum_{j=1}^{6-\nu} \frac{\partial\lambda_1(0, t_1)}{\partial\sigma_j}\frac{d\sigma_j(0)}{ds}, \quad \sum_{j=1}^{6-\nu} \frac{\partial\lambda_2(0, t_1)}{\partial\sigma_j}\frac{d\sigma_j(0)}{ds}\right) = (\bar{\mathbf{x}} - \bar{\mathbf{r}}, \bar{\mathbf{y}} - \bar{\mathbf{v}}). \tag{4.19}$$

If $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\bar{\mathbf{r}}, \bar{\mathbf{v}})$, and this must be true if $\nu = 6$, we let Γ consist of the single point $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$, or, equivalently, let $\mathfrak{d}(s) \equiv 0$, in which case (4.19) remains valid.

Let $\boldsymbol{\Theta}(\cdot\,, \cdot)$ be the function from $N \times (-1, 1)$ to E_7 defined as follows:

$$\begin{aligned}\boldsymbol{\Theta}(\delta_1, \cdots, \delta_7, s) = \boldsymbol{\Theta}(\boldsymbol{\delta}, s) = \Big(&\mathbf{f}_1(\boldsymbol{\delta}) - \mathbf{g}_1(\boldsymbol{\delta}, s), \quad \mathbf{f}_2(\boldsymbol{\delta}) \\ &+ \mathbf{f}_3(\boldsymbol{\delta}) - \mathbf{g}_2(\boldsymbol{\delta}, s), \quad \sum_{j=1}^{7} \delta_j[\mathrm{STV}\ \mathbf{u}_j - \mathrm{STV}\ \tilde{\mathbf{u}} + \alpha_j] - s\bar{\alpha}\Big),\end{aligned} \tag{4.20}$$

where the $\mathbf{f}_i(\cdot)$ are functions from N to E_3 defined as follows:

$$\mathbf{f}_1(\boldsymbol{\delta}) = \boldsymbol{\varrho}(\tau(\boldsymbol{\delta}), \boldsymbol{\delta}), \quad \mathbf{f}_2(\boldsymbol{\delta}) = \mathbf{z}(\tau(\boldsymbol{\delta}), \boldsymbol{\delta}), \quad \mathbf{f}_3(\boldsymbol{\delta}) = \hat{\mathbf{u}}(\tau(\boldsymbol{\delta}); \boldsymbol{\delta}), \tag{4.21}$$

and the $\mathbf{g}_i(\cdot\,, \cdot)$ are functions from $N \times (-1, 1)$ defined by

$$\mathbf{g}_k(\boldsymbol{\delta}, s) = \lambda_k(\mathfrak{d}(s), \tau(\boldsymbol{\delta})), \qquad k = 1 \text{ or } 2. \tag{4.22}$$

We shall show that $\boldsymbol{\Theta}(\cdot\,, \cdot)$ is continuously differentiable in $N \times (-1, 1)$.

It follows from (4.15), (4.16) and (4.21) that $\mathbf{f}_3(\cdot)$ is differentiable and that

$$\frac{\partial\mathbf{f}_3(\boldsymbol{\delta})}{\partial\delta_i} = \mathbf{u}_i(t_1) - \tilde{\mathbf{u}}(t_1). \tag{4.23}$$

It was shown in §3 that the partial derivatives $\partial\varrho(t, \boldsymbol{\delta})/\partial\delta_i$ and $\partial\mathbf{z}(t, \boldsymbol{\delta})/\partial\delta_i$ exist and are continuous functions of t and $\boldsymbol{\delta}$ in $[0, \infty) \times E_7$. Further, $\partial\mathbf{z}(\cdot\,, \cdot)/\partial t$ exists and is continuous in $[0, \infty) \times E_7$, and since $\mathbf{u}(\cdot\,, \cdot)$ is continuous in $(t_1 - \lambda_0, \infty) \times E_7$, $\partial\varrho(\cdot\,, \cdot)/\partial t$ exists and is continuous in $(t_1 - \lambda_0, \infty) \times E_7$ (see (3.25), (2.3) and footnote 2). It now follows from (4.21), (4.14), and (2.3) that $\mathbf{f}_1(\cdot)$ and $\mathbf{f}_2(\cdot)$ have continuous derivatives in N, and that, for $\boldsymbol{\delta} \in N$,

$$\begin{aligned}
\frac{\partial\mathbf{f}_1(\boldsymbol{\delta})}{\partial\delta_i} &= \left(\frac{\partial\varrho(t, \boldsymbol{\delta})}{\partial\delta_i}\right)_{t=\tau(\boldsymbol{\delta})} + [\mathbf{z}(\tau(\boldsymbol{\delta}), \boldsymbol{\delta}) + \mathbf{u}(\tau(\boldsymbol{\delta}), \boldsymbol{\delta})]\Delta_i\,, \\
\frac{\partial\mathbf{f}_2(\boldsymbol{\delta})}{\partial\delta_i} &= \left(\frac{\partial\mathbf{z}(t, \boldsymbol{\delta})}{\partial\delta_i}\right)_{t=\tau(\boldsymbol{\delta})} + \mathbf{G}(\varrho(\tau(\boldsymbol{\delta}), \boldsymbol{\delta}), \tau(\boldsymbol{\delta}))\Delta_i\,.
\end{aligned} \tag{4.24}$$

Because of the differentiability properties that we have assumed for the functions $\mathfrak{d}(\cdot)$, $\lambda_1(\cdot\,, \cdot)$, and $\lambda_2(\cdot\,, \cdot)$ it follows from (4.22) that $\lambda_1(\cdot\,, \cdot)$ and $\lambda_2(\cdot\,, \cdot)$ have continuous first derivatives in $N \times (-1, 1)$, and

$$\begin{aligned}
\frac{\partial\mathbf{g}_k(\boldsymbol{\delta}, s)}{\partial s} &= \sum_{j=1}^{6-\nu} \frac{\partial\lambda_k(\mathfrak{d}(s), \tau(\mathfrak{d}))}{\partial\sigma_j}\,\frac{d\sigma_j(s)}{ds}\,, \\
\frac{\partial\mathbf{g}_k(\boldsymbol{\delta}, s)}{\partial\delta_i} &= \frac{\partial\lambda_k(\mathfrak{d}(s), \tau(\boldsymbol{\delta}))}{\partial t}\,\Delta_i\,, \qquad k = 1 \text{ or } 2.
\end{aligned} \tag{4.25}$$

It follows from (4.20)–(4.25) that $\boldsymbol{\Theta}(\cdot\,, \cdot)$ has continuous first partial derivatives in $N \times (-1, 1)$. In addition, by virtue of (4.14), (3.28), (3.26), (3.24), (4.1), (4.8), (3.9), (3.10), (4.9) and (4.12),

$$\left(\frac{\partial\boldsymbol{\Theta}(\boldsymbol{\delta}, s)}{\partial\delta_i}\right)_{\substack{\boldsymbol{\delta}=0 \\ s=0}} = \boldsymbol{\omega}_i - \boldsymbol{\omega}_0\,. \tag{4.26}$$

Also, it follows from (4.20), (4.25), (4.14), (4.19), (4.11), (3.10), and (4.13) that

$$\left(\frac{\partial\boldsymbol{\Theta}(\boldsymbol{\delta}, s)}{\partial s}\right)_{\substack{\boldsymbol{\delta}=0 \\ s=0}} = -\sum_{j=1}^{7} \gamma_j(\boldsymbol{\omega}_j - \boldsymbol{\omega}_0). \tag{4.27}$$

If $\nu = 6$, the λ_i are functions only of t, and obvious notation changes must be made in (4.8), (4.22), and (4.25), after which (4.26) and (4.27) follow in the same way.

Now consider the vector equation

$$\boldsymbol{\Theta}(\boldsymbol{\delta}, s) = 0 \tag{4.28}$$

for the unknown $\boldsymbol{\delta}$ as a function of s. For $s = 0$, it is easily seen (see (4.20)–(4.22), (4.14)–(4.16), (3.26), (4.1) and (4.4)) that (4.28) has the solution $\boldsymbol{\delta} = 0$. Because of (4.26), the Jacobian of (4.28) at $\boldsymbol{\delta} = 0$, $s = 0$

is the determinant of the matrix whose columns are the vectors $(\omega_i - \omega_0)$. This Jacobian does not vanish since these vectors are, by hypothesis, linearly independent. Since $\Theta(\cdot\,,\cdot)$ has continuous first partial derivatives with respect to all of its arguments in a neighborhood of $(0, 0) \in E_7 \times E_1$, we can appeal to the implicit function theorem and solve (4.28) for δ as a continuously differentiable function of s in a neighborhood of $s = 0$. Say $\delta = \phi(s) = (\phi_1(s), \cdots, \phi_7(s))$. Then $\phi(0) = 0$ and $d\phi/ds$ can be obtained by differentiating (4.28) implicitly:

$$\sum_{j=1}^{7} \frac{\partial\Theta(\phi(s), s)}{\partial\delta_j} \frac{d\phi_j(s)}{ds} + \frac{\partial\Theta(\phi(s), s)}{\partial s} = 0.$$

In particular, for $s = 0$, we obtain, by virtue of (4.26) and (4.27),

$$\sum_{j=1}^{7} (\omega_j - \omega_0) \frac{d\phi_j(0)}{ds} = \sum_{j=1}^{7} \gamma_j(\omega_j - \omega_0). \tag{4.29}$$

Since the vectors $(\omega_j - \omega_0)$, $j = 1, \cdots, 7$, are linearly independent, (4.29) implies that $d\phi_j(0)/ds = \gamma_j$ for each j. Recalling that $\gamma_j > 0$ and $\phi_j(0) = 0$ for each j, we conclude that $\phi_j(s) > 0$ if s is positive and sufficiently small. Also, $\phi(s) \to 0$ as $s \to 0$.

Thus, let $\bar{s}$, $0 < \bar{s} < 1$, be sufficiently small so that $\phi(\bar{s}) \in N$, $\phi_j(\bar{s}) > 0$ for each j, and $\sum_{j=1}^{7} \phi_j(\bar{s}) < 1$. Denote $\phi(\bar{s})$ and $\phi_j(\bar{s})$ by $\bar{\delta}$ and $\bar{\delta}_j$, respectively, so that $\Theta(\bar{\delta}, \bar{s}) = 0$. Let $\bar{\tau} = \tau(\bar{\delta})$. It follows from (4.20)–(4.22) that

$$\varrho(\bar{\tau}, \bar{\delta}) = \lambda_1(\vartheta(\bar{s}), \bar{\tau}), \tag{4.30}$$

$$\mathbf{z}(\bar{\tau}, \bar{\delta}) + \hat{\mathbf{u}}(\bar{\tau}; \bar{\delta}) = \lambda_2(\vartheta(\bar{s}), \bar{\tau}), \tag{4.31}$$

$$\sum_{j=1}^{7} \bar{\delta}_j[\mathrm{STV}\, \mathbf{u}_j - \mathrm{STV}\, \tilde{\mathbf{u}} + \alpha_j] = \bar{s}\bar{\alpha}. \tag{4.32}$$

But (4.30), (4.31), (4.18), and the representation (4.3) of $H(t)$ imply that

$$h_i(\varrho(\bar{\tau}; \hat{\mathbf{u}}(\cdot\,; \bar{\delta})), \mathbf{z}(\bar{\tau}; \hat{\mathbf{u}}(\cdot\,; \bar{\delta})) + \hat{\mathbf{u}}(\bar{\tau}; \bar{\delta}), \bar{\tau}) = 0, \qquad i = 1, \cdots, \nu, \tag{4.33}$$

i.e., (see (2.17)), $\hat{\mathbf{u}}(\cdot\,; \bar{\delta}) \in \mathcal{K}(\bar{\tau})$. But it follows from (4.17), (4.15), (3.12), (4.32), the nonnegativity of the α_j and $\bar{\delta}_j$, and the relations $\bar{s} > 0$, $\bar{\alpha} < 0$, that

$$\mathrm{STV}_{[0,\bar{\tau}]}\hat{\mathbf{u}}(\cdot\,; \bar{\delta}) \leqq \mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}}(\cdot) + \bar{s}\bar{\alpha} < \mathrm{STV}_{[0,t_1]}\tilde{\mathbf{u}}(\cdot),$$

contradicting (4.2), and thereby proving Lemma 4.

5. The necessary conditions. We now prove the following theorem which provides necessary conditions for the extended variable terminal time problem.

THEOREM 1. *Let* $\tilde{\mathbf{u}}(\cdot)$ *be a nontrivial (i.e.,* $\tilde{\mathbf{u}} \not\equiv 0$*) solution of the extended variable terminal time problem, let* t_1 *be the corresponding terminal time, and let* $\tilde{\mathbf{p}}(t)$ *be the corresponding trajectory. Suppose that the points of discontinuity of* $\tilde{\mathbf{u}}(\cdot)$ *do not cluster at* t_1. *Then there exists a twice differentiable (column) vector-valued function* $\psi(\cdot)$ *from* $[0, t_1]$ *to* E_3 *such that*

$$\ddot{\psi}(t) = \left(\frac{\partial \mathbf{G}(\tilde{\mathbf{p}}(t), t)}{\partial \mathbf{r}}\right)^T \psi(t) \tag{5.1}$$

for all $t \in [0, t_1]$,

$$\max_{0 \leqq t \leqq t_1} \| \psi(t) \| = 1, \tag{5.2}$$

and

$$\int_0^{t_1} [\psi(t)]^T \, d\tilde{\mathbf{u}}(t) = \mathrm{STV}_{[0,t_1]}\, \tilde{\mathbf{u}}. \tag{5.3}$$

Let (4.3) *and* (4.4) *be a parametric representation of* $H(t)$ *in a neighborhood of* $(\tilde{\mathbf{p}}(t_1), \mathbf{z}(t_1 ; \tilde{\mathbf{u}}) + \tilde{\mathbf{u}}(t_1), t_1) = (\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1)$, *and let* T *be the hyperplane tangent to* $H(t_1)$ *at* $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$. *Then* $\psi(\cdot)$ *satisfies the following transversality conditions*:

$$\dot{\psi}(t_1)\cdot(\mathbf{x} - \bar{\mathbf{r}}) = \psi(t_1)\cdot(\mathbf{y} - \bar{\mathbf{v}}) \quad \textit{for all} \quad (\mathbf{x}, \mathbf{y}) \in T, \tag{5.4}$$

or, equivalently,

$$(\dot{\psi}(t_1), -\psi(t_1)) = \sum_{i=1}^{\nu} \mu_i \nabla h_i(\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1) \tag{5.4'}$$

for some real constants $\mu_1, \cdots, \mu_\nu$,

and

$$-\psi(t_1)\cdot\left[\mathbf{G}(\bar{\mathbf{r}}, t_1) - \frac{\partial \lambda_2(0, t_1)}{\partial t}\right] + \dot{\psi}(t_1)\cdot\left[\bar{\mathbf{v}} - \frac{\partial \lambda_1(0, t_1)}{\partial t}\right] = \dot{\psi}(t_1)\cdot[\tilde{\mathbf{u}}(t_1) - \tilde{\mathbf{u}}(t_1^-)] \geqq 0. \tag{5.5}$$

If $\| \psi(t) \| < 1$ *in some interval* $[t', t''] \subset [0, t_1]$, *then* $\tilde{\mathbf{u}}(\cdot)$ *is constant for* $t' \leqq t \leqq t''$. *If* $\tilde{\mathbf{u}}(\tau) \neq \tilde{\mathbf{u}}(\tau^-)$ *for some* $\tau \in (0, t_1]$, *then* $\| \psi(\tau) \| = 1$, *and there is a number* κ_τ *such that*

$$\tilde{\mathbf{u}}(\tau) - \tilde{\mathbf{u}}(\tau^-) = \kappa_\tau \psi(\tau), \qquad \kappa_\tau > 0. \tag{5.6}$$

If $0 = \tilde{\mathbf{u}}(0) \neq \tilde{\mathbf{u}}(0^+)$, *then* $\| \psi(0) \| = 1$, *and* $\tilde{\mathbf{u}}(0^+) = \kappa_0 \psi(0)$, *where* $\kappa_0 > 0$. *In particular, if* $D = \{t: \| \psi(t) \| = 1, 0 \leqq t \leqq t_1\}$ *is a finite set, then* $\tilde{\mathbf{u}}(\cdot)$ *is a step function whose points of discontinuity all belong to* D, *and whose jumps are given by* (5.6), *or the modification thereof if* $\tau = 0$.

Note that (5.4) is satisfied trivially if $\nu = 6$, since, in this case, T consists of the single point $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$. Also, if $\nu = 6$, $\partial\lambda_i(0, t_1)/\partial t$, $i = 1$ and 2, in (5.5) should be replaced by $d\lambda_i(t_1)/dt$.

Proof. By Lemma 4, Q does not meet the interior of V. Since $\omega_0 \in V \cap Q$, ω_0 is a boundary point of V. Now both V and Q are convex, and $V \supset W$, so that, by virtue of Lemma 3, the interior of V is not empty. Hence, there is a supporting hyperplane to V at ω_0 that separates V from Q. Let $\bar{\mathbf{n}} = (\bar{\mathbf{n}}_1, \bar{\mathbf{n}}_2, \bar{\eta}_7) \neq 0$, where $\bar{\mathbf{n}}_1 \in E_3$, $\bar{\mathbf{n}}_2 \in E_3$, and $\bar{\eta}_7 \in E_1$, be a normal to this hyperplane directed so that

$$\bar{\mathbf{n}}\cdot\delta\xi \leqq \bar{\mathbf{n}}\cdot\omega_0 \leqq \bar{\mathbf{n}}\cdot\theta \quad \text{for all} \quad \delta\xi \in V, \quad \theta \in Q. \tag{5.7}$$

By virtue of Lemma 1, relations (5.7) allow us to conclude that $\bar{\eta}_7 < 0$. Without loss of generality we shall assume that $\bar{\eta}_7 = -1$. If we let $\psi(t) = [\mathbf{p}(t; \bar{\mathbf{n}})]^T$, (5.2) and (5.3) follow from Lemma 2 (see (3.18) and (3.19)), and (5.1) follows from (3.15) and (3.1).

Let $(\mathbf{x}, \mathbf{y})$ be an arbitrary point of T. Then (see (4.6) and (4.7)) $(\mathbf{x} - \bar{\mathbf{r}}, \mathbf{y} - \bar{\mathbf{v}}, \text{STV } \tilde{\mathbf{u}}) \in Q$, and $(-\mathbf{x} + \bar{\mathbf{r}}, -\mathbf{y} + \bar{\mathbf{v}}, \text{STV } \tilde{\mathbf{u}}) \in Q$. Consequently, by virtue of (5.7) and (3.10),

$$\bar{\mathbf{n}}_1\cdot(\mathbf{x} - \bar{\mathbf{r}}) + \bar{\mathbf{n}}_2\cdot(\mathbf{y} - \bar{\mathbf{v}}) = 0.$$

Taking (3.14) into account, we obtain (5.4). The equivalence of (5.4) and (5.4′) follows at once from the definition of T (see (4.6)).

Consider the points $\delta\xi(\tilde{\mathbf{u}}, 0, \pm 1) = \omega_0 \pm \tilde{\zeta}$ of V (see (4.9), (4.10), and (3.10)). It follows from (5.7) that $\bar{\mathbf{n}}\cdot(\pm\tilde{\zeta}) \leqq 0$, i.e., $\bar{\mathbf{n}}\cdot\tilde{\zeta} = 0$, and the equality in (5.5) follows from (4.8), (4.1), and (3.14).

The final conclusions of the theorem are direct consequences of (5.2) and (5.3) (see [3, Theorem 3]).

It remains only to prove the inequality in (5.5). If $\tilde{\mathbf{u}}(t_1) = \tilde{\mathbf{u}}(t_1^-)$, the inequality is obvious. If $\tilde{\mathbf{u}}(t_1) \neq \tilde{\mathbf{u}}(t_1^-)$, then $\| \psi(t_1) \| = 1$, and, since $\| \psi(t) \| \leqq 1$ for $0 \leqq t < t_1$,

$$0 \leqq \left(\frac{d \| \psi(t) \|^2}{dt}\right)_{t=t_1} = 2\psi(t_1)\cdot\dot{\psi}(t_1). \tag{5.8}$$

Relations (5.8) and (5.6) imply the inequality in (5.5), completing the proof of Theorem 1.

The vector-valued function $(\psi(\cdot), -\dot{\psi}(\cdot))$ is analogous to the adjoint variable in the formulation of the Pontryagin maximum principle, or to the Lagrange multipliers of the classical calculus of variations. Relation (5.3) corresponds to the maximum principle itself, or to the Weierstrass E-condition.

If the set D is finite—say $D = \{\tau_1, \cdots, \tau_l\}$—then $\tilde{\mathbf{u}}(\cdot)$ is determined by $6 + l$ scalar parameters: six initial values for (5.1), and the l constants

κ_{τ_i} in (5.6). Indeed, given the values of these parameters and the initial values $\mathbf{z}(0)$, $\varrho(0)$, it is possible to "simultaneously solve" (2.3) with $\mathbf{u} = \tilde{\mathbf{u}}$ and (5.1), and determine $\tilde{\mathbf{u}}$ through (5.6).

6. The fixed terminal time problem. In this section we shall derive the necessary conditions for the extended fixed terminal time problem. Thus, let $t_1 > 0$ be fixed, and let $\tilde{\mathbf{u}}(\cdot)$ be a solution of the corresponding extended fixed terminal time problem. We shall keep the notation introduced in §3. Let $(\bar{\mathbf{r}}, \bar{\mathbf{v}})$ be defined by (4.1), so that, by hypothesis, $h_i(\bar{\mathbf{r}}, \bar{\mathbf{v}}, t_1) = 0$ for $i = 1, \cdots, \nu$, and

$$\text{(6.1)} \qquad \text{STV}_{[0,t_1]}\,\tilde{\mathbf{u}} \leqq \text{STV}_{[0,t_1]}\,\mathbf{u} \quad \text{for every} \quad \mathbf{u} \in \mathcal{K}(t_1).$$

We define T and Q as in §4 (see (4.6) and (4.7)). Corresponding to Lemma 4, we have the following proposition.

LEMMA 5. *The set Q does not meet the interior of W.*

Note that here it is not necessary to assume that $\tilde{\mathbf{u}}(\cdot)$ is continuous in $(t_1 - \lambda, t_1)$ for some $\lambda > 0$.

Proof. The derivation is almost identical to that of Lemma 4, with certain simplifications, and we shall only outline the necessary arguments. Assuming the contrary, we show that there are points $\boldsymbol{\omega}_j = \boldsymbol{\delta}\mathbf{x}(\mathbf{u}_j, \alpha_j)$, $j = 1, \cdots, 7$, such that each $\boldsymbol{\omega}_j \in W$ and $\boldsymbol{\omega}_0, \boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_7$ are the vertices of a 7-simplex that contains a point $\mathbf{q}$ of the form (4.11), with $\bar{\alpha} < 0$, in its interior. We let $\tau(\boldsymbol{\delta}) = t_1$ for all $\boldsymbol{\delta} \in E_7$, and consider solutions of (4.28) near $s = 0$, where $\boldsymbol{\Theta}(\cdot, \cdot)$ is again given by (4.20)–(4.22) and (4.16), and the function $\vartheta(\cdot)$ has all of the properties described in §4. Relations (4.26) and (4.27) can now be derived as in §4, except that (4.26) can be obtained without having to show that $\partial\varrho(t, \boldsymbol{\delta})/\partial t$ exists in a neighborhood of t_1 (since $\tau(\cdot) \equiv t_1$). The continuity of $\tilde{\mathbf{u}}$ in $(t_1 - \lambda, t_1)$ was used only in showing the existence of this derivate, and consequently the extra continuity hypothesis can here be dispensed with. It then follows as before that there is a vector $\bar{\boldsymbol{\delta}}$ possessing the same properties as in §4 such that (4.32) and (4.33), with $\bar{\tau} = t_1$, are satisfied. But these equations are inconsistent with relation (6.1), and we have a contradiction. This completes the outlined proof of Lemma 5.

We now have the following theorem.

THEOREM 2. *Let $\tilde{\mathbf{u}}(\cdot)$ be a nontrivial (i.e., $\tilde{\mathbf{u}} \not\equiv 0$) solution of the extended fixed terminal time problem, t_1 being the terminal time, and let $\tilde{\boldsymbol{\varrho}}(\cdot)$ be the corresponding trajectory. Then there exists a function $\boldsymbol{\psi}(\cdot)$ from $[0, t_1]$ to E_3 such that $\boldsymbol{\psi}(\cdot)$ and $\tilde{\mathbf{u}}(\cdot)$ satisfy all of the conditions stated in Theorem 1 with the possible exception of (5.5). If the points of discontinuity of $\tilde{\mathbf{u}}(\cdot)$ cluster at t_1, then, in addition,*

$$\text{(6.2)} \qquad \left(\frac{d\,\|\,\boldsymbol{\psi}(t)\,\|}{dt}\right)_{t=t_1} = 0, \qquad \|\,\boldsymbol{\psi}(t_1)\,\| = 1.$$

Proof. Just as the existence of a vector $\tilde{\mathbf{n}} = (\tilde{\mathbf{n}}_1, \tilde{\mathbf{n}}_2, \tilde{\eta}_7)$ which satisfies (5.7) follows from Lemma 4, it is here a consequence of Lemma 5 that there is a vector $\tilde{\mathbf{n}} = (\tilde{\mathbf{n}}_1, \tilde{\mathbf{n}}_2, \tilde{\eta}_7)$ such that

$$\tilde{\mathbf{n}}\cdot\delta\mathbf{x} \leqq \tilde{\mathbf{n}}\cdot\omega_0 \leqq \tilde{\mathbf{n}}\cdot\boldsymbol{\theta} \quad \text{for all} \quad \delta\mathbf{x} \in W \quad \text{and} \quad \boldsymbol{\theta} \in Q. \tag{6.3}$$

With the exception of (5.5), the conclusions of Theorem 1 now follow from (6.3) as in §5. Since $\| \psi(t) \| = 1$ at all points of discontinuity of $\tilde{\mathbf{u}}$, and $\| \psi(t) \|$ is differentiable (and certainly continuous) at $t = t_1$, the last sentence of Theorem 2 follows at once.

Note that if $\tilde{\mathbf{u}}(\cdot)$ is a solution of the extended *variable* terminal time problem, it is a fortiori a solution of a fixed terminal time problem. Therefore, Theorem 2 also provides necessary conditions for the case excluded in Theorem 1; i.e., if the points of discontinuity of $\tilde{\mathbf{u}}(\cdot)$ do cluster at t_1, the conclusions of Theorem 1 remain valid, with the exception that (5.5) must be replaced by (6.2).

7. Examples. Let us apply Theorems 1 and 2 to some problems that are of contemporary interest.

First consider the variable terminal time, fixed endpoint problem (sometimes referred to as the "rendezvous" problem) in which $\lambda_1(t)$ and $\lambda_2(t)$ (the functions that describe $H(t)$) represent the position and velocity, respectively, of an actual or fictitious target at the time t. The equations of motion of such a target can be written in the form

$$\frac{d\lambda_1(t)}{dt} = \lambda_2(t), \qquad \frac{d\lambda_2(t)}{dt} = \mathbf{G}(\lambda_1(t), t) + \mathbf{a}(t), \tag{7.1}$$

where $\mathbf{a}(t)$ is the nongravitational acceleration experienced by the target. Since (7.1) must hold, in particular, when $t = t_1$, relations (5.5) in this case take the form (see (4.5))

$$\psi(t_1)\cdot\mathbf{a}(t_1) = \dot{\psi}(t_1)\cdot[\tilde{\mathbf{u}}(t_1) - \tilde{\mathbf{u}}(t_1^-)] \geqq 0.$$

If $\mathbf{a}(t) \equiv 0$, i.e., if the target is in a "free-fall" trajectory, we obtain

$$\dot{\psi}(t_1)\cdot[\tilde{\mathbf{u}}(t_1) - \tilde{\mathbf{u}}(t_1^-)] = 0, \tag{7.2}$$

which implies that either $\tilde{\mathbf{u}}(t_1) = \tilde{\mathbf{u}}(t_1^-)$, or that $\| \psi(t_1) \| = 1$ and (see (5.6)) $\psi(t_1)\cdot\dot{\psi}(t_1) = 0$, i.e., relations (6.2) are satisfied.

Also consider the following three variable endpoint problems.

The first, the so-called "intercept" problem, is the problem in which the vehicle terminal position is specified (but may depend on the terminal time), and the terminal velocity is arbitrary. In this case, (4.3) can be put in the form $\mathbf{x} = \lambda_1(t)$, $\mathbf{y} = \mathbf{0}$, and

$$H(t_1) = T = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} = \lambda_1(t_1)\}.$$

The transversality condition (5.4′) in this case implies that $\psi(t_1) = 0$. Consequently, $\| \psi(t) \| < 1$ for $t' \leqq t \leqq t_1$ and some $t' < t_1$, so that, by Theorem 1, $\tilde{\mathbf{u}}(t)$ is constant for $t' \leqq t \leqq t_1$. This conclusion is valid whether the problem is with fixed or variable terminal time. For the variable terminal time problem, relations (5.5) take the form (since $\tilde{\mathbf{u}}(t_1) = \tilde{\mathbf{u}}(t_1^-)$ and $\psi(t_1) = 0$)

$$\dot{\psi}(t_1) \cdot \left[\bar{\mathbf{v}} - \frac{d\lambda_1(t_1)}{dt} \right] = 0.$$

As a second example, consider the case where the terminal velocity is specified (but may depend on the time) and the terminal position is arbitrary. Then, (4.3) can be put in the form $\mathbf{x} = \mathfrak{d}$, $\mathbf{y} = \lambda_2(t)$, and the transversality condition (5.4′) implies that $\dot{\psi}(t_1) = 0$ whether the problem is for fixed or variable terminal time. For the variable terminal time problem, (5.5) takes the form

$$\psi(t_1) \cdot \left[\mathbf{G}(\bar{\mathbf{r}}, t_1) - \frac{d\lambda_2(t_1)}{dt} \right] = 0.$$

In the third example, the "transfer to a specified orbit" problem, we shall assume that $\mathbf{G}(\mathbf{r}, t)$ is independent of t. Here, the vehicle terminal position and velocity are to be the same as the position and velocity at any point on a specified solution curve (i.e., orbit) of (2.1) with $\mathbf{T} \equiv 0$. Thus, for every $t > 0$,

$$H(t) = \left\{ (\mathbf{x}(s), \mathbf{y}(s)) : -\infty < s < \infty, \frac{d\mathbf{x}(s)}{ds} = \mathbf{y}, \right.$$
$$\left. \frac{d\mathbf{y}(s)}{ds} = \mathbf{G}(\mathbf{x}(s)), \mathbf{x}(0) = \mathbf{a}, \mathbf{y}(0) = \mathbf{b} \right\},$$

where $\mathbf{a}$ and $\mathbf{b}$ are given vectors in E_3 that specify the orbit. Then, in the notation of §4,

$$T = \{(\bar{\mathbf{r}} + \sigma\bar{\mathbf{v}}, \bar{\mathbf{v}} + \sigma\mathbf{G}(\bar{\mathbf{r}})) : -\infty < \sigma < \infty\},$$

and the transversality condition (5.4) takes the form

$$\bar{\mathbf{v}} \cdot \dot{\psi}(t_1) = \mathbf{G}(\bar{\mathbf{r}}) \cdot \psi(t_1). \tag{7.3}$$

Since H is independent of t, the functions λ_1 and λ_2 can be chosen to be independent of t, so that, for the variable terminal time problem, we have (by virtue of (5.5) and (7.3)) that (7.2) holds, i.e., either $\tilde{\mathbf{u}}(t_1) = \tilde{\mathbf{u}}(t_1^-)$ or (6.2) holds.

8. Existence and approximation theorems. In this section we shall prove that the sets $\mathfrak{R}(t)$ possess the two properties described in §2. We first prove an almost self-evident, but nevertheless interesting, lemma.

We shall say that a function $\mathbf{h}(\cdot)$ from $[0, 1]$ to E_3 is *regular* if $\mathbf{h}(\cdot)$ is continuous and piecewise smooth, and if $d\mathbf{h}/dt$ is of strong bounded variation in $[0, 1]$.

LEMMA 6. *Let $\mathbf{h}(\cdot)$ be a regular function from $[0, 1]$ to E_3, $\bar{t}$ an arbitrary positive number, and $\mathbf{z}_0$ and $\mathbf{z}_1$ arbitrary vectors in E_3. Then there exists a function $\bar{\mathbf{u}}(\cdot) \in \mathcal{B}(\bar{t})$ such that the solution $\mathbf{z}(\cdot)$, $\boldsymbol{\varrho}(\cdot)$ of (2.3), with $\mathbf{u} = \bar{\mathbf{u}}$ and initial values $\boldsymbol{\varrho}(0) = \mathbf{h}(0)$ and $\mathbf{z}(0) = \mathbf{z}_0$, satisfies the equation $\boldsymbol{\varrho}(t) = \mathbf{h}(t/\bar{t})$ for all $t \in [0, \bar{t}]$ as well as the boundary condition $\mathbf{z}(\bar{t}) + \bar{\mathbf{u}}(\bar{t}) = \mathbf{z}_1$.*

Proof. For each $t \in [0, \bar{t}]$, let

$$\boldsymbol{\zeta}(t) = \int_0^t \mathbf{G}\left(\mathbf{h}\left(\frac{s}{\bar{t}}\right), s\right) ds + \mathbf{z}_0, \tag{8.1}$$

and let the function $\bar{\mathbf{w}}(\cdot) \in \mathcal{B}(\bar{t})$ be defined as follows:

$$\bar{\mathbf{w}}(t) = \begin{cases} \left(\dfrac{d\mathbf{h}(s)}{ds}\right)_{s=\left(\frac{t}{\bar{t}}\right)^+} & \text{for} \quad 0 < t < \bar{t}, \\ 0 & \text{for} \quad t = 0, \\ \bar{t}\mathbf{z}_1 & \text{for} \quad t = \bar{t}. \end{cases}$$

Since the function $\mathbf{G}(\cdot\,,\,\cdot)$ is bounded, it follows that the function $\boldsymbol{\zeta}(\cdot)$ defined by (8.1) is of strong bounded variation in $[0, \bar{t}]$. Let $\bar{\mathbf{u}}(\cdot) = (1/\bar{t})\bar{\mathbf{w}}(\cdot) - \boldsymbol{\zeta}(\cdot)$. Then $\bar{\mathbf{u}} \in \mathcal{B}(\bar{t})$. If we set $\mathbf{z}(t) = \boldsymbol{\zeta}(t)$ and $\boldsymbol{\varrho}(t) = \mathbf{h}(t/\bar{t})$ for $t \in [0, \bar{t}]$, it can be verified directly that the functions $\mathbf{z}(\cdot)$ and $\boldsymbol{\varrho}(\cdot)$ are a solution of (2.3) with $\mathbf{u} = \bar{\mathbf{u}}$ that satisfies the initial and boundary conditions prescribed in the statement of the lemma.

Let $K(\cdot)$ be a function whose domain is $[0, \infty)$ and whose range is the class of subsets of E_3. Preserving the notation of §2, we shall in addition denote, for every $\bar{t} > 0$, by $\tilde{\mathcal{K}}(\bar{t}; K(\cdot))$ the following subset of $\mathcal{K}(\bar{t})$:

$$\tilde{\mathcal{K}}(\bar{t}; K(\cdot)) = \{\mathbf{w}(\cdot) : \mathbf{w}(\cdot) \in \mathcal{K}(\bar{t}), \boldsymbol{\varrho}(t; \mathbf{w}) \in K(t) \quad \text{for every} \quad t \in [0, \bar{t}]\}.$$

We now prove the following existence theorem.

THEOREM 3. *Let t_1 be a given positive number and let $K(\cdot)$ be a function from $[0, \infty)$ into the class of all closed subsets of E_3 with the property that there is at least one regular function $\mathbf{h}(\cdot)$ from $[0, 1]$ to E_3 such that $\mathbf{h}(t/t_1) \in K(t)$ for every $t \in [0, t_1]$, $\mathbf{h}(0) = \mathbf{r}_0$, and $(\mathbf{h}(1), \mathbf{y}) \in H(t_1)$ for some $\mathbf{y} \in E_3$. Then $\tilde{\mathcal{K}}(t_1; K(\cdot))$ is not empty, and there is an element $\tilde{\mathbf{u}}(\cdot) \in \tilde{\mathcal{K}}(t_1; K(\cdot))$ such that*

$$\mathrm{STV}_{[0,t_1]}\, \tilde{\mathbf{u}}(\cdot) = \inf_{\mathbf{u} \in \tilde{\mathcal{K}}(t_1; K(\cdot))} \mathrm{STV}_{[0,t_1]}\, \mathbf{u}(\cdot). \tag{8.2}$$

Proof. The fact that $\tilde{\mathcal{K}}(t_1; K(\cdot))$ is not empty follows immediately from Lemma 6. If $\tilde{\mathcal{K}}(t_1; K(\cdot))$ is a finite set, the theorem is trivial. Thus,

let $\mathbf{u}_n(\cdot)$, $n = 1, 2, \cdots$, be a sequence of elements of $\tilde{\mathfrak{U}}(t_1 ; K(\cdot))$ such that

$$\lim_{n\to\infty} \mathrm{STV}_{[0,t_1]}\, \mathbf{u}_n = \inf_{\mathbf{u}\in\tilde{\mathfrak{U}}(t_1;K(\cdot))} \mathrm{STV}_{[0,t_1]}\, \mathbf{u}. \tag{8.3}$$

Denote the functions $\mathbf{z}(\cdot\,; \mathbf{u}_n)$ and $\boldsymbol{\varrho}(\cdot\,; \mathbf{u}_n)$ by $\mathbf{z}_n(\cdot)$ and $\boldsymbol{\varrho}_n(\cdot)$, respectively. Because the function $\mathbf{G}(\cdot\,, \cdot)$ is bounded, the derivatives $\dot{\mathbf{z}}_n(\cdot)$ as well as the functions $\mathbf{z}_n(\cdot)$ themselves are uniformly bounded, and the $\mathbf{z}_n(\cdot)$ are equicontinuous on $[0, t_1]$. Since $\mathbf{u}_n(0) = 0$ for each n, and the numbers $\mathrm{STV}_{[0,t_1]}\mathbf{u}_n$, $n = 1, 2, \cdots$, are bounded, it follows that the functions $\mathbf{u}_n(\cdot)$ are uniformly bounded on $[0, t_1]$. Consequently, the derivatives $\dot{\boldsymbol{\varrho}}_n(\cdot)$, which exist almost everywhere in $[0, t_1]$, are uniformly bounded and the functions $\boldsymbol{\varrho}_n(\cdot)$ are themselves uniformly bounded and equicontinuous on $[0, t_1]$. Appealing to Arzela's Theorem [6, p. 122] and the Helly Selection Theorem [11, p. 222], we conclude that there are a subsequence of the $\mathbf{u}_n(\cdot)$ (which we shall continue to denote by $\mathbf{u}_n$, without loss of generality) and functions $\mathbf{u}_\infty(\cdot)$, $\mathbf{z}_\infty(\cdot)$, and $\boldsymbol{\varrho}_\infty(\cdot)$ from $[0, t_1]$ to E_3, where $\mathbf{u}_\infty$ is of strong bounded variation and $\mathbf{z}_\infty$ and $\boldsymbol{\varrho}_\infty$ are continuous, such that, for every $t \in [0, t_1]$,

$$\lim_{n\to\infty} \mathbf{u}_n(t) = \mathbf{u}_\infty(t), \quad \lim_{n\to\infty} \boldsymbol{\varrho}_n(t) = \boldsymbol{\varrho}_\infty(t), \quad \lim_{n\to\infty} \mathbf{z}_n(t) = \mathbf{z}_\infty(t). \tag{8.4}$$

Also, it is easily seen that $\lim_{n\to\infty} \mathrm{STV}\, \mathbf{u}_n \geqq \mathrm{STV}\, [\lim_{n\to\infty} \mathbf{u}_n]$, i.e., (see (8.3) and (8.4)),

$$\mathrm{STV}\, \mathbf{u}_\infty \leqq \inf_{\mathbf{u}\in\tilde{\mathfrak{U}}(t_1;K(\cdot))} \mathrm{STV}\, \mathbf{u}. \tag{8.5}$$

Now define the function $\tilde{\mathbf{u}}(\cdot)$ from $[0, t_1]$ to E_3 as follows:

$$\tilde{\mathbf{u}}(t) = \begin{cases} \mathbf{u}_\infty(t^+) & \text{for } 0 < t < t_1, \\ \mathbf{u}_\infty(t) & \text{for } t = 0 \text{ or } t = t_1. \end{cases} \tag{8.6}$$

Since $\mathbf{u}_\infty(\cdot)$ is of strong bounded variation, $\mathbf{u}_\infty(t^+)$ exists for each $t \in (0, t_1)$, and since there are at most a denumerable number of points at which a function of bounded variation is discontinuous, $\tilde{\mathbf{u}}(t) = \mathbf{u}_\infty(t)$ for almost all t in $[0, t_1]$. It is easily seen that $\mathrm{STV}\, \tilde{\mathbf{u}} \leqq \mathrm{STV}\, \mathbf{u}_\infty$, so that, by virtue of (8.5),

$$\mathrm{STV}\, \tilde{\mathbf{u}} \leqq \inf_{\mathbf{u}\in\tilde{\mathfrak{U}}(t_1;K(\cdot))} \mathrm{STV}\, \mathbf{u}. \tag{8.7}$$

It follows from (8.4) and (8.6) that $\tilde{\mathbf{u}}(0) = 0$, so that $\tilde{\mathbf{u}}(\cdot) \in \mathfrak{B}(t_1)$. In addition, because of (8.4), (8.6), and the continuity of the functions h_i, we have that $\boldsymbol{\varrho}_\infty(0) = \mathbf{r}_0$, $\mathbf{z}_\infty(0) = \mathbf{v}_0$, and $h_i(\boldsymbol{\varrho}_\infty(t_1), \mathbf{z}_\infty(t_1) + \tilde{\mathbf{u}}(t_1), t_1) = 0$ for $i = 1, \cdots, \nu$. Since the sets $K(t)$ are closed by hypothesis, it follows from (8.4) that $\boldsymbol{\varrho}_\infty(t) \in K(t)$ for all $t \in [0, t_1]$. Consequently,

if we can show that $\boldsymbol{\varrho}_\infty(\cdot) = \boldsymbol{\varrho}(\cdot\,;\tilde{\mathbf{u}})$ and that $\mathbf{z}_\infty(\cdot) = \mathbf{z}(\cdot\,;\tilde{\mathbf{u}})$, we can conclude that $\tilde{\mathbf{u}} \in \tilde{\mathfrak{R}}(t_1\,; K(\cdot))$, and (8.2) is then an immediate consequence of (8.7).

By virtue of (2.3),

$$\mathbf{z}_n(t) = \mathbf{v}_0 + \int_0^t \mathbf{G}(\boldsymbol{\varrho}_n(s), s)\, ds,$$

$$\boldsymbol{\varrho}_n(t) = \mathbf{r}_0 + \int_0^t [\mathbf{z}_n(s) + \mathbf{u}_n(s)]\, ds, \qquad 0 \leqq t \leqq t_1.$$

Since the functions $\mathbf{G}$, $\mathbf{z}_n$, and $\mathbf{u}_n$ are uniformly bounded, we can appeal to the Lebesgue dominated convergence theorem, and conclude that

$$\mathbf{z}_\infty(t) = \mathbf{v}_0 + \int_0^t \mathbf{G}(\boldsymbol{\varrho}_\infty(s), s)\, ds,$$

$$\boldsymbol{\varrho}_\infty(t) = \mathbf{r}_0 + \int_0^t [\mathbf{z}_\infty(s) + \tilde{\mathbf{u}}(s)]\, ds, \qquad 0 \leqq t \leqq t_1,$$

where we have also used the fact that $\mathbf{u}_\infty(t) = \tilde{\mathbf{u}}(t)$ for almost all $t \in [0, t_1]$. It now follows immediately (see footnote 2) that $\mathbf{z}_\infty(\cdot) = \mathbf{z}(\cdot\,;\tilde{\mathbf{u}})$ and $\boldsymbol{\varrho}_\infty(\cdot) = \boldsymbol{\varrho}(\cdot\,;\tilde{\mathbf{u}})$, completing the proof of Theorem 3.

If $K(t) = E_3$ for every $t \geqq 0$, then $\tilde{\mathfrak{R}}(t_1\,; K(\cdot)) = \mathfrak{R}(t_1)$, and it is evident that there exists a regular function $\mathbf{h}(\cdot)$ with the required properties. The existence theorem promised in §2 then follows at once from Theorem 3.

We now prove the following theorem.

THEOREM 4. *Let* $\tilde{\mathbf{u}}(\cdot) \in \mathfrak{R}(t_1)$ *for* $t_1 > 0$. *Then there exist functions* $\mathbf{u}_n(\cdot) \in \mathbf{G}(t_1)$, $n = 1, 2, \cdots$, *such that*:

(1) *the derivatives* $d\mathbf{u}_n/dt$ *are essentially bounded*;

(2) $\mathbf{u}_n(t) \to \tilde{\mathbf{u}}(t)$ *for each* $t \in [0, t_1]$ *as* $n \to \infty$; *and*

(3) $\lim_{n\to\infty} \mathrm{STV}\ \mathbf{u}_n = \mathrm{STV}\ \tilde{\mathbf{u}}$.

Proof. We first prove two lemmas.

Lemma 7. *Let* $\mathbf{w}_1(\cdot)$, $\mathbf{w}_2(\cdot)$, $\cdots$, *denote a sequence of uniformly bounded functions in* $\mathfrak{B}(t_1)$ *such that* $\mathbf{w}_n(t) \to \tilde{\mathbf{u}}(t)$ *as* $n \to \infty$ *for each* $t \in [0, t_1]$, *where* $\tilde{\mathbf{u}}(\cdot) \in \mathfrak{B}(t_1)$. *Then* $\boldsymbol{\varrho}(t; \mathbf{w}_n) \to \boldsymbol{\varrho}(t; \tilde{\mathbf{u}})$ *and* $\mathbf{z}(t; \mathbf{w}_n) \to \mathbf{z}(t; \tilde{\mathbf{u}})$ *as* $n \to \infty$ *uniformly in* $[0, t_1]$.

Proof. Let $\theta_n(\cdot)$ be the scalar-valued function on $[0, t_1]$ defined by

$$\theta_n(t) = \| \boldsymbol{\varrho}(t; \mathbf{w}_n) - \boldsymbol{\varrho}(t; \tilde{\mathbf{u}}) \| + \| \mathbf{z}(t; \mathbf{w}_n) - \mathbf{z}(t; \tilde{\mathbf{u}}) \|.$$

Then it follows from (2.3) and the boundedness of the partial derivatives $\partial G_i/\partial r_j$ that

$$\dot{\theta}_n(t) \leqq R\theta_n(t) + \| \tilde{\mathbf{u}}(t) - \mathbf{w}_n(t) \|$$

for some positive constant R and all $t \in [0, t_1]$. Now $\theta_n(0) = 0$, so that, for $0 \leqq t \leqq t_1$,

$$\theta_n(t) \leqq R \int_0^t \theta_n(\tau)\, d\tau + \int_0^{\tau_1} \| \tilde{\mathbf{u}}(\tau) - \mathbf{w}_n(\tau) \|\, d\tau.$$

It follows from the Lebesgue dominated convergence theorem that $\int_0^{t_1} \| \tilde{\mathbf{u}}(\tau) - \mathbf{w}_n(\tau) \|\, d\tau \to 0$ as $n \to \infty$, and the lemma is now an immediate consequence of Gronwall's inequality.

LEMMA 8. *For every $\epsilon > 0$ there exists a $\delta > 0$, depending only on ϵ, such that the following proposition holds: If $\mathbf{w}(\cdot)$ is any absolutely continuous function in $\mathcal{B}(t_1)$ whose derivative is essentially bounded, and $\tilde{\boldsymbol{\varrho}}$ and $\tilde{\mathbf{z}}$ are any vectors in E_3 that satisfy the inequality $\| \tilde{\boldsymbol{\varrho}} \| + \| \tilde{\mathbf{z}} \| < \delta$, then there exists an absolutely continuous function $\mathbf{u}(\cdot) \in \mathcal{B}(t_1)$ with essentially bounded derivative such that* STV $(\mathbf{u} - \mathbf{w}) < \epsilon$ *and*

$$\begin{aligned} \boldsymbol{\varrho}(t_1 ; \mathbf{u}) &= \boldsymbol{\varrho}(t_1 ; \mathbf{w}) + \tilde{\boldsymbol{\varrho}}, \\ \mathbf{z}(t_1 ; \mathbf{u}) + \mathbf{u}(t_1) &= \mathbf{z}(t_1 ; \mathbf{w}) + \mathbf{w}(t_1) + \tilde{\mathbf{z}}. \end{aligned} \tag{8.8}$$

Proof. Fix $\epsilon > 0$. Let $\tilde{G} = \sup_{\mathbf{r},t} \| \mathbf{G}(\mathbf{r}, t) \|$, $\tilde{\epsilon} = \min \{\epsilon/(7 + 2\tilde{G}), 1, t_1\}$, and $\delta = (\tilde{\epsilon})^2$. Let $\mathbf{w}(\cdot)$ be an arbitrary, fixed, absolutely continuous function in $\mathcal{B}(t_1)$, and let $\tilde{\boldsymbol{\varrho}}$ and $\tilde{\mathbf{z}}$ be arbitrary fixed vectors in E_3 such that $\| \tilde{\boldsymbol{\varrho}} \| + \| \tilde{\mathbf{z}} \| < \delta$. Define the element $\hat{\mathbf{w}}(\cdot)$ of $\mathcal{B}(t_1)$ as follows:

$$\hat{\mathbf{w}}(t) = \begin{cases} 0 & \text{for } 0 \leqq t \leqq t_1 - \tilde{\epsilon}, \\ (t - t_1 + \tilde{\epsilon})\mathbf{m}_1 & \text{for } t_1 - \tilde{\epsilon} \leqq t \leqq t_1 - \tilde{\epsilon}/2, \\ (t - t_1)\mathbf{m}_2 + \tilde{\mathbf{z}} & \text{for } t_1 - \tilde{\epsilon}/2 \leqq t \leqq t_1, \end{cases}$$

where $\mathbf{m}_1 = (4\tilde{\boldsymbol{\varrho}}/\tilde{\epsilon}^2) - \tilde{\mathbf{z}}/\tilde{\epsilon}$ and $\mathbf{m}_2 = (3\tilde{\mathbf{z}}/\tilde{\epsilon}) - 4\tilde{\boldsymbol{\varrho}}/\tilde{\epsilon}^2$. It can be immediately verified that $\hat{\mathbf{w}}(\cdot)$ is absolutely continuous, that its derivative is essentially bounded, and that

$$\hat{\mathbf{w}}(t_1) = \tilde{\mathbf{z}}, \qquad \int_0^{t_1} \hat{\mathbf{w}}(t)\, dt = \tilde{\boldsymbol{\varrho}}. \tag{8.9}$$

Also,

$$\begin{aligned} \text{STV } \hat{\mathbf{w}}(\cdot) &= \frac{\tilde{\epsilon}}{2} [\| \mathbf{m}_1 \| + \| \mathbf{m}_2 \|] \\ &\leqq \frac{4 \| \tilde{\boldsymbol{\varrho}} \|}{\tilde{\epsilon}} + 2 \| \tilde{\mathbf{z}} \| \leqq \frac{4\delta}{\tilde{\epsilon}} + 2\delta = 4\tilde{\epsilon} + 2\tilde{\epsilon}^2 < 7\tilde{\epsilon}. \end{aligned} \tag{8.10}$$

Define the absolutely continuous functions $\boldsymbol{\varrho}(\cdot)$, $\tilde{\mathbf{z}}(\cdot)$, $\boldsymbol{\zeta}(\cdot)$, and $\mathbf{u}(\cdot)$

from $[0, t_1]$ to E_3 as follows:

$$
\begin{aligned}
\bar{\boldsymbol{\varrho}}(t) &= \mathbf{r}_0 + \int_0^t [\mathbf{z}(\tau; \mathbf{w}) + \mathbf{w}(\tau) + \hat{\mathbf{w}}(\tau)]\, d\tau, \\
\bar{\mathbf{z}}(t) &= \mathbf{v}_0 + \int_0^t \mathbf{G}(\bar{\boldsymbol{\varrho}}(\tau), \tau)\, d\tau, \qquad (8.11) \\
\boldsymbol{\zeta}(t) &= \mathbf{z}(t; \mathbf{w}) - \bar{\mathbf{z}}(t), \\
\mathbf{u}(t) &= \hat{\mathbf{w}}(t) + \mathbf{w}(t) + \boldsymbol{\zeta}(t).
\end{aligned}
$$

Since the function $\mathbf{G}(\cdot\,, \cdot)$ is bounded, it follows that the derivatives of $\mathbf{z}(\cdot\,; \mathbf{w})$ and of $\bar{\mathbf{z}}(\cdot)$, and consequently of $\boldsymbol{\zeta}(\cdot)$ and of $\mathbf{u}(\cdot)$, are essentially bounded. Now (8.11) and (2.3) imply that, for $0 \leqq t \leqq t_1$,

$$\bar{\boldsymbol{\varrho}}(t) = \boldsymbol{\varrho}(t; \mathbf{u}), \qquad \bar{\mathbf{z}}(t) = \mathbf{z}(t; \mathbf{u}),$$

and (8.8) therefore follows from (8.9), (2.3) and (8.11).

Further,

$$\mathbf{u}(\cdot) - \mathbf{w}(\cdot) = \hat{\mathbf{w}}(\cdot) + \boldsymbol{\zeta}(\cdot),$$

so that (see (8.10))

$$\text{STV}\,(\mathbf{u} - \mathbf{w}) \leqq \text{STV}\,\hat{\mathbf{w}} + \text{STV}\,\boldsymbol{\zeta} < 7\tilde{\epsilon} + \text{STV}\,\boldsymbol{\zeta}. \qquad (8.12)$$

Since $\hat{\mathbf{w}}(t) = 0$ for $0 \leqq t \leqq t_1 - \tilde{\epsilon}$, it is a consequence of (8.11) and (2.3) that, in this interval, $\boldsymbol{\varrho}(t; \mathbf{w}) = \bar{\boldsymbol{\varrho}}(t)$ and $\mathbf{z}(t; \mathbf{w}) = \bar{\mathbf{z}}(t)$; i.e., $\boldsymbol{\zeta}(t) = 0$ for $0 \leqq t \leqq t_1 - \tilde{\epsilon}$. Because $\boldsymbol{\zeta}(\cdot)$ is absolutely continuous,

$$\text{STV}\,\boldsymbol{\zeta} = \int_0^{t_1} \| \dot{\boldsymbol{\zeta}}(\tau) \|\, d\tau = \int_{t_1 - \tilde{\epsilon}}^{t_1} \| \dot{\boldsymbol{\zeta}}(\tau) \|\, d\tau.$$

But $\| \dot{\boldsymbol{\zeta}} \| \leqq 2\tilde{G}$ (see (8.11) and (2.3)), so that $\text{STV}\,\boldsymbol{\zeta} \leqq 2\tilde{G}\tilde{\epsilon}$; i.e., (see (8.12)),

$$\text{STV}\,(\mathbf{u} - \mathbf{w}) < (7 + 2\tilde{G})\tilde{\epsilon} \leqq \epsilon.$$

This completes the proof of Lemma 8.

We now turn to the proof of Theorem 4. Since $\tilde{\mathbf{u}}(\cdot) \in \mathfrak{BC}(t_1)$, $\tilde{\mathbf{u}}$ is of bounded variation and continuous from the right in $(0, t_1)$. Consequently, $\tilde{\mathbf{u}}(\cdot)$ has the representation $\tilde{\mathbf{u}}(\cdot) = \tilde{\mathbf{u}}_1(\cdot) + \tilde{\mathbf{u}}_2(\cdot)$, where $\tilde{\mathbf{u}}_1(\cdot)$ and $\tilde{\mathbf{u}}_2(\cdot)$ are in $\mathfrak{B}(t_1)$, $\tilde{\mathbf{u}}_1(\cdot)$ is continuous, and $\tilde{\mathbf{u}}_2(\cdot)$ is the jump function of $\tilde{\mathbf{u}}(\cdot)$. Let $\tau_1, \tau_2, \cdots$, denote the points of discontinuity of $\tilde{\mathbf{u}}(\cdot)$ (or, equivalently, of $\tilde{\mathbf{u}}_2(\cdot)$). Without loss of generality we shall assume that they are infinite in number. Let $\bar{\mathbf{v}}_i = \tilde{\mathbf{u}}_2(\tau_i) - \tilde{\mathbf{u}}_2(\tau_i^-)$ if $\tau_i \neq 0$; $\bar{\mathbf{v}}_i = \tilde{\mathbf{u}}_2(0^+) - \tilde{\mathbf{u}}_2(0)$ if $\tau_i = 0$. Then

$$\text{STV}\,\tilde{\mathbf{u}} = \text{STV}\,\tilde{\mathbf{u}}_1 + \text{STV}\,\tilde{\mathbf{u}}_2, \quad \text{STV}\,\tilde{\mathbf{u}}_2 = \sum_{i=1}^{\infty} \| \bar{\mathbf{v}}_i \| < \infty.$$

Let $s(\cdot\,;\tau)$ denote the unit step function at $t = \tau$:

$$s(t;\tau) = \begin{cases} 0 & \text{for } t < \tau, \\ 1 & \text{for } t > \tau, \end{cases}$$

$$s(\tau;\tau) = \begin{cases} 1 & \text{if } \tau \neq 0, \\ 0 & \text{if } \tau = 0. \end{cases}$$

For each positive integer n, define the absolutely continuous functions $\mathbf{w}_n'(\cdot)$, $\mathbf{w}_n''(\cdot) \in \mathcal{B}(t_1)$ as follows:

$$\mathbf{w}_n'(t) = \tilde{\mathbf{u}}_1\left(\frac{it_1}{n}\right) + \left[\frac{nt}{t_1} - i\right]\left[\tilde{\mathbf{u}}_1\left(\frac{(i+1)t_1}{n}\right) - \tilde{\mathbf{u}}_1\left(\frac{it_1}{n}\right)\right]$$

$$\text{for } \frac{it_1}{n} \leqq t \leqq (i+1)\frac{t_1}{n}, i = 0, \cdots, n-1;$$

$$\mathbf{w}_n''(\cdot) = \sum_{i=1}^{n} \bar{\mathbf{v}}_i\, s(\cdot\,;\tau_i).$$

Let $\hat{\mathbf{w}}_n(\cdot) = \mathbf{w}_n'(\cdot) + \mathbf{w}_n''(\cdot)$. It is clear that $\hat{\mathbf{w}}_n(t) \to \tilde{\mathbf{u}}(t)$ uniformly in $[0, t_1]$ as $n \to \infty$, that each $\hat{\mathbf{w}}_n(\cdot) \in \mathcal{B}(t_1)$, and that

$$\text{(8.13)} \quad \text{STV } \hat{\mathbf{w}}_n = \text{STV } \mathbf{w}_n' + \sum_{i=1}^{n} \| \bar{\mathbf{v}}_i \| \leqq \text{STV } \tilde{\mathbf{u}}_1 + \text{STV } \tilde{\mathbf{u}}_2 = \text{STV } \tilde{\mathbf{u}}.$$

For every $\epsilon > 0$ and positive integer j, define open intervals $I_{j,\epsilon}$ as follows:

$$I_{j,\epsilon} = (0, \epsilon) \quad \text{if} \quad \tau_j = 0, \qquad I_{j,\epsilon} = (\tau_j - \epsilon, \tau_j) \quad \text{if} \quad \tau_j > 0,$$

and let $r(\cdot\,;\tau_j, \epsilon)$ denote the absolutely continuous, real-valued functions defined as follows: if $\tau_j > 0$,

$$r(t;\tau_j,\epsilon) = \begin{cases} 0 & \text{for } 0 \leqq t \leqq \tau_j - \epsilon, \\ \dfrac{1}{\epsilon}(t - \tau_j + \epsilon) & \text{for } \tau_j - \epsilon \leqq t \leqq \tau_j, \\ 1 & \text{for } \tau_j \leqq t \leqq t_1; \end{cases}$$

and if $\tau_j = 0$,

$$r(t;\tau_j,\epsilon) = \begin{cases} \dfrac{t}{\epsilon} & \text{for } 0 \leqq t \leqq \epsilon, \\ 1 & \text{for } \epsilon \leqq t \leqq t_1. \end{cases}$$

Let $\epsilon_1, \epsilon_2, \cdots$, be a strictly decreasing sequence of positive numbers, with $\epsilon_n < 1/n$ for each n, such that, for every n, the intervals $I_{1,\epsilon_n}, \cdots, I_{n,\epsilon_n}$ are mutually disjoint and all contained in $[0, t_1]$. Let the absolutely con-

tinuous function $\mathbf{w}_n(\cdot) \in \mathcal{B}(t_1)$ be defined as follows:

$$\mathbf{w}_n(\cdot) = \mathbf{w}_n'(\cdot) + \sum_{i=1}^{n} r(\cdot\,; \tau_i, \epsilon_n)\bar{\mathbf{v}}_i .$$

It is easily seen that the derivative of $\mathbf{w}_n(\cdot)$ is essentially bounded, that

$$\mathbf{w}_n(t) = \hat{\mathbf{w}}_n(t) \quad \text{for} \quad t \notin \bigcup_{i=1}^{n} I_{i,\epsilon_n} ;$$

$$\| \mathbf{w}_n(t) - \hat{\mathbf{w}}_n(t) \| < \| \bar{\mathbf{v}}_i \| \quad \text{for} \quad t \in I_{i,\epsilon_n} ; \qquad i = 1, \cdots, n;$$

and that (see (8.13)),

$$\text{(8.15)} \qquad \text{STV } \mathbf{w}_n \leqq \text{STV } \mathbf{w}_n' + \sum_{i=1}^{n} \| \bar{\mathbf{v}}_i \| \leqq \text{STV } \tilde{\mathbf{u}}.$$

Now, $\| \bar{\mathbf{v}}_i \| \to 0$ as $i \to \infty$; i.e., for any fixed $\bar{\epsilon} > 0$, there is an $i' > 0$ such that $\| \bar{\mathbf{v}}_i \| < \bar{\epsilon}$ for all $i > i'$. Consequently, it follows from (8.14) that $\| \mathbf{w}_n(t) - \hat{\mathbf{w}}_n(t) \| < \bar{\epsilon}$ for all $t \notin \bigcup_{i=1}^{i'} I_{i,\epsilon_n}$. Since, for each i, $I_{i,\epsilon_n} \supset I_{i,\epsilon_{n+1}}$ and $\bigcap_{n=1}^{\infty} I_{i,\epsilon_n} = \varnothing$, we conclude that $(\mathbf{w}_n(t) - \hat{\mathbf{w}}_n(t)) \to 0$ as $n \to \infty$ for every $t \in [0, t_1]$. Recalling that $\hat{\mathbf{w}}_n(t) \to \tilde{\mathbf{u}}(t)$, we obtain that

$$\text{(8.16)} \qquad \lim_{n\to\infty} \mathbf{w}_n(t) = \tilde{\mathbf{u}}(t) \quad \text{for all} \quad t \in [0, t_1].$$

Further, since $\mathbf{w}_n(0) = 0$ for each n, it follows from (8.15) that the $\mathbf{w}_n(\cdot)$ are uniformly bounded.

Appealing to Lemmas 7 and 8, we can assert that there exist absolutely continuous functions $\mathbf{u}_n(\cdot) \in \mathcal{B}(t_1)$ with essentially bounded derivatives such that

$$\varrho(t_1\,; \mathbf{u}_n) = \varrho(t_1\,; \tilde{\mathbf{u}}),$$

$$\mathbf{z}(t_1\,; \mathbf{u}_n) + \mathbf{u}_n(t_1) = \mathbf{z}(t_1\,; \tilde{\mathbf{u}}) + \tilde{\mathbf{u}}(t_1),$$

and

$$\text{STV } (\mathbf{u}_n - \mathbf{w}_n) \to 0 \quad \text{as} \quad n \to \infty .$$

Thus, each $\mathbf{u}_n \in \mathcal{G}(t_1)$, and $(\mathbf{w}_n(t) - \mathbf{u}_n(t)) \to 0$ as $n \to \infty$ for all $t \in [0, t_1]$, which, by virtue of (8.16), implies that

$$\text{(8.17)} \qquad \lim_{n\to\infty} \mathbf{u}_n(t) = \tilde{\mathbf{u}}(t) \quad \text{for all} \quad t \in [0, t_1].$$

Now STV $\mathbf{u}_n \leqq$ STV $\mathbf{w}_n$ + STV $(\mathbf{u}_n - \mathbf{w}_n)$. Therefore (see (8.15)), $\limsup_{n\to\infty}$ STV $\mathbf{u}_n \leqq$ STV $\tilde{\mathbf{u}}$. But it follows from (8.17) that $\liminf_{n\to\infty}$ STV $\mathbf{u}_n \geqq$ STV $\tilde{\mathbf{u}}$. Consequently, $\lim_{n\to\infty}$ STV $\mathbf{u}_n$ exists and is equal to STV $\tilde{\mathbf{u}}$. This completes the proof of Theorem 4.

According to Theorems 1 and 2, there is good reason to expect that a solution $\tilde{\mathbf{u}}(\cdot)$ of the extended variational problem is a step function. The result of Theorem 4, as well as the method for constructing the approximat-

ing functions in the proof thereof, together with (2.5), indicate that, loosely speaking, *a minimum-fuel thrust program generally consists of a finite number of impulses.*

9. Bounded thrust problem and limit theorem. In this section we shall consider the original fixed terminal time variational problem described in §2 in the presence of the additional constraint that $\|\mathbf{F}(t)\| \leqq \mu$ for all t, where $\mu < \infty$ is a given positive constant. This constraint is a mathematical representation of the physical fact that the magnitude of the thrust vector of a rocket engine is always limited.

The problem may be precisely formulated as follows. Preserving the notation of §2, for given positive numbers t_1 and μ, let

$$\hat{\mathfrak{F}}(t_1, \mu) = \{\mathbf{F}(\cdot) : \mathbf{F} \in \mathfrak{F}(t_1),\ \|\mathbf{F}(t)\| \leqq \mu \quad \text{for all} \quad t \in [0, t_1]\}.$$

Then we shall consider the problem of finding (for given t_1 and μ) an element $\hat{\mathbf{F}}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$ such that $M(t_1; \hat{\mathbf{F}}) \geqq M(t_1; \mathbf{F})$ for every $\mathbf{F} \in \hat{\mathfrak{F}}(t_1, \mu)$. We shall refer to this problem as the *μ-bounded problem.* (It is to be understood in this section that all problems are fixed terminal time.)

We shall show that the μ-bounded problem always has a solution if μ is sufficiently large; that, in a certain sense, the solutions of the μ-bounded problem tend to a solution of the extended fixed terminal time problem when $\mu \to \infty$; and that the necessary conditions for the latter problem can, in a sense, be obtained by passing to the limit in the necessary conditions for the μ-bounded problem.

We shall first prove the existence theorem.

THEOREM 5. *If μ is sufficiently large, then $\hat{\mathfrak{F}}(t_1, \mu)$ is not empty and there exists an element $\hat{\mathbf{F}}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$ such that $M(t_1; \hat{\mathbf{F}}) \geqq M(t_1; \mathbf{F})$ for every $\mathbf{F} \in \hat{\mathfrak{F}}(t_1, \mu)$; i.e., the μ-bounded problem admits a solution.*

Proof. Let us show that $\hat{\mathfrak{F}}(t_1, \mu)$ is not empty for μ sufficiently large.

According to Theorems 3 and 4, there exists a function $\mathbf{u}(\cdot) \in \mathfrak{G}(t_1)$ with essentially bounded derivatives. Let STV $\mathbf{u} = \alpha$ and let $\| d\mathbf{u}(t)/dt \| < \beta < \infty$ for almost all $t \in [0, t_1]$. If $\mathbf{F}(\cdot)$ is now defined by means of relations (2.5) and (2.6), it follows that $\mathbf{F}(\cdot) \in \mathfrak{F}(t_1)$ and $\|\mathbf{F}(t)\| \leqq M_0\beta$ for almost all $t \in [0, t_1]$. Changing the values of $\mathbf{F}(t)$ for t in a set of measure zero if necessary, we conclude that $\mathbf{F}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$ whenever $\mu \geqq M_0\beta$.

Denoting r_i by x_{i+3} and $\dot{r}_i$ by x_i, where $i = 1, 2$, or 3, and $(-M)$ by x_7, (2.1) and (2.2) may be rewritten as follows:

$$\begin{aligned} \dot{x}_i(t) &= x_{i-3}(t), && i = 4, 5, 6, \\ \dot{x}_i(t) &= G_i(x_4, x_5, x_6, t) - \frac{T_i(t)}{x_7(t)}, && i = 1, 2, 3, \\ \dot{x}_7(t) &= \frac{\|\mathbf{T}(t)\|}{A}, \quad \mathbf{T} = (T_1, T_2, T_3). \end{aligned} \tag{9.1}$$

In order to show the existence of an element $\hat{\mathbf{F}}$ in $\mathfrak{F}(t_1, \mu)$ that maximizes $M(t_1; \mathbf{F})$, we shall first consider a system which, in a sense, is more general than (9.1), and which, following Warga [12], we shall refer to as the relaxed system. Namely, we consider the following system of equations for the scalar variables $y_1, \cdots, y_7$:

$$
\begin{aligned}
\dot{y}_i(t) &= y_{i-3}(t), && i = 4, 5, 6,\\
\dot{y}_i(t) &= G_i(y_4, y_5, y_6, t) - \frac{T_i(t)}{y_7(t)}, && i = 1, 2, 3,\\
\dot{y}_7(t) &= \frac{\| \mathbf{T}(t) \| + \tilde{T}(t)}{A}, \quad \mathbf{T} = (T_1, T_2, T_3).
\end{aligned} \tag{9.2}
$$

If $\mathbf{F}(\cdot)$ is a measurable function from $[0, \infty)$ to E_3 and $\tilde{F}(\cdot)$ is a summable function from $[0, \infty)$ to E_1 such that the inequality $\int_0^\infty [\| \mathbf{F}(t)\| + \tilde{F}(t)]\, dt < AM_0$ holds, we shall denote by $(y_1(t; \mathbf{F}, \tilde{F}), \cdots, y_7(t; \mathbf{F}, \tilde{F})) = \mathbf{y}(t; \mathbf{F}, \tilde{F})$ the solution for $t \geqq 0$ of (9.2), with $\mathbf{T}(t) = \mathbf{F}(t)$ and $\tilde{T}(t) = \tilde{F}(t)$, that assumes the initial values $(y_1(0; \mathbf{F}, \tilde{F}), \cdots, y_3(0; \mathbf{F}, \tilde{F})) = \mathbf{v}_0$, $(y_4(0; \mathbf{F}, \tilde{F}), \cdots, y_6(0; \mathbf{F}, \tilde{F})) = \mathbf{r}_0$, $y_7(0; \mathbf{F}, \tilde{F}) = -M_0$. Let $\mathfrak{F}_R(t_1, \mu)$ denote the class of all pairs $(\mathbf{F}(\cdot), \tilde{F}(\cdot))$ such that

(1) $\mathbf{F}(\cdot)$ is a measurable function from $[0, t_1]$ to E_3,

(2) $\tilde{F}(\cdot)$ is a summable function from $[0, t_1]$ into $[0, \infty)$,

(3) $\int_0^{t_1} [\| \mathbf{F}(t)\| + \tilde{F}(t)]\, dt < AM_0$,

(4) $\| \mathbf{F}(t)\| + \tilde{F}(t) \leqq \mu$ for all $t \in [0, t_1]$, and

(5) $h_i(y_4(t_1; \mathbf{F}, \tilde{F}), \cdots, y_6(t_1; \mathbf{F}, \tilde{F}), y_1(t_1; \mathbf{F}, \tilde{F}), \cdots, y_3(t_1; \mathbf{F}, \tilde{F}), t_1) = 0$ for $i = 1, \cdots, \nu$.

Then we shall consider the *relaxed μ-bounded problem* which consists in finding an element $(\hat{\mathbf{F}}(\cdot), \bar{F}(\cdot)) \in \mathfrak{F}_R(t_1, \mu)$ such that $y_7(t_1; \hat{\mathbf{F}}, \bar{F}) \leqq y_7(t_1; \mathbf{F}, \tilde{F})$ for every $(\mathbf{F}(\cdot), \tilde{F}(\cdot)) \in \mathfrak{F}_R(t_1, \mu)$. According to a result of Warga [12, Theorem 3.3], such a minimizing element always exists so long as $\mathfrak{F}_R(t_1, \mu)$ is not empty, and we shall show below that this set is not empty whenever μ is sufficiently large.

Indeed, it follows at once from (9.1) and (9.2) that, if $\tilde{F}_0(\cdot)$ denotes the function which vanishes identically, then, for every measurable function $\mathbf{F}(\cdot)$ from $[0, t_1]$ to E_3 with $\int_0^{t_1} \|\mathbf{F}(t)\|\, dt < AM_0$,

$$
\begin{aligned}
(y_1(t; \mathbf{F}, \tilde{F}_0), \cdots, y_3(t; \mathbf{F}, \tilde{F}_0)) &= \dot{\mathbf{r}}(t; \mathbf{F}),\\
(y_4(t; \mathbf{F}, \tilde{F}_0), \cdots, y_6(t; \mathbf{F}, \tilde{F}_0)) &= \mathbf{r}(t; \mathbf{F}),\\
y_7(t; \mathbf{F}, \tilde{F}_0) &= -M(t; \mathbf{F}),
\end{aligned}
$$

for every $t \in [0, t_1]$. It then follows at once that $\mathbf{F}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$ if and only if $\mathbf{F}(\cdot), \bar{F}_0(\cdot) \in \mathfrak{F}_R(t_1, \mu)$. Since we have shown that $\hat{\mathfrak{F}}(t_1, \mu)$ is not empty for μ sufficiently large, we conclude that $\mathfrak{F}_R(t_1, \mu)$ is not empty for μ large enough.

We shall prove below that if $(\hat{\mathbf{F}}(\cdot), \bar{F}(\cdot))$ is a solution of the relaxed μ-bounded problem (and we have shown that such a solution exists if μ is sufficiently large), then $\bar{F}(t) = \bar{F}_0(t) = 0$ for almost all $t \in [0, t_1]$. By what was said above, this will imply that $\hat{\mathbf{F}}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$ and that

$$M(t_1; \hat{\mathbf{F}}) = -y_7(t_1; \hat{\mathbf{F}}, \bar{F}_0) \geqq -y_7(t_1; \mathbf{F}, \bar{F}_0) = M(t_1; \mathbf{F})$$

for every $\mathbf{F}(\cdot) \in \hat{\mathfrak{F}}(t_1, \mu)$; i.e., $\hat{\mathbf{F}}(\cdot)$ is a solution of the μ-bounded problem.

It easily follows from the Pontryagin maximum principle [8, Chap. 1] that if $(\hat{\mathbf{F}}(\cdot), \bar{F}(\cdot))$ is a solution of the relaxed μ-bounded problem, then there exist a twice differentiable function $\boldsymbol{\psi}(\cdot)$ from $[0, t_1]$ to E_3 and an absolutely continuous function $\psi(\cdot)$ from $[0, t_1]$ to E_1, with $\| \boldsymbol{\psi}(\cdot)\| + \psi(\cdot)$ not vanishing identically, such that

$$\frac{d^2\boldsymbol{\psi}(t)}{dt^2} = \left(\frac{\partial \mathbf{G}(y_4(t; \hat{\mathbf{F}}, \bar{F}), \cdots, y_6(t; \hat{\mathbf{F}}, \bar{F}), t)}{\partial \mathbf{r}}\right)^T \boldsymbol{\psi}(t), \qquad 0 \leqq t \leqq t_1; \tag{9.3}$$

$$\frac{d\boldsymbol{\psi}(t)}{dt} = -[y_7(t; \hat{\mathbf{F}}, \bar{F})]^{-2}[\boldsymbol{\psi}(t) \cdot \hat{\mathbf{F}}(t)] \text{ for almost all } t, \qquad 0 \leqq t \leqq t_i; \tag{9.4}$$

$$\psi(t_1) \leqq 0; \tag{9.5}$$

$$\psi(t)\frac{\|\hat{\mathbf{F}}(t)\| + \bar{F}(t)}{A} - \frac{\boldsymbol{\psi}(t)\cdot\hat{\mathbf{F}}(t)}{y_7(t; \hat{\mathbf{F}}, \bar{F})}$$

$$= \max_{\substack{\mathbf{V}\in E_3, V \geqq 0 \\ \|\mathbf{V}\| + V \leqq \mu}} \left[\psi(t)\frac{\|\mathbf{V}\| + V}{A} - \frac{\boldsymbol{\psi}(t)\cdot\mathbf{V}}{y_7(t; \hat{\mathbf{F}}, \bar{F})}\right], \text{ for almost all } t, \tag{9.6}$$

$$0 \leqq t \leqq t_1.$$

It follows from (9.6) that, for almost all $t \in [0, t_1]$, either $\bar{F}(t) = 0$ or $\|\boldsymbol{\psi}(t)\| = 0$. Hence, if the zeros of $\|\boldsymbol{\psi}(\cdot)\|$ are isolated, then $\bar{F}(t) = 0$ for almost all t. If the zeros of $\| \boldsymbol{\psi}(\cdot)\|$ in $[0, t_1]$ are not isolated, then $\boldsymbol{\psi}(t) \equiv 0$. For, if t_2 is an accumulation point of zeros of $\| \boldsymbol{\psi}(\cdot)\|$, then $\boldsymbol{\psi}(t_2) = \dot{\boldsymbol{\psi}}(t_2) = 0$, which, because of the uniqueness of solutions of (9.3), implies that $\boldsymbol{\psi}(t) \equiv 0$. If $\boldsymbol{\psi}(t) \equiv 0$, it follows from (9.4) and (9.5) and the fact that $\boldsymbol{\psi}$ and ψ cannot both vanish identically, that $\psi(t) = \text{const.} < 0$, and (9.6) then implies that $\| \hat{\mathbf{F}}(t)\| = \bar{F}(t) = 0$ almost everywhere in $[0, t_1]$.

According to the Pontryagin maximum principle as applied to (9.1), if $\hat{\mathbf{F}}(\cdot)$ is any solution of the μ-bounded problem, there exist a twice differentiable function $\boldsymbol{\psi}(\cdot)$ from $[0, t_1]$ to E_3 and an absolutely continuous function $\psi(\cdot)$ from $[0, t_1]$ to E_1, with $\| \boldsymbol{\psi}(\cdot)\| + \psi(\cdot)$ not vanishing identically,

such that

$$\ddot{\psi}(t) = \left(\frac{\partial \mathbf{G}(\mathbf{r}(t; \hat{\mathbf{F}}), t)}{\partial \mathbf{r}}\right)^T \psi(t), \qquad 0 \leqq t \leqq t_1; \tag{9.7}$$

$$\dot{\psi}(t) = -[M(t; \hat{\mathbf{F}})]^{-2}[\psi(t)\cdot\hat{\mathbf{F}}(t)] \quad \text{for almost all } t, \quad 0 \leqq t \leqq t_1; \tag{9.8}$$

$$\psi(t_1) \leqq 0; \tag{9.9}$$

$$\frac{\psi(t)\,\|\hat{\mathbf{F}}(t)\|}{A} + \frac{\psi(t)\cdot\hat{\mathbf{F}}(t)}{M(t; \hat{\mathbf{F}})} = \max_{\substack{\mathbf{V}\in E_3 \\ \|\mathbf{V}\|\leqq\mu}} \left[\frac{\psi(t)\,\|\mathbf{V}\|}{A} + \frac{\psi(t)\cdot\mathbf{V}}{M(t; \hat{\mathbf{F}})}\right], \text{ for almost all } t, \quad 0 \leqq t \leqq t_1; \tag{9.10}$$

$$\dot{\psi}(t_1)\cdot[\mathbf{x} - \mathbf{r}(t_1; \hat{\mathbf{F}})] = \psi(t_1)\cdot[\mathbf{y} - \dot{\mathbf{r}}(t_1; \hat{\mathbf{F}})] \quad \text{for all} \quad (\mathbf{x}, \mathbf{y}) \in T, \tag{9.11}$$

where T is the hyperplane tangent to $H(t_1)$ at $(\mathbf{r}(t_1; \hat{\mathbf{F}}), \dot{\mathbf{r}}(t_1; \hat{\mathbf{F}}))$. We shall say that such a pair $(\psi(\cdot), \psi(\cdot))$ is an adjoint function which corresponds to $\hat{\mathbf{F}}(\cdot)$.

It follows from (9.10) that, for almost all $t \in [0, t_1]$,

$$\hat{\mathbf{F}}(t) = \begin{cases} 0 & \text{if} \quad \|\psi(t)\| < \tilde{\psi}(t), \\ \alpha\psi(t), \text{ where } 0 \leqq \alpha \leqq \dfrac{\mu}{\|\psi(t)\|}, & \text{if} \quad \|\psi(t)\| = \tilde{\psi}(t) \neq 0, \\ \dfrac{\mu\psi(t)}{\|\psi(t)\|} & \text{if} \quad 0 \neq \|\psi(t)\| > \tilde{\psi}(t), \end{cases} \tag{9.12}$$

$$\|\hat{\mathbf{F}}(t)\| = \mu \qquad \text{if} \quad 0 = \|\psi(t)\| > \tilde{\psi}(t),$$

where

$$\tilde{\psi}(t) = -\frac{M(t; \hat{\mathbf{F}})\psi(t)}{A}. \tag{9.13}$$

It is easily verified that $\tilde{\psi}(\cdot)$ is differentiable on $[0, t_1]$ and that (see (9.8) and (2.2))

$$d\tilde{\psi}(t) = \begin{cases} 0 & \text{if} \quad \|\psi(t)\| - \tilde{\psi}(t) \leqq 0, \\ \dfrac{\mu}{AM(t; \hat{\mathbf{F}})}\,[\|\psi(t)\| - \tilde{\psi}(t)] & \text{if} \quad \|\psi(t)\| - \tilde{\psi}(t) \geqq 0. \end{cases} \tag{9.14}$$

Hence, $d\tilde{\psi}(t)/dt \geqq 0$ for all t, $0 \leqq t \leqq t_1$.

Thus, let $\hat{\mathbf{F}}(\cdot)$ be a solution of the μ-bounded problem, let $(\psi(\cdot), \psi(\cdot))$ be a corresponding adjoint function, and let $\tilde{\psi}(\cdot)$ be defined by (9.13). It follows as before that if $\psi(t) \not\equiv 0$ on $[0, t_1]$, then the zeros of $\|\psi(\cdot)\|$ are isolated. Further, if $\psi(t) \equiv 0$ on $[0, t_1]$, then (see (9.8), (9.9), (9.13) and (9.12)) $\psi(t) = \text{const.} < 0$, $\tilde{\psi}(t) > 0$ for all $t \in [0, t_1]$, and $\hat{\mathbf{F}}(t) = 0$ almost

everywhere in $[0, t_1]$. Note that $(\alpha\psi(\cdot), \alpha\psi(\cdot))$ is an adjoint function corresponding to $\hat{\mathbf{F}}(\cdot)$ for every $\alpha > 0$. Thus, if $\{t:\hat{\mathbf{F}}(t) \neq 0,\ 0 \leqq t \leqq t_1\}$ is of positive measure, then $\psi(t) \neq 0$ for some $t \in [0, t_1]$, and, multiplying $(\psi(\cdot), \psi(\cdot))$ by a suitable positive constant if necessary, we may assume that

$$\max_{0 \leqq t \leqq t_1} \| \psi(t) \| = 1. \tag{9.15}$$

Adjoint functions $(\psi(\cdot), \psi(\cdot))$ that satisfy (9.15) will be said to be *normalized.*

Let $\mathbf{F}_0(\cdot)$ denote the function from $[0, t_1]$ to E_3 defined by $\mathbf{F}_0(t) = 0$ for all t, $0 \leqq t \leqq t_1$.

We now prove the following limit theorem.

THEOREM 6. *Let $\tilde{\mu}_1, \tilde{\mu}_2, \cdots$, be a sequence of positive numbers such that $\tilde{\mu}_i \to \infty$ as $i \to \infty$. Then, if $\mathbf{F}_0(\cdot) \notin \mathfrak{F}(t_1)$, there exist a subsequence $\mu_1, \mu_2, \cdots$, of the $\tilde{\mu}_i$, solutions $\mathbf{F}_i(\cdot)$ of the μ_i-bounded problems, normalized adjoint functions $(\bar{\psi}_i(\cdot), \bar{\psi}_i(\cdot))$ corresponding to the $\mathbf{F}_i$, and functions $\tilde{\mathbf{u}}(\cdot)$ and $\psi(\cdot)$ from $[0, t_1]$ to E_3 such that $\psi(\cdot)$ is twice differentiable and $\tilde{\mathbf{u}}(\cdot)$ is a solution of the extended fixed terminal time problem,*

$$\bar{\psi}_i(t) \to \psi(t) \text{ as } i \to \infty, \tag{9.16}$$

$$\frac{-M(t; \mathbf{F}_i)\bar{\psi}_i(t)}{A} \to 1 \text{ as } i \to \infty, \tag{9.17}$$

$$\int_0^t \frac{\mathbf{F}_i(s)}{M(s; \mathbf{F}_i)}\, ds \to \tilde{\mathbf{u}} \text{ as } i \to \infty, \tag{9.18}$$

$$\mathbf{r}(t; \mathbf{F}_i) \to \varrho(t; \tilde{\mathbf{u}}) \quad \text{as} \quad i \to \infty, \tag{9.19}$$

$$\dot{\mathbf{r}}(t; \mathbf{F}_i) \to \mathbf{z}(t; \tilde{\mathbf{u}}) + \tilde{\mathbf{u}}(t) \quad \text{as} \quad i \to \infty, \tag{9.20}$$

$$M(t_1; \mathbf{F}_i) \to M_0 \exp\left[-A^{-1}\,\mathrm{STV}\,\tilde{\mathbf{u}}\right] \quad \text{as} \quad i \to \infty, \tag{9.21}$$

where (9.16), (9.17), *and* (9.19) *hold for every $t \in [0, t_1]$ and the convergence is uniform with respect to $t \in [0, t_1]$; and* (9.18) *and* (9.20) *hold for almost all $t \in [0, t_1]$ including 0, t_1, and the points of continuity of $\tilde{\mathbf{u}}(\cdot)$. Also, $\tilde{\mathbf{u}}(\cdot)$ and $\psi(\cdot)$ satisfy* (5.1)–(5.4), *where $\tilde{\varrho}(t) = \varrho(t; \tilde{\mathbf{u}})$, and $\bar{\mathbf{r}}$, $\bar{\mathbf{v}}$, and T are defined as in the statement of Theorem* 1.

Proof. According to Theorem 5, solutions of the μ-bounded problem exist when $\mu \geqq \mu^*$, where $\mu^* < \infty$ is a sufficiently large positive constant. For every $\mu \geqq \mu^*$, let $\hat{\mathbf{F}}_\mu(\cdot) \in \mathfrak{F}(t_1)$ be a solution of the μ-bounded problem. Since $\mathbf{F}_0(\cdot) \notin \mathfrak{F}(t_1)$, $\{t:\hat{\mathbf{F}}_\mu(t) \neq 0,\ 0 \leqq t \leqq t_1\}$ is of positive measure for every $\mu \geqq \mu^*$, and, by virtue of the immediately preceding discussion, there exists a normalized adjoint function, which we shall denote by $(\psi_\mu(\cdot), \psi_\mu(\cdot))$, corresponding to each $\hat{\mathbf{F}}_\mu$ with $\mu \geqq \mu^*$.

Let $\mathbf{u}_\mu(\cdot) = \phi(\hat{\mathbf{F}}_\mu(\cdot))$; i.e.,

$$\mathbf{u}_\mu(t) = \int_0^t \frac{\hat{\mathbf{F}}_\mu(s)}{M(s; \hat{\mathbf{F}}_\mu)}\, ds, \qquad 0 \leqq t \leqq t_1. \tag{9.22}$$

It follows from the discussion in §2 that, for each $\mu \geqq \mu^*$, $\mathbf{u}_\mu(\cdot) \in \mathcal{G}(t_1)$, and (see (2.6), (2.10), and (2.14))

$$\text{STV } \mathbf{u}_\mu(\cdot) = A \ln\left[\frac{M_0}{M(t_1; \hat{\mathbf{F}}_\mu)}\right]. \tag{9.23}$$

Since $\hat{\mathbf{F}}_{\mu^*}(\cdot) \in \mathfrak{F}(t_1, \mu)$ when $\mu \geqq \mu^*$,

$$M(t_1; \hat{\mathbf{F}}_\mu) \geqq M(t_1; \hat{\mathbf{F}}_{\mu^*}) \quad \text{for} \quad \mu \geqq \mu^*. \tag{9.24}$$

By virtue of (9.23) and (9.24), we can conclude that STV $\mathbf{u}_\mu \leqq$ STV $\mathbf{u}_{\mu^*}$ for all $\mu \geqq \mu^*$.

The functions $\psi_\mu(\cdot)$ are uniformly bounded by definition, and satisfy (9.7) with $\hat{\mathbf{F}}$ replaced by $\hat{\mathbf{F}}_\mu$. Since $\mathbf{G}(\cdot, \cdot)$ has bounded first partial derivatives the functions $\ddot{\psi}_\mu(\cdot)$ are also uniformly bounded. But if the ψ_μ and the $\ddot{\psi}_\mu$ are uniformly bounded, then the functions $\dot{\psi}_\mu(\cdot)$ must also possess this property. Consequently, the functions ψ_μ and $\dot{\psi}_\mu$ for $\mu \geqq \mu^*$ are equicontinuous as well as uniformly bounded.

Just as was done in the proof of Theorem 3, we can now show with the aid of (2.9) that there exist functions $\tilde{\mathbf{u}}(\cdot)$, $\psi(\cdot)$ and a subsequence $\mu_1, \mu_2, \cdots$, of the μ_i such that $\tilde{\mathbf{u}} \in \mathcal{K}(t_1)$, and, denoting ψ_{μ_i} by $\bar{\psi}_i$ and $\hat{\mathbf{F}}_{\mu_i}$ by F_i, such that:

(a) (9.16) and (9.19) are satisfied uniformly in $[0, t_1]$;

(b) (9.18) and (9.20) hold for almost all $t \in [0, t_1]$ including 0, t_1 and the points of continuity of $\tilde{\mathbf{u}}(\cdot)$;

(c) (5.1), (5.2), and (5.4) are satisfied; and

(d) $\lim_{i\to\infty}$ STV $\mathbf{u}_{\mu_i}$ exists and

$$\text{STV } \tilde{\mathbf{u}} \leqq \lim_{i\to\infty} \text{STV } \mathbf{u}_{\mu_i}. \tag{9.25}$$

We now shall verify (9.21). Let $\hat{\mathbf{u}}(\cdot) \in \mathcal{K}(t_1)$ be a solution of the extended fixed terminal time problem. Then (see Theorem 4) there exist functions $\hat{\mathbf{u}}_i(\cdot) \in \mathcal{G}(t_1)$, $i = 1, 2, \cdots$, and positive constants $M_1, M_2, \cdots$, such that $\lim_{n\to\infty}$ STV $\hat{\mathbf{u}}_n =$ STV $\hat{\mathbf{u}}$, and $\| d\hat{\mathbf{u}}_i(t)/dt \| < M_i < \infty$ for almost all $t \in [0, t_1]$ and every $i = 1, 2, \cdots$. Setting $\tilde{\mathbf{F}}_n = \Psi(\hat{\mathbf{u}}_n)$, we conclude on the basis of (2.5) and (2.6) that $\| \tilde{\mathbf{F}}_n(t) \| \leqq M_0 M_n$ almost everywhere in $[0, t_1]$, i.e., (modifying the values of $\tilde{\mathbf{F}}_n(t)$ for t in a set of measure zero, if necessary) $\tilde{\mathbf{F}}_n \in \mathfrak{F}(t_1, \mu)$ for every $\mu \geqq M_0 M_n$. Consequently, for each $n = 1, 2, \cdots$, there is an integer I, depending on n, such that $M(t_1; \mathbf{F}_i)$

$\geqq M(t_1 ; \tilde{\mathbf{F}}_n)$ for every $i \geqq I(n)$, or (see (9.23)), STV $\mathbf{u}_{\mu_i} \leqq$ STV $\hat{\mathbf{u}}_n$ for every $i \geqq I(n)$. Therefore (see (9.25)),

$$\text{STV } \tilde{\mathbf{u}} \leqq \lim_{i \to \infty} \text{STV } \mathbf{u}_{\mu_i} \leqq \lim_{i \to \infty} \text{STV } \hat{\mathbf{u}}_i = \text{STV } \hat{\mathbf{u}}.$$

But $\hat{\mathbf{u}} \in \mathfrak{R}(t_1)$ and $\hat{\mathbf{u}}$ is a solution of the extended fixed terminal time problem, so that STV $\hat{\mathbf{u}} \leqq$ STV $\tilde{\mathbf{u}}$. Therefore, STV $\tilde{\mathbf{u}} =$ STV $\hat{\mathbf{u}}$, and $\tilde{\mathbf{u}}$ is a solution of this problem also. Consequently, (9.25) is actually an equality, and (9.21) now follows at once from (9.23).

We now verify (9.17). Denote $\psi_{\mu_i}(\cdot)$ by $\bar{\psi}_i(\cdot)$, and let

$$\tilde{\psi}_i(t) = -\frac{M(t; \mathbf{F}_i)\bar{\psi}_i(t)}{A}, \qquad 0 \leqq t \leqq t_1, \qquad i = 1, 2, \cdots. \tag{9.26}$$

We shall show that $\tilde{\psi}_i(0) \to 1$ as $i \to \infty$, and that $\tilde{\psi}_i(t) \leqq 1$ for every $i = 1, 2, \cdots$, and $t \in [0, t_1]$. Since each $\tilde{\psi}_i$ is differentiable, and $d\tilde{\psi}_i(t)/dt \geqq 0$ for each i and t, this will imply that $\tilde{\psi}_i(t) \to 1$ uniformly in $[0, t_1]$ as $i \to \infty$, i.e., that (9.17) holds uniformly in $[0, t_1]$.

Let $E_i = \{t : \| \bar{\psi}_i(t) \| > \tilde{\psi}_i(t),\ 0 \leqq t \leqq t_1\}$, and let $| E_i |$ denote the Lebesgue measure of E_i. According to (9.12), $\|\mathbf{F}_i(t)\| = \mu_i$ when $t \in E_i$, so that (see (2.2)) $0 \leqq M(t_1 ; \mathbf{F}_i) \leqq M_0 - \mu_i | E_i |/A$. Consequently,

$$\mu_i | E_i | \leqq AM_0, \qquad i = 1, 2, \cdots, \tag{9.27}$$

and, since $\lim_{i \to \infty} \mu_i = \infty$, $| E_i | \to 0$ as $i \to \infty$.

Now suppose that $\tilde{\psi}_i(\bar{t}) > 1$ for some $\bar{t} \in [0, t_1]$ and some $i = 1, 2, \cdots$. Since $\tilde{\psi}_i(\cdot)$ satisfies (9.14) with $\mathbf{F}$ replaced by $\mathbf{F}_i$, μ by μ_i, and ψ by $\bar{\psi}_i$, and $\| \bar{\psi}_i(t) \| \leqq 1$ for every $t \in [0, t_1]$, it follows that $\tilde{\psi}_i(t) = \tilde{\psi}_i(\bar{t}) > 1$ for every $t \in [0, t_1]$, and, by virtue of (9.12), that $\mathbf{F}_i(t) = 0$ for almost all $t \in [0, t_1]$. This implies that $\mathbf{F}_0 \in \mathfrak{F}(t_1)$, which is a contradiction, so that $\tilde{\psi}_i(t) \leqq 1$ for every i and t.

Because

$$M(t; \mathbf{F}_i) \geqq M(t_1 ; \mathbf{F}_i) \geqq M(t_1 ; \hat{\mathbf{F}}_{\mu^*}) = M^*$$

(see (9.24)), and $\| \bar{\psi}_i(t) \| \leqq 1$ for each i and $t \in [0, t_1]$, it follows from (9.14) that, for every i and t,

$$\frac{d\tilde{\psi}_i(t)}{dt} \leqq \frac{\mu_i}{AM^*}[1 - \tilde{\psi}_i(t)]c_{E_i}(t),$$

where $c_{E_i}(t)$ is the characteristic function of E_i. Thus,

$$\frac{d\tilde{\psi}_i(t)}{dt} = \frac{\mu_i}{AM^*}[1 - \tilde{\psi}_i(t)]c_{E_i}(t) - \tilde{\phi}_i(t), \qquad 0 \leqq t \leqq t_1. \tag{9.28}$$

where $\tilde{\phi}_i(t) \geqq 0$ for all $t \in [0, t_1]$. But the solution of (9.28) is given by

$$\tilde{\psi}_i(t) = 1 - \left\{\exp\left[-\int_0^t \frac{\mu_i}{AM^*} c_{E_i}(s)\, ds\right]\right\} \cdot \left\{1 - \tilde{\psi}_i(0) + \int_0^t \tilde{\phi}_i(s) \exp\left[\int_0^s \frac{\mu_i}{AM^*} c_{E_i}(\tau)\, d\tau\right] ds\right\},$$

so that

$$\tilde{\psi}_i(t) \leqq 1 - [1 - \tilde{\psi}_i(0)] \exp\left[-\frac{\mu_i}{AM^*} |E_i|\right] \leqq 1 - [1 - \tilde{\psi}(0)]A^*, \qquad 0 \leqq t \leqq t_1, \tag{9.29}$$

where $A^* = \exp[-M_0/M^*] > 0$ (see (9.27)). We shall show that (9.29) implies that $\tilde{\psi}_i(0) \to 1$ as $i \to \infty$.

For each i, let τ_i be any value in $[0, t_1]$ such that $\| \bar{\Psi}_i(\tau_i)\| = 1$. (Such values exist by definition of the $\bar{\Psi}_i = \Psi_{\mu_i}$.) If $\tilde{\psi}_i(\tau_i) = 1$, it follows from (9.29) and the fact that $\tilde{\psi}_i(0) \leqq 1$ that $\tilde{\psi}_i(0) = 1$. Now suppose that $\tilde{\psi}_i(\tau_i) < 1$. Then $\tau_i \in E_i$, and let (τ_i', τ_i'') be the largest open interval contained in E_i which contains τ_i in its closure. We shall suppose that i is sufficiently large so that $(\tau_i', \tau_i'') \neq (0, t_1)$. Then, $\| \bar{\Psi}_i(\tau_i')\| = \tilde{\psi}_i(\tau_i')$ and/or $\| \bar{\Psi}_i(\tau_i'')\| = \tilde{\psi}_i(\tau_i'')$. Without loss of generality, we shall assume that $\| \bar{\Psi}_i(\tau_i')\| = \tilde{\psi}_i(\tau_i')$. Since $0 \leqq \tau_i - \tau_i' \leqq |E_i|$, and there is a constant $K > 0$ such that $\| d\bar{\Psi}_i(t)/dt \| < K$ for each $i = 1, 2, \cdots$ and $t \in [0, t_1]$ we have that

$$0 \leqq 1 - \| \bar{\Psi}_i(\tau_i')\| = \| \bar{\Psi}_i(\tau_i)\| - \| \bar{\Psi}_i(\tau_i')\| = \int_{\tau_i'}^{\tau_i} \frac{d}{dt} \| \bar{\Psi}_i(t)\|\, dt \leqq \int_{\tau_i'}^{\tau_i} \left\| \frac{d}{dt} \bar{\Psi}_i(t) \right\| dt < K|E_i|.$$

But $|E_i| \to 0$ as $i \to \infty$, so that $1 - \| \bar{\Psi}_i(\tau_i')\| = 1 - \tilde{\psi}_i(\tau_i')$ will be nonnegative and arbitrarily small if i is sufficiently large. Consequently we can conclude, by virtue of (9.29), that $\lim_{i\to\infty} [1 - \tilde{\psi}_i(0)]A^* = 0$, i.e., that $\lim_{i\to\infty} \tilde{\psi}_i(0) = 1$. This completes the verification of (9.17).

It only remains to prove that (5.3) holds.

Let $G_i = \{t: \|\bar{\Psi}_i(t)\| \geqq \tilde{\psi}_i(t),\ 0 \leqq t \leqq t_1\}$. It follows from (9.22) and (9.12) that $d\mathbf{u}_{\mu_i}(s)/dt = 0$ when $s \notin G_i$, $0 \leqq s \leqq 1$, and that, for almost all $s \in G_i$, either $d\mathbf{u}_{\mu_i}(s)/dt$ is a nonnegative scalar multiple of $\bar{\Psi}_i(s)$, or $\bar{\Psi}_i(s) = 0$. Hence,

$$\int_0^{t_1} [\bar{\Psi}_i(t)]^T \frac{d\mathbf{u}_{\mu_i}(t)}{dt}\, dt = \int_{G_i} \| \bar{\Psi}_i(t)\| \cdot \left\| \frac{d\mathbf{u}_{\mu_i}(t)}{dt} \right\| dt. \tag{9.30}$$

If $t \in G_i$, $1 \geqq \| \bar{\Psi}_i(t)\| \geqq \tilde{\psi}_i(t) \geqq \tilde{\psi}_i(0)$, so that, by virtue of (9.30),

$$\tilde{\psi}_i(0) \int_{G_i} \left\| \frac{d\mathbf{u}_{\mu_i}(t)}{dt} \right\| dt \leqq \int_0^{t_1} [\bar{\Psi}_i(t)]^T \frac{d\mathbf{u}_{\mu_i}(t)}{dt}\, dt \leqq \int_{G_i} \left\| \frac{d\mathbf{u}_{\mu_i}(t)}{dt} \right\| dt. \tag{9.31}$$

Since (9.25) has been shown to be an equality, we have, by virtue of (2.14), that

$$\int_{\sigma_i} \left\| \frac{d\mathbf{u}_{\mu_i}(t)}{dt} \right\| dt = \int_0^{t_1} \left\| \frac{d\mathbf{u}_{\mu_i}(t)}{dt} \right\| dt = \text{STV } \mathbf{u}_{\mu_i} \to \text{STV } \tilde{\mathbf{u}} \quad \text{as} \quad i \to \infty. \tag{9.32}$$

Also, $\tilde{\psi}_i(0) \to 1$ as $i \to \infty$, so that (9.31) and (9.32) yield that

$$\int_0^{t_1} [\bar{\psi}_i(t)]^T \frac{d\mathbf{u}_{\mu_i}(t)}{dt}\, dt \to \text{STV } \tilde{\mathbf{u}} \quad \text{as} \quad i \to \infty. \tag{9.33}$$

Finally, it follows from (9.16), (9.18), (9.22) and the Helly-Bray Theorem [6, p. 288] that

$$\int_0^{t_1} [\bar{\psi}_i(t)]^T \frac{d\mathbf{u}_{\mu_i}(t)}{dt}\, dt = \int_0^{t_1} [\bar{\psi}_i(t)[^T\, d\mathbf{u}_{\mu_i}(t) \to \int_0^{t_1} [\psi(t)]^T\, d\tilde{\mathbf{u}}(t) \quad \text{as} \quad i \to \infty, \tag{9.34}$$

and (5.3) is now an immediate consequence of (9.33) and (9.34).

COROLLARY 1. *Let $\tilde{\mathbf{u}}$ be a solution of the extended fixed terminal time problem, and suppose that $\tilde{\mathbf{u}} \neq \mathbf{F}_0$. Then, if $\hat{\mathbf{F}}_\mu$ is any solution of the μ-bounded problem (for every μ sufficiently large), we have that*

(1) $$M(t_1\,; \hat{\mathbf{F}}_\mu) \to M_0 \exp[-A^{-1}\,\text{STV } \tilde{\mathbf{u}}] \quad \textit{as} \quad \mu \to \infty.$$

Further, if $\tilde{\mathbf{u}}(\cdot)$ is the unique solution of the extended fixed terminal time problem, and $\mathbf{u}_\mu$ is given by (9.22), then as $\mu \to \infty$,

(2) $$\mathbf{u}_\mu(t) \to \tilde{\mathbf{u}}(t),$$

(3) $$\dot{\mathbf{r}}(t; \hat{\mathbf{F}}_\mu) \to \mathbf{z}(t; \tilde{\mathbf{u}}) + \tilde{\mathbf{u}}(t),$$

(4) $$\mathbf{r}(t; \hat{\mathbf{F}}_\mu) \to \boldsymbol{\varrho}(t; \tilde{\mathbf{u}}),$$

where (4) holds uniformly in $[0, t_1]$, and (2) and (3) hold for almost all $t \in [0, t_1]$ including $0, t_1$, and the points of continuity of $\tilde{\mathbf{u}}(\cdot)$. If, in addition the function $\psi(\cdot)$ that satisfies relations (5.1)–(5.4) is unique, and (ψ_μ, ψ_μ) is any normalized adjoint function corresponding to $\hat{\mathbf{F}}_\mu$, then the following limits exist uniformly in $[0, t_1]$ as $\mu \to \infty$:

(5) $$\psi_\mu(t) \to \psi(t),$$

(6) $$-M(t; \hat{\mathbf{F}}_\mu)\psi_\mu(t)/A \to 1.$$

The proof of the corollary is straightforward and is therefore omitted.

REFERENCES

[1] D. F. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworths, London, 1963.

[2] G. M. EWING, *A fundamental problem of navigation in free space*, Quart. Appl. Math., 18 (1961), pp. 355–362.

[3] L. W. NEUSTADT, *Optimization, a moment problem, and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[4] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, this Journal, 3 (1965), pp. 191–205.

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

[6] L. M. GRAVES, *The Theory of Functions of Real Variables*, 2nd ed. McGraw-Hill, New York, 1956.

[7] C. L. SIEGEL, *Vorlesungen über Himmelsmechanik*, Springer-Verlag, Berlin, 1956.

[8] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[9] J. WARGA, *Necessary conditions for minimum in relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 129–145.

[10] A. N. KOLMOGOROV AND S.V. FOMIN, *Elements of the Theory of Functions and Functional Analysis*, vol. 1, Graylock, Rochester, New York, 1957.

[11] I. P. NATANSON, *Theory of Functions of a Real Variable*, vol. 1, Ungar, New York, 1955.

[12] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[13] B. H. BILLIK, *Some optimal low-acceleration rendezvous maneuvers*, AIAA J., 2 (1964), pp. 510–516.

[14] J. S. MEDITCH, *Synthesis of a class of linear feedback minimum energy controls*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 376–378.

[15] L. W. NEUSTADT, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[16] J. WARGA, *Variational problems with unbounded controls*, this Journal, 3 (1965), to appear.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

ON STABILITY IN CONTROL SYSTEMS*

EMILIO ROXIN†

1. Introduction. An axiomatic foundation of the theory of control systems was developed recently, based upon the notion of attainable set [2], [7], [8], [9]. Starting from a set of basic axioms, one proves that the properties of the so defined systems (called sometimes "generalized dynamical systems" or "generalized control systems") are in accordance with those of commonly known control systems. The main advantage of this approach lies in the fact that concepts like invariance, recurrence, stability, etc., are introduced in their greatest generality, showing their intrinsic nature.

The relation of these systems with those defined by contingent equations were studied in [10]. A way of defining generalized control systems locally, on a closed subset of the phase space, was given in [12].

In the present paper, definitions of different kinds of stability for generalized control systems are given, similar to those known for classical dynamical systems (see, for example, [6]). Practically every kind of stability for dynamical systems corresponds to a strong and a weak similar property in the case of control systems. This has already been mentioned in a communication of the author [11].

It should be noted that the relationship of different kinds of stability of control systems with some "Lyapunov functions" was already studied, in a few cases, by Zubov [14]; here it is not treated, but it is, obviously, a good subject for further investigations.

2. Definition of general control systems. Consider as phase space X a complete, locally compact metric space. Elements of X will be denoted by small letters $(x, y, \cdots)$, subsets of X by capitals $(Y, F, A, \cdots)$. Let also denote:

(i) $\rho(x, y)$ the distance between the points $x, y \in X$,

(ii) $\rho(A, x) = \rho(x, A) = \inf \{\rho(x, y); y \in A\}$ (distance between the point x and the set A),

(iii) $\beta(A, B) = \sup \{\rho(x, B); x \in A\}$ ("deviation" of the set A from the set B),

(iv) $\alpha(A, B) = \alpha(B, A) = \max \{\beta(A, B), \beta(B, A)\}$ (distance between the sets A, B in the Hausdorff pseudo-metric),

* Received by the editors February 19, 1965, and in revised form April 22, 1965.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island. Now at Department Matematicas, Facultad de Ciencias Exactas, Universidad de Buenos Aires, Buenos Aires, Argentina. This work was supported in part by the National Aeronautics and Space Administration under Contract No. NGR 46-002-015 and in part by the United States Air Force through the Air Force Office of Scientific Research under Contract No. AF-AFOSR-693-64.

(v) $\gamma(A, B) = \inf\{\rho(x, B); x \in A\} = \inf\{\rho(x, y); x \in A, y \in B\}$,
(vi) $S_\epsilon(A) = \{x \in X; \rho(x, A) < \epsilon\}$ (ϵ-neighborhood of the set A).

The independent variable t (which will be called time) may be assumed to take all real values or all nonnegative values ($t \in R$ or $t \in R^+$, respectively). Generally, only $t \in R^+$ will be considered, but in most cases the difference is irrelevant.

A control system will be assumed given by its "attainability function" $F(x_0, t_0, t)$, which corresponds to the set of all points attainable, at time t, from x_0 at time t_0.

The following axioms are assumed to hold:

(I) $F(x_0, t_0, t)$ is a closed nonempty subset of X, defined for every $x_0 \in X$, $t_0 \leqq t$.

(II) $F(x_0, t_0, t_0) = \{x_0\}$ for every $x_0 \in X$, $t_0 \in R$.

(III) For any $t_0 \leqq t_1 \leqq t_2$,

$$F(x_0, t_0, t_2) = \bigcup_{x_1 \in F(x_0, t_0, t_1)} F(x_1, t_1, t_2).$$

(IV) For any $x_1 \in X$, $t_0 \leqq t_1$, there exists some $x_0 \in X$ such that $x_1 \in F(x_0, t_0, t_1)$.

(V) For each $x_0 \in X$, $t_0 \leqq t_1$, $\epsilon > 0$, there is $\delta > 0$ such that $|t - t_1| < \delta$ implies

$$\alpha(F(x_0, t_0, t), F(x_0, t_0, t_1)) < \epsilon.$$

(VI) For each $x_0 \in X$, $t \leqq \tau$, $\epsilon > 0$, there is $\delta > 0$ such that

$$\rho(x_0, y_0) < \delta, \quad |t - t'| < \delta, \quad |\tau - \tau'| < \delta, \quad t' \leqq \tau',$$

imply

$$\beta(F(y_0, t', \tau'), F(x_0, t, \tau)) < \epsilon.$$

It was shown in [9] how the behavior of the control system can be satisfactorily derived from these axioms. In the case when the control system is only defined on a closed subset of the space X, the axioms have to be modified as pointed out in [12].

The following properties, proved in [9], will be needed.

The attainability function $F(x, t, \tau)$ can be extended backwards, i.e., for $\tau < t$ (in [9] this extension was denoted by G). The properties of this backward extension are almost the same as for the forward part, the main exception being that the continuity of $F(x, t, \tau)$ in τ (axiom V) may fail and F may become unbounded (finite escape time backwards).

DEFINITION 2.1. A mapping $u: I \to X$, defined in some interval $I = [t_0, t_1]$ and such that

$$t_0 \leqq \tau_0 \leqq \tau_1 \leqq t_1$$

implies

$$u(\tau_1) \in F(u(\tau_0), \tau_0, \tau_1),$$

is called a *motion* of the control system F; the corresponding curve in X-space, a *trajectory*.

The continuity of a motion follows from its definition and axioms I–VI.

A motion, $u_1 : [t_a, t_b] \to X$, is a *prolongation* of the motion, $u_2 : [t_c, t_d] \to X$, if $[t_a, t_b] \supset [t_c, t_d]$ and $u_1(t) = u_2(t)$ for $t \in [t_c, t_d]$.

In [8] the following properties are proved.

THEOREM 2.1. *If $x_1 \in F(x_0, t_0, t_1)$, there exists a motion $u(t)$ of the control system, such that $u(t_0) = x_0$, $u(t_1) = x_1$.*

THEOREM 2.2. *If the motions $u_i(t)$, $i = 1, 2, 3, \cdots$, of a control system are all defined in an interval $[t_0, t_1]$ (or $[t_0, +\infty)$), and if $\lim_{i \to \infty} u_i(t_0) = x_0$, then some subsequence $u_{i_k}(t)$ converges to a certain motion $u_0(t)$ and the convergence is uniform in any finite interval.*

Finally, the notation,

$$F(A, t_0, t) = \bigcup_{x \in A} F(x, t_0, t),$$

will be used. If A is compact, then $F(A, t_0, t)$ is also compact for every $t \geqq t_0$.

3. Strong stability.

DEFINITION 3.1. The set $A \subset X$ is called *strongly positively invariant* with respect to a certain control system, if for any $x_0 \in A$, $t_0 \leqq t$, the relation $F(x_0, t_0, t) \subset A$ holds. If A consists of a single point, it will also be called a strong point of rest.

Note. If the control system is defined only in the closed subset $Y \subset X$, then A must be assumed to belong to the interior of Y, at a positive distance from its boundary.

DEFINITION 3.2. The strongly positively invariant set $A \subset X$ is called *strongly stable* if, for every $\epsilon > 0$ and $t_0 \geqq 0$, there is $\delta = \delta(\epsilon, t_0) > 0$ such that $\rho(x_0, A) < \delta$ implies

$$F(x_0, t_0, t) \subset S_\epsilon(A)$$

for all $t \geqq t_0$.

This stability will also be called *strong Lyapunov stability*. Now, as in classical dynamical systems (see, for example, [13], [1], [6]), it is possible to define the following stability-type properties, which for simplicity are denoted by numbers followed by "s" (in order to indicate that it is a stability of the "strong" type). The properties are:

(1s) The strong Lyapunov stability according to Definition 3.2.

(2s) The same Definition 3.2, but with $\delta(\epsilon, t_0) = \delta(\epsilon)$ independent of t_0 (uniform strong stability).

(3s) For every $t_0 \geqq 0$ there exists a $\delta_0(t_0) > 0$ such that for any motion with $u(t_0) = x_0 \in S_{\delta_0}(A)$,

$$\lim_{t \to +\infty} \rho(u(t), A) = 0$$

holds (quasi-asymptotic strong stability).

(4s) Property (3s) with δ_0 independent of $t_0 \geqq 0$.

(5s) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ such that $\rho(x_0, A) < \delta_0$ implies

$$\lim_{t \to +\infty} \beta(F(x_0, t_0, t), A) = 0$$

(i.e., property (3s) uniformly for all motions $u(t)$ starting at (x_0, t_0)).

(6s) Property (5s) with δ_0 independent of $t_0 \geqq 0$.

(7s) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ such that

$$\lim_{t \to +\infty} \beta(F(S_{\delta_0}(A), t_0, t), A) = 0$$

(i.e., property (5s) uniformly in $x \in S_{\delta_0}(A)$; quasi-equi-asymptotic strong stability).

(8s) Property (7s) with δ_0 independent of $t_0 \geqq 0$.

(9s) There is $\delta_0 > 0$ such that

$$\lim_{\tau \to +\infty} \beta(F(S_{\delta_0}(A), t_0, t_0 + \tau), A) = 0$$

uniformly for all $t_0 \geqq 0$ (uniform quasi-equi-asymptotic strong stability).

The relations between these properties are indicated in Fig. 1. The two groups of properties 1-2 and 3-9 are independent, as the following example shows.

Example 3.1. Let $X = R$ and the control system be defined in Fig. 2, where the motions $u(t)$ are given graphically (this characterizes them sufficiently well, the decrease for $t \to +\infty$ may, for instance, be taken exponentially). It should be noted that through $x_0 = 0$ there are infinitely many different motions for every t_0. Axioms I–VI are satisfied, as it is easy to verify.

The set $A = \{x: x < 0\}$ is positively strongly invariant and satisfies property (9s), but it does not satisfy (1s). Therefore, the two groups of properties in Fig. 1 are independent.

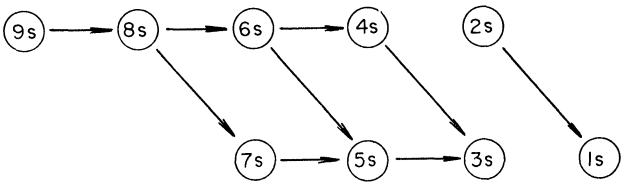

FIG. 1

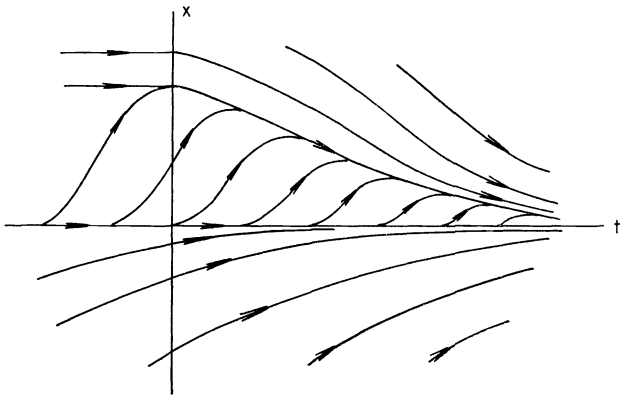

FIG. 2

It may be noted that if the definitions are not restricted to $t_0 \geqq 0$, but taken for all $t_0 \in R$, then property (9s) is not satisfied any more, but property (8s) is.

In this example, the set A is not closed. Indeed, for a compact A we can prove:

THEOREM 3.1. *For a compact, positively strongly invariant set, property* (7s) *implies* (1s).

Proof. Let A be compact, positively strongly invariant and satisfy property (7s). Then, for every $t_0 \geqq 0$ and $\epsilon > 0$, there are $\delta_0 > 0$ and $t_1 \geqq t_0$ such that

$$\beta(F(S_{\delta_0}(A), t_0, t), A) < \epsilon$$

for all $t \geqq t_1$.

If A is a single point, it follows from axiom VI that there is $\delta_1 > 0$ such that, for all t in the interval $[t_0, t_1]$,

$$\beta(F(S_{\delta_1}(A), t_0, t), A) < \epsilon. \tag{3.1}$$

Taking $\delta = \min(\delta_1, \delta_2)$, this value satisfies property (1s).

If A is not a single point, the existence of δ_1 satisfying (3.1) can be proved as follows. Take for every $x \in A$ a value $\delta_x > 0$ such that $\beta(F(S_{\delta_x}(x), t_0, t), A) < \epsilon$ uniformly in $t_0 \leqq t \leqq t_1$. A is covered by a finite collection, $A \subset \bigcup_i S_{\delta_{x_i}}(x_i)$, $i = 1, 2, \cdots, p$. Then $\bigcup S_{\delta_{x_i}}$ is a neighborhood of A and there is some δ_1 satisfying

$$S_{\delta_1}(A) \subset \bigcup_i S_{\delta_{x_i}}(x_i),$$

and therefore

$$F(S_{\delta_1}(A), t_0, t) \subset S_\epsilon(A)$$

for all $t_0 \leqq t \leqq t_1$.

THEOREM 3.2. *Properties* (2s) *and* (3s) *together imply* (5s).

Proof. Let $A \subset X$ be positively strongly invariant and satisfy properties (2s) and (3s). Let $t_0 \geqq 0$ be given and $\delta_0 = \delta_0(t_0)$ be the same as in the definition of property (3s). It will be proved that the same δ_0 satisfies (5s).

Assuming, indeed, the contrary, there is some $x_0 \in S_{\delta_0}(A)$ and there is a sequence $t_i \to +\infty$ such that

$$\beta(F(x_0, t_0, t_i), A) > a > 0, \qquad i = 1, 2, 3, \cdots . \tag{3.2}$$

As A satisfies (2s), there is $\delta > 0$ such that $\rho(x, A) < \delta$ implies $\beta(F(x, t, \tau), A) < a$ for all $\tau \geqq t \geqq t_0$. According to (3.2) there is a motion $u_1(t)$ through (x_0, t_0) such that

$$\rho(u_1(t_1), A) > a,$$

and, therefore,

$$\rho(u_1(t), A) \geqq \delta$$

for all $t \in [t_0, t_1]$. In the same way there is, for each $i = 2, 3, \cdots$, a motion $u_i(t)$ such that $u_i(t_0) = x_0$ and $\rho(u_i(t_i), A) > a$, and, therefore,

$$\rho(u_i(t), A) \geqq \delta$$

for all $t \in [t_0, t_i]$. By Theorem 2.2 some subsequence of $u_i(t)$ converges to a limit motion $u_0(t)$ for all $t \geqq t_0$, which therefore satisfies

$$\rho(u_0(t), A) \geqq \delta$$

for all $t \geqq t_0$, contrary to property (3s).

The same proof applies to the following.

THEOREM 3.3. *Properties* (2s) *and* (4s) *together imply* (6s).

For compact sets the following stronger results are valid.

THEOREM 3.4. *If $A \subset X$ is conditionally compact* (*i.e., the closure of A is compact*), *positively strongly invariant and satisfies properties* (2s) *and* (3s), *then A also satisfies* (7s).

Proof. Let $t_0 \geqq 0$, $\delta_0(t_0)$ be defined according to property (3s), and $\delta_0 > \eta > 0$. It will be proved that η satisfies the requirement of property (7s).

Assuming the contrary, there are $\epsilon > 0$, $x_i \in S_\eta(A)$, and $t_i \to +\infty$, $i = 1, 2, 3, \cdots$, such that

$$\beta(F(x_i, t_0, t_i), A) > \epsilon > 0, \qquad i = 1, 2, 3, \cdots .$$

As the closure of $S_\eta(A)$ may be assumed compact, the proof coincides essentially with the preceding one, taking

$$u_i(t_0) = x_i$$

and

$$\rho(u_i(t_i), A) > \epsilon.$$

Therefore

$$\rho(u_i(t), A) \geqq \delta > 0$$

for all $t \in [t_0, t_i]$, δ being related to ϵ by property (2s). By compactness, $x_i \to x_0 \in S_{\delta_0}(A)$ may be assumed, so that there is some limit motion $u_0(t)$, for which

$$\rho(u_0(t), A) \geqq \delta$$

for all $t \geqq t_0$, contradicting property (3s).

In the same way one proves the following.

THEOREM 3.5. *If $A \subset X$ is conditionally compact, positively strongly invariant and satisfies properties* (2s) *and* (4s), *then A also satisfies property* (8s).

The definitions (1s) to (9s) should, of course, be such that no two of them turn out to be identical (to imply each other). This is obvious in many cases, because it is known for classical dynamical systems (which are a special case of control systems, the strong stability being for them the common stability). For less obvious typical cases, two examples are given here.

Example 3.2. $X = R$ and the control system is an ordinary dynamical system whose motions are given in Fig. 3. The set $\{0\}$ satisfies properties (2s) and (7s), but not (4s) .

Example 3.3. $X = R^2$ and polar coordinates ρ, θ are used. With the auxiliary function $h(s)$ given in Fig. 4a, the equation of the motions are given by

$$\dot{\theta} = |\sqrt{|\theta|\pi - \theta^2}| \operatorname{sgn} \theta$$

(so that $\theta = \theta(t)$ are given in Fig 4b), and

$$\rho(t) = [e^{-t} + h(\theta)] \cdot \frac{\rho(t_0)}{e^{-t_0} + h(\theta(t_0))} .$$

The motions starting at $\rho(0) = \rho_0$, $\theta(0) = 0$, lie on the funnel-shaped surface of equation

$$\rho = \rho_0[e^{-t} + h(\theta)],$$

drawn in Fig. 4c.

For every motion, $\rho(t) \to 0$, so that the solution $\rho \equiv 0$ satisfies property (3s). In spite of this, the attainable set from $\rho(0) = \rho_0$, $\theta(0) = 0$, which for $t \geqq \pi$ is the cross-section of the above mentioned surface, does not tend

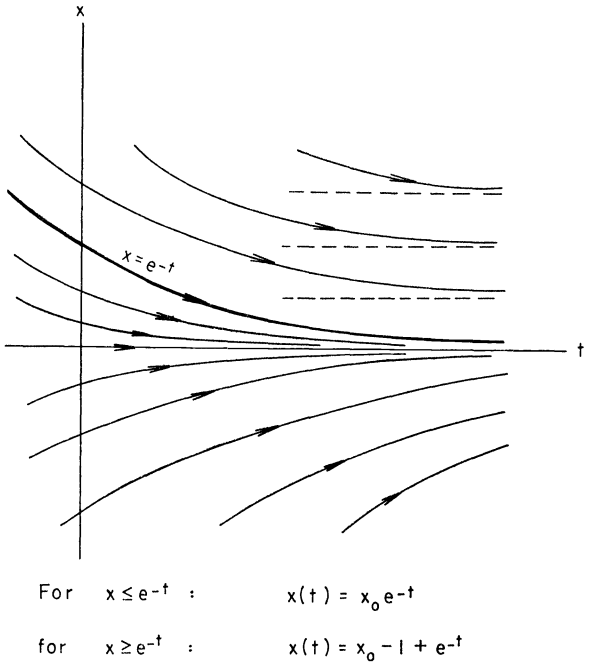

FIG. 3

to zero because for $\theta = \theta^* = \pi/2$,

$$\rho^* = \rho_0[e^{-t} + 1] \to \rho_0 \qquad \text{for} \quad t \to +\infty.$$

Therefore property (3s) does not imply (5s).

Example 3.4. Let $X = R^2$ and the motions be defined by

$$x = k \cos \alpha,$$

$$y = ke^{-t} \sin \alpha,$$

$$\alpha = \arctan (t + c), \qquad \text{or} \quad \alpha = \pm \frac{\pi}{2}.$$

Here α is taken mod 2π and k and c are constants determined by the initial conditions. This system satisfies property (5s) but not (7s) (see Fig. 5). All motions tend to A = the origin, but

$$\lim \beta(F(S_\delta(A), t_0, t), A) = \delta > 0.$$

4. Weak stability.

DEFINITION 4.1. The set $A \subset X$ is called *weakly positively invariant* with respect to a certain control system if for every $x_0 \in A$, $t_0 \geqq 0$,

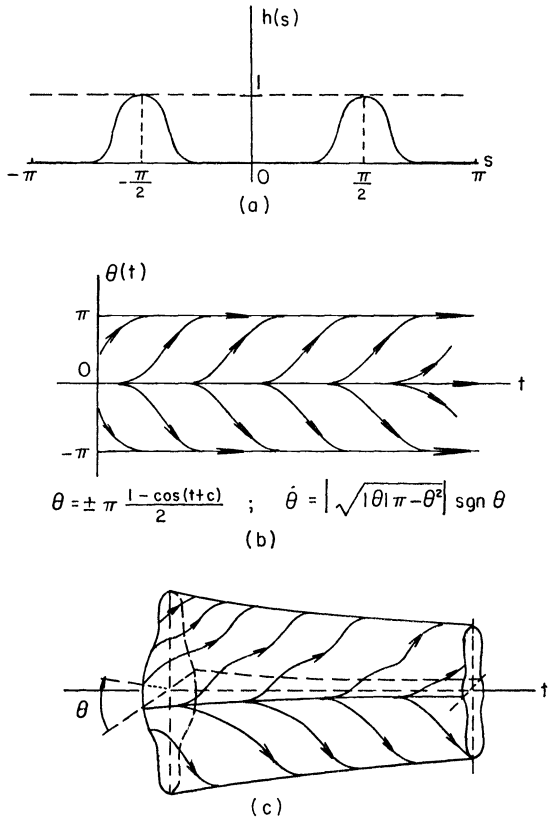

FIG. 4

there exists some motion $u(t)$ such that $u(t_0) = x_0$ and $u(t) \in A$ for all $t \geqq t_0$. If A consists of a single point, it also will be called a weak point of rest.

Note. If the control system is defined only on the closed subset $Y \subset X$, then the motion $u(t)$ should be defined (not empty) for all $t \geqq t_0$. For the stability properties defined below, A is assumed to belong to the interior of Y, at a finite distance from ∂Y.

THEOREM 4.1. (Barbashin [2]). *Necessary and sufficient for the weak positive invariance of a closed set A is the condition*

$$F(x_0, t_0, t) \cap A \neq \varnothing,$$

for every $x_0 \in A$, $t \geqq t_0$, where $\varnothing$ is the empty set.

DEFINITION 4.1. The weakly positively invariant set $A \subset X$ is called weakly stable if, for every $\epsilon > 0$ and $t_0 \geqq 0$, there is $\delta = \delta(\epsilon, t_0) > 0$ such

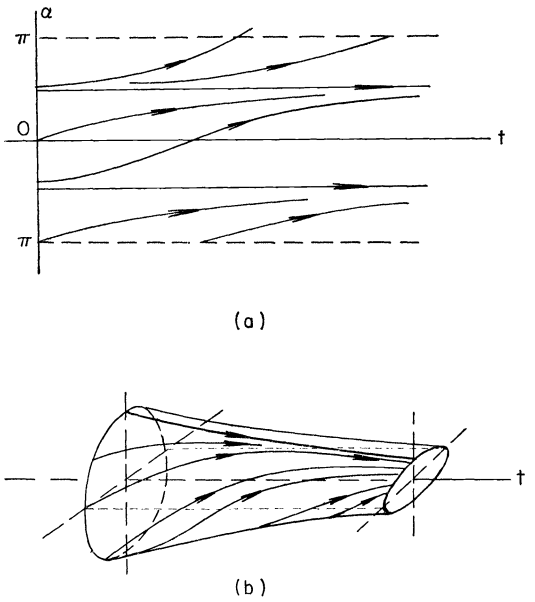

FIG. 5

that $\rho(x_0, A) < \delta$ implies the existence of some motion $u(t)$ with $u(t_0) = x_0$ and $\rho(u(t), A) < \epsilon$ for all $t \geqq t_0$.

This kind of stability will be called also *weak Lyapunov stability.*

Now, as in the preceding section, the following stability properties are defined; the "w" indicates that they correspond to the weak type.

(1w) The weak Lyapunov stability according to Definition 4.1.

(2w) The same Definition 4.1, but with $\delta(\epsilon, t_0)$ independent of $t_0 \geqq 0$ (uniform weak stability).

(3w) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ such that $\rho(x_0, A) < \delta_0$ implies

$$\lim_{t \to +\infty} \gamma(F(x_0, t_0, t), A) = 0,$$

where $\gamma(A, B) = \inf \{\rho(a, b); a \in A, b \in B\}$.

(4w) Property (3w) with δ_0 independent of $t_0 \geqq 0$.

(5w) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ such that if $\rho(x_0, A) < \delta_0$, there is some motion $u(t)$ with $u(t_0) = x_0$ and

$$\lim_{t \to +\infty} \rho(u(t), A) = 0$$

(quasi-asymptotic weak stability).

(6w) Property (5w) with δ_0 independent of $t_0 \geqq 0$.

(7w) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ and for every $\epsilon > 0$ there is

$T = T(t_0, \epsilon)$ such that $\rho(x_0, A) < \delta_0$ implies the existence of a motion $u(t)$ with $u(t_0) = x_0$ and $\lim \rho(u(t), A) = 0$ for $t \to +\infty$, in such a way that $\rho(u(t), A) < \epsilon$ for all $t \geqq t_0 + T$ (quasi-equi-asymptotic weak stability).

(8w) Property (7w) with δ_0 independent of $t_0 \geqq 0$.

(9w) Property (8w) with $T = T(\epsilon)$ independent of $t_0 \geqq 0$ (uniform quasi-equi-asymptotic weak stability).

Note. For a strong stable compact set A and any finite interval $[t_1, t_2]$, it was proved in [9] that a value $\delta(\epsilon)$ can be taken such that the stability condition of Definition 3.2 is satisfied for all $t_0 \in [t_1, t_2]$. This is similar to the classical dynamical systems. The following example shows, however, that this is not true for the weak stability.

Example 4.1. Let $X = R^2$ and the motions of the control system be defined by:

(a) In the solid pyramidal cone $t > 0$, $|x| < t - y$, $|x| < 2y - t$, the motions are given by

$$\frac{dx}{dt} = \frac{x}{t}, \qquad \frac{dy}{dt} = \frac{y}{t}.$$

(b) Outside that cone,

$$\frac{dx}{dt} = \frac{dy}{dt} = 0.$$

(c) On the boundary of that cone, the tangent to the motion $(dx/dt, dy/dt)$ at any point is required to belong to the convex hull of the set of tangents at infinitely nearby points, plus the vector $\dot{x} = \pm 1$, $\dot{y} = 0$.

This way, the motions are really defined by a contingent equation (see [10]) and are shown in Fig. 6. At the points of the boundary of the pyramidal cone, the solutions are not unique. It is easy to verify that the origin $x = y = 0$ is weakly positively invariant and satisfies property (1w). On the other hand, there is no $\delta_0(\epsilon, t_0)$ valid for all $0 < t_0 < T$ for any $T > 0$, $\epsilon > 0$.

This example can be easily modified in such a way that it applies to properties (3w), (5w), and (7w) (the only thing to do is to change conveniently the motions outside the pyramidal cone). Therefore, it makes sense to define the properties:

(1*w) For every finite interval $[t_1, t_2] \subset R^+$, there is $\delta_0 > 0$ such that the condition of property (1w) is satisfied for all $t_0 \in [t_1, t_2]$,

(3*w) Similarly for property (3w).

(5*w) Similarly for property (5w).

(7*w) Similarly for property (7w), for both $\delta_0(t_0)$ and $T(\epsilon, t_0)$.

The relations between all these properties are given in Fig. 7.

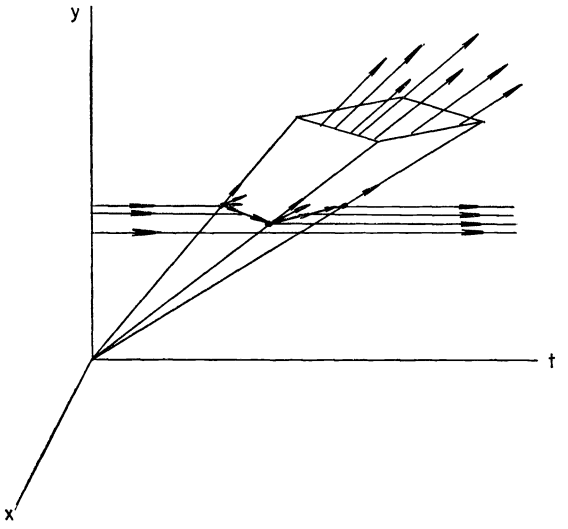

FIG. 6

Example 3.1 (Fig. 2) is valid also for the weak stability; $x = 0$ is a weak point of rest which satisfies property (9w) but not (1w). This proves the independence of both groups of properties in Fig. 7.

Similarly, Example 3.2 shows that property (7w) does not imply (4w).

The following example shows that property (3w) does not imply (5w).

Example 4.2. Let $X = R^2$ and

$$x = k \cos \alpha(t), \qquad y = ke^{-t} \sin \alpha(t).$$

Here, k is a constant determined by the initial conditions, and the functions $\alpha(t)$ are given (mod 2π) in Fig. 8a. It is to be noted that $\alpha(t) \equiv 0$ is an admissible function, from which other curves branch off.

The motions lie on tubes which become more flat as $t \to +\infty$, but the attainable set from any point of the tube-surface is, for sufficiently large t, the whole cross section of the tube-surface; therefore, its minimal distance to the origin tends to zero (property (3w)).

5. Finite stability. In the preceding two sections, the properties number 1 and 2 correspond to the common (Lyapunov) stability, and those numbered 3 to 9, to the quasi-asymptotic stabilities. Assuming both to hold, one obtains the very important asymptotic stabilities. As in control systems there is no assumption about uniqueness of the motion $u(t)$ through each point (x_0, t_0) (which restricts so much the classical dynamical systems); there can be defined even stronger stabilities than the asymptotic ones, by requiring that the motions $u(t)$ not only tend to, but actually reach, the

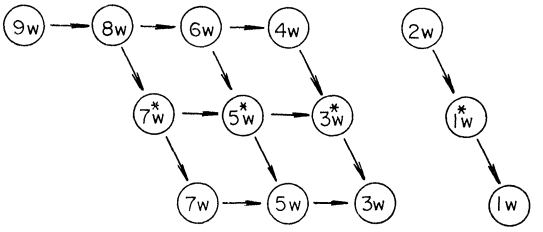

FIG. 7

invariant set A in finite time. This type of stability will be called finite stability; it can be defined for the strong and for the weak stability, and like the asymptotic one, it will be split up into the quasi-finite plus the Lyapunov stability.

Once the main idea is established, the development is quite straightforward. Even some examples given above can be slightly changed so that they apply to the finite stabilities.

Finite stabilities of the strong type (here A is a strongly positively invariant set).

(10s) For every $t_0 \geqq 0$ there exists a $\delta_0(t_0) > 0$ such that for every motion $u(t)$ with $u(t_0) = x_0 \in S_{\delta_0}(A)$, there is a finite value $\tau_f > 0$ such that $u(t_0 + \tau_f) \in A$ (and therefore $u(t) \in A$ for all $t > t_0 + \tau_f$). In general, τ_f depends on the motion $u(t)$. (This is the quasi-finite-strong stability.)

(11s) Property (10s) with δ_0 independent of $t_0 \geqq 0$.

(12s) For every $t_0 \geqq 0$, there is $\delta_0(t_0) > 0$ such that $x_0 \in S_{\delta_0}(A)$ implies the existence of $\tau_f = \tau_f(x_0, t_0) > 0$ such that

$$F(x_0, t_0, t_0 + \tau_f) \subset A,$$

and therefore $F(x_0, t_0, t) \subset A$ for all $t > t_0 + \tau_f$.

(13s) Property (12s) with δ_0 independent of t_0.

(14s) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ and a finite $\tau_f(t_0) > 0$ such that

$$F(S_{\delta_0}(A), t_0, t_0 + \tau_f) \subset A.$$

This is the quasi-equi-finite strong stability.

(15s) Property (14s) with δ_0 independent of t_0.

(16s) Property (15s) with τ_f independent of t_0 (uniform quasi-equi-finite strong stability).

Obviously the following implications hold:

$$(10s) \Rightarrow (3s); \quad (11s) \Rightarrow (4s); \quad (12s) \Rightarrow (5s); \quad (13s) \Rightarrow (6s);$$

$$(14s) \Rightarrow (7s); \quad (15s) \Rightarrow (8s); \quad (16s) \Rightarrow (9s).$$

Fig. 9 shows the implications between the stabilities of this last group.

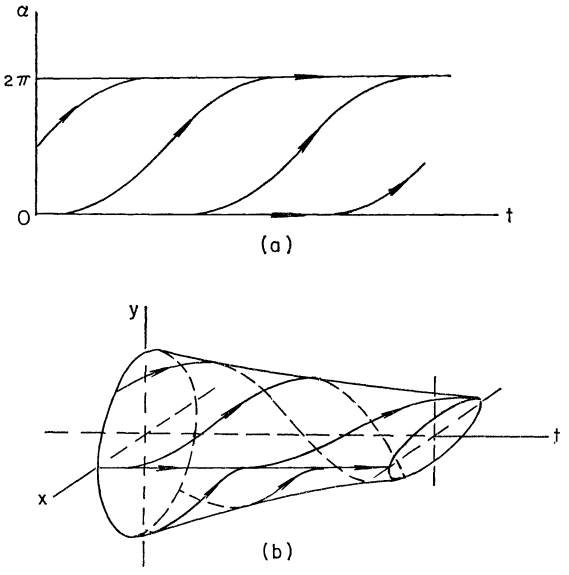

FIG. 8

Finite stabilities of the weak type (here A is a weakly positively invariant set).

(10w) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ such that if $x_0 \in S_{\delta_0}(A)$, there is a motion $u(t)$ with $u(t_0) = x_0$ and $u(t_0 + \tau_f) \in A$ for some finite $\tau_f > 0$ (and, therefore, this motion can be prolonged indefinitely in A). This is the quasi-finite weak stability.

(10*w) Property (10w), and for any finite interval $[t_1, t_2] \in R^+$, $\delta_0(t_0)$ can be taken to hold uniformly for all $t_0 \in [t_1, t_2]$.

(11w) Property (10w), with δ_0 independent of $t_0 \geqq 0$.

(12w) For every $t_0 \geqq 0$ there is $\delta_0(t_0) > 0$ and some value τ_f, $0 < \tau_f = \tau_f(t_0)$ such that $x_0 \in S_{\delta_0}(A)$ implies the existence of a motion $u(t)$ with $u(t_0) = x_0$ and $u(t_0, \tau_f) \in A$ (quasi-equi-finite weak stability).

(12*w) Property (12w), and for any finite interval $[t_1, t_2] \in R^+$, $\delta_0(t_0)$ can be taken to hold uniformly for all $t_0 \in [t_1, t_2]$.

(13w) Property (12w), with δ_0 independent of $t_0 \geqq 0$.

(14w) Property (13w), and $\tau_f = \tau_f(\delta_0)$ independent of $t_0 \geqq 0$ (uniform quasi-equi-finite weak stability).

Obviously, the following implications hold:

$$(10w) \Rightarrow (5w); \quad (10^*w) \Rightarrow (5^*w); \quad (11w) \Rightarrow (6w); \quad (12w) \Rightarrow (7w).$$

$$(12^*w) \Rightarrow (7^*w); \quad (13w) \Rightarrow (8w); \quad (14w) \Rightarrow (9w).$$

Fig. 10 shows the implications between the stabilities of this last group.

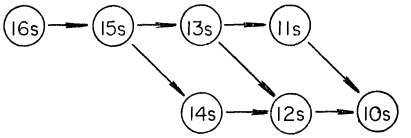

FIG. 9

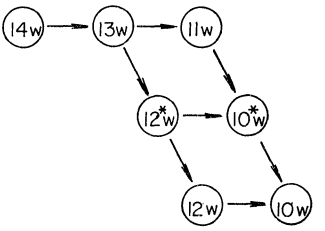

FIG. 10

Remarks about the finite stabilities. The importance of motions arriving at the origin (supposed to be a positively weakly invariant set) in a finite time plays an important role in control theory. Therefore, the stabilities of the finite type have already been used, without special denomination, by numerous authors (for example, Kalman [3], Lee and Markus [5], LaSalle [4]).

The region of attraction for the finite weak stability corresponds to what is known as the domain of controllability [5]. It may be noted that most asymptotically stable systems of the real physical world are, indeed, finitely stable.

The strong type of finite stability has not been used, apparently, but a rather trivial example shows that it can appear even in the simple case of:

$$\dot{x} = -2\sqrt{|x|}\cdot(2 + u)\cdot\operatorname{sgn} x$$

with the control $u(t)$ restricted by $|u| \leqq 1$. In this equation, the extreme values of $u(t)$ correspond to the motions for $u = 1$:

$$|x(t)| = \begin{cases} [\sqrt{|x_0|} + 3t_0 - 3t]^2 & \text{for } t \leqq t_0 + \dfrac{\sqrt{|x_0|}}{3}, \\ 0 & \text{for } t \geqq t_0 + \dfrac{\sqrt{|x_0|}}{3}; \end{cases}$$

for $u = -1$:

$$|x(t)| = \begin{cases} [\sqrt{|x_0|} + t_0 - t]^2 & \text{for } t \leqq t_0 + \sqrt{|x_0|}, \\ 0 & \text{for } t \geqq t_0 + \sqrt{|x_0|}. \end{cases}$$

Of course, the definitions given above do not solve any specific problem, but they may help to treat systematically cases which appear frequently in applications.

REFERENCES

[1] H. A. Antosiewicz, *A survey of Liapunov's second method*, Contributions to the Theory of Nonlinear Oscillations, vol. IV, Princeton University Press, Princeton, 1958, pp. 141–166.

[2] E. A. Barbashin, *On the theory of generalized dynamical systems*, Učem. Zap. Moskov. Gos. Univ., 135 (1949), pp. 110–133.

[3] R. E. Kalman, *Contributions to the theory of control*, Bol. Soc. Mat. Mexicana, (1960), pp. 102–119.

[4] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.

[5] E. B. Lee and L. Markus, *Optimal control for non-linear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 38–58.

[6] J. L. Massera, *Contributions to stability theory*, Ann. of Math., 64 (1956), pp. 182–205.

[7] E. Roxin, *Axiomatic theory of control systems*, RIAS Tech. Rep. 62-12, 1962.

[8] ———, *Axiomatic foundation of the theory of control systems*, Second Int. Conf. of IFAC, Basel, 1963.

[9] ———, *Stability in general control systems*, J. Differential Equations, 1 (1965), pp. 115–150.

[10] ———, *On generalized dynamical systems defined by contingent equations*, Ibid., 1 (1965), pp. 188–205.

[11] ———, *Stabilität in allgemeinen Regelungssystemen*, Third Conference on Nonlinear Oscillations, Berlin, 1964.

[12] ———, *Local definition of generalized control systems*, to be published.

[13] T. Yoshizawa, *Asymptotic behaviour of solutions of ordinary differential equations near sets*, RIAS Tech. Rep. 61-5, 1961.

[14] V. I. Zubov, *Methods of A. M. Liapunov and Their Application*, Izdat. Leningrad University, 1957, English transl. Noordhoff, Groningen, 1964.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

ON THE EXISTENCE OF LYAPUNOV FUNCTIONS FOR THE PROBLEM OF LUR'E*

K. R. MEYER†

Introduction. This paper is an extension of the work of Yacubovich and Kalman on the existence of Lyapunov functions for the problem of Lur'e. The primary result of this paper is the removal of the unnecessary hypothesis of complete controllability and complete observability from the theorem of Kalman. These hypotheses have been used either explicitly or implicitly by many authors working in this field. Indeed, the change of coordinates introduced by Lur'e, the so-called Lur'e transformations, can be made only if the system is completely controllable.

The first section contains a summary of elementary results and definitions from linear algebra and control theory.

The proofs of these preliminaries are elementary and can be found in [1], [2], and [3].

The second section contains the extensions of the lemma of Kalman-Yacubovich. The proof of the first lemma follows very closely the proof as given by Kalman in [2].

The third section contains a few applications of the lemmas developed in the second section.

1. Preliminaries. Let A be a real $n \times n$ matrix and b, c two real n-vectors (column). Let E^n be Euclidean n-space. Denote by $A(z)$ the characteristic matrix of A, that is, $A(z) = zI - A$, where I is the identity matrix and z is a scalar complex variable and let $A(z)^{-1} = \{A(z)\}^{-1}$. Let $'$ denote the transpose, * the conjugate transpose and $|\ \ |$ the determinant. Thus $|A(z)|$ is the characteristic polynomial of A. The subspaces of E^n generated by the vectors b, Ab, $\cdots$ will be denoted by $[A, b]$. The orthogonal complement of $[A, b]$ in E^n will be denoted by $[A, b]^0$. Let the dimension of $[A, b]$ be p.

LEMMA A. *In general,*

$$[A, b]^0 = \{x \in E^n : x'A^k b = 0,\ k = 0, 1, 2, \cdots\}$$
$$= \{x \in E^n : x'(\exp At)b \equiv 0 \text{ for all } t \in (-\infty, \infty)\}$$
$$= \{x \in E^n : x'A(z)^{-1}b \equiv 0 \text{ for any set of } z \text{ having finite limit point}\},$$

* Received by the editors January 14, 1965, and in revised form April 30, 1965.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island. This research was supported in part by the National Aeronautics and Space Administration under Grant No. NGR-40-002-015, in part by the United States Army at Durham through the Army Research Office under Contract No. DA-31-124-ARO-D 270, and in part by the United States Air Force through the Air Force Office of Scientific Research under Grant No. AF-AFOSR-693-64.

and if all the characteristic roots of A have negative real parts then

$$[A, b]^0 = \{x \in E^n : \operatorname{Re} x'A(i\omega)^{-1}b \equiv 0 \text{ for all real } \omega\}.$$

One says the pair (A, b) is *completely controllable* provided $[A, b] = E^n$ and the pair (A, c') is *completely observable* if (A', c) is completely controllable.

LEMMA B. *There exists a basis for E^n such that*

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix},$$

where A_1, A_2, and A_3 are $p \times p$, $p \times (n-p)$ and $(n-p) \times (n-p)$ matrices respectively, b_1 is a p-vector and (A_1, b_1) is completely controllable.

LEMMA C. *Let (A, b) be completely controllable and let $\tilde{g}(z) = g_1 + g_2 z + \cdots + g_n z^{n-1}$ be any polynomial with real coefficients of degree less than n. Then there exists a real n-vector g such that $g'A(z)^{-1}b = \tilde{g}(z)\{|A(z)|\}^{-1}$.*

LEMMA D. *Let (A, b) be completely controllable and k any real n-vector. Let $k'A(z)^{-1}b = p(z)\{|A(z)|\}^{-1}$. Then the degree of the greatest common divisor of $p(z)$ and $|A(z)|$ is equal to the dimension of $[A', k]^0$.*

A rational function $f(z)$ is said to be a *positive real function* provided $\operatorname{Re} f(z_0) \geqq 0$ whenever z_0 is not a pole of $f(z)$ and $\operatorname{Re} z_0 \geqq 0$.

2. The main lemmas. The extension of the Kalman-Yacubovich lemma will require several steps. The first lemma is a slight extension of the lemma as given by Kalman [2] and the proof of this lemma follows very closely his proof. We obtain the additional information that B is positive definite and that (A, q') is completely observable.

LEMMA 1. *Let A be an $n \times n$ real matrix all of whose characteristic roots have negative real parts, let τ be a nonnegative real number and let b, k be two real n-vectors. Assume (A, b) is completely controllable. If the function*

$$T(z) = \tau + 2k'A(z)^{-1}b \tag{1.1}$$

is a positive real function then there exist two $n \times n$ real symmetric matrices B and D and a real n-vector q such that

(a) $A'B + BA = -qq' - D$,
(b) $Bb - k = \sqrt{\tau}q$,
(c) (A, q') *is completely observable,*
(d) B *is positive definite and D is positive semidefinite,*
(e) *if $i\omega_0$, ω_0 real, is a zero of $-q'A(z)^{-1}b + \sqrt{\tau}$, then it is a zero of $b'A(-z)^{-1}DA(z)^{-1}b$, and*
(f) *all the zeros of $-q'A(z)^{-1}b + \sqrt{\tau}$ are in the closed left halfplane.*

Proof. Let $m(z) = A(z)^{-1}b$ and $\psi(z) = |A(z)|$. Since $T(z)$ is positive

real,

$$0 \leqq \tau + m(i\omega)^* k + k' m(i\omega) = \frac{\eta(i\omega)}{\psi(i\omega)\psi(-i\omega)}. \tag{1.2}$$

Clearly $\eta(z)$ is an even polynomial with real coefficients and hence its zeros are symmetric about both the real axis and the imaginary axis. Since Re $\eta(i\omega) \geqq 0$ for all real ω, the zeros of $\eta(z)$ on the imaginary axis are of even multiplicity. Thus $\eta(z) = \theta(z)\theta(-z)$, where $\theta(z)$ is a real polynomial all of whose zeros have nonpositive real parts.

Let $\theta(z) = \theta_1(z)\theta_2(z)$, where all the zeros of $\theta_1(z)$ have negative real parts, all the zeros of $\theta_2(z)$ are pure imaginary, and the leading coefficient of $\theta_2(z)$ is one. Let ϵ_0 be the greatest lower bound of $\theta_1(i\omega)\theta_1(-i\omega)$ taken over all real ω. Since $\theta_1(z)$ has no pure imaginary zeros, $\epsilon_0 > 0$. Let α be a real positive number such that $\alpha^2 \leqq \epsilon_0$ and $\alpha^2 \neq \theta_1(\lambda_i)\theta_1(-\lambda_i)$, $i = 1, \cdots, n$, where λ_i is a zero of $\psi(z)$. If $\theta_1(z)$ is a constant, take $\alpha = 0$. Consider $\Gamma(z) = \theta_2(z)\theta_2(-z)[\theta_1(z)\theta_1(-z) - \alpha^2]$. By the definition of α and Γ it follows that (i) $\Gamma(i\omega) \geqq 0$ for all real ω, and (ii) the greatest common divisor of $\Gamma(z)$ and $\psi(z)\psi(-z)$ is one.

Since $\Gamma(z)$ is an even polynomial and Re $\Gamma(i\omega) \geqq 0$ for all real ω, there exists a polynomial $\nu(z)$ with real coefficients all of whose zeros have nonpositive real parts such that $\Gamma(z) = \nu(z)\nu(-z)$. Define the vector g such that $g' A(i\omega)^{-1} b = \alpha\theta_2(z)\{\psi(z)\}^{-1}$. Thus

$$\begin{aligned} 0 &\leqq \tau + m(i\omega)^* k + k' m(i\omega) - m(i\omega)^* g g' m(i\omega) \\ &= \frac{\Gamma(i\omega)}{\psi(i\omega)\psi(-i\omega)} \\ &= \frac{\nu(i\omega)\nu(-i\omega)}{\psi(i\omega)\psi(-i\omega)}. \end{aligned} \tag{1.3}$$

The formal degree of $\nu(z)$ is n and its leading coefficient is $\sqrt{\tau}$ and so

$$\frac{\nu(z)}{\psi(z)} = -\frac{\mu(z)}{\psi(z)} + \sqrt{\tau},$$

where μ is real and of degree less than or equal to $n - 1$. The vector q is then defined by $\mu(z)\{\psi(z)\}^{-1} = q' m(z)$. By construction, $\mu(z)$ and $\psi(z)$ have greatest common divisor one and so (A, q') is completely observable. Thus property (c) holds. Define $D = gg'$; since by construction the pure imaginary zeros of $g' m(z)$ and $-q' m(z) + \sqrt{\tau}$ are the same, property (e) holds.

Now define

$$B = \int_0^\infty e^{A't}\{qq' + D\}e^{At}\, dt,$$

and so, $A'B + BA = -qq' - D$. Since (A, q') is completely observable, B is positive definite. From (1.3) it follows that

$$\begin{aligned} m^*(i\omega)k + k'm(i\omega) &= m^*(i\omega)Dm(i\omega) + (-q'm(i\omega) + \sqrt{\tau})(-m^*(i\omega)q + \sqrt{\tau}) - \tau \\ &= m^*(i\omega)\{qq' + D\}m(i\omega) - \sqrt{\tau}(q'm(i\omega) + m^*(i\omega)q) \\ &= b'Bm(i\omega) + m^*(i\omega)Bb - \sqrt{\tau}(q'm(i\omega) + m^*(i\omega)q) \end{aligned}$$

and hence Re $\{Bb - k - \sqrt{\tau}q\}'m(i\omega) = 0$ and so $Bb - k = \sqrt{\tau}q$.

The next step is the removal of the assumption that (A, b) be completely controllable. This is done with the following lemma.

LEMMA 2. *Let A be a real $n \times n$ matrix all of whose characteristic roots have negative real parts; let τ be a real nonnegative number and let b, k be two real n-vectors. If*

$$T(z) = \tau + 2k'A(z)^{-1}b$$

is a positive real function then there exist two $n \times n$ real symmetric matrices B, D and a real n-vector q such that

(a) $A'B + BA = -qq' - D$,
(b) $Bb - k = \sqrt{\tau}q$,
(c) *D is positive semidefinite and B is positive definite,*
(d) $\{x \in E^n : x'Dx = 0\} \cap [A', q]^0 = \{0\}$,
(e) $q \notin [A, b]^0$, *and*
(f) *if $i\omega$, ω real, is a zero of $-q'A(z)^{-1}b + \sqrt{\tau}$, then it is a zero of $b'A(-z)^{-1}DA(z)^{-1}b$.*

Proof. Choose a coordinate system for E^n such that

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix}, \qquad k = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix},$$

where A_1, A_2, A_3 are $p \times p$, $p \times (n - p)$, $(n - p) \times (n - p)$ matrices, respectively, b_1, k_1 are p-vectors, k_2 is an $(n - p)$-vector, and such that (A_1, b_1) is completely controllable. Clearly if A has all characteristic roots with negative real parts then so do A_1 and A_3. If we partition B, D and q in the same way, i.e.,

$$B = \begin{pmatrix} B_1 & B_2 \\ B_2' & B_3 \end{pmatrix}, \qquad D = \begin{pmatrix} D_1 & 0 \\ 0 & D_3 \end{pmatrix}, \qquad q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix},$$

we find that we must solve the following set of matrix equations:

(1) $A_1'B_1 + B_1A_1 = -q_1q_1' - D_1$,
(2) $A_2'B_1 + A_3'B_2' + B_2'A_1 = -q_2q_1'$,

(3) $A_2'B_2 + A_3'B_3 + B_2'A_2 + B_3A_3 = -q_2q_2' - D_3$,
(4) $B_1b_1 - k_1 = \sqrt{\tau}q_1$,
(5) $B_2'b_1 - k_2 = \sqrt{\tau}q_2$.

By hypothesis, $\tau + 2k_1'A_1(z)^{-1}b_1$ is a positive real function and so by Lemma 1 there exists a solution to (1) and (4) and by Lemma 1(c) the condition of Lemma 2(e) is satisfied. Also by Lemma 1(e) the condition of Lemma 2(f) is satisfied. Now let us consider (2) and (5). Since B_1 and q_1 are known by Lemma 1 these two equations have only B_2' and q_2 as unknowns. We can solve (2) for B_2' in terms of q_2 by the formula

$$B_2' = \int_0^\infty e^{A_3't}\{q_2\, q_1' - A_2'\, B_1\}\, e^{A_1t}\, dt,$$

and then substitute this into (5) to obtain

$$Rq_2 = \left\{\int_0^\infty e^{A_3't}\, q_1\, e^{A_1t}\, b\, dt - \sqrt{\tau}I\right\} q_2 = k + \int_0^\infty e^{A_3't}A_2'B_1\, e^{A_1t}\, b\, dt.$$

Since the right hand side of the above is known, we can solve for q_2 provided the matrix in the braces, R, is nonsingular. There is no loss of generality in assuming that A_3' is in triangular form and so $e^{A_3't}$ is in triangular form. A typical term from the diagonal of R is then

$$\begin{aligned}\int_0^\infty e^{\lambda_i t}q_1\, e^{A_1t}\, b_1\, dt - \sqrt{\tau} &= q_1(-\lambda_i\, I - A_1)^{-1}\, b_1 - \sqrt{\tau} \\ &= q_1'\, A_1(-\lambda_1)^{-1}\, b_1 - \sqrt{\tau}.\end{aligned}$$

But this term is not zero since $-\lambda_i$ is in the open right halfplane and by Lemma 1(f), we know that the zeros of $q_1A_1(z)^{-1}b_1 - \sqrt{\tau}$ are in the closed left halfplane. Thus R is nonsingular and q_2 and B_2 are determined.

Now choose D_3 to be any positive definite matrix. It is clear then that (5) has a solution and that (d) is satisfied.

Since B satisfies $A'B + BA = -qq' - D$, it must be of the form

$$B = \int_0^\infty e^{A't}qq'\, e^{At}\, dt + \int_0^\infty e^{A't}\, De^{At}\, dt.$$

If x_0 is such that $x_0Bx_0 = 0$, then $x_0e^{A't}q \equiv 0$ and $x_0Dx_0 = 0$; and thus by (d), $x_0 = 0$. Hence B is positive definite.

The converse of this lemma is true also. The proof of the converse as given in [2] does not depend on complete controllability and complete observability.

In some critical cases the following lemma is useful. This lemma is in essence due to Yacubovich [5] and was implicitly used by Meyer in [6].

LEMMA 3. *Let A be a $2n \times 2n$ real matrix with simple distinct pure im-*

aginary characteristic roots $\pm i\omega_j$, $j = 1, \cdots, n$. *If the residues of* $k'A(z)^{-1}b$ *at each* $\pm i\omega_j$ *are positive then there exists a positive definite matrix* B *such that*

$$A'B + BA = 0 \quad \textit{and} \quad Bb - k = 0.$$

This lemma follows at once by making a change of coordinates so that A is diagonal. In this coordinate system B is chosen to be diagonal also.

Using the same procedure as used in the proof of Lemma 2 one can extend the lemma of [4, p. 115] as follows.

LEMMA 4. *Let* A *be a real* $n \times n$ *matrix all of whose characteristic roots have negative real parts,* τ *be a nonnegative number and* b, k *be any two real* n*-vectors. If*

$$\tau + 2 \operatorname{Re} k'A(i\omega)^{-1}b > 0$$

for all real ω, *then there exist two real positive definite matrices* B *and* D *and a real* n*-vector* q *such that*

(a) $A'B + BA = -qq' - D$,
(b) $Bb - k = \sqrt{\tau}q$.

This lemma is almost the same as the lemma given by Yacubovich [7].

3. Applications. The lemmas developed in §2 can be applied to many different systems that have been considered in the literature. Let us consider the so-called direct control system. The equations are

$$\begin{aligned} \dot{x} &= Ax - b\phi(\sigma), \\ \sigma &= c'x, \end{aligned} \tag{3.1}$$

where A is a real $n \times n$ matrix, b, x and c are real n-vectors and $\phi(\sigma)$ is a continuous scalar function of the scalar σ such that $\sigma\phi(\sigma) > 0$ for all $\sigma \neq 0$. The vector x and the scalar σ are functions of the real variable t, time, and $\dot{x}$ is the derivative of x with respect to t. Let us assume also that through each point in E^n there exists a unique trajectory of (3.1).

THEOREM 1. *If all the characteristic roots of* A *have negative real parts and if there exist two nonnegative constants* α *and* β, $\alpha + \beta > 0$, *such that*

$$T(z) = (\alpha + \beta z)c'A(z)^{-1}b \tag{3.2}$$

is a positive real function then all solutions of (3.1) *are bounded, the trivial solution* $x = 0$ *is stable, and moreover if* $\alpha \neq 0$ *the trivial solution is asymptotically stable in the large.*

If, in the case where $\alpha = 0$, *all the characteristic roots of the matrix* $A - \mu bc'$ *have negative real parts for all* $\mu > 0$ *then the trivial solution* $x = 0$ *of* (3.1) *is asymptotically stable.*

Proof. Using the relation $zI = A(z) + A$ we obtain

$$T(z) = \beta c'b + 2 \operatorname{Re}\left(\frac{\alpha c + \beta A'c}{2}\right)' A(z)^{-1} b,$$

and thus by Lemma 2 there exist a real n-vector q and two positive symmetric matrices B and D such that

$$A'B + BA = -qq' - D, \quad Bb - \left(\frac{\alpha c + \beta A'c}{2}\right) = \sqrt{\beta c'b}\, q,$$

and moreover B is definite. Thus

$$V = x'Bx + \beta \int_0^\sigma \phi(\sigma)\, d\sigma \tag{3.3}$$

is a positive definite function and tends to ∞ as $|x| \to \infty$. The derivative $\dot{V}$ of V along the trajectories of (3.1) is given by

$$\begin{aligned} -\dot{V} &= -x'(A'B + BA)\, x + 2\left(Bb - \frac{\alpha}{2} c - \frac{\beta}{2} A'c\right)' x\phi(\sigma) \\ &\qquad + \beta c'b\phi(\sigma)^2 + \alpha\sigma\phi(\sigma) \\ &= x'Dx + (\sqrt{\tau}\, \phi(\sigma) + q'x)^2 + \alpha\sigma\phi(\sigma). \end{aligned} \tag{3.4}$$

Note that $\alpha\sigma\phi(\sigma)$ has been added and subtracted from $\dot{V}$ and that $\tau = \beta c'b$.

Clearly $-\dot{V}$ is also nonnegative and hence, by the well known theorems of Lyapunov theory all solutions are bounded and the origin is stable. In order to prove asymptotic stability we must show that no solution remains in the set where $-\dot{V} = 0$. Let $\alpha \neq 0$ and assume there exists a solution $x(t)$ of (3.1) such that $x(0) = x_0$ and $x(t)$ remains in the set where $-\dot{V} = 0$. But if $\dot{V} = 0$ then $\sigma = 0$, and thus, such a solution is a solution of $\dot{x} = Ax$. Hence $x(t) = (\exp At)x_0$. From the second term we obtain $q'(\exp At)x_0 \equiv 0$. Also, $x_0 Dx_0 = 0$ and so by Lemma 2(d), $x_0 = 0$.

In general we cannot conclude more than stability in the case where $\alpha = 0$, but if the linear system $\dot{x} = \{A - \mu bc'\}x$ is asymptotically stable for all $\mu > 0$ then (3.1) is asymptotically stable in the large also. In order to rule out solutions that remain in the set where $-\dot{V} = 0$, we must be sure that there is no solution such that $\sqrt{\tau}\phi(\sigma(t)) = -q'x(t)$.

If $\tau \neq 0$ then a solution of (3.1) that remains in the set where $\dot{V} = 0$ must satisfy the linear equation $\dot{x} = \{A + \tau^{-1/2} bq'\}x$. By Lemma 2(e) there exists a nonnegative integer m such that $q'b = q'Ab = \cdots = q'A^{m-1}b = 0$ and $q'A^m b \neq 0$. Hence if $\tau = 0$ there exists an m such that a solution of (3.1) that remains in the set where $-\dot{V} = 0$ must satisfy $\dot{x} = \{A - (q'A^m b)^{-1} bq'A^{m+1}\}x$.

As we have seen, a solution that remains in the set where $-\dot{V} = 0$ is

a solution of the linear constant coefficient differential equation. Let us assume that there exists a nontrivial solution $x(t)$ of (3.1) that remains in the set where $-\dot{V} = 0$. We can assume $\sigma(t) \not\equiv 0$ since if $\sigma \equiv 0$ we could repeat the previous argument. Since $x(t)$ is a solution of a linear equation and is bounded for all t, then $x(t)$ must be of the form

$$x(t) = \sum_{j=-N}^{N} v_j\{\exp i\omega_j t\},$$

where the v_j are n-vectors such that $v_{-j} = \bar{v}_j$ and ω_j are real scalars such that $\omega_{-j} = -\omega_j$. Clearly $\phi(\sigma(t))$ must be of the form

$$\phi(\sigma(t)) = \sum_{j=-N}^{N} a_j\{\exp i\omega_j t\},$$

where the a_j are scalars such that $a_j = -a_{-j}$. By substituting these forms into (3.1) one obtains

$$v_j = -a_j A(i\omega_j)^{-1}b.$$

Thus, by the well known formula from the theory of almost periodic functions,

$$\lim_{t\to\infty} \frac{1}{T}\int_0^T \sigma(t)\phi(\sigma(t))\,dt = -\sum_{j=-N}^{N} |\, a_j\,|^2 c'A(i\omega_j)^{-1}\,b > 0.$$

We shall have a contradiction once we prove the following remark.

Let the characteristic roots of the matrix $A - \mu bc'$ have negative real parts for all $\mu > 0$. If $i\omega_j$, ω_j real, is a characteristic root of $A + \tau^{-1/2}bq'$ when $\tau \neq 0$ or of $A - (q'A^m b)^{-1}bq'A^{m+1}$ when $\tau = q'b = \cdots = q'A^{m-1}b = 0$ and $q'A^m b \neq 0$, then $\operatorname{Im} c'A(i\omega_j)^{-1}b = 0$ *and* $c'A(i\omega_j)^{-1}b \geqq 0$.

We shall consider only the case where $\tau \neq 0$, since the other case is very similar. Since $\alpha = 0$ we may take $\beta = 1$. Then

$$qq' + D = -(A'B + BA) = A^*(i\omega_j)B + BA(i\omega_j),$$

and

$$|\, q'A(i\omega_j)^{-1}b\,|^2 + b'A^*(i\omega_j)^{-1}DA(i\omega_j)^{-1}b = 2\operatorname{Re} b'BA(i\omega_j)^{-1}b.$$

Now the characteristic polynomial of $A + \tau^{-1/2}bq'$ is

$$|\, A(z)\,|\{1 - \tau^{-1/2}q'A(z)^{-1}b\}$$

and so

$$\tau = \tau^{-1/2}q'A(i\omega_j)^{-1}b = b'BA(i\omega_j)^{-1}b - \tfrac{1}{2}c'AA(i\omega_j)^{-1}b.$$

Since $-\sqrt{\tau} + q'A(i\omega_j)^{-1}b = 0$ by Lemma 2(f), $b'A(i\omega_j)^{-1}DA(i\omega_j)^{-1}b = 0$.

Thus

$$\tau + 2 \operatorname{Re} c'AA(i\omega_j)^{-1}b = \operatorname{Re} i\omega_j c'A(i\omega_j)^{-1}b = 0$$

or

$$\operatorname{Im} c'A(i\omega_j)^{-1}b = 0.$$

Since the linear system $\dot{x} = \{A - \mu bc'\}x$ is asymptotically stable for all $\mu > 0$, the theorem of Nyquist gives $c'A(i\omega_j)^{-1}b \geqq 0$.

The above theorem can be modified several ways:

(i) If the matrix A has some characteristic roots on the imaginary axis then Lemmas 2 and 3 can be used to prove asymptotic stability in a manner similar to that found in [5] and [6]. In particular, we have the following.

THEOREM 1′. *If A has $2s$ simple, distinct, nonzero pure imaginary characteristic roots, the characteristic root zero of multiplicity p where $p = 0, 1, 2$, and all other characteristic roots having negative real parts, then* (3.1) *is asymptotically stable in the large, provided:*

(1) *there exist two nonnegative constants α and β, $\alpha + \beta > 0$, such that $T(z) = (\alpha + z\beta)c'A(z)^{-1}b$ is a positive real function, and if $i\omega$, ω real, is a characteristic root of A, then the residue of $(\alpha + z\beta)c'A(z)^{-1}b$ at $i\omega$ is positive;*

(2) *if $p = 2$, then $\lim_{z\to 0} z^2 c'A(z)^{-1}b \neq 0$,*

(3) *when $\alpha = 0$, the characteristic roots of $A - \mu bc'$ have negative real parts for all $\mu > 0$,*

(4) *if A is singular and $\alpha = 0$ then $\int_0^\sigma \phi(\tau)\, d\tau \to \infty$ as $|\sigma| \to \infty$.*

In order to prove this theorem one first changes coordinates such that the system (3.1) takes the form

$$\begin{aligned}
\dot{x}_1 &= A_1 x_1 - b_1\phi(\sigma),\\
\dot{x}_2 &= A_2 x_2 - b_2\phi(\sigma),\\
\dot{x}_3 &= A_3 x_3 - b_3\phi(\sigma),\\
\sigma &= c_1' x_1 + c_2' x_2 + c_3' x_3,
\end{aligned}$$

where x_1, b_1, c_1 are r-vectors, x_2, b_2, c_2 are $2s$-vectors and A_1, A_2 are $r \times r$, $2s \times 2s$ matrices, respectively. The vectors x_3, c_3 and b_3 are p-vectors and A_3 is a $p \times p$ matrix, where $p = 0, 1, 2$. The characteristic roots of A_1 all have negative real parts, the characteristic roots of A_2 are all simple nonzero pure imaginary numbers and the characteristic root of A_3 is zero. The matrix $A_3 = (0)$ if $p = 1$ and $A_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ if $p = 2$. Let

$$V = x_1' B_1 x + x_2' B_2 x_2 + x_3' B_3 x_3 + \beta \int_0^\sigma \phi(\tau)\, d\tau,$$

where B_1 is given by Lemma 2 as in the above and B_2 is given by Lemma 3 and $B_3 = 0$ if $p = 0$, $B_3 = \alpha$ if $p = 1$, $B_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ if $p = 2$. Thus, B_1, B_2, and B_3 are $r \times r$, $2s \times 2s$ and $p \times p$ symmetric matrices, respectively, and V is positive definite. One can proceed as before with only very minor changes in the argument.

(ii) If $\phi(\sigma)$ is restricted so that $0 < \sigma\phi(\sigma) < k\sigma^2$ for $\sigma \neq 0$, then instead of adding and subtracting $\alpha\sigma\phi(\sigma)$ from $-\dot{V}$ one can subtract $\alpha\phi(\sigma) \cdot (\sigma - k^{-1}\phi(\sigma))$. The proof carries over and the theorem remains the same except that $c'A(i\omega)^{-1}b$ is replaced by $c'A(i\omega)^{-1}b + k^{-1}$ (see [9]).

(iii) Let us make the change of variables $y(t) = e^{-\lambda t}x(t)$, where $x(t)$ is a solution of (3.1) and λ is any real number such that $\lambda > \operatorname{Re} \lambda_i$, $i = 1, \cdots, n$, and λ_i, $i = 1, \cdots, n$, are the characteristic roots of A. Note that λ may be positive or negative and the characteristic roots of A may have positive or negative real parts. Then $y(t)$ satisfies the equation

$$\dot{y} = (A - \lambda I)y - be^{-\lambda t}\phi(e^{\lambda t}c'y). \tag{3.5}$$

Let $V = y'By$ and then the derivative of V along the trajectories of (3.5) is

$$-\dot{V} = -y'\{(A - \lambda I)'B + B(A - \lambda I)\}y + 2\{Bb - \tfrac{1}{2}c\}'ye^{-\lambda t}\phi(e^{\lambda t}c'y) + c'ye^{-\lambda t}\phi(e^{\lambda t}c'y).$$

As before there exists a B such that V is positive definite and $-\dot{V} \geqq 0$ for all y, provided

$$T_1(z) = c'(A - \lambda I)(z)^{-1}b = c'A(z + \lambda)^{-1}b$$

is a positive real function. Thus $y(t)$ is bounded and the bound depends only on $\| y_0 \|$.

THEOREM 2. *If λ is as defined above and $T_1(z) = c'A(z + \lambda)^{-1}b$ is a positive real function, then there exists a nonnegative monotone scalar function $K(\cdot)$ such that $\| x(t) \| \leqq K(\| x_0 \|)e^{\lambda t}$, where $x(t)$ is the solution of (3.1) such that $x(0) = x_0$.*

REFERENCES

[1] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[2] ———, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. U. S. A., 49 (1963), pp. 201–205.

[3] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128–151.

[4] S. LEFSCHETZ, *Stability of Nonlinear Control Systems*, Academic Press, New York, 1964.

[5] V. A. YACUBOVICH, *Absolute stability of nonlinear control systems in the critical cases*, Avtomat. i Telemeh, 24 (1963), pp. 293–303, 717–731.

[6] K. R. MEYER, *On a system of equations in automatic control theory*, Contributions to Differential Equations, 3 (1964), pp. 163–173.

[7] V. A. YACUBOVICH, *The solution of certain matrix inequalities in automatic control theory*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1304–1307.

[8] K. R. MEYER, *Liapunov functions for the problem of Lur'e*, Proc. Nat. Acad. Sci. U. S. A., 53 (1965), pp. 501–503.

[9] M. A. AIZERMAN AND F. R. GANTMACHER, *Absolute Stability of Regulator Systems*, Holden-Day, San Francisco, 1964.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

A METHOD FOR COMPUTING LEAST SQUARES ESTIMATORS THAT KEEP UP WITH THE DATA*

ARTHUR ALBERT AND ROBERT W. SITTLER†

1. Introduction. The classical technique of least squares appears over and over again in a multitude of contemporary applications. For example, when one observes the values of a time function $z(t)$ at various (discrete) instants of time (say at $t = 1, 2, \cdots, n$) and one tries to find the best fit to $z(t)$ by a function of the form

$$\sum_{i=1}^{m} x_i \eta_i(t)$$

(where the $\eta_i(t)$ are specified), the weights x_i are customarily determined by the method of least squares. The classical (discrete time) Fourier analysis is exactly of this form when the functions η_i are sinusoidal. If the $\eta_i(t)$ are polynomials, the problem reduces to that of finding the best polynomial fit (of a given degree) to the time history $z(t)$. These are but two of the many possibilities, since the choice of the η-functions is completely arbitrary and at the disposal of the investigator. Once one decides on the "family" of functions to be used, least square theory addresses itself to the problem of finding those coefficients x_i which yield up a "best" fit in the sense of minimizing the mean-square residual error

$$E_n(x_1, \cdots, x_m) = \sum_{t=1}^{n} \left[z(t) - \sum_{i=1}^{m} x_i \eta_i(t) \right]^2. \tag{1.1}$$

The method of least squares also appears in a statistical setting. If one assumes that the observations $z(t)$ are of the form

$$z(t) = \sum_{i=1}^{m} x_i \eta_i(t) + v(t), \qquad t = 1, 2, \cdots, n, \tag{1.2}$$

where $v(t)$ is a zero mean uncorrelated stochastic process (white noise), the functions $\eta_i(t)$ are known, and the parameters $x_1, x_2, \cdots, x_m$ are to be estimated, this formulation leads to the classical problem of linear regression when one seeks to choose those estimators for the x's which are unbiased, linear in the data, and have minimum variance. The well known Gauss-Markov theorem shows that the least squares estimates for the x's have these properties. Further, if the noise is assumed to be Gaussian, then least squares estimators are also maximum likelihood estimators and

* Received by the editors February 8, 1964, and in final revised form May 28, 1965.
† ARCON Corporation, 803 Massachusetts Avenue, Lexington, Massachusetts.

have minimum variance in the class of all unbiased estimators (be they linear in the data or not). In addition, they enjoy many desirable decision theoretic properties when the loss function is Euclidean-distance.

Let us now examine the problem of actually determining the least squares estimator more closely. We begin by adopting the notational convenience of matrix algebra. Let Z be the n-dimensional (column) vector whose components are $z(1), z(2), \cdots, z(n)$, let H be the $n \times m$ matrix whose (i,j)th element is $\eta_j(i)$, $i = 1, \cdots, n, j = 1, 2, \cdots, m$, and we let X be the m-dimensional (column) vector whose components are $x_1, x_2, \cdots, x_m$.

The squared residual error of (1.1) is now expressible as a Euclidean distance, or equivalently, as an inner product:

$$E_n(X) = \sum_{t=1}^{n} [z(t) - \sum_{i=1}^{m} x_i \eta_i(t)]^2 = \| Z - HX \|^2 = (Z - HX)^t (Z - HX) \tag{1.3}$$

where A^t is the transpose of A. It is well known that $E_n(X)$ has a minimum at X if and only if X satisfies the so-called normal equations,

$$H^t H X = H^t Z, \tag{1.4}$$

so that the problem of choosing the weights $x_1, x_2, \cdots, x_m$ which yield the best fit to the observations $z(1), z(2), \cdots, z(n)$ (in the sense of minimizing (1.1)) is equivalent to the problem of finding a solution to (1.4).

Suppose $x_1, x_2, \cdots, x_m$ are chosen to minimize (1.1) and suppose that an additional observation, $z(n+1)$, is taken. How does one modify the current least squares coefficients to take into account the new data point? In terms of the linear equations (1.4) we notice that a new data point $z(n+1)$ adds an additional component to Z and requires that a new row be adjoined to the matrix H. In Theorem 2.4, we will show that $\hat{X}_{n+1}$ (the least squares estimate based upon $n+1$ observations) is related to $\hat{X}_n$ (the least squares estimate based upon n observations) by a first order difference scheme of the form

$$\hat{X}_{n+1} = \hat{X}_n + K_{n+1}[z(n+1) - h_{n+1}^t \hat{X}_n],$$

where $z(n+1)$ is the $(n+1)$st observation, h_{n+1}^t is the new row of the H matrix

$$h_{n+1} = \begin{bmatrix} \eta_1(n+1) \\ \eta_2(n+1) \\ \vdots \\ \eta_m(n+1) \end{bmatrix},$$

and K_{n+1} is an m-dimensional vector which is defined recursively in terms of the vectors $h_1, \cdots, h_{n+1}$.

In §3 we show how the problem of minimizing (1.3) subject to linear constraints can be reduced to the problem of minimizing (1.3) subject to no constraints (but with a new H-matrix), and following Theorem 3.1, we indicate how the constrained least squares estimates can be computed iteratively.

In many applications (particularly regression analysis and the analysis of variance) the minimal squared residual error,

$$E_n = \inf_X \| Z_n - H_n X \|^2,$$

is a quantity of prime interest. In §4 we ask the question: What happens to E_n when a new observation is taken? In Theorem 4.1 we will show that E_{n+1} (the residual based upon $n + 1$ observations) is related to E_n by a recursion of the form

$$E_{n+1} = E_n + k_{n+1}[z(n + 1) - h_{n+1}^t \hat{X}_n]^2,$$

where $\{k_{n+1}\}$ are a sequence of scalars which can be computed iteratively in a convenient manner. In §5 we extend the results of §2 and §4 to the case of weighted least squares and in so doing, we will arrive at an alternate approach to the recursive calculation of constrained least squares estimators (and their associated residual errors) in which constraints are treated as "infinitely reliable" observations in a weighted least squares recursion.

It seems natural that these results may be applied to situations where computations must be performed in "real time" as the data flow in. For example, the computation of satellite orbit parameters is often of this nature. The sequential analysis of variance is yet another such application.

Our methods and results were prompted by Kalman's work (mainly [5], where he pointed out that least squares theory could be viewed as a limiting case of his essentially Bayesian approach). The development in §§2–4 is intended to be elementary. It is, therefore, lengthier (and hopefully more instructive) than it might be if economy of exposition were our paramount interest. For example, some of the results of §2 are scattered throughout the literature in the context of the theory of generalized inverses (see [2], [3], [4], and [6]), but we have rederived them as purely analytic properties of matrices, without any references to their geometric interpretation in terms of generalized inverses and projections. On the other hand, §5 depends crucially on several established results in the theory of generalized inverses whose proofs are not included because of their great length.

2. How to update the last least squares estimate. Let Z be an n-dimensional vector and let H be an $n \times m$ matrix. In the last section it was shown that the m-vector $\hat{X}$ minimizes $\| Z - HX \|^2$ if and only if $\hat{X}$ is a solution

of the system of so-called normal equations

$$H^tHX = H^tZ, \tag{2.1}$$

where H^t is the transpose of H.

Equation (2.1) always has a solution in X. In fact there is a (unique) value of X of minimum norm which satisfies (2.1). This fact will be stated as Theorem 2.1. A preliminary lemma is required.

LEMMA 1. *For any matrix* H,

(a) *the matrices*

$$P(H) = \lim_{\epsilon \to 0+} HH^t(HH^t + \epsilon I)^{-1}$$

and

$$\lim_{\epsilon \to 0+} H(H^tH + \epsilon I)^{-1}H^t$$

always exist and are equal;

(b) $P^2(H) = P(H)$;

(c) $H^tP(H) = H^t$;

(d) *the matrices*

$$\lim_{\epsilon \to 0+} (HH^t + \epsilon I)^{-1}H$$

and

$$\lim_{\epsilon \to 0+} H(H^tH + \epsilon I)^{-1}$$

exist and are equal;

(e) *for any vector* X, *if* $HX \equiv HP(H^t)X = 0$, *then* $P(H^t)X = 0$.

Proof. Since

$$H^t(HH^t + \epsilon I) = (H^tH + \epsilon I)H^t$$

(I will always denote the identity matrix of the appropriate dimension), and since HH^t and H^tH are nonnegative definite, it follows that $(HH^t + \epsilon I)$ and $(H^tH + \epsilon I)$ are positive definite when $\epsilon > 0$ and so have inverses. Hence

$$H^t(HH^t + \epsilon I)^{-1} = (H^tH + \epsilon I)^{-1}H^t$$

for every positive ϵ, so that if either of the limits in (a) or either of the limits in (d) exist, they are equal.

(a) HH^t is symmetric, so there are an orthogonal matrix T and a diagonal matrix D such that $T(HH^t)T^t = D$, where $D = \text{diag}\ (\lambda_1, \lambda_2, \cdots, \lambda_n)$.

(In the remainder of this section, T, D, and the matrix

$U = \text{diag}\,(v_1, v_2, \cdots, v_n)$ where

$$v_i = \begin{cases} 1 & \text{if } \lambda_i \neq 0, \\ 0 & \text{if } \lambda_i = 0, \end{cases} \qquad i = 1, 2, \cdots, n,$$

will remain as defined here.) Thus,

$$(HH^t)(HH^t + \epsilon I)^{-1} = T^t D(D + \epsilon I)^{-1} T \to T^t U T$$

as $\epsilon \to 0$, so that $P(H)$ exists and is equal to $T^t U T$.

(b) Since $U^2 = U$ and $UD = DU = D$, $P^2(H) = P(H)$, and

(c) $HH^t P(H) = T^t DUT = HH^t$, so that for any vector ,

$$\| H^t(I - P(H))Y \|^2 = [H^t(I - P(H))Y]^t[H^t(I - P(H))Y]$$
$$= [(I - P(H))\, Y]^t[HH^t(I - P(H))Y] = 0,$$

so that $H^t(I - P(H)) = 0$.

(d) As stated in the beginning of the proof, if either of the limits in (d) exists, then both exist and are equal. Replacing H by H^t in (c), we see that $H = HP(H^t)$ so that by (a),

$$(HH^t + \epsilon I)^{-1}H = (HH^t + \epsilon I)^{-1}HP(H^t)$$
$$= (HH^t + \epsilon I)^{-1} \lim_{\delta \to 0} HH^t(HH^t + \delta I)^{-1}H = T^t(D + \epsilon I)^{-1}UTH.$$

But

$$\lim_{\epsilon \to 0} (D + \epsilon I)^{-1}U$$

exists and equals $V = \text{diag}\,(v_1, v_2, \cdots, v_n)$, where

$$v_i = \begin{cases} 0 & \text{if } \lambda_i = 0, \\ \dfrac{1}{\lambda_i} & \text{otherwise.} \end{cases}$$

It follows that

$$\lim_{\epsilon \to 0} (HH^t + \epsilon I)^{-1}H = T^t V TH$$

exists.

(e) If $HP(H^t)X = 0$, then for every $\epsilon > 0$,

$$0 = [HP(H^t)X]^t[H(H^tH + \epsilon I)^{-1}X] = [P(H^t)X]^t[H^tH(H^tH + \epsilon I)^{-1}X].$$

Letting $\epsilon \to 0$, we see that

$$[P(H^t)X]^t[P(H^t)X] = \| P(H^t)X \|^2 = 0.$$

(The geometrically minded reader may have noticed that $P(H)$ is the operator which projects onto the range of H.)

THEOREM 2.1. *Let H be an $n \times m$ matrix and let Z be an n-vector. Among the class of vectors which satisfy the equation*

$$H^tHX = H^tZ \tag{2.1}$$

(and hence minimize $\| Z - HX \|^2$), the vector

$$\hat{X} = \lim_{\epsilon \to 0+} (H^tH + \epsilon I)^{-1}H^tZ \tag{2.2}$$

is the unique vector of minimum norm.

Proof. From Lemma 1(d), $\hat{X}$ always exists.

$$H^tH\hat{X} = H^tP(H)Z = H^tZ,$$

by Lemma 1(a, c), so that $\hat{X}$ satisfies (2.1).

Let Y be any solution to (2.1). Then

$$Y = P(H^t)Y + [I - P(H^t)]Y$$

Since $HP(H^t) = H$, we see that

$$H^tZ = H^tHY = H^tH[P(H^t)Y],$$

so that $P(H^t)Y$ satisfies (2.1). Since $P^2(H^t) = P(H^t)$,

$$[P(H^t)Y]^t[I - P(H^t)]Y = Y^t[P(H^t) - P^2(H^t)]Y = 0,$$

so that $P(H^t)Y$ is orthogonal to $[I - P(H^t)]Y$. Thus

$$\| Y \|^2 = \| P(H^t)\, Y \|^2 + \| (I - P(H^t)Y \|^2 \geqq \| P(H^t)Y \|^2$$

with strict inequality holding unless $Y = P(H^t)Y$.

Since $P(H^t)\, Y$ and $P(H^t)\hat{X}$ satisfy (2.1),

$$H^tHP(H^t)[\, Y - \hat{X}] = 0.$$

Since

$$\begin{aligned} \| HP(H^t)[\, Y - \hat{X}] \|^2 &= [HP(H^t)(\, Y - \hat{X})]^t[HP(H^t)(\, Y - \hat{X})] \\ &= [P(H^t)(Y - \hat{X})]^t[H^tHP(H^t)(Y - \hat{X})] = 0, \end{aligned}$$

it follows that $HP(H^t)(\, Y - \hat{X}) = 0$. From Lemma 1(e), we conclude $P(H^t)\hat{X} = P(H^t)\, Y$. But using Lemma 1(d), (c), with H replaced by H^t throughout,

$$P(H^t)\hat{X} = \lim_{\epsilon \to 0+} P(H^t)H^t(HH^t + \epsilon I)^{-1}Z = \lim_{\epsilon \to 0+} H^t(HH^t + \epsilon I)^{-1}Z = \hat{X},$$

so that $P(H^t)\, Y = \hat{X}$. Thus, we have shown that if Y is a solution to (2.1), then $\| Y \| \geqq \| \hat{X} \|$, with strict inequality holding unless $Y = \hat{X}$.

Let us hereafter use the notation

$$\hat{X}(\epsilon) = (H^tH + \epsilon I)^{-1}H^tZ. \tag{2.3}$$

$\hat{X}(\epsilon)$ has a useful statistical interpretation. If Z is a random variable of the form

$$Z = HX + V, \tag{2.4}$$

where H is a (known) matrix, X is a zero mean normal random vector whose covariance is $(1/\epsilon)I$ and V is a zero mean normal random vector whose covariance is the identity matrix, then

$$\hat{X}(\epsilon) = E(X \mid Z), \tag{2.5}$$

where $E(X \mid Z)$ is the conditional expectation of X given Z.

If one thinks of Z as a set of observations and X as a non-observable quantity which must be estimated on the basis of the Z's, it is natural to wonder how the acquisition of a new piece of data will affect the current estimate of X. In [5] Kalman has examined this question in detail and in great generality. We will specialize his results to the case at hand (and for the sake of self containment, prove them here). Since $\hat{X}$ (defined in (2.2)) has an interpretation as the *least squares* estimator of X, and since $\hat{X} = \lim_{\epsilon \to 0} \hat{X}(\epsilon)$, we will be able to derive a similar recursion for least squares estimators by passing to the limit appropriately.

THEOREM 2.2. (Kalman). *Let* $v(1), v(2). \cdots$ *be independent identically distributed normal random variables with zero mean and unit variance. Let* X *be an* m*-dimensional vector-valued normal random variable with zero mean and covariance* $(1/\epsilon)I$. *Let* $h_1, h_2, \cdots$ *be a known sequence of* m*-dimensional vectors; let*

$$z(n) = h_n{}^tX + v(n), \qquad n = 1, 2, \cdots;$$

let H_n *be the* $n \times m$ *matrix whose* j*th row vector is* $h_j{}^t$, $j = 1, 2, \cdots, n$; *let*

$$Z_n = \begin{bmatrix} z(1) \\ z(2) \\ \vdots \\ z(n) \end{bmatrix};$$

and let

$$\hat{X}_n(\epsilon) = E(X \mid Z_n), \qquad n = 1, 2, \cdots. \tag{2.6}$$

Then

$$\hat{X}_n(\epsilon) = \hat{X}_{n-1}(\epsilon) + K_n(\epsilon)[z(n) - h_n{}^t\hat{X}_{n-1}(\epsilon)], \tag{2.7}$$

where $\hat{X}_0(\epsilon) = 0$, $n = 1, 2, \cdots$; $K_n(\epsilon)$ *is an m-vector of the form*

$$K_n(\epsilon) = \frac{\Sigma_{n-1}(\epsilon)h_n}{1 + h_n{}^t\Sigma_{n-1}(\epsilon)h_n}; \tag{2.8}$$

and $\Sigma_n(\epsilon)$ *is the covariance of* $X - \hat{X}_n(\epsilon)$. $\Sigma_n(\epsilon)$ *satisfies a first order recursion,*

$$\Sigma_n(\epsilon) = \Sigma_{n-1}(\epsilon) - \frac{[\Sigma_{n-1}(\epsilon)h_n][\Sigma_{n-1}(\epsilon)h_n]^t}{1 + h_n{}^t\Sigma_{n-1}(\epsilon)h_n}, \quad n = 1, 2, \cdots,$$

$$\Sigma_0(\epsilon) = \frac{1}{\epsilon} I, \tag{2.9a}$$

and is given in closed form by

$$\Sigma_n(\epsilon) = \frac{1}{\epsilon}[I - (H_n{}^tH_n + \epsilon I)^{-1}H_n{}^tH_n]. \tag{2.9b}$$

Proof. For the purpose of this proof, we will fix $\epsilon > 0$ and suppress it in the notation so that for the time being $\hat{X}_n(\epsilon)$ becomes $\hat{X}_n$ and $\Sigma_n(\epsilon)$ becomes Σ_n.

Let

$$\tilde{z}(n) = z(n) - h_n{}^t\hat{X}_{n-1},$$

and

$$\tilde{X}_n = X - \hat{X}_{n-1}.$$

Since

$$E\hat{X}_{n-1}Z_{n-1}^t = EE(XZ_{n-1}^t \mid Z_{n-1}) = E(XZ_{n-1}^t),$$

it follows that $E\tilde{X}_nZ_{n-1}^t = 0$, so that $\tilde{X}_n$ and Z_{n-1} are also independent. So then are $\tilde{z}(n)$ and Z_{n-1}. Thus,

$$\hat{X}_n = E(X \mid Z_n) = E(X \mid Z_{n-1}, \tilde{z}(n))$$

$$= E(X \mid Z_{n-1}) + E(X \mid \tilde{z}(n)) = \hat{X}_{n-1} + \left\{\frac{E\tilde{z}(n)X}{E\tilde{z}^2(n)}\right\}\tilde{z}(n).$$

(The intermediate step is a well known property of normal distributions. We also use the fact that for any zero mean vectors having a joint normal distribution, $E(X \mid Y) = E(X\,Y^t)E(Y\,Y^t)^{-1}Y$.)

Let

$$\Sigma_{n-1} = E\tilde{X}_n\tilde{X}_n{}^t. \tag{2.10a}$$

Since

$$\tilde{z}(n) = h_n{}^t\tilde{X}_n + v(n),$$

it is easy to see that

$$E\tilde{z}^2(n) = h_n{}^t\Sigma_{n-1}h_n + 1. \tag{2.10b}$$

Since $\tilde{X}_n$ and Z_{n-1} are independent,

$$E\tilde{X}_n\hat{X}_{n-1}^t = 0, \tag{2.10c}$$

so that

$$E\tilde{z}(n)X = E\tilde{z}(n)\tilde{X}_n = \Sigma_{n-1}h_n . \tag{2.10d}$$

Thus,

$$\hat{X}_n = \hat{X}_{n-1} + K_n[z(n) - h_n{}^t\hat{X}_{n-1}],$$

where $K_n = K_n(\epsilon)$ is defined in (2.8). We establish (2.9) by noting that

$$\hat{X}_n = \hat{X}_{n-1} + E(X \mid \tilde{z}(n)),$$

so that

$$\tilde{X}_n = \tilde{X}_{n+1} + E(X \mid \tilde{z}(n)).$$

The second term on the right is proportional to $\tilde{z}(n) = h_n{}^t\tilde{X}_n + v(n)$. In fact, $E(X \mid \tilde{z}(n)) = K_n\tilde{z}_n$. Since $\tilde{X}_{n+1}$ is independent of Z_n, it is also independent of $\tilde{z}(n)$. Thus,

$$\Sigma_{n-1} = E\tilde{X}_n\,\tilde{X}_n{}^t = E\tilde{X}_{n+1}\,\tilde{X}_{n+1}^t + K_n\,K_n{}^tE\tilde{z}_n{}^2 = \Sigma_n + \frac{[\Sigma_{n-1}\,h_n][\Sigma_{n-1}\,h_n]^t}{1 + h_n{}^t\Sigma_{n-1}h_n}.$$

To establish (2.9b), we proceed as follows. From (2.10),

$$\Sigma_n(\epsilon) = E\tilde{X}_{n+1}\tilde{X}_{n+1}^t = E\tilde{X}_{n+1}(X - \hat{X}_n(\epsilon))^t = E\tilde{X}_{n+1}X^t$$

(since $E\tilde{X}_{n+1}\hat{X}_n{}^t(\epsilon) = 0$). Since

$$\hat{X}_n(\epsilon) = (H_n{}^tH_n + \epsilon I)^{-1}H_n{}^tZ_n$$

by (2.3), and since

$$Z_n = H_nX + V_n ,$$

we see that

$$\Sigma_n(\epsilon) = E(X - \hat{X}_n(\epsilon))X^t = \frac{1}{\epsilon}\,I - \frac{1}{\epsilon}\,(H_n{}^tH_n + \epsilon I)^{-1}H_n{}^tH_n .$$

The recursion (2.7)–(2.9) has a more general interpretation which we now state.

THEOREM 2.3. *Let h_1, h_2, $\cdots$ be a sequence of m-dimensional vectors, and let H_n be the $n \times m$ matrix whose jth row vector is $h_j{}^t$. Let $z(1), z(2), \cdots$ be a sequence of real numbers and let Z_n be the n-vector whose jth component is $z(j)$. For each n, let*

$$\hat{X}_n(\epsilon) = (H_n{}^tH_n + \epsilon I)^{-1}H_n{}^tZ_n , \qquad n = 1, 2, \cdots$$

Then

$$\hat{X}_n(\epsilon) = \hat{X}_{n-1}(\epsilon) + K_n(\epsilon)[z(n) - h_n{}^t\hat{X}_{n-1}(\epsilon)], \tag{2.11}$$

where $K_n(\epsilon)$ *is defined in* (2.8) *and* (2.9).

In words, $\hat{X}_{n-1}(\epsilon)$ is the solution of the equation

$$(H_{n-1}^t H_{n-1} + \epsilon I)X = H_{n-1}^t Z_{n-1}$$

and $\hat{X}_n(\epsilon)$ is the solution of the equation

$$(H_n{}^t H_n + \epsilon I)X = H_n{}^t Z_n ,$$

where H_n is obtained by adjoining a new row $(h_n{}^t)$ to H_{n-1} and Z_n is obtained by adjoining a new component to Z_{n-1}.

Proof. We need only point out that if each $z(n)$ is a random variable of the form $z(n) = h_n{}^t X + v(n)$, where the $v(n)$ are independent, zero mean normal variates with unit variance and X is a zero mean normal vector with covariance of $(1/\epsilon)I$, then the conditional expectation of X given Z_n is exactly

$$H_n{}^t(H_n H_n{}^t + \epsilon I)^{-1} Z_n = (H_n{}^t H_n + \epsilon I)^{-1} H_n{}^t Z_n .$$

By Theorem 2.2, the (purely algebraic) relation (2.7) holds between $E(X \mid Z_n)$ and $E(X \mid Z_{n-1})$, and so (2.7) must also hold between $(H_n{}^t H_n + \epsilon I)^{-1} H_n{}^t Z_n$ and $(H_{n-1}^t H_{n-1} + \epsilon I)^{-1} H_{n-1}^t Z_{n-1}$ for any Z_n and any H_n.

By letting ϵ tend to zero in (2.11), we can now establish a similar recursion between least squares estimators. We remind the reader that a vector $\hat{X}$ minimizes the Euclidean distance $\| Z - HX \|$ if and only if $\hat{X}$ satisfies the (so-called "normal") equation, $H^t H\hat{X} = H^t Z$.

THEOREM 2.4. *Let* $z(1), z(2), \cdots$ *be a sequence of real numbers* (*observations*), *and let* $h_1, h_2, \cdots$ *be a sequence of m-dimensional vectors. Let* H_n *be the* $n \times m$ *matrix whose* j*th row is* $h_j{}^t$, $j = 1, 2, \cdots, n$, *and let* Z_n *be the n-vector whose components are* $z(1), z(2), \cdots, z(n)$. *For each n, let* $\hat{X}_n$ *be the m-dimensional vector which is the minimum norm solution* (*in* X) *of the equation* $H_n{}^t H_n X = H_n{}^t Z_n$, $n = 1, 2, \cdots$. *Then*

$$\hat{X}_n = \hat{X}_{n-1} + K_n[z(n) - h_n{}^t\hat{X}_{n-1}], \qquad \hat{X}_0 = 0, \tag{2.12}$$

where

$$K_n = \begin{cases} \dfrac{A_{n-1}h_n}{h_n{}^t A_{n-1}\, h_n} & \text{if } h_n \text{ is not a linear combination of } h_1, \cdots, h_{n-1}, \\ \dfrac{B_{n-1}\, h_n}{1 + h_n{}^t B_{n-1}\, h_n} & \text{otherwise;} \end{cases} \tag{2.13}$$

$$(2.14)\qquad B_n = \begin{cases} B_{n-1} - \dfrac{[B_{n-1}\,h_n][A_{n-1}\,h_n]^t + [A_{n-1}\,h_n][B_{n-1}\,h_n]^t}{h_n{}^t A_{n-1}\,h_n} \\ \quad + \dfrac{[1 + h_n{}^t B_{n-1}\,h_n]}{(h_n{}^t A_{n-1} h_n)^2}\,[A_{n-1}\,h_n][A_{n-1}\,h_n]^t & \text{if } h_n \text{ is not a linear combination of } h_1, h_2, \cdots, h_{n-1}, \\ B_{n-1} - \dfrac{[B_{n-1}\,h_n][B_{n-1}\,h_n]^t}{h_n{}^t B_{n-1}\,h_n + 1} & \text{otherwise;} \end{cases}$$

A_n satisfies the recursion,

$$(2.15)\qquad A_n = \begin{cases} A_{n-1} - \dfrac{[A_{n-1}\,h_n][A_{n-1}\,h_n]^t}{h_n{}^t A_{n-1}\,h_n} & \text{if } h_n \text{ is not a linear combination of } h_1, h_2, \cdots, h_{n-1}, \\ A_{n-1} & \text{otherwise;} \end{cases}$$

A_0 is the $m \times m$ identity matrix, and B_0 is the $m \times m$ null matrix. For every value of n,

$$(2.16)\qquad A_n = I - H_n{}^t H_n B_n = I - B_n H_n{}^t H_n\,.$$

Proof. We begin the proof by asserting that $\Sigma_n(\epsilon)$ defined in (2.9) can be written as

$$(2.17)\qquad \Sigma_n(\epsilon) = \frac{1}{\epsilon}[A_n + \epsilon B_n + o(\epsilon)],$$

where the term $o(\epsilon)/\epsilon$ converges to zero as $\epsilon \to 0$, and where A_n and B_n satisfy the first halves of (2.15) and (2.14), respectively, if $A_{n-1}h_n \neq 0$, and the second halves otherwise. This assertion is true for $n = 0$, and is verified by induction, to be true for all n.

Inserting (2.16) into (2.8), we find that

$$K_n(\epsilon) = \begin{cases} \dfrac{A_{n-1}\,h_n}{h_n{}^t A_{n-1}\,h_n} + o(1) & \text{if } A_{n-1}\,h_n \neq 0, \\ \dfrac{B_{n-1}\,h_n}{1 + h_n{}^t B_{n-1}\,h_n} + o(1) & \text{if } A_{n-1}\,h_n = 0, \end{cases}$$
$$= K_n + o(1)$$

where K_n is defined in (2.13) and $o(1)$ tends to zero as $\epsilon \to 0$.

Thus, by Theorem 2.3,

$$\hat{X}_n(\epsilon) = \hat{X}_{n-1}(\epsilon) + [K_n + o(1)][z(n) - h_n{}^t \hat{X}_{n-1}(\epsilon)]$$

for every n, where

$$\hat{X}_n(\epsilon) = (H_n{}^t H_n + \epsilon I)^{-1} H_n{}^t Z_n\,.$$

By Theorem 2.1, $\hat{X}_n = \lim_{\epsilon \to 0} \hat{X}_n(\epsilon)$ for every n.

Thus,

$$\hat{X}_n = \hat{X}_{n-1} + K_n[z(n) - h_n{}^t\hat{X}_{n-1}].$$

It is easy to show (either by induction or by referring to Theorem 2.5) that $A_{n-1}h_n = 0$ if and only if h_n is a linear combination of $h_1, \cdots, h_{n-1}$. This establishes the first part of Theorem 2.4.

Now, to establish (2.16), let

$$D_n = \text{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_m), \tag{2.18}$$

where the λ_i are the eigenvalues of $H_n{}^tH_n$, let

$$V_n = \text{diag}\,(v_1, v_2, \cdots, v_m), \tag{2.19}$$

where

$$v_i = \begin{cases} 1/\lambda_i & \text{if } \lambda_i \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.20}$$

and let

$$U_n = D_nV_n. \tag{2.21}$$

If T_n is the orthogonal matrix which reduces $H_n{}^tH_n$ to the diagonal form D_n, then

$$H_n{}^tH_n = T_n{}^tD_nT_n, \tag{2.22}$$

so from (2.9b),

$$\Sigma_n(\epsilon) = \frac{1}{\epsilon} T_n{}^t[I - (D_n + \epsilon I)^{-1}D_n]T_n. \tag{2.23}$$

Thus,

$$\lim_{\epsilon\to 0} \epsilon\Sigma_n(\epsilon) = T_n{}^t(I - U_n)T_n. \tag{2.24}$$

Combining (2.24) and (2.17), we see that

$$A_n = T_n{}^t(I - U_n)T_n. \tag{2.25}$$

Furthermore,

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}[\epsilon\Sigma_n(\epsilon) - A_n] = B_n = T_n{}^tV_nT_n. \tag{2.26}$$

But

$$H_n{}^tH_nB_n = T_n{}^tD_nV_nT_n = T_n{}^tU_nT_n = I - A_n = T_n{}^tV_nD_nT_n = B_nH_n{}^tH_n,$$

thereby establishing (2.16).

The matrices A_n, defined in (2.15), have a useful interpretation. If one

were to orthogonalize the set of vectors $h_1, h_2, \cdots, h_n$ by the Gram-Schmidt orthogonalization, (in the given order) the jth orthogonal vector in the set would be nonzero if and only if h_j were linearly independent of $h_1, h_2, \cdots, h_{j-1}$. If the nonzero vectors in the orthogonalized set are normalized and the entire set of zero and unit vectors denoted by $\{\varphi_j\}_{j=1}^{n}$, then it turns out that

$$A_{j-1}h_j = \| A_{j-1}h_j \| \varphi_j .$$

In fact, $I - A_j$ is, for each j, the operator that projects any vector onto the subspace spanned by $h_1, h_2, \cdots, h_j$. This we state as follows.

THEOREM 2.5. *Let $h_1, h_2, \cdots, h_n, \cdots$ be a sequence of m-dimensional vectors, and let φ_j be the jth orthonormal vector obtained by the Gram-Schmidt orthogonalization,*

$$\varphi_j = \begin{cases} 0 & \text{if } h_j \text{ is a linear combination of } h_1, h_2, \cdots, h_{j-1}, \\ \dfrac{\left[h_j - \sum_{k=1}^{j-1} \varphi_k(\varphi_k{}^t h_j)\right]}{\left\| h_j - \sum_{k=1}^{j-1} \varphi_k(\varphi_k{}^t h_j) \right\|}, & \text{otherwise.} \end{cases}$$

Let A_n be a sequence of matrices defined inductively by the recursion

$$A_{n+1} = \begin{cases} A_n - \dfrac{[A_n h_{n+1}][A_n h_{n+1}]^t}{h_{n+1}^t A_n h_{n+1}} & \text{if } h_{n+1} \text{ is not a linear combination of } h_1, \cdots, h_n, \\ A_n, & \text{otherwise,} \end{cases}$$

where A_0 is the identity matrix.

Then for every $n \geqq 1$,

$$I - A_n = \sum_{j=1}^{n} \varphi_j \varphi_j{}^t,$$

so that for any vector y, $(I - A_n)y$ is the projection of y onto the linear manifold spanned by $h_1, \cdots, h_n$.

Proof. We proceed by induction. The assertion is true for $n = 1$. If the assertion is true up to n, then $A_n h_{n+1} = 0$ if and only if

$$h_{n+1} = \sum_{j=1}^{n} \varphi_j(\varphi_j{}^t h_{n+1}),$$

which is true if and only if h_{n+1} is a linear combination of the vectors $h_1, \cdots, h_n$. Under the induction hypothesis,

$$A_n h_{n+1} = h_{n+1} - \sum_{j=1}^{n} \varphi_j(\varphi_j{}^t h_{n+1}).$$

Thus

$$A_n - A_{n+1} = \begin{cases} 0 & \text{if } h_{n+1} \text{ is a linear combination of the vectors } h_1, \cdots, h_n, \\ \varphi_{n+1}\varphi_{n+1}^t, & \text{otherwise,} \end{cases}$$

so that

$$(I - A_{n+1}) = (I - A_n) + \varphi_{n+1}\varphi_{n+1}^t,$$

proving the assertion.

3. Least squares with constraints. In the last section, we showed how a least squares estimator can be computed in "real time" as the data roll in. In many applications, it is necessary that the least squares estimator be computed subject to certain linear constraints. That is to say, instead of choosing X to minimize

$$\| Z_n - H_n X \|^2 = \sum_{j=1}^{n} [z(j) - h_j^t X]^2, \tag{3.1}$$

one wishes to choose X subject to constraints of the form

$$g_i^t X = w(i), \qquad i = 1, 2, \cdots, k, \tag{3.2}$$

so as to minimize (3.1). (Here, the g_i are a sequence of given vectors and $w(i)$ are given scalars.)

We will show that the problem of minimizing (3.1) subject to constraints of the form (3.2) can be reduced to the problem of minimizing an expression of the form (3.1) (with different h vectors) subject to no constraints. Furthermore, the relationship between the "new" h-vectors and the "old" ones is a straightforward (easily computed) one.

We begin by pointing out that the problem of minimizing $\| Z - HX \|^2$ subject to the constraints $GX = W$ (where G is a matrix and W is a prescribed vector) is reducible to one where W is zero.

For, let X_0 be any solution of the equation $GX = W$. Then X minimizes

$$\| Z - HX \|^2 = \| (Z - HX_0) - H(X - X_0) \|^2$$

subject to the constraint $GX = W$ if and only if $Y = X - X_0$ minimizes $\| Z^* - H Y \|^2$ subject to the constraint $G Y = 0$ (where $Z^* = Z - HX_0$). So, if Y^* minimizes $\| Z^* - H Y \|^2$ subject to $G Y = 0$, then $X^* = Y^* + X_0$ minimizes $\| Z - HX \|^2$ subject to $GX = W$. For this reason, we will always take W equal to zero in the sequel.

The constraints $GX = 0$ dictate that X must be orthogonal to the rows of G. Let the row vectors of G be denoted by $g_1^t, g_2^t, \cdots, g_k^t$. If we orthogonalize these vectors (as in Theorem 2.5) and then choose an orthonormal basis

for the remainder of the space (call this basis $\varphi_{k+1}, \cdots, \varphi_m$), then the set of X's satisfying $GX = 0$ is the same as the set of X's which are linear combinations of $\varphi_{m+1}, \cdots, \varphi_k$.

To put it another way, the set of X's satisfying $GX = 0$ is the same as the set of X's which are orthogonal to $g_1, g_2, \cdots, g_k$, which, in turn, is the same as the set of X's of the form $X = A\,Y$, where $I - A$ is the projection onto the linear subspace spanned by $g_1, g_2, \cdots, g_k$. This argument is the substance of the next theorem.

THEOREM 3.1. *Let H_n by the $n \times m$ matrix whose jth row vector is h_j^t, $j = 1, 2, \cdots, n$, and let G_k be the $k \times m$ matrix whose jth row vector is g_j^t, $j = 1, 2, \cdots, k$.*

Let $I - A_k$ be the matrix which projects onto the linear subspace spanned by $g_1, g_2, \cdots, g_k$, and let $\bar{H}_n$ be the $n \times m$ matrix whose jth row vector is $\bar{h}_j^t = (A_k h_j)^t$, $j = 1, 2, \cdots, n$.

Let $\hat{Y}_n$ be the vector of minimum norm among those which minimize $\| Z_n - \bar{H}_n Y \|$ (i.e., the minimum norm solution of the normal equations $\bar{H}_n{}^t \bar{H}_n Y = \bar{H}_n{}^t Z_n$).

Then $\hat{X}_n = \hat{Y}_n$ minimizes $\| Z_n - H_n X \|$ subject to the constraints $G_n X = 0$.

Proof. As was mentioned above, the set of X's for which $G_k X = 0$ is the same as the set of X's which are of the form $X = A_k Y$ for some Y, where $I - A_k$ is the projection onto the set $g_1, g_2, \cdots, g_k$ and is given inductively by Theorem 2.5. Thus

$$\inf_{G_k X = 0} \| Z_n - H_n X \| = \inf_Y \inf_{X = A_k Y} \| Z_n - H_n X \| = \inf_Y \| Z_n - H_n A_k Y \|.$$

But $\bar{H}_n = H_n A_k$, so that if Y minimizes $\| Z_n - \bar{H}_n Y \|$, then $X = A_k Y$ minimizes $\| Z_n - H_n X \|$ subject to the constraints $G_k X = 0$. By Theorem 2.1,

$$\hat{Y}_n = \lim_{\epsilon \to 0} (\bar{H}_n{}^t \bar{H}_n + \epsilon I)^{-1} \bar{H}_n{}^t Z_n$$

is the vector of minimum norm among those which minimize $\| Z_n - \bar{H}_n Y \|$. By Lemma 1d,

$$\hat{Y}_n = \lim_{\epsilon \to 0} \bar{H}_n{}^t (\bar{H}_n \bar{H}_n{}^t + \epsilon I)^{-1} Z_n = \lim_{\epsilon \to 0} A_k{}^t H_n{}^t (\bar{H}_n \bar{H}_n{}^t + \epsilon I)^{-1} Z_n .$$

Since A_k is a projection,

$$A_k{}^t = A_k = A_k{}^2,$$

so that

$$A_k \hat{Y}_n = \lim_{\epsilon \to 0} A_k{}^2 H_n{}^t (\bar{H}_n \bar{H}_n{}^t + \epsilon I)^{-1} Z_n = \hat{Y}_n .$$

Thus, $X = \hat{Y}_n$ minimizes $\| Z_n - H_n X \|$ subject to $G_k X = 0$, as asserted.

From the point of view of "real time" computation, one would proceed as follows. In advance of the data acquisition, one would use the constraint vectors, $g_1, g_2, \cdots, g_k$ to compute the matrices $A_1, A_2, \cdots$, etc. According to Theorem 2.5, $A_0 = I$, and for $j = 0, \cdots, k-1$,

$$A_{j+1} = \begin{cases} A_j & \text{if } g_{j+1} \text{ is a linear combination of } g_1, \cdots, g_j, \\ A_j - \dfrac{[A_j g_{j+1}]\,[A_j g_{j+1}]^t}{g_{j+1}^t A_j g_{j+1}}, & \text{otherwise.} \end{cases}$$

This procedure terminates when A_k is obtained. Then, in "real time" (i.e., as the data occurs) one applies A_k to the regression vectors h_n to obtain the regression vectors $\bar{h}_n$, and $\hat{X}_n$ is obtained recursively according to the iterations (2.12)–(2.14), where h_n is replaced throughout by $\bar{h}_n$.

4. Residual errors. After one has computed a least squares estimate (either constrained or unconstrained) it is often desired to evaluate the residual error. That is to say, one wishes to evaluate the expression

$$E_n = \inf_X \| Z_n - H_n X \|^2$$

in the unconstrained case and

$$\bar{E}_n = \inf_{GX=0} \| Z_n - H_n X \|^2 = \inf_Y \| X_n - \bar{H}_n Y \|^2$$

in the constrained case.

This type of computation is at the very heart of linear regression and the analysis of variance. In general, a model of the form $Z_n = H_n X + V_n$ is assumed for the observations Z_n (where V_n is the vector of measurement errors), and one compares this model with one that stipulates that a set of linear relationships (of the form $GX = 0$) exists among the components of X. In order to decide which model fits best, the residuals

$$E_n = \inf_X \| Z_n - H_n X \|^2 \quad \text{and} \quad \bar{E}_n = \inf_{GX=0} \| Z_n - H_n X \|^2$$

must be compared and if their ratio is "unduly" large (or small) one model is accepted in favor of the other.

The obvious computational approach is not entirely satisfactory if a running (real time) evaluation of E_n and/or $\bar{E}_n$ are/is required. By this, we mean that the technique of substituting the least squares estimator for X into the expression $\| Z_n - H_n X \|^2$ becomes increasingly tedious as the number of observations (and hence the dimensionality of Z_n) grows increasingly large.

If one wishes to perform a sequential analysis of variance, it is evident that a computational method which allows E_{n+1} (the residual error based

upon $n + 1$ observations) to be computed in a convenient fashion from E_n (the residual error based upon n observations) will be of great practical importance.

In this section we will prove the following.

THEOREM 4.1. *Let $z(1), z(2), \cdots$ be a sequence of real numbers and let $h_1, h_2, \cdots$ be a sequence of m-dimensional vectors. Let H_n be the $n \times m$ matrix whose jth row is h_j^t, $j = 1, 2, \cdots, n$, and let Z_n be the n-vector whose components are $z(1), z(2), \cdots, z(n)$. For each n, let*

$$E_n = \inf_X \|Z_n - H_nX\|^2 = \| Z_n - H_n\hat{X}_n \|^2,$$

where $\hat{X}_n$ is the minimum norm solution of (2.1). Then, for every n,

$$E_{n+1} = E_n + \begin{cases} \dfrac{[z(n+1) - h_{n+1}^t\hat{X}_n]^2}{1 + h_{n+1}^t B_n h_{n+1}}, & \text{if } h_{n+1} \text{ is a linear combination of } h_1, h_2, \cdots, h_n, \\ 0, & \text{otherwise,} \end{cases}$$

where $\hat{X}_n$ and B_n are defined inductively in Theorem 2.4.

Before embarking upon the proof of this theorem, we point out that $\bar{E}_n$ (the residual squared error in the constrained case) is computed in exactly the same fashion. For by Theorem 3.1,

$$\bar{E}_n = \inf_{GX=0} \| Z_n - H_nX \|^2 = \inf_X \| Z_n - \bar{H}_nX \|^2,$$

where $\bar{H}_n$ is the $n \times m$ matrix whose jth row is $\bar{h}_j^t$ (defined in Theorem 3.1). Hence, the results of Theorem 4.1 apply to the constrained minimization if H_n and its rows h_j^t are replaced throughout by $\bar{H}_n$ and $\bar{h}_j^t$, respectively.

Proof. The proof follows directly from the recursions (2.12)–(2.15). Indeed, since we can write H_n as a partitioned matrix

$$H_n = \begin{bmatrix} H_{n-1} \\ \hline h_n^t \end{bmatrix},$$

we see from (2.12), that

$$H_n\hat{X}_n = \begin{bmatrix} H_{n-1}\hat{X}_{n-1} \\ \hline h_n^t\hat{X}_{n-1} \end{bmatrix} + \begin{bmatrix} H_{n-1}K_n \\ \hline h_n^tK_n \end{bmatrix}[z(n) - h_n^t\hat{X}_{n-1}]. \tag{4.1}$$

Thus,

$$Z_n - H_n\hat{X}_n = \begin{bmatrix} Z_{n-1} - H_{n-1}\hat{X}_{n-1} \\ \hline [z(n) - h_n^t\hat{X}_{n-1}] \end{bmatrix} - \begin{bmatrix} H_{n-1}K_n \\ \hline h_n^tK_n \end{bmatrix}[z(n) - h_n^t\hat{X}_{n-1}]. \tag{4.2}$$

We now treat two cases.

Case 1. *h_n is not a linear combination of $h_1, h_2, \cdots, h_{n-1}$.* In this case,

by (2.13),

$$K_n = \frac{A_{n-1}h_n}{h_n{}^t A_{n-1}h_n}.$$

Since $A_{n-1}h_j = 0$ for $j = 1, 2, \cdots, n-1$ (see Theorem 2.5), we find that $H_{n-1}K_n = 0$. Consequently,

$$Z_n - H_n\hat{X}_n = \begin{bmatrix} Z_{n-1} - H_{n-1}\hat{X}_{n-1} \\ \hline 0 \end{bmatrix},$$

so that

$$E_n = \| Z_n - H_n\hat{X}_n \|^2 = \| Z_{n-1} - H_{n-1}\hat{X}_{n-1} \|^2 = E_{n-1},$$

which establishes Case 1.

Case 2. h_n *is a linear combination of* $h_1, h_2, \cdots, h_{n-1}$. In this case $A_{n-1}h_n = 0$ (see Theorem 2.5), so that

$$K_n = \frac{B_{n-1}h_n}{1 + h_n{}^t B_{n-1}h_n}.$$

By (2.2) of Theorem 2.1 and Lemma 1a,

$$H_{n-1}\hat{X}_{n-1} = P(H_{n-1})Z_{n-1},$$

so that by (4.2),

$$\begin{aligned} \| Z_n - H_n\hat{X}_n \|^2 &= \| Z_{n-1} - H_{n-1}\hat{X}_{n-1} \|^2 \\ &\quad - 2(Z_{n-1} - H_{n-1}\hat{X}_{n-1})^t H_{n-1}K_n[z(n) - h_n{}^t\hat{X}_{n-1}] \\ &\quad + K_n{}^t H_n{}^t H_n K_n[z(n) - h_n{}^t\hat{X}_{n-1}]^2 \\ &\quad + (1 - h_n{}^t K_n)^2[z(n) - h_n{}^t\hat{X}_{n-1}]^2 \end{aligned} \tag{4.3}$$

$$\begin{aligned} &= \| Z_{n-1} - H_{n-1}\hat{X}_{n-1} \|^2 \\ &\quad - 2Z_{n-1}^t(I - P(H_{n-1}))^t H_{n-1}K_n[z(n) - h_n{}^t\hat{X}_{n-1}] \\ &\quad + \frac{h_n{}^t B_{n-1}H_{n-1}^t H_{n-1}B_{n-1}h_n + 1}{(1 + h_n{}^t B_{n-1}h_n)^2}[z(n) - h_n{}^t\hat{X}_{n-1}]^2. \end{aligned} \tag{4.4}$$

By Lemma 1c,

$$(I - P(H_{n-1}))^t H_{n-1} = 0. \tag{4.5}$$

Furthermore, it is easy to show, using (2.22) and (2.26), that

$$B_n H_n{}^t H_n B_n = B_n. \tag{4.6}$$

Combining (4.3)–(4.6), we then see that for Case 2,

$$E_n = \| Z_n - H_n\hat{X}_n \|^2 = E_{n-1} + \frac{[z(n) - h_n{}^t\hat{X}_{n-1}]^2}{1 + h_n{}^t B_{n-1}h_n}.$$

5. Weighted and constrained least squares. In §3 we described a method whereby the computation of constrained least squares estimators could be reduced to the computation of unconstrained least squares estimators. An essential feature of the technique was the necessity to specify the constraints in advance of the data acquisition so that the vectors $\bar{h}_j$ could be computed from the vectors h_j either before or during the "real time" computation period (see Theorem 3.1).

Under certain circumstances this method has shortcomings. In some situations, one may wish to collect all the data and, after all the data is in, compute the least squares estimator subject to a succession of progressively more restrictive conditions. By examining the associated residual errors, one can then make judgments about the reasonableness of the constraints in the light of the available data. Since the method described in Theorem 3.1 requires that the constraints be incorporated into the real time computation scheme, there is no way of experimenting with various constraints unless one is willing to carry on as many parallel computations as there are sets of constraints (i.e., one for every G matrix).

Another pertinent objection to our method centers about the fact that every h-vector must be modified if one wishes to compute a running-value of the constrained least squares estimator. There is, however, another way of doing things that overcomes this difficulty and we introduce this method by considering an estimation problem in which one observes a sequence of real random variables, $z(1), z(2), \cdots, z(n)$, where each observation is of the form

$$z(j) = h_j{}^t X + v(j), \qquad j = 1, 2, \cdots, n.$$

The vectors h_j are known, while the $v(j)$ are independent and normally distributed with variances σ_j^2.

It is well known that the minimum variance (in each component) unbiased estimator of X is, in this case, the *weighted* least squares estimator of X(i.e., the value of X which minimizes

$$E_n(X) = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} [z(j) - h_j{}^t X]^2. \tag{5.1}$$

But

$$E_n(X) = \sum_{j=1}^{n} \left[\frac{z(j)}{\sigma_j} - \frac{h_j{}^t}{\sigma_j} X \right]^2 = \sum_{j=1}^{n} [z^*(j) - h_j^{*t} X]^2, \tag{5.2}$$

where

$$z^*(j) = \frac{z(j)}{\sigma_j}$$

and

$$h_j^* = \frac{h_j}{\sigma_j}.$$

Thus $\hat{X}_n$, the value of X which minimizes $E_n(X)$, satisfies the recursion (2.12)–(2.15) (with stars affixed to the z's and h's). Replacing $z^*(j)$ by $z(j)/\sigma_j$ and h_j^* by h_j/σ_j, we obtain the appropriate recursion for $\hat{X}_n$. This we state as a theorem.

THEOREM 5.1. *Let $z(1), z(2), \cdots$ be a sequence of real numbers and let $h_1, h_2, \cdots$ be a sequence of m-dimensional vectors. Define*

$$\begin{aligned} \hat{X}_n &= \hat{X}_{n-1} + K_n[z(n) - h_n{}^t\hat{X}_{n-1}], \\ \hat{X}_0 &= 0, \end{aligned} \tag{5.3a}$$

and

$$E_n = E_{n-1} + \begin{cases} 0 & \text{if } h_n \text{ is not a linear combination of } h_1, \cdots, h_{n-1}, \\ \dfrac{(z(n) - h_n{}^t\hat{X}_{n-1})^2}{\sigma_n{}^2 + h_n{}^tB_{n-1}h_n}, & \text{otherwise,} \end{cases} \tag{5.3b}$$

$$E_0 = 0,$$

where

$$K_n = \begin{cases} \dfrac{A_{n-1}h_n}{h_n{}^tA_{n-1}h_n} & \text{if } h_n \text{ is not a linear combination of } h_1, h_2, \cdots, h_{n-1} \\ \dfrac{B_{n-1}h_n}{\sigma_n{}^2 + h_n{}^tB_{n-1}h_n}, & \text{otherwise,} \end{cases} \tag{5.4}$$

$$A_n = \begin{cases} A_{n-1} - \dfrac{[A_{n-1}h_n][A_{n-1}h_n]^t}{h_n{}^tA_{n-1}h_n} & \text{if } h_n \text{ is not a linear combination of } h_1, \cdots, h_{n-1}, \\ A_{n-1}, & \text{otherwise,} \end{cases} \tag{5.5}$$

$$B_n = \begin{cases} B_{n-1} - \dfrac{[B_{n-1}\,h_n][A_{n-1}\,h_n]^t + [A_{n-1}\,h_n][B_{n-1}\,h_n]^t}{h_n{}^tA_{n-1}\,h_n} \\ \quad + \dfrac{\sigma_n{}^2 + h_n{}^tB_{n-1}\,h_n}{(h_n{}^tA_{n-1}\,h_n)^2}[A_{n-1}\,h_n][A_{n-1}\,h_n]^t & \text{if } h_n \text{ is not a linear combination of } h_1, \cdots, h_{n-1}, \\ B_{n-1} - \dfrac{[B_{n-1}\,h_n][B_{n-1}\,h_n]^t}{\sigma_n{}^2 + h_n{}^tB_{n-1}\,h_n}, & \text{otherwise,} \end{cases} \tag{5.6}$$

with the initial conditions

$$A_0 = I, \qquad B_0 = 0.$$

Then for every n, $\hat{X}_n$ *minimizes*

$$\sum_{j=1}^{n} \frac{1}{\sigma_j^2} (z(j) - h_j^t X)^2,$$

and

$$E_n = \sum_{j=1}^{n} \frac{1}{\sigma_j^2} (z(j) - h_j^t \hat{X}_n)^2$$

is the associated (minimal) residual squared error, provided the σ_j^2 *are all positive.*

The minimum variance property of weighted least squares estimators forms ample justification for weighting each term in the sum (5.1) by its "reliability" (the reciproval of its variance). If one were to push this idea to the limit, one would be tempted to assign infinite weight to "perfectly reliable" observations. To be explicit, suppose that one knew with certainty that X satisfied the constraint

$$h_j^t X = z(j)$$

for some j. This is like saying that

$$z(j) = h_j^t X + v(j),$$

where $v(j)$ has mean zero and variance zero. In the light of the preceding discussion, it seems reasonable to think that the constraint can be treated like a fictitious observation with zero variance, in so far as least squares theory is concerned, and that the presence of a set of consistent constraints can be incorporated into an unconstrained least squares model by passing to the limit suitable (i.e., by letting appropriate variances approach zero).

This idea is indeed valid. In fact, we will show that constraints can be treated as fictitious observations having zero variance in so far as the recursions (5.3)–(5.6) are concerned. The proof of this fact is quite intricate and relies heavily on certain properties of matrix pseudo-inverses. So, we digress momentarily to state, without proof, the results which will be needed. A full discussion (with proofs) of these results can be found in [1].

DEFINITION. For any rectangular matrix H, the matrix

$$H^\dagger = \lim_{\epsilon \to 0} (H^t H + \epsilon I)^{-1} H^t$$

always exists and is called the *pseudo-inverse* of H.

The pseudo-inverse of H enjoys the following properties:

P1. $HH^\dagger$ is the operator which projects onto the range of H.

P2. $H^\dagger H$ is the operator which projects onto the range of H^t.
P3. For any vector Z, the vector $\hat{X} = H^\dagger Z$ is the minimum norm solution of the equation

$$H^t H X = H^t Z,$$

and hence is the vector of minimum norm among those which minimize $\| Z - HX \|^2$.
P4. $Y = HX$ if and only if $Y = HH^\dagger Y$ and $X = H^\dagger Y + (I - H^\dagger H)V$ for some V.
P5. The ranges of the matrices $H^\dagger$, H^t and $H^t H$ are identical.
P6. $(H^t H)^\dagger = H^\dagger (HH^t)^\dagger H = H^t (HH^t)^\dagger (H^t)^\dagger$.
P7. $HH^\dagger H = H$.
P8. $(H^t H)^\dagger H^t x = 0$ if and only if $H^t x = 0$.
P9. $(H^t H)^\dagger y = 0$ if and only if $Hy = 0$.
P10. $H^\dagger = (H^t H)^\dagger H^t = H^t (HH^t)^\dagger$.
P11. If H is a projection, $H^\dagger = H$.
(The ambitious reader can amuse himself by deriving P1–P11 from the results of Lemma 1 in §2.)

We point out that $H^\dagger$ is exactly the so-called Penrose-pseudo-inverse of H. This is so because the Penrose-pseudo-inverse (cf. [6]) of a matrix H is the *unique* solution of the equations,

(a) $HXH = H$,
(b) $XHX = X$,
(c) $(HX)^t = HX$,
(d) $(XH)^t = XH$.

$X = H^\dagger$ satisfies (a) by P7. By P1 and P2, $H^\dagger H$ and $HH^\dagger$ are projections, therefore are symmetric; so (c) and (d) hold.

Finally, by P10,

$$(H^\dagger H)H^\dagger = (H^\dagger H)H^t (HH^t)^\dagger$$

and by P2,

$$H^\dagger H H^t = H^t.$$

Therefore,

$$(H^\dagger H)H^\dagger = H^\dagger,$$

establishing (b).

In the remainder of this section, our program consists of nine main steps.

First, we will show that

$$\hat{X}_n(\lambda) = B_n(\lambda) \left[\frac{1}{\lambda^2} G^t W + H_n{}^t Z_n \right]$$

minimizes

$$\sum_{j=1}^{n} [z(j) - h_j{}^t X]^2 + \frac{1}{\lambda^2} \sum_{j=1}^{k} [w(j) - g_j{}^t X]^2,$$

where Z_n is the n-vector whose components are $z(j)$, H_n is the $n \times m$ matrix whose row vectors are $h_j{}^t$, $j = 1, \cdots, n$, W is the k-vector whose components are $w(j)$, G is the $k \times m$ matrix whose rows are $g_j{}^t$, and

$$B_n(\lambda) = \left[\frac{1}{\lambda^2} G^t G + H_n{}^t H_n\right]^\dagger.$$

Second, we will point out that $B_n(\lambda)$ satisfies a recursion like (2.14)–(2.15), with the initial conditions replaced by

$$B_0(\lambda) = \lambda^2 [G^t G]^\dagger, \qquad A_0 = I - G^\dagger G.$$

Third, we will prove that $B_n(\lambda) = B_n + O(\lambda^2)$ as $\lambda \to 0$, where B_n satisfies a recursion similar to (2.14)–(2.15) with the initial conditions $B_0 = 0$, $A_0 = I - G^\dagger G$.

Fourth, we will show that

$$\bar{B}_n = (\bar{H}_n{}^t \bar{H}_n)^\dagger$$

(where $\bar{H}_n = H_n(I - G^\dagger G)$) satisfies the same recursion as B_n with the same initial conditions. Hence, $B_n = \bar{B}_n$ for every n, so that

$$B_n(\lambda) = (\bar{H}_n{}^t \bar{H}_n)^\dagger + O(\lambda^2) \qquad \text{as} \quad \lambda \to 0.$$

Fifth, we then conclude that

$$\hat{Y}_n(\lambda) = B_n(\lambda) H_n{}^t Z_n$$

converges to

$$\hat{Y}_n = (\bar{H}_n{}^t \bar{H}_n)^\dagger \bar{H}_n{}^t Z_n ;$$

and by Theorem 3.1, this means that $\hat{Y}_n(\lambda)$, which minimizes

$$\sum_{j=1}^{n} [z(j) - h_j{}^t Y]^2 + \frac{1}{\lambda^2} \sum_{j=1}^{k} [0 - g_j{}^t Y]^2,$$

converges to a vector which minimizes

$$\sum_{j=1}^{n} [z(j) - h_j{}^t Y]^2$$

subject to the constraints $GY = 0$.

Sixth, this permits us to establish the more general result, namely, that

$$\hat{X}_n(\lambda) = B_n(\lambda) \left[H_n{}^t Z_n + \frac{1}{\lambda^2} G^t W\right]$$

converges to a value which minimizes

$$\sum_{j=1}^{n} [z(j) - h_j{}^t X]^2$$

subject to the constraints $GX = W$.

Seventh, we will show that the residual error $E_n(\lambda)$, which is associated with $\hat{X}_n(\lambda)$, converges to the residual error E_n associated with $\hat{X}_n$.

Eighth, we then show that a recursion for $\hat{X}_n$ is obtainable by writing the recursion for $\hat{X}_n(\lambda)$ and setting $\lambda = 0$ throughout. The recursion for E_n is obtained from the recursion for $E_n(\lambda)$ in the same way.

Ninth, finally, we will extend the results to the general case of constrained, weighted least squares.

We begin by defining the matrix

$$F_n(\lambda) = \begin{bmatrix} \frac{1}{\lambda} G \\ \hline H_n \end{bmatrix},$$

and the vector

$$U_n(\lambda) = \begin{bmatrix} \frac{1}{\lambda} W \\ \hline Z_n \end{bmatrix}.$$

Then, the vector

$$\hat{X}_n(\lambda) = [F_n{}^t(\lambda) F_n(\lambda)]^\dagger F_n{}^t(\lambda) U_n(\lambda) \tag{5.7}$$

minimizes

$$\| U_n(\lambda) - F_n(\lambda) X \|^2. \tag{5.8}$$

(We have used P3 and P10.) But (5.8) is exactly equal to

$$\| Z_n - H_n X \|^2 + \frac{1}{\lambda^2} \| W - GX \|^2, \tag{5.9}$$

and (5.7) is exactly the same as

$$\hat{X}_n(\lambda) = \left[\frac{1}{\lambda^2} G^t G + H_n{}^t H_n\right]^\dagger \left[\frac{1}{\lambda^2} G^t W + H_n{}^t Z_n\right]. \tag{5.10}$$

If we let

$$B_n(\lambda) = \left[\frac{1}{\lambda^2} G^t G + H_n{}^t H_n\right]^\dagger, \tag{5.11}$$

where $B_0(\lambda) = \lambda^2 [G^t G]^\dagger$, we arrive at the first important result:

LEMMA 2. *The vector*

$$\hat{X}_n(\lambda) = B_n(\lambda) \left[\frac{1}{\lambda^2} G^t W + H_n{}^t Z_n\right]$$

minimizes

$$\| Z_n - H_n X \|^2 + \frac{1}{\lambda^2} \| W - GX \|^2.$$

In [1, Theorem 4.2] it is shown that $B_n(\lambda)$ satisfies a recursion which is exactly like the one described in equations (2.14)–(2.15), except for some differences in notation. For the sake of future reference, we will spell out the recursion:

LEMMA 3. *If*

$$B_n(\lambda) = \left[\frac{1}{\lambda^2} G^t G + H_n{}^t H_n\right]^\dagger,$$

and if the row vectors of G and H_n are, respectively, $g_1{}^t$, $g_2{}^t$, $\cdots$, $g_k{}^t$; $h_1{}^t$, $h_2{}^t$, $\cdots$, $h_n{}^t$, then for every n,

$$(5.12)\quad B_n(\lambda) = \begin{cases} B_{n-1}(\lambda) - \dfrac{[B_{n-1}(\lambda)h_n][A_{n-1}\,h_n]^t + [A_{n-1}\,h_n][B_{n-1}(\lambda)h_n]^t}{h_n{}^t A_{n-1}\,h_n} \\ \qquad + \left[\dfrac{1 + h_n{}^t B_{n-1}(\lambda)h_n}{(h_n{}^t A_{n-1}\,h_n)^2}\right][A_{n-1}\,h_n][A_{n-1}\,h_n]^t \\ \qquad\qquad \text{if } h_n \text{ is not a linear combination of } g_1, g_2, \cdots, g_k, h_1, \cdots, h_{n-1}, \\ B_{n-1}(\lambda) - \dfrac{[B_{n-1}(\lambda)h_n][B_{n-1}(\lambda)h_n]^t}{1 + h_n{}^t B_{n-1}(\lambda)h_n}, \quad \text{otherwise,} \end{cases}$$

where $I - A_n$ is the projection onto the subspace spanned by $g_1, \cdots, g_k, h_1, \cdots, h_n$, and where A_n is defined inductively,

$$(5.13)\quad A_n = \begin{cases} A_{n-1} - \dfrac{[A_{n-1}\,h_n][A_{n-1}\,h_n]^t}{h_n{}^t A_{n-1}\,h_n} & \text{if } h_n \text{ is not a linear combination of } g_1, \cdots, g_k, h_1, \cdots, h_{n-1}, \\ A_{n-1}, & \text{otherwise,} \end{cases}$$

with the initial conditions

$$B_0(\lambda) = \lambda(G^t G)^\dagger, \qquad A_0 = I - G^\dagger G.$$

Since A_n is a projection, $A_n = A_n{}^t = A_n{}^2$ and so

$$(5.14)\qquad \| A_n h \|^2 = (A_n h)^t (A_n h) = h^t A_n h.$$

We will use this fact later on.

It is easy to show by induction that $B_n(\lambda)$ converges to a limit and we state this as follows.

LEMMA 4. $B_n(\lambda) = B_n + O(\lambda^2)$ *as* $\lambda \to 0$, *where* B_n *satisfies the recursion* (5.12) *with the initial condition* $B_0 = 0$ *and* A_n *satisfies the recursion* (5.13) *with the initial condition* $A_0 = I - G^\dagger G$.

On the other hand, if we let $\bar{h}_j = (I - G^\dagger G)h_j$, $j = 1, 2, \cdots$, and let $\bar{H}_n$ be the $n \times m$ matrix whose rows are $\bar{h}_1{}^t, \bar{h}_2{}^t, \cdots, \bar{h}_n{}^t$ (i.e., $\bar{H}_n = \bar{H}_n(I - G^\dagger G)$) and if we let

$$\bar{B}_n = (\bar{H}_n{}^t \bar{H}_n)^\dagger, \tag{5.15}$$

then, again by [1, Theorem 4.2], we find that for every n, $\bar{B}_n$ satisfies the recursion (2.14) with B_{n-1} replaced by $\bar{B}_{n-1}$, h_n replaced by $\bar{h}_n$, and A_{n-1} replaced by $\bar{A}_{n-1}$, where $I - \bar{A}_{n-1}$ is the projection onto the subspace spanned by $\bar{h}_1, \cdots, \bar{h}_{n-1}$. $\bar{A}_n$ satisfies the recursion (2.15) with A_{n-1} replaced by $\bar{A}_{n-1}$ and h_n replaced by $\bar{h}_n$. The initial conditions are $\bar{B}_0 = 0$, $\bar{A}_0 = I$. We are now in a position to prove:

LEMMA 5. *For every* n,

$$\left[\frac{1}{\lambda^2} G^t G + H_n{}^t H_n\right]^\dagger = (\bar{H}_n{}^t \bar{H}_n)^\dagger + O(\lambda^2) \qquad \text{as } \lambda \to 0, \tag{5.16}$$

where $\bar{H}_n = H_n(I - G^\dagger G)$.

Proof. We will show that $\bar{B}_n$ (defined by (5.15)) satisfies the same recursion as B_n (defined in Lemma 4). To do this it suffices to prove that
(a) $\bar{h}_n$ is a linear combination of $\bar{h}_1, \cdots, \bar{h}_{n-1}$ if and only if h_n is a linear combination of $g_1, \cdots, g_k, h_1, \cdots, h_{n-1}$,
(b) $\bar{A}_{n-1}\bar{h}_n = A_{n-1}h_n$ and $\bar{h}_n{}^t \bar{A}_{n-1}\bar{h}_n{}^t = h_n{}^t A_{n-1} h_n$, where A_n satisfies (5.13),
(c) $\bar{B}_{n-1}\bar{h}_n = \bar{B}_{n-1}h_n$, and
(d) $\bar{h}_n{}^t \bar{B}_{n-1}\bar{h}_n = h_n{}^t \bar{B}_{n-1} h_n$.

(a) $\bar{h}_n$ is a linear combination of $\bar{h}_1, \cdots, \bar{h}_{n-1}$ if and only if

$$\bar{h}_n = \bar{H}_{n-1}^t Y_1, \tag{5.17}$$

for some Y_1. The last is true if

$$(I - G^\dagger G)h_n = (I - G^\dagger G)H_{n-1}^t Y_1. \tag{5.18}$$

Since $I - G^\dagger G$ is a projection, we have $(I - G^\dagger G)^\dagger = (I - G^\dagger G)$ by P11 and so, by P4, (5.18) holds if and only if

$$h_n = (I - G^\dagger G)H_{n-1}^t Y_1 + G^\dagger G Y_2 \tag{5.19}$$

for some Y_2. But (5.19) holds if and only if

$$h_n = H_{n-1}^t Y_1 + G^\dagger G Y_3 \tag{5.20}$$

for some Y_3. Equation (5.20) says that h_n is a linear combination of a vector in the range of H_{n-1}^t and a vector in the range of $G^\dagger G$. By *P2*, $G^\dagger G$

is the projection onto the range of G^t so that the range of $G^\dagger G$ coincides with the range of G^t. Combining the results of (5.17)–(5.20), $\bar{h}_n$ is a linear combination of $\bar{h}_1, \cdots, \bar{h}_{n-1}$ if and only if h_n is a linear combination of vectors in the ranges of H^t_{n-1} and G^t, i.e., if and only if h_n is a linear combination of the vectors $h_1, \cdots, h_{n-1}, g_1, \cdots, g_k$, which span the ranges of the matrices H^t_{n-1} and G^t.

(b) $$\bar{A}_0\bar{h}_1 = \bar{h}_1 = (I - G^\dagger G)h_1 = A_0 h_1,$$

and it is easy to show by induction that $\bar{A}_n\bar{h}_{n+1} = A_n h_{n+1}$ for every n. A_n and $\bar{A}_n$ are projections so that by (5.14),

$$(h_n{}^t A_{n-1} h_n) = \| A_{n-1}h_n \|^2 = \| \bar{A}_{n-1}\bar{h}_n \|^2 = (\bar{h}_n{}^t \bar{A}_{n-1}\bar{h}_n).$$

(c) By definition, $\bar{B}_n = (\bar{H}_n{}^t \bar{H}_n)^\dagger$ and by P6,

$$\bar{B}_n(I - G^\dagger G) = \bar{H}_n{}^\dagger (\bar{H}_n \bar{H}_n{}^t)^\dagger \bar{H}_n (I - G^\dagger G).$$

Since $\bar{H}_n = H_n(I - G^\dagger G)$ and $(I - G^\dagger G)^2 = (I - G^\dagger G)$, we see that $\bar{H}_n(I - G^\dagger G) = \bar{H}_n$, so that $\bar{B}_n(I - G^\dagger G) = \bar{B}_n$. Consequently, $\bar{B}_n\bar{h}_{n+1} = \bar{B}_n h_{n+1}$ and

(d) $$\bar{h}^t_{n+1}\bar{B}_n\bar{h}_{n+1} = \bar{h}^t_{n+1}\bar{B}_n h_{n+1} = (\bar{B}_n\bar{h}_{n+1})^t h_{n+1}$$
$$= (\bar{B}_n h_{n+1})^t h_{n+1} = h^t_{n+1}\bar{B}_n h_{n+1}.$$

This establishes the lemma.

Actually, a more general result follows from the last lemma. Although we do not use it in the sequel, it is of general interest.

THEOREM 5.2. *If A and B are nonnegative definite matrices then*

$$(5.21)\quad \left(A + \frac{1}{\lambda^2} B\right)^\dagger = [(I - B^\dagger B)A(I - B^\dagger B)]^\dagger + O(\lambda^2) \qquad \text{as } \lambda \to 0.$$

Proof. Write $A = H^t H$, where $H = H^t = A^{1/2}$, and $B = G^t G$, where $G = G^t = B^{1/2}$. By Lemma 5,

$$\left(A + \frac{1}{\lambda^2} B\right)^\dagger = [(I - G^\dagger G)(H^t H)(I - G^\dagger G)]^\dagger + O(\lambda^2).$$

Since

$$G^\dagger G = [B^{1/2}]^\dagger B^{1/2}$$

is the projection onto the range of $B^{1/2}$ and since the range of $B^{1/2}$ coincides with the range of B, we see that $G^\dagger G = B^\dagger B$. This establishes the assertion.

To return to the mainstream, let $\hat{Y}_n(\lambda) = B_n(\lambda) H_n{}^t Z_n$. Lemma 2 tells us that $\hat{Y}_n(\lambda)$ minimizes

$$\| Z_n - H_n Y \|^2 + \frac{1}{\lambda^2} \| GY \|^2,$$

while Lemma 5 tells that

$$\hat{Y}_n(\lambda) \to \hat{Y}_n = \bar{B}_n H_n{}^t Z_n \tag{5.22}$$

as $\lambda \to 0$. Since $\bar{B}_n(I - G^\dagger G) = \bar{B}_n$ (see part (c) of the proof of Lemma 5), it follows that $\bar{B}_n \bar{H}_n{}^t = \bar{B}_n H_n{}^t$, so that by P10 and (5.15),

$$\hat{Y}_n = \bar{B}_n \bar{H}_n{}^t Z_n = \bar{H}_n{}^\dagger Z_n . \tag{5.23}$$

By Theorem 3.1 and P3, $\hat{Y}_n$ minimizes $\| Z_n - H_n Y \|^2$ subject to the constraints $GY = 0$. Thus, we obtain:

LEMMA 6. *Let*

$$\hat{Y} = \lim_{\lambda \to 0} \hat{Y}(\lambda),$$

where

$$\hat{Y}(\lambda) = \left(H^t H + \frac{1}{\lambda^2} G^t G \right)^\dagger H^t Z$$

minimizes

$$\| Z - HY \|^2 + \frac{1}{\lambda^2} \| GY \|^2.$$

Then $\hat{Y}$ minimizes $\| Z - HY \|^2$ subject to the constraints $GY = 0$.

It is now relatively easy to prove the more general result for the case of inhomogeneous constraints.

LEMMA 7. *Let*

$$\hat{X} = \lim_{\lambda \to 0} \hat{X}(\lambda),$$

where

$$\hat{X}(\lambda) = \left(H^t H + \frac{1}{\lambda^2} G^t G \right)^\dagger \left(H^t Z + \frac{1}{\lambda^2} G^t W \right) \tag{5.24}$$

minimizes

$$\| Z - HX \|^2 + \frac{1}{\lambda^2} \| W - GX \|^2.$$

Then $\hat{X}$ minimizes $\| Z - HX \|^2$ subject to the constraints $GX = W$, provided the constraints are consistent (i.e., provided that there is at least one solution to the constraint equation).

Proof. It suffices to show that

$$\left(H^t H + \frac{1}{\lambda^2} G^t G \right)^\dagger \left(H^t H + \frac{1}{\lambda^2} G^t G \right) G^\dagger = G^\dagger. \tag{5.25}$$

For, if (5.25) holds, then since (by P7)

$$G^tGG^\dagger = (GG^\dagger G)^t = G^t, \tag{5.25a}$$

we have

$$\frac{1}{\lambda^2}\left(H^tH + \frac{1}{\lambda^2}G^tG\right)^\dagger G^t = \left[I - \left(H^tH + \frac{1}{\lambda^2}G^tG\right)^\dagger H^tH\right]G^\dagger.$$

Consequently,

$$\begin{aligned}\hat{X}(\lambda) &= \left(H^tH + \frac{1}{\lambda^2}G^tG\right)^\dagger \left(H^tZ + \frac{1}{\lambda^2}G^tW\right) \\ &= \left(H^tH + \frac{1}{\lambda^2}G^tG\right)^\dagger (H^tZ^*) + X_0,\end{aligned} \tag{5.26}$$

where

$$Z^* = Z - HX_0 \tag{5.27}$$

and $X_0 = G^\dagger W$. Hence, $\hat{X}(\lambda) \to \hat{Y} + X_0$, where $\hat{Y}$ minimizes $\| Z^* - HY \|^2$ subject to $GY = 0$. But in §3, we showed that $\hat{X}$ minimizes $\| Z - HX \|^2$ subject to $GX = W$ if and only if $\hat{X} = \hat{Y} + X^*$, where X^* is any vector satisfying the constraints, and $\hat{Y}$ minimizes $\| (Z - HX^*) - HY \|^2$ subject to $GY = 0$. Since

$$\hat{X} = \lim_{\lambda \to 0} \hat{X}(\lambda)$$

is precisely of this form, the lemma follows.

To prove (5.25), we point out that the range of $G^\dagger$ is the same as the range of G^tG (by P5), so that every column of $G^\dagger$ is in the range of G^tG. By P2,

$$\left(H^tH + \frac{1}{\lambda^2}G^tG\right)^\dagger \left(H^tH + \frac{1}{\lambda^2}G^tG\right) = R(\lambda)$$

is the projection onto the range of

$$H^tH + \frac{1}{\lambda^2}G^tG,$$

which contains the range of G^tG. Consequently, the columns of $G^\dagger$ are (and hence $G^\dagger$ is) left unchanged by the multiplication by $R(\lambda)$.

The corresponding result for residual errors follows easily.

LEMMA 8. *Let $\hat{X}(\lambda)$ and $\hat{X}$ be as defined in Lemma 7, let*

$$E(\lambda) = \| Z - H\hat{X}(\lambda)\|^2 + \frac{1}{\lambda^2}\| W - G\hat{X}(\lambda)\|^2, \tag{5.27}$$

and let $E = \| Z - H\hat{X} \|^2$. *If the constraint equation* $GX = W$ *has a solution, then*

$$\lim_{\lambda \to 0} E(\lambda) = E.$$

Proof. It suffices to show that

$$\frac{1}{\lambda^2} \| W - G\hat{X}(\lambda) \|^2 \to 0 \qquad \text{as } \lambda \to 0.$$

By (5.24), (5.25), (5.25a), and the fact that $W = GG^\dagger W$ (by P_4),

$$\begin{aligned} W - G\hat{X}(\lambda) &= G\left(H^tH + \frac{1}{\lambda^2} G^tG\right)^\dagger H^tHG^\dagger W \\ &\quad - G\left(H^tH + \frac{1}{\lambda^2} G^tG\right)^\dagger H^tZ \\ &= G\left(H^tH + \frac{1}{\lambda^2} G^\dagger G\right)^\dagger H^t[HG^\dagger W - Z]. \end{aligned} \tag{5.28}$$

Thus, by Lemma 5,

$$W - G\hat{X}(\lambda) = G(\bar{H}^t\bar{H})^\dagger H^t[HG^\dagger W - Z] + O(\lambda^2) \qquad \text{as } \lambda \to 0. \tag{5.29}$$

By P6,

$$(\bar{H}^t\bar{H})^\dagger = \bar{H}^tM = (I - G^\dagger G)H^tM,$$

where $M = (\bar{H}\bar{H}^t)^\dagger(\bar{H}^t)^\dagger$, so that $G(\bar{H}^t\bar{H})^\dagger = 0$ (by P7). Whence, $W - G\hat{X}(\lambda) = O(\lambda^2)$, so that $\| W - G\hat{X}(\lambda) \|^2 = o(\lambda^2)$.

The results established so far allow us to derive a recursion for a constrained least squares estimator and its associated residual error.

THEOREM 5.3. *Let* S *be a finite set of integers, let* $h_1, h_2, \cdots$ *be a sequence of* m*-dimensional vectors, let* $z(1), z(2), \cdots$ *be a sequence of real numbers and let* $\hat{X}_n$ *and* E_n *be defined inductively by* (5.3)–(5.6). *Let*

$$S_n = \{1, 2, \cdots, n\} \cap S$$

and

$$T_n = \{1, 2, \cdots, n\} - S_n.$$

If

$$\sigma_j^2 = \begin{cases} 0 & \text{for } j \in S, \\ 1 & \text{for } j \notin S, \end{cases}$$

then for every n, $\hat{X}_n$ *minimizes*

$$\sum_{j \in T_n} [z(j) - h_j{}^tX]^2$$

subject to the constraints

$$h_j{}^t X = z(j), \qquad j \in S_n ;$$

and

$$E_n = \sum_{j \in T_n} [z(j) - h_j{}^t \hat{X}_n]^2$$

is the associated residual error, provided that the set of vectors $\{h_j \mid j \in S\}$ *are linearly independent.*

Proof. Let $\hat{X}_n(\lambda)$ and $E_n(\lambda)$ be defined by the recursion (5.3)–(5.6) with

$$\sigma_j^2 = \begin{cases} \lambda^2 & \text{if } j \in S, \\ 1 & \text{if } j \notin S. \end{cases}$$

By Theorem 2.4, $\hat{X}_n(\lambda)$ is the unique vector of minimum norm among those which minimize

$$\sum_{j \in T_n} [z(j) - h_j{}^t X]^2 + \sum_{j \in S_n} \left[\frac{z(j)}{\lambda} - \frac{h_j{}^t}{\lambda} X \right]^2 ; \tag{5.30}$$

and by Theorem 4.1, $E_n(\lambda)$ is its associated residual error. Let Z_n be the vector whose components are $z(j)$, $j \in T_n$, let H_n be the matrix whose row vectors are $h_j{}^t$, $j \in T_n$, let W_n be the vector whose components are $z(j)$, $j \in S_n$, and let G_n be the matrix whose row vectors are $h_j{}^t$, $j \in S_n$. Let

$$F_n(\lambda) = \begin{bmatrix} \frac{1}{\lambda} G_n \\ H_n \end{bmatrix}, \qquad U_n(\lambda) = \begin{bmatrix} \frac{1}{\lambda} W_n \\ Z_n \end{bmatrix}. \tag{5.31}$$

Since $\hat{X}_n(\lambda)$ is the unique vector of minimum norm among those which minimize (5.30), we have by Theorem 2.1 that

$$\hat{X}_n(\lambda) = \lim_{\epsilon \to 0} (F_n{}^t(\lambda) F_n(\lambda) + \epsilon I)^{-1} F_n{}^t(\lambda) U_n(\lambda) ; \tag{5.32}$$

and by the definition of $F_n{}^\dagger(\lambda)$, we see then that

$$\hat{X}_n(\lambda) = F_n{}^\dagger(\lambda) U_n(\lambda). \tag{5.33}$$

By P10,

$$\hat{X}_n(\lambda) = [F_n{}^t(\lambda) F_n(\lambda)]^\dagger F_n{}^t(\lambda) U_n(\lambda), \tag{5.34}$$

or equivalently,

$$\hat{X}_n(\lambda) = \left[\frac{1}{\lambda^2} G_n{}^t G_n + H_n{}^t H_n \right]^\dagger \left[\frac{1}{\lambda^2} G_n{}^t W_n + H_n{}^t Z_n \right].$$

Now, consider the recursions (5.3)–(5.6) for $\hat{X}_n(\lambda)$. If we examine the

associated recursion for B_n, we see by [1, Theorem 4.2] that

$$B_n = B_n(\lambda) = \left[\frac{1}{\lambda^2} G_n{}^t G_n + H_n{}^t H_n\right]^\dagger$$

for every n and that $I - A_n$ is the projection of h_n onto the subspace spanned by $h_1, \cdots, h_{n-1}$.

By Lemma 7,

$$\hat{X}_n = \lim_{\lambda \to 0} \hat{X}_n(\lambda)$$

exists and minimizes $\| Z_n - H_n X \|^2$ subject to $G_n X = W_n$. Furthermore,

$$\hat{X}_n = \lim_{\lambda \to 0} \{X_{n-1}(\lambda) + K_n(\lambda)[z(n) - h_n{}^t X_{n-1}(\lambda)]\},$$

and $K_n(\lambda)$ is defined by (5.4) with

$$\sigma_n{}^2 = \begin{cases} 1 & \text{if } n \in T_n, \\ \lambda^2 & \text{if } n \in S_n. \end{cases}$$

It therefore suffices to show that

$$\lim_{\lambda \to 0} h_n{}^t B_{n-1}(\lambda) h_n > 0 \tag{5.35}$$

if $n \in S_n$ and h_n is a linear combination of $h_1, \cdots, h_{n-1}$.

For, if (5.35) holds, then

$$\hat{X}_n = \lim_{\lambda \to 0} \hat{X}_n(\lambda)$$

satisfies the same recursion as $\hat{X}_n(\lambda)$ with $\lambda = 0$, and E_n satisfies the same recursion as $E_n(\lambda)$ with $\lambda = 0$. To prove (5.35), we proceed as follows: By Lemma 5,

$$\lim_{\lambda \to 0} h_n{}^t B_{n-1}(\lambda) h_n = \bar{h}_n{}^t (\bar{H}_{n-1}^t \bar{H}_{n-1})^\dagger \bar{h}_n, \tag{5.36}$$

where

$$\bar{H}_{n-1} = H_{n-1}(I - G_{n-1}^\dagger G_{n-1}), \tag{5.37}$$

and $\bar{h}_n = (I - G_{n-1}^\dagger G_{n-1}) h_n$. Furthermore, by the proof of Lemma 5, h_n is a linear combination of $h_1, \cdots, h_{n-1}$ if and only if there is a vector y such that

$$\bar{h}_n = \bar{H}_{n-1}^t y. \tag{5.38}$$

But by P8, $(\bar{H}_{n-1}^t \bar{H}_{n-1})^\dagger \bar{H}_{n-1} y = 0$ only if $\bar{H}_{n-1} y = 0$, so that $(\bar{H}_{n-1}^t \bar{H}_{n-1})^\dagger \bar{h}_n = 0$ only if $\bar{h}_n = 0$. But $\bar{h}_n = 0$ if and only if $h_n = G_{n-1}^\dagger G_{n-1} h_n$, i.e., if and only if h_n is in the range of G_{n-1}^t (which is spanned by the set $\{h_j \mid j \in S_{n-1}\}$).

This cannot happen if $n \in S_n$, since we have assumed that $\{h_j \mid j \in S\}$ is a linearly independent set and hence has no linearly dependent subset. Thus, $(\bar{H}_{n-1}^t \bar{H}_{n-1})^\dagger \bar{h}_n \neq 0$, which implies

$$\bar{h}_n (\bar{H}_{n-1}^t \bar{H}_{n-1})^\dagger \bar{h}_n > 0. \tag{5.39}$$

Combining (5.36) and (5.39) we obtain (5.35).

The general case of weighted least squares subject to constraints follows in the obvious way.

THEOREM 5.4. *Let S be a finite set of integers, and let $\hat{X}_n$ and E_n be as defined in Theorem* 5.3 *except that*

$$\sigma_j^2 = \begin{cases} 0 & \text{if } j \in S, \\ \tau_j^2 > 0 & \text{if } j \notin S. \end{cases}$$

Then for every n, $\hat{X}_n$ minimizes

$$\sum_{j \in T_n} \frac{1}{\tau_j^2} [z(j) - h_j^t X]^2$$

subject to the constraints

$$h_j^t X = z(j), \qquad j \in S_n,$$

and

$$E_n = \sum_{j \in T_n} \frac{1}{\tau_j^2} [z(j) - h_j^t \hat{X}_n]^2$$

is the associated residual error, provided that the vectors in the set $\{h_j \mid j \in S\}$ are linearly independent.

Proof. By letting $h_j^* = h_j/\tau_j$ and $z^*(j) = z(j)/\tau_j$ for $j \notin S$, we reduce the problem of minimizing

$$\sum_{j \in T_n} \frac{1}{\tau_j^2} [z(j) - h_j^t X]^2$$

subject to the constraints, to that of minimizing

$$\sum_{j \in T_n} [z^*(j) - h_j^{*t} X]^2$$

subject to the same constraints. The solution to this problem along with its residual error is given (in terms of the $z^*(j)$'s and h_j^*'s) by Theorem 5.3. The present result is obtained when the $z^*(j)$'s are replaced by $z(j)/\tau_j$ and the h_j^*'s are replaced by h_j/τ_j.

This result completely justifies the intuitive feeling that we described at the beginnning of §5. Constraints can indeed be treated as though they were observations having zero variance in so far as the least squares recursion is concerned.

The time and order of introduction of the linear constraints is immaterial provided only that they are noncontradictory (the linear independence assumption guarantees this). Thus, one can collect all the data and introduce the constraints afterwards, one at a time. On the other hand, if one wishes a running ("real time") estimate subject to the constraints $g_j{}^t X = w(j), j = 1, 2, \cdots, k$, one would let S be the set $\{1, 2, \cdots, k\}$. In advance of the actual data acquisition, one would compute $\hat{\hat{X}}_1, \cdots, \hat{\hat{X}}_k$ according to (5.3)–(5.6) with $\sigma_j^2 = 0$, $h_j = g_j$, and $z(j) = w(j)$, $j = 1, 2, \cdots, k$. The terminal values, $\hat{X}_k$, $\hat{\hat{B}}_k$ and $\hat{\hat{A}}_k$, would be stored; and then, when the actual data were taken, the iteration (5.3)–(5.6) would again be employed (this time with positive σ's).

Compare this method with the method of Theorem 3.1 in a problem of ordinary (unweighted) least squares, subject to the constraints $g_j{}^t X = w(j)$, $j = 1, \cdots, k$. Under the method of Theorem 3.1, one must first find a solution, $\bar{X}$, to the constraint equations, so that the constraints can be reduced to homogeneous ones (see the beginning of §3). Then, *every* regression vector h_j has to be modified (either before or during the data acquisition) and the iteration is carried out with every h_j replaced by $\bar{h}_j$ and every $z(j)$ replaced by $\bar{z}(j) = z(j) - h_j{}^t \bar{X}$.

The method of Theorem 5.4, on the other hand, would have no effect on the "real time" computation (2.12)–(2.15) except through the initial conditions on $\hat{X}_0$, B_0, and A_0. Whereas the unconstrained least squares estimator starts with $\hat{X}_0 = 0$, $B_0 = 0$, and $A_0 = I$, the constrained estimator "starts" with $\hat{X}_0 = \hat{\hat{X}}_k$, $B_0 = \hat{\hat{B}}_k$, and $A_0 = \hat{\hat{A}}_k$. All subsequent steps of the iteration are carried out in strict accordance with (2.12)–(2.15) except that the iterations are conditioned on whether or not h_n is a linear combination of $g_1, g_2, \cdots, g_k, h_1, \cdots, h_{n-1}$.

REFERENCES

[1] A. Albert, *An introduction and beginner's guide to matrix pseudo inverses*, Tech. Report, ARCON, Lexington, Massachusetts, 1964.

[2] A. Ben-Israel and A. Charnes, *Contributions to the theory of generalized inverses*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 667–700.

[3] G. G. den Broeder, Jr., and A. Charnes, *Contributions to the theory of generalized inverses for matrices*, Purdue University, Lafayette, Indiana, 1957; republished as ONR Res. Memo No. 39 Technological Institute, Northwestern University, Evanston, Illinois, 1962.

[4] C. A. Desoer and B. H. Whalen, *A note on pseudo inverses*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 442–448.

[5] R. E. Kalman, *New methods and results in linear prediction and filtering theory*, Tech. Report 61-1, RIAS, Baltimore, 1961.

[6] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

WEAKENED HYPOTHESES FOR THE VARIATIONAL PROBLEM CONSIDERED BY HESTENES*

T. GUINN†

1. Introduction. Recently Hestenes [1] has obtained necessary conditions for a generalization of the classic problem of Bolza. Since the extensions are of such a nature as to require reformulation of the method of proof, this problem will be referred to here as the problem of Hestenes.

The proof given by Hestenes for this problem is in both variational and control formulations and the equivalence of these is shown. Since in the problem of Hestenes, inequality constraints which jointly involve both control and state variables are given explicitly, necessary conditions are derived for piecewise continuous control variables. Here it is shown that if these constraints are not stated explicitly but instead are given by describing the properties of a certain region, necessary conditions can be obtained under somewhat weaker hypotheses. As a consequence, the resulting necessary conditions hold only almost everywhere and the development is considerably more complicated. The results include as a special case those obtained by the Pontryagin school [2] for a less general problem under the assumption that the controls are bounded and measurable.

Since the proof here closely parallels that of Hestenes, we avoid repetition by assuming that the reader has in hand a copy of [1]. All references will be to that paper unless otherwise stated.

2. Formulation of the problem. The problem considered is that of minimizing a function

$$I_0(x) = g_0(b) + \int_{t^1}^{t^2} L_0(t, x(t), u(t), b)\, dt,$$

in a class of arcs

$$x: x^i(t), u^k(t), b^\sigma, \quad t^1 \leqq t \leqq t^2; i = 1, \cdots, n; k = 1, \cdots, m; \sigma = 1, \cdots, r;$$

satisfying differential equations

$$\dot{x}^i(t) = f^i(t, x(t), u(t), b), \tag{2.1}$$

* Received by the editors December 1, 1964, and in revised form June 15, 1965.

† Solid Fluid Physics Department, Missile and Space Systems, Douglas Aircraft Company, Incorporated, Santa Monica, California. Now at Department of Mathematics, Michigan State University, East Lansing, Michigan. This work is part of a dissertation submitted in partial satisfaction of the requirements for the Ph.D. degree in mathematics at the University of California, Los Angeles. This research was supported in part by the United States Army Research Office (Durham) and in part by Douglas Aircraft Company, Incorporated.

a set of initial and terminal conditions

$$t^s = T^s(b), \qquad x^i(t^s) = X^{is}(b), \qquad s = 1, 2, \tag{2.3}$$

and a set of isoperimetric relations

$$I_\gamma(x) \leqq 0, \quad 1 \leqq \gamma \leqq p'; \qquad I_\gamma(x) = 0, \quad p' < \gamma \leqq p, \tag{2.4}$$

where

$$I_\gamma(x) = g_\gamma(b) + \int_{t^1}^{t^2} L_\gamma(t, x(t), u(t), b)\, dt.$$

Note this corresponds to the problem given in [1, §2] except for the inequalities (2.2) which we have here deleted.

We assume that $T^s(b)$, $X^{is}(b)$, $g(b)$, $g_\gamma(b)$ have continuous partial derivatives with respect to b^σ.

Let R be a region in (t, x, u, b)-space which is convex in x and b for each t. Set $h^0 = L_0$, $h^i = f^i$, $h^{n+\gamma} = L_\gamma$, and assume the following hold for $j = 0, 1, \cdots, n + p$:

(2.5)

(a) $h^j(t, x, u, b)$ is defined on R,

(b) h^j is continuous in x, u, and b for fixed t and locally integrable in t for fixed x, u, and b,

(c) the partial derivatives of h^j with respect to x^i, u^k, b^σ exist and satisfy (b),

(d) for each function $u(t)$ for which $h^j(t, x, u, b)$ is locally integrable for fixed x and b there is a locally integrable function $S(t)$ such that

$$| h^j_{x^i}(t, x, u(t), b) | \leqq S(t)$$

and

$$| h^j_{b^\sigma}(t, x, u(t), b) | \leqq S(t)$$

hold for $(t, x, u(t), b)$ on R.

That under these assumptions the system (2.1) has solutions with properties used subsequently is shown in [3].

Now let $u(t)$ be a function satisfying (d). A point $\bar{t}$ will be called an *ordinary point* for $u(t)$, or interchangeably for $h^j(t, \bar{x}, u(t), \bar{b})$, if it is a point of definition of $u(t)$ and

$$\lim_{\delta \to 0} \frac{1}{\delta} \int_{\bar{t}}^{\bar{t}+\delta} h^j(t, \bar{x}, u(t), \bar{b})\, dt = h^j(\bar{t}, \bar{x}, u(\bar{t}), \bar{b}), \qquad j = 0, 1, \cdots, n + p.$$

We now distinguish an arc in R as

$$x_0 : x_0{}^i(t), u_0{}^i(t), b_0{}^\sigma, \qquad t^1 \leqq t \leqq t^2,$$

where $u_0(t)$ is a function defined on $t^1 \leqq t \leqq t^2$ for which

$$h^j(t, x, u_0(t), b_0), \qquad j = 0, 1, \cdots, n + p,$$

satisfy (2.5) as functions of t and x, and $x_0{}^i(t)$ is a solution of

$$\dot{x}^i = f^i(t, x, u_0(t), b_0), \qquad t^s = T^s(b_0), \qquad s = 1, 2.$$

Let R_0 be a subset of R containing x_0. Let N be a neighborhood of x_0 in R and let M_0 be the intersection of the projections of R_0 and N on (t, x, b)-space. We assume here that R_0 has the property that there exists a set of functions $U_0(t, x, b)$ defined for every point of M_0 such that

(2.6)
(i) $U_0{}^k(t, x_0(t), b_0) = u_0(t)$,
(ii) $(t, x, U_0(t, x, b), b)$ is in R_0 for almost all t,
(iii) the functions

$$r_j{}^k(t) = \frac{\partial U_0{}^k}{\partial x^j}(t, x_0(t), b_0),$$

$$A_j{}^i(t) = f^i_{xj} + f^i_{u^k} r_j{}^k,$$

$$B_{\gamma j}(t) = L_{\gamma x^j} + L_{\gamma u^k} r_j{}^k,$$

where $x^{n+\sigma} = b^\sigma$, exist and together with h^j, $j = 1, \cdots, n + p$, satisfy (2.5b).

Furthermore, we require that for every point $(\bar{t}, \bar{x}, \bar{u}, \bar{b})$ in R_0, except possibly for $\bar{t}$ in a set of linear measure zero, there are a neighborhood M of $(\bar{t}, \bar{x}, \bar{b})$ in (t, x, b)-space and a set of functions $U(t, x, b)$ satisfying (ii) and (iii) for which $U^k(\bar{t}, \bar{x}, \bar{b}) = \bar{u}^k$ and $\bar{t}$ is an ordinary point for U^k.

An element (t, x, u, b) will be called *admissible* if it is in R_0. A function $u(t)$ will be called *admissible* if it is defined for $t^1 \leqq t \leqq t^2$, $h^j[t, x, u(t), b]$ is integrable for fixed x and b, and solutions $x(t)$ of

$$\dot{x}^i = f^i[t, x, u(t), b], \qquad t^s = T^s(b), \qquad t^1 \leqq t \leqq t^2,$$

are such that $[t, x(t), u(t), b]$ are in R_0.

If $x(t)$ is a solution of (2.1) for admissible $u = u(t)$ such that $x(\bar{t}) = \bar{x}$, where $\bar{t}$ is an ordinary point for $f^j[t, \bar{x}, u(t), b]$, then $\bar{t}$ is also an ordinary point for $f^j[t, x(t), u(t), b]$, as one readily verifies. Also for any admissible $u(t)$, the set of ordinary points has full measure.

An arc

$$x: x^i(t), u^k(t), b^\sigma, \qquad t^1 \leqq t \leqq t^2,$$

will be called *admissible* if its elements $[t, x(t), u(t), b]$ are in R_0. We denote by B the class of admissible arcs satisfying (2.1), (2.3) and (2.4).

3. The basic theorem. The reader is referred to [1, §5] which will be followed now in detail. No changes are required through to the statement of Theorem 5.1 except that differential equations in $q_{\gamma i}$ and P_{ij} are satisfied only almost everywhere. It is clear the functions F_ρ are integrable along x_0.

Then under the hypothesis of our §2, we have:

THEOREM 3.1. *The conclusions of Theorem* 5.1 *hold except that* (ii) *holds at all ordinary points* t *of* $u_0(t)$ *on* $t^1 \leqq t \leqq t^2$, *and* (iii) *holds if* t^1 *and* t^2 *are ordinary points.*

We now turn to [1, §7] for the proof. Here K is the class of vectors as in [1, (7.1)] except that now t must be an ordinary point for $u_0(t)$ rather than a point of continuity. The remaining proof except for Lemma 7.1 is unchanged, keeping in mind the new definition of K. Hence if Lemma 7.1 still holds, our Theorem 3.1 has been proved.

Following the proof of Lemma 7.1, the functions $U_j(t, x, b)$ exist by hypothesis. The construction through [1, (7.7)] is unchanged. That the functions $f_\rho(\epsilon)$ are continuous is clear. What must now be shown to complete the proof is that the functions $f_\rho(\epsilon)$ have the required partial derivatives. To show this consider

$$
\begin{aligned}
&|\, f_\rho(\epsilon) - f_\rho(0) - k_j{}^\rho \epsilon_j - \bar{k}_\sigma{}^\rho b^\sigma(\epsilon) \,|/|\, \epsilon \,| \\
&\qquad \leqq |\, G_\rho(b(\epsilon)) - G_\rho(0) - \bar{k}_\sigma{}^\rho b^\sigma(\epsilon) \,|/|\, \epsilon \,| \\
&\qquad\quad + \sum_{j=1}^{N} \left| \int_{T_j}^{T_j+\epsilon_j} (F_\rho(t, x(t, \epsilon), U_j(t, x, \epsilon), b(\epsilon)) \right. \\
&\qquad\quad \left. - F_\rho(t, x_0(t), u_0(t), b(\epsilon))) \, dt - k_j{}^\sigma \epsilon_j \right| \Big/ |\, \epsilon \,| \qquad (3.1) \\
&\qquad\quad + \sum_{j=1}^{N} \left| \int_{T_j+\epsilon_j}^{T_{j+1}} (F_\rho(t, x(t, \epsilon), u_0(t), b(\epsilon)) \right. \\
&\qquad\quad \left. - F_\rho(t, x_0(t), u_0(t), b(\epsilon))) \, dt \right| \Big/ |\, \epsilon \,|,
\end{aligned}
$$

where $k_j{}^\rho \epsilon_j$ in the second term on the right hand side is not summed on j and $T_{N+1} = t^2$. On the interval $(T_j, T_j + \epsilon_j)$, $x(t, \epsilon)$ is the solution of

$$\dot{x}^i = f^i(t, x, U_j(t, x, \epsilon), b(\epsilon)),$$

and, on the interval $(T_j + \epsilon_j, T_{j+1})$, is the solution of

$$\dot{x}^i = f^i(t, x, u_0(t), b(\epsilon)),$$

where $x(t^1, \epsilon) = X^1(b(\epsilon))$ and the initial conditions for each interval are chosen to be the final value of the solution on the previous interval.

A term from the first sum is dominated by

$$
\begin{aligned}
&\int_{T_j}^{T_j+\epsilon_j} \Big| F_\rho(t, x(t, \epsilon), U_j(t, x, \epsilon), b(\epsilon)) \\
&\qquad - F_\rho(t, x_0(t), U_j(t, x, \epsilon), b(\epsilon)) \Big| \, dt \Big/ |\epsilon| \\
&\qquad + \int_{T_j}^{T_j+\epsilon_j} \Big| F_\rho(t, x_0(t), U_j(t, x, \epsilon), b(\epsilon)) \\
&\qquad - F_\rho(t_j, x_0(t_j), u_j, b(\epsilon)) \Big| \, dt \Big/ |\epsilon| \\
&\qquad + \int_{T_j}^{T_j+\epsilon_j} \Big| F_\rho(t, x_0(t), u_0(t), b(\epsilon)) \\
&\qquad - F_\rho(t_j, x_0(t_j), u_0(t_j), b(\epsilon)) \Big| \, dt \Big/ |\epsilon| \equiv I_1 + I_2 + I_3 .
\end{aligned} \tag{3.2}
$$

Next a term from the second sum is dominated by

$$
\begin{aligned}
&\int_{t_j}^{t^2} \Big| \int_0^1 \frac{\partial}{\partial x^i} F_\rho(t, x_0(t) + \theta(x(t, \epsilon) - x_0(t)), u_0(t), b(\epsilon)) \, d\theta \\
&\qquad \cdot (x^i(t, \epsilon) - x_0^{\,i}(t)) \Big| \, dt \Big/ |\epsilon| \equiv I_4 ,
\end{aligned} \tag{3.3}
$$

where we use that $x(t, 0) = x_0(t)$. Now by [3, Theorem 4.1], $x(t, \epsilon)$ has a bounded difference quotient with respect to ϵ at $\epsilon = 0$. Hence using (7.8), where $u_0(t) = U_0(t, x_0(t), u_0(t))$, we have immediately that $I_4 = 0$ at $\epsilon = 0$. Also I_1 can be represented as an integral of the same form as I_4 with upper limit $T_j + \epsilon_j$ and $u_0(t)$ replaced by $U_j(t, x, \epsilon)$. Estimating

$$
\frac{\partial F}{\partial x^i}(t, x_0(t), U_j(t, x, \epsilon), b(\epsilon))
$$

by the function $S(t)$ given by (2.5d) and again using [3, Theorem 4.1] gives $I_1 = 0$ at $\epsilon = 0$. Now also $I_2 = I_3 = 0$ since t_j is an ordinary point for both $u_0(y)$ and $U_j(t, x, \epsilon)$. That the first term of the right hand side is zero follows from the fact that $G_\rho(b)$ has continuous partial derivatives.

Hence $f_\rho(\epsilon)$ has a differential at $\epsilon = 0$ and the proof of the lemma follows from the definition of a derived set given in [1, §6].

If the hypothesis of local integrability with respect to t is replaced by joint continuity in (t, x, u, b), the set R_0 is given by [1, (2.2)], and the admissible functions $u(t)$ are taken to be piecewise continuous, the above reduces to the problem of Hestenes.

4. A simplification. If we consider the case which corresponds to constraints [1, (2.2)] being functions of t and u alone the situation is considerably simplified. In (2.5), partial derivatives with respect to u^k are no longer required. Also R_0 need only be a subregion of R with the property that if $(\bar{t}, \bar{x}, \bar{u}, \bar{b})$ is in R_0, except possibly for $\bar{t}$ in a set of measure zero, there is a function

$$u^k(t), \qquad \bar{t} - \delta \leqq t \leqq \bar{t} + \delta,$$

for some $\delta > 0$ such that for $|t - \bar{t}| < \delta$ the following hold for $j = 0, 1, \cdots, n + p$:

(4.1)
- (i) $(t, x, u(t), b)$ is in R_0 for $|x - x_0| < \delta$, $|b - \bar{b}| < \delta$,
- (ii) for fixed x and b, the function $h^j(t, x, u(t), b)$ as well as its partial derivatives with respect to x^i and b^σ are integrable in t,
- (iii) $u(\bar{t}) = \bar{u}$,
- (iv) $\bar{t}$ is an ordinary point for $h^j(t, \bar{x}, u(t))$.

Here we do not need to distinguish the arc x_0 since admissible functions are independent of x. Hence the functions $r_j^k(t)$ given by (2.6) do not appear and the resulting development is accordingly simplified as are also the estimates given in (3.2) and (3.3) above.

In the case corresponding to constraints [1, (2.2)] being functions of u alone, the hypotheses (4.1) are trivially satisfied by the function $u(t) \equiv \bar{u}$. This is the case considered by the Pontryagin school in [2] where continuity in t was assumed for the functions L_0, f^i, constraints of the form (2.4) did not appear, and admissible controls were bounded and measurable. They also assumed the range space for admissible controls to be Hausdorff. Here no restrictions are made directly on the controls or their range space. All restrictions on controls relate to their effect on the functions in which they appear.

REFERENCES

[1] M. R. Hestenes, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[2] V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, *The theory of optimal processes. I The maximum principle*, Izv. Akad. Nauk SSSR Ser. Mat. 24 (1960), pp. 3–24.

[3] T. Guinn and M. R. Hestenes, *An embedding theorem for ordinary differential equations*, to appear.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

VARIATIONAL PROBLEMS WITH UNBOUNDED CONTROLS*

J. WARGA†

1. Introduction. Let

$$\frac{dx}{dt} = \dot{x}(t) = g(x(t), r(t)) \qquad \text{a. e. in } T = [t_0, t_1], \tag{1.1.1}$$

$$(x(t_0), x(t_1)) \in B, \tag{1.1.2}$$

$$x(t) \in A, \qquad t \in T, \tag{1.1.3}$$

be a given system of ordinary differential equations and restrictive conditions, where $x = (x^1, \cdots, x^n)$, $t_0 < t_1$, $r(t)$ is a function from T to some arbitrary set R, $g(x, r)$ is defined for all $(x, r) \in V \times R$, V is an open set in Euclidean n-space E_n, A is a closed set in V, and B is a closed set in $V \times V$. We shall say that a sequence $S = \{(r_j(t), x_j(t))\}_{j=1}^{\infty}$ is an *approximate solution* of system (1.1) if the $x_j(t)$ are absolutely continuous, the $r_j(t)$ and the $x_j(t)$ satisfy (1.1.1) for each j, the points $x_j(t_0)$ and $x_j(t_1)$ converge to points $x_S(t_0)$ and $x_S(t_1)$, respectively, as $j \to \infty$, $(x_S(t_0), x_S(t_1)) \in B$, and for every positive ϵ the $x_j(t)$ are in the ϵ-neighborhood of A for all t and all sufficiently large j.

We consider the problem of determining an approximate solution of (1.1) that yields $x_S{}^1(t_1) = \inf x_{S'}{}^1(t_1)$ among all approximate solutions S'. This problem has been considered by Young [13], [14] and McShane [3], [4], [5] in the case where R and A are Euclidean spaces, and by Warga [9], [10], [11], [12] and Gamkrelidze [2] in the case where R is a compact Hausdorff space (and, in particular, a compact Euclidean set). Special systems were considered by Neustadt [6], [7].

In the present note we are concerned with problems defined by (1.1) in the case where R is not necessarily a Euclidean or a compact Hausdorff space and where the admissible curves may have unbounded derivatives. In such problems it appears appropriate to introduce "solutions" $x(t)$ that may be discontinuous functions of t. We shall define a class of problems which admit a "minimizing parametric solution" (Assumptions 2.1 and 3.1 and inequalities (2.2)) and we shall indicate (Theorem 3.2 and §3.3) how this minimizing parametric solution can be used to determine a minimizing approximate solution of (1.1). We shall also apply previous results [9], [10] to derive necessary conditions satisfied by a minimizing

* Received by the editors April 5, 1965, and in revised form June 29, 1965.

† Research and Advanced Development Division, AVCO Corporation, 201 Lowell Street, Wilmington, Massachusetts.

parametric solution in the case $A = E_n$. The more general necessary conditions for the case where

$$A = \{x \mid a^j(x) \leqq 0, j = 1, \cdots, m\}$$

can be derived in a straightforward manner by applying the results of [11] and [12].

Our approach is basically parametric. Its domain of applicability is limited to problems in which the "competing" admissible curves $x(t)$ are of uniformly bounded lengths.

After these investigations were completed, the author received a report of Rishel [8] on a related problem. Rishel derives necessary conditions for minimum for cases where (in our notation)

$$r = (u, v), \qquad g(x, r) = g_1(x, u) + vg_2(u),$$

the vector control $u(t)$ is bounded and restricted to a set U, the control $v(t)$ is a scalar nonnegative measure (which simulates the "limit" of a scalar unbounded control), and a minimizing solution is a priori assumed to exist and to satisfy a certain condition.

2. The parametrization. Henceforth, whenever we refer to a function that is differentiable a.e. in T we shall mean an *absolutely continuous* function. This convention will apply, in particular, to solutions of differential equations.

We first state certain preliminary assumptions.

Assumption 2.1.

(2.1.1) There exists an approximate solution $S = \{(\bar{r}_j(t), \bar{x}_j(t))\}_{j=1}^{\infty}$ of system (1.1).

(2.1.2) There exist positive constants c and c' and a compact set D in V such that

$$\int_{t_0}^{t_1} |\, g(x(t), r(t))\,|\, dt \leqq c \quad \text{and} \quad x(t) \in D, \qquad t \in T,$$

provided $r(t)$ and $x(t)$ satisfy (1.1.1), $x^1(t_1) \leqq \bar{x}_s^1(t_1) + c'$, $x(t)$ is in the c'-neighborhood of A, and $(x(t_0), x(t_1))$ is in the c'-neighborhood of B. Here $|\, g\,|$ denotes the Euclidean length of g.

(2.1.3) There exists a constant l such that

$$|\, g(x, r) - g(\bar{x}, r)\,| \leqq l|\, x - \bar{x}\,|$$

for all $(r, x, \bar{x}) \in R \times V \times V$.

We can easily verify that $c + t_1 - t_0$ is an upper bound of the lengths of the "competing" admissible curves $(t, x(t))$ in the (t, x)-space.

Let now $v(x, r): V \times R \to E_1$ be a positive and continuous function of

x for every $r \in R$, and let positive constants c_1 and c_2 be such that

$$c_1 \leqq \frac{v(x, r)}{1 + |g(x, r)|} \leqq c_2 \qquad \text{on } V \times R. \tag{2.2}$$

We consider the following system of differential equations and restrictive conditions:

$$\begin{aligned} \dot{x}(t) &= g(x(t), r(t)) && \text{a. e. in } T, \\ \dot{\theta}(t) &= \frac{1}{a} v(x(t), r(t)) && \text{a. e. in } T, \\ \dot{a}(t) &= 0 && \text{a. e. in } T, \end{aligned} \tag{2.3.1}$$

$$\theta(t_0) = 0, \quad \theta(t_1) = 1, \quad (x(t_0), x(t_1)) \in B, \quad x(t) \in A, \quad t \in T. \tag{2.3.2}$$

This system is formally obtained by adjoining two differential equations and two boundary conditions to (1.1). We cannot, however, assert as yet that this system is equivalent to (1.1) since we have not proved that the function $v(x(t), r(t))$ is measurable for every $r(t)$ and $x(t)$ that satisfy (1.1.1). We cannot even assert at present that this system has any solutions. We observe, however, that if there exists a solution $x(t)$ satisfying the conditions of Assumption (2.1.2), then

$$a(t) = a(t_0) = a = \int_{t_0}^{t_1} v(x(\tau), r(\tau))\, d\tau;$$

hence

$$0 < c_3 = c_1(t_1 - t_0) \leqq a \leqq c_2(t_1 - t_0) + cc_2 = c_4\,. \tag{2.4}$$

Furthermore, $\theta(t)$ is strictly increasing, hence t is an increasing function of θ on [0, 1], say $t = \tau(\theta)$. Let $x(\tau(\theta)) = \xi(\theta)$, and $r(\tau(\theta)) = \rho(\theta)$. Thus system (2.3) is equivalent to the system

$$\begin{aligned} \frac{d\xi}{d\theta} = \xi'(\theta) &= \frac{a g(\xi(\theta), \rho(\theta))}{v(\xi(\theta), \rho(\theta))} && \text{a. e. in } [0, 1], \\ \frac{d\tau}{d\theta} = \tau'(\theta) &= \frac{a}{v(\xi(\theta), \rho(\theta))} && \text{a. e. in } [0, 1], \\ \frac{da}{d\theta} = a'(\theta) &= 0 && \text{a. e. in } [0, 1], \end{aligned} \tag{2.5.1}$$

$$(\xi(0), \xi(1)) \in B, \quad \tau(0) = t_0\,, \quad \tau(1) = t_1\,, \quad \xi(\theta) \in A, \quad \theta \in [0,1]. \tag{2.5.2}$$

This last system has uniformly bounded right hand sides if we adjoin to it the inequalities (2.4).

3. The relaxed parametric solutions. We have introduced the system (2.5) primarily for purposes of exposition and to motivate the procedure that follows.

Let system (1.1) be given and let Assumption 2.1 be satisfied. Let $v(x, r)$ be a continuous and positive function on V for every $r \in R$, satisfying inequalities (2.2). We require:

Assumption 3.1. There exists a constant c_5 such that

$$\left| \frac{g(x, r)}{v(x, r)} - \frac{g(\bar{x}, r)}{v(\bar{x}, r)} \right| \leqq c_5 \, | x - \bar{x} | ,$$

$$\left| \frac{1}{v(x, r)} - \frac{1}{v(\bar{x}, r)} \right| \leqq c_5 \, | x - \bar{x} | ,$$

$$| v(x, r) - v(\bar{x}, r) | \leqq c_5 \, | x - \bar{x} | ,$$

for all $(r, x, \bar{x}) \in R \times V \times V$.

Let now

$$G(x) = \left\{ \left(\frac{g(x, r)}{v(x, r)}, \frac{1}{v(x, r)} \right) | \, r \in R \right\} \subset E_{n+1} , \qquad x \in V,$$

and let $F(x)$ be the convex closure of $G(x)$. We shall call the function $(\xi(\theta), \tau(\theta))$: $[0, 1] \to E_{n+1}$ a *relaxed parametric solution* of system (1.1) if there exists a number a such that $c_3 \leqq a \leqq c_4$,

$$\frac{1}{a} (\xi'(\theta), \tau'(\theta)) \in F(\xi(\theta)) \qquad \text{a. e. in } [0, 1],$$

$\xi(\theta) \in A$ in $[0, 1]$, $\tau(0) = t_0$, $\tau(1) = t_1$, and $(\xi(0), \xi(1)) \in B$. A *minimizing parametric solution* of system (1.1) is a relaxed parametric solution that minimizes $\xi^1(1)$.

THEOREM 3.2. *Let Assumptions* 2.1 *and* 3.1 *be satisfied and let* $v(x, r)$ *be continuous on* V *for every* $r \in R$ *and satisfy* (2.2), *where* c_1 *and* c_2 *are positive constants. Then there exists a minimizing parametric solution* $(\xi(\theta), \tau(\theta))$ *of* (1.1). *Furthermore, there exist an approximate solution* $S = \{r_j(t), x_j(t)\}_{j=1}^{\infty}$ *of* (1.1) *and a sequence* $\{\tau_j(\theta)\}_{j=1}^{\infty}$ *of increasing functions on* $[0, 1]$ *such that* $\tau_j(0) = t_0$, $\tau_j(1) = t_1$, $j = 1, 2, \cdots$,

$$| \xi(\theta) - x_j(\tau_j(\theta)) | + | \tau(\theta) - \tau_j(\theta) | \underset{j \to \infty}{\to} 0 \qquad \textit{uniformly on } [0, 1],$$

and $\xi^1(1) = x_S^1(t_1) \leqq x_{S'}^1(t_1)$ *for every approximate solution* S' *of* (1.1).

Proof. We first observe that if $\bar{r}(t)$: $T \to R$ and $\bar{x}(t)$: $T \to V$ satisfy (1.1.1) then for every positive integer j there exist a *piecewise constant*

control $\bar{r}_j(t): T \to R$ and a function $\bar{x}_j(t): T \to V$ that satisfy the relations

$$\dot{\bar{x}}_j(t) = g(\bar{x}_j(t), \bar{r}_j(t)) \qquad \text{a. e. in } T,$$

$$| \bar{x}(t) - \bar{x}_j(t) | \leqq \frac{1}{j}, \qquad t \in T.$$

The proof of this statement is almost identical with the proof of [9, Theorem 2.2, p. 113].

We next consider a sequence $S_1, S_2, \cdots$ of approximate solutions of (1.1) such that $x^1_{S_j}(t_1)$ converges to $\inf_S x_S{}^1(t_1)$ and all the elements of the S_j satisfy the conditions of Assumption (2.1.2). Because of our previous remark, we may assume that each control in $S_1, S_2, \cdots$ is piecewise constant. We can construct, with elements of $S_1, S_2, \cdots$, a sequence $S^* = \{(r_j(t), x_j(t))\}_{j=1}^{\infty}$ that is an approximate solution of (1.1) and such that $x^1_{S^*}(t_1) = \inf_{S'} x^1_{S'}(t_1)$.

Now let

$$\bar{\theta}_j(t) = \int_{t_0}^{t} v(x_j(\tau), r_j(\tau))\, d\tau, \qquad j = 1, 2, \cdots; \quad t \in T,$$

and let $a_j = \bar{\theta}_j(t_1)$, $\theta_j(t) = \bar{\theta}_j(t)/a_j$. The functions $\bar{\theta}_j(t)$ and $\theta_j(t)$ are defined for each j and are continuous since the $r_j(t)$ are piecewise constant, the function $v(x, r)$ is continuous in x, $x_j(t) \in D$ for all $t \in T$ and $j = 1, 2, \cdots$, and $v(x_j(t), r_j(t))$ is dominated by the integrable function $c_2(1 + | g(x_j(t), r_j(t)) |)$. Furthermore, the $\theta_j(t)$ are strictly increasing by (2.2), and $\theta_j(t_0) = 0$ and $\theta_j(t_1) = 1$. Thus the mapping $\theta_j(t): T \to [0, 1]$ has an inverse $\tau_j(\theta): [0, 1] \to T$ that is continuous and strictly increasing.

We now observe that the functions $r_j(t)$, $x_j(t)$, $\theta_j(t)$, and $a_j(t) = a_j$ satisfy (2.3.1). Thus the functions $\rho_j(\theta) = r_j(\tau_j(\theta))$, $\xi_j(\theta) = x_j(\tau_j(\theta))$, $\tau_j(\theta)$, and $a_j(\theta) = a_j$ satisfy (2.5.1); hence

$$\frac{1}{a_j} (\xi_j'(\theta), \tau_j'(\theta)) \in F(\xi_j(\theta)) \text{ a. e. in } [0, 1].$$

Furthermore, $\xi_j(\theta) \in D$, $t_0 \leqq \tau_j(\theta) \leqq t_1$, and $c_3 \leqq a_j \leqq c_4$ for $j = 1, 2, \cdots$ and $\theta \in [0, 1]$, $\xi_j(\theta)$ is arbitrarily close to A for every θ and every sufficiently large j, and

$$(\xi_j(0), \xi_j(1)) = (x_j(t_0), x_j(t_1)) \to (x_{S^*}(t_0), x_{S^*}(t_1)) \in B.$$

It follows, therefore, from [9, Theorem 3.1, p. 119] that there exists a relaxed parametric solution $(\xi(\theta), \tau(\theta))$ of (1.1) and $(\xi(\theta), \tau(\theta))$ is the uniform limit on $[0, 1]$ of a sequence $\{(\xi_{j_i}(\theta), \tau_{j_i}(\theta))\}_{i=1}^{\infty}$, where $j_1, j_2, \cdots$ is some subsequence of positive integers. Let us, for the sake of simplicity, redefine our subscripts so that $r_i(t)$, $x_i(t)$, $\xi_i(\theta)$, and $\tau_i(\theta)$ denote the old

$r_{j_i}(t), \cdots, \tau_{j_i}(\theta)$. Then

$$| \xi(\theta) - \xi_j(\theta) | + | \tau(\theta) - \tau_j(\theta) |$$
$$= | \xi(\theta) - x_j(\tau_j(\theta)) | + | \tau(\theta) - \tau_j(\theta) | \underset{j\to\infty}{\to} 0 \qquad \text{uniformly on } [0,1].$$

The new sequence $S = \{(r_j(t), x_j(t))\}_{j=1}^{\infty}$, being an infinite subsequence of S^*, is an approximate solution of (1.1). Finally,

$$\xi^1(1) = x_S{}^1(t_1) = x_{S^*}^1(t_1) = \inf_{S'} x_{S'}^1(t_1).$$

It now remains to show that $(\xi(\theta), \tau(\theta))$ is a minimizing parametric solution of (1.1). We observe that, by [9, Theorem 3.1, p. 119], there exists a minimizing parametric solution $(\xi^*(\theta), \tau^*(\theta))$ of (1.1). Furthermore, by [9, Theorem 2.2, p. 113], $(\xi^*(\theta), \tau^*(\theta))$ is the uniform limit on [0, 1] of a sequence $\{(\xi_j{}^*(\theta), \tau_j{}^*(\theta))\}_{j=1}^{\infty}$ of solutions of (2.5.1) (corresponding to piecewise constant controls $\rho_1{}^*(\theta), \rho_2{}^*(\theta), \cdots$). Since (2.5.1) is equivalent to (2.3.1) to every $(\rho_j{}^*(\theta), \xi_j{}^*(\theta), \tau_j{}^*(\theta))$ on [0, 1] there corresponds a solution $(r_j{}^*(t), x_j{}^*(t), \theta_j{}^*(t))$ of (2.3.1) on $[t_0, t_1]$, and $\xi_j{}^*(0) = x_j{}^*(t_0)$ and $\xi_j{}^*(1) = x_j{}^*(t_1), j = 1, 2, \cdots$. We have

$$\xi^{*1}(1) = \lim_{j\to\infty} \xi_j{}^{*1}(1) = \lim_{j\to\infty} x_j{}^{*1}(t_1) \geqq x_S{}^1(t_1) = \xi^1(1),$$

since $\{(r_j{}^*(t), x_j{}^*(t))\}_{j=1}^{\infty}$ is an approximate solution of (1.1), as can be easily verified. Furthermore, $\xi^{*1}(1) \leqq \xi^1(1)$ since $(\xi^*(\theta), \tau^*(\theta))$ is a minimizing parametric solution of (1.1). It follows that $\xi^{*1}(1) = \xi^1(1)$ and thus we see that $(\xi(\theta), \tau(\theta))$ is a minimizing parametric solution of (1.1).

This concludes the proof of the theorem.

3.3 Let $(\xi(\theta), \tau(\theta))$ be a minimizing parametric solution of (1.1), and let $\mathfrak{A}$ be the union of all open subintervals of [0, 1] on which $\tau(\theta)$ is constant. Then $\tau(\theta)$ is strictly increasing on $\mathfrak{B} = [0, 1] - \mathfrak{A}$ and, on that set, θ is a continuous and increasing function of τ. Therefore $x(t) = \xi(\theta(t))$ is a continuous function of t on $\mathfrak{B}_t = \tau(\mathfrak{B})$.

If $\mathfrak{A}$ is empty then $\tau(\theta)$ is a continuous one-to-one mapping of [0, 1] onto $[t_0, t_1]$; hence $x(t)\colon [t_0, t_1] \to V$ is continuous. This need not be the case when $\mathfrak{A}$ is nonempty. In either case, however, once a minimizing parametric solution $(\xi(\theta), \tau(\theta))$ of (1.1) is known, the construction of [9, Theorem 2.2, p. 113] yields a sequence of solutions $S = \{(r_j(t), x_j(t))\}_{j=1}^{\infty}$ of (1.1.1) and a sequence $\{\tau_j(\theta)\}_{j=1}^{\infty}$ that approximate $(\xi(\theta), \tau(\theta))$ in the sense of Theorem 3.2. The sequence S is an approximate solution of (1.1).

This brings us to the next topic, namely, the derivation of "constructive" necessary conditions satisfied by minimizing parametric solutions. When the set A is defined by the simultaneous inequalities $a^j(x) \leqq 0$,

$j = 1, \cdots, m$, the results of [12] are directly applicable to the system (2.5.1), (2.5.2). Since these results are, however, relatively complicated, we shall limit ourselves here to restating the pertinent conditions for the special case $A = E_n$.

4. Proper representations. The necessary conditions derived in [10, Theorem 6.1, p. 142] apply to the minimizing parametric solution of (1.1) if $A = E_n$ and there exists a proper representation $f(x, \sigma)$ of $F(x)$ [10, Definition 2.1, p. 130]. We shall indicate two ways in which such *proper representations* can be constructed.

4.1. *The Filippov representation.* Let S be a compact set in some Euclidean space and let $f(x, \sigma): V \times S \to E_{n+1}$ be continuous and satisfy the following conditions:

$$F(x) = \{f(x, \sigma) \mid \sigma \in S\}, \qquad x \in V, \tag{4.1.1}$$

(4.1.2) the partial derivatives $\partial f^i(x, \sigma)/\partial x^j$, $i = 1, \cdots, n + 1$; $j = 1, \cdots, n$, exist, they are continuous functions of x for each σ, and they are uniformly bounded.

Then it easily follows from our previous assumptions and from a lemma of Filippov [1, p. 78] that $f(x, \sigma)$ is a proper representation of $F(x)$.

4.2. *The Young representation.* Let U be a compact Hausdorff space (in particular, a compact Euclidean set), and let $h(x, u): V \times U \to E_{n+1}$ be continuous and such that

$$\{h(x, u) \mid u \in U\} = \text{closure of } G(x), \qquad x \in V.$$

Assume, furthermore, that the partial derivatives $\partial h^i(x, u)/\partial x^j$, $i = 1, \cdots, n + 1; j = 1, \cdots, n$, exist, that they are continuous functions of x, uniformly in u, and that they are uniformly bounded. Finally, let S be the class of probability measures σ defined on Borel subsets of U, and let

$$f(x, \sigma) = \int_U h(x, u)\, d\sigma, \qquad (x, \sigma) \in V \times S.$$

Then it follows from [9, Theorem 4.1, p. 124] that $f(x, \sigma)$ is a proper representation of $F(x)$.

As a special case of the Filippov representation we may mention the Gamkrelidze representation [2]. Let U be a compact set in some Euclidean space E_m, and let $h(x, u)$ satisfy the same conditions as in §4.2. Let

$$S = \Big\{\sigma \mid \sigma = (u_1, \cdots, u_{n+2}, p_1, \cdots, p_{n+2}), u_i \in U \ (i = 1, \cdots, n + 2),$$
$$p_i \geqq 0 \quad (i = 1, \cdots, n + 2), \quad \sum_{j=1}^{n+2} p_j = 1\Big\}.$$

We let

$$f(x, \sigma) = \sum_{j=1}^{n+2} p_j h(x, u_j).$$

Since, by a known theorem of Carathéodory, $\{f(x, \sigma) \mid \sigma \in S\}$ is the convex closure of $G(x)$, we conclude that the Gamkrelidze representation is a Filippov representation.

5. Necessary conditions. Let $A = E_n$, let $f(x, \sigma)$ be a proper representation of $F(x)$, and let $(\xi(\theta), \tau(\theta))$ be a minimizing parametric solution of (1.1). Furthermore, let $\phi(p) = (\phi^1(p), \cdots, \phi^n(p))$, $\psi(p) = (\psi^1(p), \cdots, \psi^n(p))$, $\chi(p) = (\phi(p), \psi(p))$, and let $\chi(p)$ be a continuously differentiable mapping from some Euclidean, compact, and convex set C to B, such that $(\xi(0), \xi(1)) \in \chi(C)$. Then we can easily deduce from [10, Theorem 6.1, p. 142] that either there exists a point $\bar{p} \in C$ such that

$$\xi(0) = \phi(\bar{p}), \qquad \xi(1) = \psi(\bar{p}), \quad \text{and} \quad \psi_p{}^1(\bar{p})\bar{p} = \min_{p \in C} \psi_p{}^1(\bar{p})p$$

(where $\psi_p{}^1$ is the gradient of ψ^1), or there exist constants $a > 0$ and γ, scalar functions $z^i(\theta)$, $i = 1, \cdots, n+1$, on $[0, 1]$, a function $\sigma(\theta) : [0, 1] \to S$, and a point $\bar{p} \in C$ such that

$$\frac{d(\xi(\theta), \tau(\theta))}{d\theta} = af(\xi(\theta), \sigma(\theta)) \qquad \text{a. e. in } [0, 1], \tag{5.1}$$

$$\sum_{j=1}^{n+1} |z^j(\theta)| \not\equiv 0, \qquad 0 \leqq \theta \leqq 1, \tag{5.2}$$

$$\frac{dz^i(\theta)}{d\theta} = -a \sum_{j=1}^{n+1} z^j(\theta) \frac{\partial f^j(\xi(\theta), \sigma(\theta))}{\partial x_i} \qquad \text{a. e. in } [0, 1], \quad i = 1, \cdots, n, \tag{5.3}$$

$$\frac{dz^{n+1}(\theta)}{d\theta} = 0,$$

$$\sum_{j=1}^{n+1} z^j(\theta) f^j(\xi(\theta), \sigma(\theta)) = \min_{\sigma \in S} \sum_{j=1}^{n+1} z^j(\theta) f^j(\xi(\theta), \sigma) = 0 \qquad \text{a. e. in } [0, 1], \tag{5.4}$$

$$\xi(0) = \phi(\bar{p}), \qquad \xi(1) = \psi(\bar{p}), \qquad \gamma \geqq 0, \quad \text{and}$$

$$a \cdot \bar{p} = \min_{p \in C} a \cdot p, \quad \text{where} \tag{5.5}$$

$$a = \gamma \psi_p{}^1(\bar{p}) - \sum_{j=1}^{n} z^j(1) \psi_p{}^j(\bar{p}) + \sum_{j=1}^{n} z^j(0) \phi_p(\bar{p}).$$

6. Examples. We shall illustrate the application of our results with two examples.

6.1. *A continuous minimizing curve with an unbounded control.* Let (1.1) be of the form

$$\begin{aligned} &\dot{x}_1 = r^2, \qquad \dot{x}_2 = x_1 r, \\ &x_1(0) = x_2(0) = 0, \qquad x_2(t_1) = L > 0, \end{aligned} \tag{6.1.1}$$

where subscripts, instead of superscripts, denote the components of x and $0 \leqq r < \infty$. We shall prove that this system admits an ordinary minimizing solution with the (unbounded) control

$$\bar{r}(t) = \left(\frac{L}{3t_1}\right)^{1/3} t^{-1/3}.$$

We observe that (6.1.1) has a solution $y(t)$ corresponding to the constant control

$$r(t) = \tilde{r} = (2L)^{1/3} t_1^{-2/3}.$$

This yields

$$y_1(t) = \tilde{r}^2 t, \qquad y_2(t) = \tfrac{1}{2}\tilde{r}^3 t^2.$$

Clearly $| x_1(t) | \leqq y_1(t_1) = \tilde{r}^2 t_1$ for all solutions $x(t)$ such that $x_1(t_1) \leqq y_1(t_1)$; hence

$$\int_0^{t_1} (r^2(t) + x_1(t) r(t))\, dt = x_1(t_1) + L \leqq \tilde{r}^2 t_1 + L$$

for all such solutions. Thus Assumption 2.1 is satisfied.

We can easily verify that the function $v(x, r) = (r + 1)^2$ satisfies (2.2). Thus (2.5) is of the form

$$\begin{aligned} &\xi_1' = \frac{a r^2}{(r+1)^2}, \quad \xi_2' = \frac{a \xi_1 r}{(r+1)^2}, \quad \tau' = \frac{a}{(r+1)^2}, \quad a' = 0, \\ &\xi_1(0) = 0, \quad \xi_2(0) = 0, \quad \xi_2(1) = L, \quad \tau(0) = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \tau(1) = t_1. \end{aligned} \tag{6.1.2}$$

Let now $u = r/(r + 1)$. Then

$$\xi_1' = a u^2, \quad \xi_2' = a \xi_1 u(1 - u), \quad \tau' = a(1 - u)^2, \quad a' = 0. \tag{6.1.3}$$

Since $0 \leqq r < \infty$, it follows that $0 \leqq u < 1$. We shall obtain the closure of $G(x)$ by choosing u from the closed interval [0, 1]. Assumption 3.1 is clearly satisfied.

Let now $f(x, \sigma)$ be the Young representation of $F(x)$, where $h(x, u)$ is

defined by the right hand sides of (6.1.3), $U = [0, 1]$, and the bases of the topology are open subintervals of [0, 1].

The necessary conditions of §5 imply that, if $(\xi_1(\theta), \xi_2(\theta), \tau(\theta))$ is a minimizing parametric solution of (6.1.1), and if $\sigma(\theta)$ is the corresponding "control", then there exist absolutely continuous functions $z_i(\theta)$, $i = 1, 2, 3$, and a positive constant a such that:

$$
\begin{aligned}
&\xi_1' = a\int_0^1 u^2\,d\sigma \qquad \text{a. e. in } [0, 1],\\
&\xi_2' = a\xi_1\int_0^1 u(1-u)\,d\sigma \qquad \text{a. e. in } [0, 1],\\
&\tau' = a\int_0^1 (1-u)^2\,d\sigma \qquad \text{a. e. in } [0, 1],\\
&a' = 0,\\
&\xi_1(0) = \xi_2(0) = 0, \quad \tau(0) = 0, \quad \xi_2(1) = L, \quad \tau(1) = t_1;
\end{aligned}
\tag{6.1.4}
$$

$$
\begin{aligned}
&z_1' = -az_2\int_0^1 u(1-u)\,d\sigma, \quad z_2' = z_3' = 0, \qquad \text{a. e. on } [0, 1],\\
&\qquad\qquad |z_1(\theta)| + |z_2| + |z_3| \neq 0 \qquad \text{on } [0, 1];
\end{aligned}
\tag{6.1.5}
$$

$$
\begin{aligned}
&\min_{0\leqq u\leqq 1} H(\xi, z, u)\\
&= \min_{0\leqq u\leqq 1}\{(z_1 - z_2\xi_1 + z_3)u^2 + (z_2\xi_1 - 2z_3)u + z_3\} = 0
\end{aligned}
\tag{6.1.6}
$$

a.e. in [0, 1].

Relation (6.1.6) yields (setting $u = 0$ and $u = 1$)

$$z_1(\theta) \geqq 0 \quad \text{on } [0, 1] \text{ and } \quad z_3 \geqq 0. \tag{6.1.7}$$

Relations (6.1.4), (6.1.5) and (6.1.6) yield

$$\xi_1 z_1' + z_2\xi_2' = 0 \quad \text{and} \quad z_1\xi_1' + z_2\xi_2' + z_3\tau' = 0 \qquad \text{a.e. on } [0, 1]; \tag{6.1.8}$$

hence $(z_1\xi_1)' + 2z_2\xi_2' + z_3\tau' = 0$ a.e. on [0, 1] and

$$z_1(1)\xi_1(1) + 2z_2\xi_2(1) + z_3t_1 = 0. \tag{6.1.9}$$

If $z_2 = 0$ then, by (6.1.7) and (6.1.9), $z_3 = z_1(1)\xi_1(1) = 0$ and, by (6.1.5) and (6.1.6), $z_1 > 0$ and σ is a.e. concentrated at $u = 0$. This, however, implies $\xi_2(1) = 0 \neq L$, contradicting (6.1.4). Thus $z_2 \neq 0$.

If $z_3 = 0$ then, by (6.1.6), $z_1u^2 + z_2\xi_1(u - u^2)$ must be nonnegative for small positive u, and this implies $z_2\xi_1 \geqq 0$ a.e. in [0, 1]. From (6.1.9) we deduce

$$z_1(1)\xi_1(1) + 2z_2L = 0;$$

hence $z_2 < 0$ and $\xi_1 \leqq 0$ a.e. in $[0, 1]$. Since $\xi_1(0) = 0$ and $\xi_1'(\theta) \geqq 0$, we see that $\xi_1'(\theta) = 0$ a.e. in $[0, 1]$, and this implies that σ is a.e. concentrated at 0 and $\xi_2(1) = 0 \neq L$, contrary to (6.1.4). Thus $z_3 \neq 0$; hence $z_3 > 0$.

Relation (6.1.9) implies $z_2 < 0$ since $z_1(1)\xi_1(1) \geqq 0$ and $t_1 > 0$. Thus $z_1 - z_2\xi_1 + z_3 \geqq 0$ and $H(\xi, z, u)$ has a unique minimum with respect to u in $[0, 1]$; this minimum is achieved at

$$\bar{u} = \frac{z_3 - \frac{1}{2}z_2\,\xi_1}{z_1 - z_2\,\xi_1 + z_3},$$

which belongs to $[0, 1]$ since $z_1 \geqq 0$, $z_3 > 0$, and $z_2\xi_1 \leqq 0$. Thus the measure σ is always concentrated at $\bar{u}$. Furthermore, since $\min_u H(\xi, z, u) = 0$, $H(\xi, z, u)$ is a perfect square and it follows that

$$(z_2\xi_1 - 2z_3)^2 = 4z_3(z_1 - z_2\xi_1 + z_3); \tag{6.1.10}$$

hence $\bar{u} = 2z_3/(2z_3 - z_2\xi_1)$. It follows now from (6.1.4) that

$$\frac{d\xi_1}{d\tau} = \left(\frac{\bar{u}}{1 - \bar{u}}\right)^2 = \frac{4z_3^2}{z_2^2\xi_1^2};$$

hence

$$\frac{1}{3}\,\xi_1^3(\theta) = \frac{4z_3^2}{z_2^2}\,\tau(\theta).$$

Furthermore,

$$\frac{d\xi_2}{d\xi_1} = \xi_1\,\frac{1 - \bar{u}}{\bar{u}} = -\frac{z_2}{2z_3}\,\xi_1^2;$$

hence

$$\xi_2(\theta) = -\frac{z_2}{6z_3}\,\xi_1^3(\theta) = -\frac{2z_3}{z_2}\,\tau(\theta)$$

and

$$L = -\frac{2z_3}{z_2}\,t_1.$$

We thus verify that

$$x_1 = 3^{1/3}L^{2/3}t_1^{-2/3}t^{1/3}$$

and $x_2 = Lt/t_1$ are continuous functions of t. Furthermore, $\bar{u} = \bar{r}/(\bar{r} + 1)$ implies

$$\bar{r} = \frac{\bar{u}}{1 - \bar{u}} = -\frac{2z_3}{z_2\,x_1} = \frac{L}{t_1\,x_1} = 3^{-1/3}\left(\frac{L}{t_1}\right)^{1/3} t^{-1/3}.$$

Thus (6.1.1) admits an ordinary minimizing solution with the unbounded

control

$$\bar{r}(t) = 3^{-1/3}\left(\frac{L}{t_1}\right)^{1/3} t^{-1/3}.$$

We next consider a problem for which the minimizing control is a "delta-function".

6.2. *A minimizing curve with a single jump discontinuity.* We wish to minimize the value of $\phi(x(t_1))$ subject to the conditions

$$\begin{aligned} \dot{x}(t) &= k(x(t)) + r \qquad \text{a.e. in } [t_0, t_1], \\ \int_{t_0}^{t_1} |r(t)|\, dt &= L > 0, \\ x(t_0) &\in B_0, \end{aligned} \tag{6.2.1}$$

where $t_0 < t_1$, $x = (x^1, \cdots, x^n)$, $r \in R = E_n$, $|r|$ is the Euclidean norm (length) of r, B_0 is a compact set in E_n, $k(x): E_n \to E_n$ is continuous and continuously differentiable in E_n, and $\phi(x)$ is continuous and continuously differentiable in E_n. We shall also assume that

(a) $|k(x)| \leqq c$ in E_n for some constant c, and

(b) the Jacobian matrix $k_x = (\partial k^i/\partial x^j)$, $i, j = 1, \cdots, n$, is either positive definite in E_n or negative definite in E_n.

Assumption (a) and the continuous differentiability of $k(x)$ are clearly sufficient to insure the existence of a solution of (6.2.1). Let now $x^*(t; x_0)$ be the solution of the system $\dot{x}^*(t) = k(x^*(t))$ a.e. in $[t_0, t_1]$, $x^*(t_0) = x_0$. We shall prove that there exists a minimizing parametric solution $(\xi(\theta), \tau(\theta))$, $0 \leqq \theta \leqq 1$, with the following properties: either

$\xi(1)$ is a stationary point of $\phi(x)$, i.e., $\partial\phi/\partial x^j = 0$ at $\xi(1)$, $j = 1, \cdots, n$, or

k_x is negative definite and

$$\xi(\tau) = x^*(\tau; \xi(0)), \qquad \tau(\theta) = t_0 + a\theta \qquad \text{for } 0 \leqq \theta \leqq \theta_1 = \frac{t_1 - t_0}{a},$$

$$\xi(\theta) = x^*(t_1; \xi(0)) + \frac{\theta - \theta_1}{1 - \theta_1} L\nu(1), \qquad \tau(\theta) = t_1 \qquad \text{for } \theta_1 < \theta \leqq 1,$$

where $a = t_1 - t_0 + L$ and $\nu(1)$ is a vector of length 1, or

k_x is positive definite and

$$\xi(\theta) = \xi(0) + \frac{\theta}{\theta_1} L\nu(0), \qquad \tau(\theta) = t_0 \qquad \text{for } 0 \leqq \theta \leqq \theta_1 = \frac{L}{a},$$

$$\xi(\tau) = x^*(\tau; \xi(0) + L\nu(0)), \qquad \tau(\theta) = t_0 + a(\theta - \theta_1) \qquad \text{for } \theta_1 \leqq \theta \leqq 1,$$

where $a = t_1 - t_0 + L$ and $\nu(0)$ is a vector of length 1.

Such a parametric solution corresponds to a "solution" of (6.2.1) in which $r(t) = L\nu\delta(t - t')$, where ν is some unit vector, $t' = t_0$ (or t_1) if k_x is positive (or negative) definite, and $\delta(t)$ is the Dirac δ-function.

We shall now proceed to prove our assertion. Let $r = r_0\nu$, where $r_0 = |r|$ and ν is a unit vector. Our problem is equivalent to that of minimizing $x^0(t_1)$ subject to the conditions

$$\begin{aligned} \dot{x}^0(t) &= 0 \qquad \text{a.e. in } [t_0, t_1], \\ \dot{x}(t) &= k(x(t)) + r_0(t)\nu(t) \qquad \text{a.e. in } [t_0, t_1], \\ \dot{x}^{n+1}(t) &= r_0(t) \qquad \text{a.e. in } [t_0, t_1], \\ x^0(t_1) &= \phi(x(t_1)), \quad x(t_0) \in B_0, \quad x^{n+1}(t_0) = 0, \quad x^{n+1}(t_1) = L. \end{aligned} \tag{6.2.2}$$

We can easily verify that all solutions $(x^0(t), x(t), x^{n+1}(t))$ of (6.2.2) are uniformly bounded and that Assumption 2.1 is satisfied. We can also verify that the function $v(x, r) = |r| + 1 = r_0 + 1$ satisfies (2.2). Let $u_0 = r_0/(r_0 + 1)$. Then (2.5) is of the form

$$\begin{aligned} &(\xi^0)' = 0, \quad \xi' = a(1 - u_0)k(\xi) + au_0\nu, \quad (\xi^{n+1})' = au_0, \\ &\tau' = a(1 - u_0), \quad a' = 0 \quad \text{a.e. in } [0, 1], \end{aligned} \tag{6.2.3}$$

$$\xi^0(1) = \phi(\xi(1)), \quad \xi(0) \in B_0, \quad \xi^{n+1}(0) = 0, \quad \xi^{n+1}(1) = L, \quad \tau(0) = t_0, \quad \tau(1) = t_1, \tag{6.2.4}$$

where $u_0(\theta)$ and $\nu(\theta)$ denote $u_0(\tau(\theta))$ and $\nu(\tau(\theta))$.

Let now $\tilde{\xi} = (\xi^0, \xi^1, \cdots, \xi^{n+1}, \tau)$. We shall use a Filippov representation to replace system (6.2.3). We let

$$S = \{(u_0, \nu, u_1) \mid 0 \leqq u_0 \leqq 1, 0 \leqq u_1 \leqq 1, |\nu| = 1\},$$

and replace (6.2.3) by

$$\begin{aligned} &(\xi^0)' = 0, \quad \xi' = a(1 - u_0)k(\xi) + au_1u_0\nu, \quad (\xi^{n+1})' = au_0, \\ &\tau' = a(1 - u_0), \quad a' = 0 \quad \text{a.e. in } [0, 1]. \end{aligned} \tag{6.2.5}$$

As the (u_0, ν, u_1) range over S, the right hand sides of (6.2.5) span the convex hull of the corresponding set of (6.2.3). The necessary conditions of §5 yield the conclusion that either $\partial\phi/\partial x^i = 0$, $i = 1, \cdots, n$, at $x = \xi(1)$ or there exist a constant $a > 0$, an absolutely continuous function

$$\tilde{z}(\theta) = (z^0(\theta), z^1(\theta), \cdots, z^{n+2}(\theta)), \qquad 0 \leqq \theta \leqq 1,$$

and a control $u(\theta) = (u_0(\theta), \nu(\theta), u_1(\theta))$ such that:

$$\begin{aligned} &(z^0)' = 0, \quad z' = -a(1 - u_0)k_x^T(\xi)z, \quad (z^{n+1})' = (z^{n+2})' = 0 \\ &\text{a.e. in } [0, 1], \end{aligned} \tag{6.2.6}$$

$$\sum_{j=0}^{n+2} |z^j| \neq 0 \qquad \text{on } [0, 1],$$

(where k_x^T is the transpose of the matrix $k_x = (\partial k^i/\partial x^j)$, $i, j = 1, \cdots, n$);

$$H(\tilde{\xi}(\theta), \tilde{z}(\theta), u(\theta)) = \min_{0 \leqq u_0 \leqq 1, 0 \leqq u_1 \leqq 1, |\nu|=1} H(\tilde{\xi}(\theta), \tilde{z}(\theta), u) = 0 \quad \text{a.e. in } [0, 1], \tag{6.2.7}$$

where

$$H(\tilde{\xi}, \tilde{z}, u) = (-z \cdot k(\xi) + u_1 z \cdot \nu + z^{n+1} - z^{n+2})u_0 + z \cdot k(\xi) + z^{n+2}; \quad z^0(0) = z^0 = 0. \tag{6.2.8}$$

Let us henceforth assume that $\xi(1)$ is not a stationary point of $\phi(x)$. Then, by (6.2.6), either $|z(\theta)| = 0$ on $[0, 1]$ or $|z(\theta)| \neq 0$ on $[0, 1]$. If $|z(\theta)| = 0$ then, by (6.2.6), (6.2.7) and (6.2.8),

$$\min_{0 \leqq u_0 \leqq 1} (z^{n+1} - z^{n+2})u_0 + z^{n+2} = 0 \qquad \text{a.e. in } [0, 1]$$

and z^{n+1} and z^{n+2} are unequal constants. This implies that either $z^{n+1} - z^{n+2} > 0$ and $u_0(\theta) = 0$ a.e. in $[0, 1]$ or $z^{n+1} - z^{n+2} < 0$ and $u_0(\theta) = 1$ a.e. in $[0, 1]$; hence $\xi^{n+1}(1) = 0$ or $\tau(1) = t_0$, both contrary to (6.2.4). Thus $|z(\theta)| \neq 0$ on $[0, 1]$.

Since $0 \leqq u_0(\theta) \leqq 1$ and $|z(\theta)| \neq 0$, it follows that

$$\min_{|\nu|=1, 0 \leqq u_1 \leqq 1} u_0 u_1 z \cdot \nu = -|z| u_0$$

and

$$u_1(\theta) = 1 \quad \text{and} \quad \nu(\theta) = -\frac{z(\theta)}{|z(\theta)|} \qquad \text{a. e. in } [0, 1]. \tag{6.2.9}$$

Relation (6.2.7) yields

$$H(\tilde{\xi}(\theta), \tilde{z}(\theta), u(\theta)) = \min_{0 \leqq u_0 \leqq 1} \{(-z(\theta) \cdot k(\xi(\theta)) - |z(\theta)| + z^{n+1} - z^{n+2})u_0 + z(\theta) \cdot k(\xi(\theta)) + z^{n+2}\} = 0 \qquad \text{a.e. in } [0, 1]. \tag{6.2.10}$$

This last relation implies that, except for a set of measure 0, $u_0(\theta) = 0$ if

$$\alpha(\theta) = -z(\theta) \cdot k(\xi(\theta)) - |z(\theta)| + z^{n+1} - z^{n+2} > 0$$

and $u_0(\theta) = 1$ if $\alpha(\theta) < 0$.

We observe that, by (6.2.5), (6.2.6), and (6.2.9),

$$\frac{d\alpha(\theta)}{d\theta} = \frac{\alpha z^T(\theta) k_x(\xi(\theta)) z(\theta)}{|z(\theta)|}.$$

We can now conclude, in view of Assumption (b), that $d\alpha(\theta)/d\theta > 0$ if k_x is positive definite and $d\alpha(\theta)/d\theta < 0$ if k_x is negative definite.

We have already observed that $u_0(\theta)$ cannot be always 0 or always 1 since, by (6.2.4) and (6.2.5),

$$a \int_0^1 u_0(\theta)\, d\theta = L \neq 0$$

and

$$a \int_0^1 (1 - u_0(\theta))\, d\theta = t_1 - t_0 \neq 0.$$

It follows that $\alpha(\theta)$ changes sign at a unique point θ_1 in $(0, 1)$; hence, by (6.2.10), $u_0(\theta) = 0$ a.e. in $[0, \theta_1]$ and $u_0(\theta) = 1$ a.e. in $[\theta_1, 1]$ if k_x is negative definite, and $u_0(\theta) = 1$ a.e. in $[0, \theta_1]$ and $u_0(\theta) = 0$ a.e. in $[\theta_1, 1]$ if k_x is positive definite. Our assertion now follows directly from relations (6.2.4), (6.2.5), (6.2.6), and (6.2.9).

REFERENCES

[1] A. V. Filippov, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[2] R. V. Gamkrelidze, *On sliding optimal states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245.

[3] E. J. McShane, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.

[4] ———, *Necessary conditions in generalized-curve problems of the calculus of variations*, Ibid., 7 (1940), pp. 1–27.

[5] ———, *Existence theorems for Bolza problems in the calculus of variations*, Ibid., 7 (1940), pp. 28–61.

[6] L. W. Neustadt, *Optimization, a moment problem, and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[7] ———, *A general theory of minimum-fuel space trajectories*, this Journal, 3 (1965), pp. 317–356.

[8] R. W. Rishel, *An extended Pontryagin principle for control systems whose control laws contain measures*, this Journal, 3 (1965), pp. 191–205.

[9] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[10] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129–145.

[11] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[12] ———, *Unilateral variational problems with several inequalities*, Michigan Math. J., to appear.

[13] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Compt. Rend. Soc. Sci. et Lettres Varsovie, Cl. III., 30 (1937), pp. 212–234.

[14] ———, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239–258.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

A REMARK ON COMPLETE CONTROLLABILITY*

HSIN CHU†

1. Introduction. Consider a plant defined by

$$\begin{aligned} \dot{X} &= AX + bU, \\ Y &= c \cdot X, \end{aligned} \tag{1}$$

where b and c are constant n-dimensional column vectors. The plant has an input U and an output Y. In this note, we show that the plant is always completely controllable (or completely observable) for every nonzero vector b (or c) if and only if $n = 2$ and the characteristic roots of A are complex. We also give a different and somewhat seemingly shorter proof of the following well known result of Kalman's: the plant is completely controllable (or completely observable) for some b (or c) if and only if in the Jordan canonical form of A no two blocks are associated with the same eigenvalue.

For the terminology used here, consult [5], [6], or [8].

2. On $\dot{X} = AX$. Consider the principal part

$$\dot{X} = AX, \tag{2}$$

where $A = (a_{ij})$ is an $n \times n$ real constant matrix and X is a column vector, $X^t = (x_1, x_2, \cdots, x_n)$, where x_i, $i = 1, 2, \cdots, n$, are differentiable functions of t. Let W be the volume of a parallelepiped, defined by X, AX, $\cdots$, $A^{n-1}X$, and the order of these vectors is their orientation. Then

$$W = D(X, AX, \cdots, A^{n-1}X)$$

is the determinant of these column vectors X, AX, $\cdots$, $A^{n-1}X$. It is obvious that $W = D(X, \dot{X}, \cdots, X^{(n-1)})$.

LEMMA. *Along every trajectory of* (2), *we have*

$$W(X) = Ce^{\mathrm{Tr}(A)\,t},$$

where C can be determined by any initial values X_0, t_0 and Tr (A) *is the trace of A.*

Proof. This result is a direct consequence of the Jacobi identity (see [3, vol. II]).

THEOREM 1. *Along every trajectory of* (2), $W(X) \neq 0$, *except at* $X = 0$,

* Received by the editors May 18, 1965.

† University of Alabama Research Institute, Huntsville, Alabama. This work was supported by Contract NAS8-1646 with the George C. Marshall Space Flight Center, Huntsville, Alabama.

if and only if $n = 2$, the characteristic roots of A are complex, and $W(X)$ is either positive definite or negative definite.

Proof. Let $W(X) = 0$ except at $X = 0$. Suppose $n > 2$. There exists a nonsingular matrix P over the field of real numbers such that

$$PAP^{-1} = A' = \begin{pmatrix} A_1' & A_2' \\ 0_r & A_3' \end{pmatrix},$$

where 0_r is some $r \times r$ zero matrix such that $1 \leqq r \leqq [n/2]$, where $[n/2]$ is the integer part of $n/2$. Let $X' = PX$ be a linear transformation. Then $\dot{X}' = (PAP^{-1})X'$. Choose a column vector X_0' such that its first r coordinates are not zero and its remaining $n - r$ coordinates are all zero. Then the last r coordinates of all the vectors, $A'X_0'$, $(A')^2X_0'$, $\cdots$, $(A')^{n-1}X_0'$, are zero. It follows that $W(X_0') = 0$. Let $X_0 = P^{-1}X_0'$, which is not zero. We have $W(X_0) = |P|^{-1}W(X') = 0$, a contradiction! Let $n = 2$ and the characteristic roots of A be real. Then there exists a nonsingular matrix P over the field of real numbers such that

$$PAP^{-1} = A' = \begin{pmatrix} \alpha_1 & k \\ 0 & \alpha_2 \end{pmatrix}.$$

By the same argument as above, we can find a nonzero vector X_0 such that $W(X_0) = 0$. Consequently, if $W(X) \not\equiv 0$, except at $X = 0$, then $n = 2$ and the characteristic roots of A are complex. If $n = 2$ and the characteristic roots of A are complex, then there exists a nonsingular matrix P such that

$$PAP^{-1} = A' = \begin{pmatrix} \mu & \nu \\ -\nu & \mu \end{pmatrix},$$

where μ and ν are real numbers with $\nu > 0$.

Let $X' = PX$ be a linear transformation and X', $(X')^t = (x_1', x_2')$, be any vector. We have

$$W(X') = (X', A'X') = \begin{pmatrix} x_1' & \mu x_1' + \nu x_2' \\ x_2' & \nu x_1' + \mu x_2' \end{pmatrix} = -\nu((x_1')^2 + (x_2')^2),$$

and

$$W(X) = -|P|^{-1}\nu((x_1')^2 + (x_2')^2).$$

It follows that $W(X)$ is positive definite if $|P| < 0$ and $W(X)$ is negative definite if $|P| > 0$.

COROLLARY 1. *A plant* (1) *is always completely controllable* (*or completely observable*) *for every nonzero vector* b (*or* c) *if and only if* $n = 2$ *and the characteristic roots of* A *are complex.*

COROLLARY 2. *The function* W *is not a Lyapunov function unless* $n = 2$ *and the characteristic roots of* A *are complex with negative real parts.*

Proof. This follows by the lemma, Theorem 1, and the fact that $\mathrm{Tr}\,(A) < 0$ in this case.

COROLLARY 3. *The singularities of zero vectors of* $\dot{X} = AX$, *where* $X \in R^2$, *can be classified by* W, $\mathrm{Tr}\,(A)$ *and* $|A|$, *which is the determinant of* A, *as follows:*

(a) $|A| < 0$: *saddle point,*

(b) $|A| > 0$, $\mathrm{Tr}\,(A) < 0$, W *is either positive definite or negative definite: stable focus,*

(c) $|A| > 0$, $\mathrm{Tr}\,(A) < 0$, W *is either positive semidefinite or negative semidefinite: stable node,*

(d) $|A| > 0$, $\mathrm{Tr}\,(A) > 0$, W *is either positive definite or negative definite: unstable focus,*

(e) $|A| > 0$, $\mathrm{Tr}\,(A) > 0$, W *is either positive semidefinite or negative semidefinite: unstable node,*

(f) $|A| > 0$, $\mathrm{Tr}\,(A) = 0$: *center,*

(g) $|A| = 0$: *degenerate point.*

The following theorem is a modified version of a well known result of Kalman's (see [5, Theorem 25]). Here, we offer a different proof.

THEOREM 2. *Along every trajectory of* (2), $W(X) = 0$ *if and only if the minimal polynomial of* A *is not equal to the characteristic polynomial of* A.

Proof. If $W(X) = 0$, then $A, AX, \cdots, A^{n-1}X$ are linearly dependent for all X. If the minimal polynomial of A is equal to the characteristic polynomial of A, it is known that there always exists a nonzero vector X_0 whose minimal polynomial coincides with the characteristic polynomial of A (e.g., see [3, vol. I, p. 180, Theorem 2]). Consequently, $A, AX_0, \cdots, A^{n-1}X_0$ are linearly dependent, a contradiction.

Conversely, if the minimal polynomial of A is not equal to the characteristic polynomial of A, let

$$\psi(\lambda) = \lambda^p + b_1\lambda^{p-1} + \cdots + b_p$$

be the minimal polynomial of A, where $p \leqq n - 1$, and we have $\psi(A) = 0$ or $\psi(A)X = 0$ for all X. It follows that $A, AX, \cdots, A^pX$ are linearly dependent and so are $A, AX, \cdots, A^{p-1}X, \cdots, A^{n-1}X$. Consequently, $W(X) = 0$ for all X.

COROLLARY 4. *A plant* (1) *is completely controllable* (*or completely observable*) *for some* b (*or* c) *if and only if in the Jordan canonical form of* A *no two blocks are associated with the same eigenvalue.*

REFERENCES

[1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[2] G. BIRKHOFF AND S. MACLANE, *A Survey of Modern Algebra*, Macmillan, New York, 1953.

[3] F. R. GANTMACHER, *Matrix Theory*, Chelsea, New York, 1959, vols. I and II.

[4] N. JACOBSON, *Lectures in Abstract Algebra*, Van Nostrand, Princeton, 1953, vols. I and II.

[5] R. E. KALMAN, *On the general theory of control systems*, Proceedings of First International Congress on Automatic Control, Eyre & Sprottswoode Ltd., London, 1961.

[6] ———, *Mathematical description of linear dynamical systems*, this Journal, 2 (1963), pp. 152–192.

[7] J. P. LASALLE AND S. LEFSCHETZ, *Stability of Lyapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[8] S. LEFSCHETZ, *Recent advances in the stability of non-linear controls*, SIAM Rev., 7 (1965), pp. 1–12.

[9] L. S. PONTRYAGIN, *Ordinary Differential Equations*, Addison-Wesley, Reading, Massachusetts, 1962.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

THE APPLICATION OF LYAPUNOV'S SECOND METHOD TO INTERCONNECTED SYSTEMS*

F. N. BAILEY†

Summary. As many engineering systems are made up of an interconnection of simple subsystems, it is natural to attempt to utilize this interconnection structure in developing system analysis techniques. In this paper interconnection information is used in conjunction with properties of the individual subsystems to obtain sufficient conditions for asymptotic stability-in-the-large. This method may be applied to a broad class of linear, nonlinear and time-varying interconnected systems as indicated in the following three steps.

First, the Lyapunov functions and comparison equations are found for the individual subsystems. Second, the comparison equations are interconnected, following the interconnections in the original system, into a system of comparison equations. This system is linear, with constant coefficients and of order equal to the number of subsystems in the original interconnected system. Finally, the stability of the null solution of this auxiliary system is examined. If it is asymptotically stable, then the null solution of the interconnected system is asympotically stable-in-the-large. An example illustrates the use of interconnection information in determining the sufficient conditions for asymptotic stability-in-the-large for a ninth-order interconnected system.

1. Introduction. It has long been realized that in the analysis of complex, high dimensional systems, a straightforward application of general mathematical tools would often become bogged down in the welter of detail. For this reason the full promise of many potentially valuable mathematical tools has never been realized. In attempts to overcome or to circumvent these difficulties, it has frequently been found that the more efficient procedures are those which depend strongly on special properties or structural features present in the particular system under study. However, such approaches have generally ignored a basic structural feature of complex systems, namely, the interconnection structure. Many complex, high order systems are actually composite systems—interconnections of a large number of relatively simple subsystems into a complex whole—with special structural features which might be used to advantage in analysis.

The value of interconnection structure has been recognized by Kron [1] who has developed procedures for determining the behavior of certain types of high order composite systems through an analysis of each of the individual, low order subsystems and their interconnections. This sub-

* Received by the editors February 12, 1965, and in revised form June 18, 1965.

† Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota.
This research was supported in part by the National Science Foundation under Grant No. GP-540.

stitution of many low order (and hopefully easier) problems for one high order problem can be of value in cases where difficulties in analysis (computer time, etc.) depend on problem order to a power higher than unity. Although Kron's work has never become popular, the basic concept of studying a complex system through an analysis of its components (subsystems) and their interconnections is quite appealing since the implied piece-by-piece analysis through the subsystems might avoid the formidable difficulties that are often encountered in a straightforward attack.

This paper describes an application of this "composite system approach" to stability problems or, more specifically, to the problem of determining criteria for asymptotic stability-in-the-large (sometimes termed global asymptotic stability) of interconnected systems containing linear, nonlinear, and time-varying elements. The innovation lies in the use of Lyapunov's second method *in conjunction with interconnection information* to simplify the stability analysis of composite systems. Lyapunov functions for the individual subsystems are found and then "interconnected", following the existing interconnections between the subsystems, to obtain a vector Lyapunov function applicable to the composite system. By using this approach the manifold difficulties commonly encountered in a direct application of Lyapunov's second method to the original composite system can be largely circumvented. In §7 the suggested procedure is applied in determining stability criteria for a ninth-order, nonlinear, time-varying composite system.

2. Notation and definitions. The vector notation used is similar to that employed by Hahn [2] or Cesari [3]. Let E^n denote the n-dimensional Euclidean space of n vectors, $x = \text{col}\,[x_1, x_2, \cdots, x_n]$, where the x_i are real numbers or real valued functions on the interval $T = [0, \infty)$ of the real line. The transpose of x is denoted by x' and the inner product is defined as

$$(x, y) = x'y = \sum_{i=1}^{n} x_i y_i .$$

The norm of a vector in E^n is the Euclidean norm $\| x \| = (x, x)^{1/2}$ and if P is an $m \times n$ matrix of real elements, then

$$\| P \| = \min \{\alpha \mid \alpha \| x \| \geqq \| Px \| \quad \text{for all} \quad x \in E^n\}.$$

A useful metric on E^n is $d(x, y) = \| x - y \|$ and when limits and continuity are mentioned the implied topology is taken with respect to this metric. For any subset A of E^n the distance from x to A is

$$d(x, A) = \inf_{y \in A} d(x, y),$$

and for any $\epsilon > 0$,

$$S_\epsilon(A) = \{x \mid d(x, A) < \epsilon\}.$$

If $f(x)$ is defined on $R \subset E^n$ and $B \subset R$, then $f(B) = \{f(x) | x \in B\}$. For any clearly defined $t_0 \in T$, let T_0 denote the set $[t_0, \infty)$.

The differential equation $\dot{x} = f(x, t)$ is normally an nth order vector differential equation with $x(t)$ and $f(x, t)$ denoting n-vectors defined on T and $E^n \times T$, respectively. It is generally assumed that all differential equations satisfy conditions sufficient to guarantee the existence, uniqueness, and continuity of all solutions in t, x_0, and t_0 (continuity from the inside is implied at the boundary of any closed region).

A continuous, real valued function $v(x)$ on E^n with continuous first partial derivatives is said to be positive definite [positive semidefinite] if $v(0) = 0$ and $v(x) > 0 [v(x) \geqq 0]$ for all $x \neq 0$. A continuous, real valued function $v(x, t)$ on $E^n \times T$ with continuous first partial derivatives is said to be positive definite [positive semidefinite] if $v(0, T) = 0$ and there is a positive definite [positive semidefinite] function $w(x)$ such that $v(x, t) \geqq w(x)$ on $E^n \times T$. A continuous, real valued function $v(x)$ or $v(x, t)$ with continuous first partial derivatives is said to be negative definite [negative semidefinite] if $-v(x)$ or $-v(x, t)$ is positive definite [positive semidefinite]. *If* $v(x)$ is a definite or semidefinite scalar function on E^n then ∇v is the vector $\operatorname{col}\left[\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}, \cdots, \frac{\partial v}{\partial x_n}\right]$.

3. Stability and Lyapunov's second method. Following common practice, all further discussion of stability will refer to the stability of the null solution, $x = 0$, of the vector ordinary differential equation (equation of perturbed motion [4]),

$$\dot{x} = f(x, t), \qquad x(t_0) = x_0, \tag{3.1}$$

where x is an n-vector and $f(0, T) = 0$. Most questions concerning stability of equilibrium points or stability of a given motion can be formulated in the manner of (3.1) by a simple coordinate transformation [2].

The following definitions are standard [2].

DEFINITION 3.1. The null solution, $x = 0$, of (3.1) is *stable*[1] if for every $\epsilon > 0$ there is a $\delta > 0$ such that $x(T_0 ; S_\delta(0), t_0) \subset S_\epsilon(0)$.

DEFINITION 3.2. The null solution, $x = 0$, of (3.1) is *asymptotically stable* if it is stable and there is some $\gamma > 0$ such that

$$\lim_{t \to \infty} \| x(t; x_0, t_0) \| = 0$$

for every $x_0 \in S_\gamma(0)$.

DEFINITION 3.3. The null solution, $x = 0$, of (3.1) is *asymptotically stable-*

[1] In Definitions 3.1, 3.2, and 3.3 the initial time t_0 is not clearly defined. However it is easily shown [2] that if any of these properties hold for one initial time then they must also hold for all future (greater) initial times.

in-the-large (abbreviated ASL) if it is stable and for every $x_0 \in E^n$,

$$\lim_{t\to\infty} \| x(t; x_0, t_0)\| = 0.$$

Lyapunov's second (or direct) method, abbreviated LSM, is simply a method of determining certain qualitative features, such as stability, of the solutions of (3.1) without actually solving the equations. In recent years LSM has received considerable attention in both the control engineering and mathematics literature and a wide variety of results are available [2], [5], [6], [7]. In addition, there is a generalization of LSM following Conti [8], [9] who views the positive definite function v (here called a Lyapunov function[2]) as the dependent variable in a first-order auxiliary equation (Brauer [9] uses the more descriptive term *comparison equation*). For example, let $v(x, t)$ be a positive definite function on $E^n \times T$ and $\dot{v}$ be the total derivative of v with respect to (3.1). If there is a function $\omega(v, t)$, with $\omega(0, T) = 0$, such that along the solutions of (3.1),

$$\dot{v} \leqq \omega(v, t), \tag{3.2}$$

then under quite weak conditions it can be shown that the null solution of (3.1) has the same stability and boundedness properties as the null solution of the first-order (scalar) auxiliary differential equation

$$\dot{r} = \omega(r, t). \tag{3.3}$$

This reduces the problem of determining the stability of an nth order system to that of determining the stability of a first-order system—a very significant simplification.

While the auxiliary equation is an important generalization of the earlier formulations of LSM [the earlier versions of LSM are the special cases where $\omega(v, t)$ is required to be a constant] the application of this new formulation requires the choice of two functions $v(x, t)$ and $\omega(v, t)$. For high order systems this choice will generally be very difficult and a major limitation on the practical value of LSM. (As in the earlier theory, the important theorems generally give only sufficient conditions for stability so that negative results yield little useful information.) The remaining sections of this paper describe a procedure for simplifying the choice of the two functions $v(x, t)$ and $\omega(v, t)$ by applying the composite system approach suggested above. Through an effective use of interconnection information this approach circumvents some of the current obstacles to the application of LSM and comparison equations in high order systems.

[2] This terminology is not uniform. Many authors call a positive definite function a Lyapunov function only if its derivative meets certain requirements. In this paper the terms *Lyapunov function* and *positive definite function* are interchangeable.

4. Composite systems, transfer systems, and models. As mentioned in the introduction, a composite system is a complex system composed of interconnections of simpler subsystems. The basic building blocks of the composite systems considered here will be called transfer systems.

DEFINITION 4.1. A *transfer system* is an input-output device whose terminal variables may be characterized by relations of the form

$$\dot{x} = f(x, t, u(t)), \tag{4.1}$$

$$y = h(x(t), t), \tag{4.2}$$

where $x(t)$ is an n-dimensional state vector, $u(t)$ is a p-dimensional input vector, and $y(t)$ is a q-dimensional output vector.

The terminal relations (4.1) and (4.2) characterizing the transfer system are called the *transfer system model.* A composite system can now be defined as an interconnection of transfer systems.

DEFINITION 4.2. Consider a set of m transfer systems, S_i, $i = 1, \cdots, m$. A *composite system* is an interconnection of these transfer systems so that for the ith transfer system the (vector) input u_i is given as

$$u_i = \sum_{j=1}^{m} B_{ij} y_j + G_i u, \qquad i = 1, \cdots, m,$$

where y_j is the (vector) output of the jth transfer system, u is an external (vector) input to the composite system and B_{ij}, G_i are constant matrices. (Note that only time-invariant linear interconnections are allowed.)

The partitioned matrix

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ B_{m1} & B_{m2} & \cdots & B_{mm} \end{bmatrix}, \tag{4.3}$$

where the submatrices B_{ij}, $i, j = 1, \cdots, m$, are the same as those used in Definition 4.2, will be termed the composite system interconnection matrix or simply the interconnection matrix since it indicates the interconnection structure of the composite system.

It should be pointed out that this sort of interconnection implies the usual system theory assumption [10] that *the individual transfer system models are not affected by the various types of interconnections; that is, there is no "loading" effect of one system on another.* While this in itself greatly simplifies the mechanisms through which instability of the composite system can occur it is a reasonable assumption for a large number of interconnected physical systems. The external input u is included to emphasize the fact that the composite system itself might be a transfer system in a larger composite.

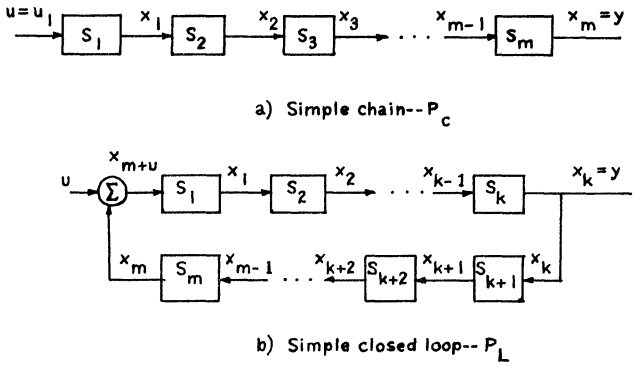

FIG. 4.1. *Two simple composite systems*

In this paper the individual transfer system models may be linear, nonlinear and/or time-varying but will be assumed to have the special form

$$\dot{x}_i = f_i(x_i, t) + D_i u_i, \tag{4.4}$$

$$y_i = H_i x_i, \tag{4.5}$$

for $i = 1, \cdots, m$, where D_i and H_i are matrices (the subscript i denotes the ith transfer system in a given composite system). When these transfer systems are interconnected,

$$D_i u_i = \sum_j D_i B_{ij} H_j x_j + D_i G_i u,$$

and the composite system model takes on the form

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, t) + C_{11}x_1 + C_{12}x_2 + C_{13}x_3 + \cdots + C_{1m}x_m + K_1 u,\\
\dot{x}_2 &= f_2(x_2, t) + C_{21}x_1 + C_{22}x_2 + C_{23}x_3 + \cdots + C_{2m}x_m + K_2 u,\\
&\quad\cdot\quad\cdot\quad\cdot\\
\dot{x}_m &= f_m(x_m, t) + C_{m1}x_1 + C_{m2}x_2 + C_{m3}x_3 + \cdots + C_{mm}x_m + K_m u,\\
y &= h(x_1, x_2, \cdots, x_m, t),
\end{aligned} \tag{4.6}$$

where

$$C_{ij} = D_i B_{ij} H_j, \qquad K_i = D_i G_i,$$

and y is the composite system output. If the individual transfer system models are n_ith order, the composite system model will be nth order where $n = \sum_{i=1}^{m} n_i$. By defining a composite system state vector

$$x = \text{col}\,[x_{11}, \cdots, x_{1n_1}, x_{21}, \cdots, x_{2n_2}, \cdots, x_{m1}, \cdots, x_{mn_m}],$$

the composite system model can be further refined to the form

$$\dot{x} = f(x, t) + Cx + Ku, \tag{4.7}$$

$$y = h(x, t), \tag{4.8}$$

where x is the composite system state vector, f is a column vector of the f_i's, C is the partitioned matrix of elements C_{ij}, and K is the partitioned (column) matrix of elements K_i. Since $C_{ij} = D_i B_{ij} H_j$, the matrix C also indicates the interconnection structure of the composite system.

As examples, consider the two simple composite systems shown in Fig. 4.1. Here it is assumed, for simplicity, that $H = I$ (the output of each transfer system is its state vector x). For the simple chain P_c, the matrix C has the form

$$C_c = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ C_{21} & 0 & 0 & \cdots & 0 & 0 \\ 0 & C_{32} & 0 & \cdots & 0 & 0 \\ 0 & 0 & C_{43} & \cdots & 0 & 0 \\ & \cdot & & \cdot & \cdot & \\ 0 & 0 & 0 & \cdots & C_{m,m-1} & 0 \end{bmatrix},$$

and for the simple closed loop P_L, the matrix C has the form

$$C_L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & C_{1m} \\ C_{21} & 0 & 0 & \cdots & 0 & 0 \\ 0 & C_{32} & 0 & \cdots & 0 & 0 \\ 0 & 0 & C_{43} & \cdots & 0 & 0 \\ & \cdot & & \cdot & \cdot & \\ 0 & 0 & 0 & \cdots & C_{m,m-1} & 0 \end{bmatrix}.$$

More general interconnections will result in more general patterns of nonzero elements (submatrices) in the matrix C.

5. Properties of transfer systems. In keeping with the composite system approach outlined in the introduction, the basic interconnection structure has been developed in §4. The next step is a study of significant properties of the subsystems, the transfer systems. In the following section interconnection structure and transfer system properties will be combined in the development of stability criteria for the composite system.

At this point it is necessary to restrict attention to a special, but very important, class of transfer systems. The members of this class are related by the property of exponential stability-in-the-large.

DEFINITION 5.1. The null solution of (3.1) is said to be *exponentially stable-in-the-large* (ESL) if there are two positive constants α and β such

that

$$\| x(t; x_0, t_0) \| \leqq \beta \| x_0 \| e^{-\alpha(t - t_0)},$$

for all $(x_0, t_0) \in E^n \times T$.

DEFINITION 5.2. A vector function $f(x, t)$ on $E^n \times T$ is said to be in *class B* (denoted $f \in B$) if, for all $(x, t) \in E^n \times T$, $f(x, t)$ is continuous in x and t and has continuous first partial derivatives with respect to $x_1, x_2, \cdots, x_n$ such that $| \partial f_i / \partial x_j | < L$, where L is a constant and $i, j = 1, \cdots, n$.

The following theorem brings out an important feature of equations having a null solution which is ESL.

THEOREM 5.1. *Assume that* $f(x, t) \in B$ *in* (3.1). *Then* (3.1) *is ESL if, and only if, there is a positive definite function* $v(x, t)$ *such that*

(a) $c_1 \| x \|^2 \leqq v(x, t) \leqq c_2 \| x \|^2$,

(b) $\dot{v}(x, t) \leqq -c_3 \| x \|^2$,

(c) $\| \nabla v \| \leqq c_4 \| x \|$.

Proof. See [11, p. 59].

DEFINITION 5.3. A system modeled by (4.4) and (4.5) will be said to be in *class E* if the unforced model [(4.4) with $u(t) = 0$] is ESL and $f(x, t) \in B$.

The practical importance of class E is indicated in the following theorem.

THEOREM 5.2. *The following ordinary differential equations are ESL:*

(a) *the linear constant coefficient equation* $\dot{x} = Ax$, *where* A *is stable* (*i.e., all of its eigenvalues have negative real parts*);

(b) *the linear time-varying equation* $\dot{x} = A(t)x$, *where* $A(t)$ *is continuous and bounded and* $\dot{x} = A(t)x + u(t)$ *has the property that bounded* u *implies bounded* x;

(c) *the equation* $\dot{x} = f(x, t)$, *where* $f(x, t) \in B$, $f(0, T) = 0$ *and* $x'f(x, t) \leqq -c_3 \| x \|^2 < 0$ *for all* $(x, t) \in E^n \times T$ *with* $x \neq 0$.

Proof. (a) is a well known result [2]. A proof of (b) is given in [12, p. 518]. Part (c) is easily verified using the Lyapunov function,

$$v(x) = \tfrac{1}{2} \| x \|^2,$$

so that

$$\dot{v}(x, t) = x'f(x, t) \leqq -c_3 \| x \|^2.$$

The desired result then follows immediately from Theorem 5.1.

The list given in Theorem 5.2 is far from exhaustive. A more detailed description of class E and a consideration of other interesting classes of transfer systems will be the subject of future study.

One additional property of systems in class E should be noted here. It can be shown [13] that for such systems it is possible to define a gain η as a ratio of a bound on the size (norm) of the state vector x to a bound

on the size (norm) of the input u. Roughly speaking (neglecting transients),

$$\eta = \frac{\sup_{t \geqq t_0} \| x \|}{\sup_{t \geqq t_0} \| u \|}.$$

An upper bound on η can be estimated with the aid of a simple extension of some of the concepts of LSM. A complete discussion of this gain and its estimation may be found in [13]. However, the following theorem (also proved in [13]) is included here because it leads to an interesting interpretation of results obtained in §6.

THEOREM 5.3. *A transfer system in class E has a gain,*

$$\eta \leqq \left(\frac{c_4}{c_3}\right)\left(\frac{c_2}{c_1}\right)^{1/2} \| D \|,$$

where c_1, c_2, c_3, *and* c_4 *are the constants noted in Theorem* 5.1.

Note that the accuracy of the gain estimate depends on the constants c_1, c_2, c_3, c_4 and thus on the particular Lyapunov function used in making the estimate.

6. Stability of composite systems. The concepts developed in earlier sections can now be applied to the central problem, stability of the composite system. However, before attacking the most complex situation with general interconnections the salient features of the composite system approach will be illustrated by first considering the two simple systems shown in Fig. 4.1. While both of these systems are shown with an input, only the stability of unforced composite systems ($u = 0$) will be considered.

The following lemmas will be of major importance in this section.

LEMMA 6.1. *Let* $x(t; x_0, t_0)$ *be a solution of the differential inequality*[3]

$$\dot{x} \leqq Ax, \tag{6.1}$$

with $x(t_0; x_0, t_0) = x_0$, *and let* $y(t; y_0, t_0)$ *be a solution of the differential equation*

$$\dot{y} = Ay. \tag{6.2}$$

If all the elements a_{ij}, $i, j = 1, \cdots, n$, *of* A *are nonnegative, and* $x_0 = y_0$, *then* $x(t; x_0, t_0) \leqq y(t; y_0, t_0)$ *for all* $t \in T_0$.

Proof. See [14].

LEMMA 6.2. *Let* B *be a matrix with negative diagonal elements and non-*

[3] Throughout this section, the notation $x \leqq y$, where x and y are n-vectors, means that $x_i \leqq y_i$ for $i = 1, \cdots, n$.

negative off-diagonal elements. If $x(t; x_0, t_0)$ *and* $y(t; y_0, t_0)$ *are solutions of*

$$\dot{x} \leqq Bx, \tag{6.3}$$

$$\dot{y} = By, \tag{6.4}$$

and $x_0 = y_0$, *then* $x(t; x_0, t_0) \leqq y(t; y_0, t_0)$ *for all* $t \in T_0$.

Proof.[4] Let $-d$ be the smallest of the diagonal elements of B. Application of the transformations $v = e^{+dt}x$ and $w = e^{+dt}y$ changes (6.3) and (6.4) to

$$\dot{v} \leqq (B + dI)v,$$

$$\dot{w} = (B + dI)w,$$

and $B + dI$ has all nonnegative elements. Then by Lemma 6.1, since $v_0 = w_0$, it follows that $v(t; v_0, t_0) \leqq w(t; w_0, t_0)$ and thus $x(t; x_0, t_0) \leqq y(t; y_0, t_0)$ for all $t \in T_0$.

LEMMA 6.3. *The null solution of the system of linear differential equations*

$$\begin{aligned} \dot{x}_1 &= -a_1x_1 + b_1x_n, \\ \dot{x}_2 &= -a_2x_2 + b_2x_1, \\ &\cdots \\ \dot{x}_n &= -a_nx_n + b_nx_{n-1}, \end{aligned}$$

with a_i *and* b_i *real,* $a_i > 0$ *and* $b_i \geqq 0$, $i = 1, \cdots, n$, *is ASL if and only if*

$$\prod_{i=1}^{n} \frac{b_i}{a_i} < 1.$$

A proof of this lemma is given in the Appendix.

LEMMA 6.4. *If* $a > 0$ *and* $b \geqq 0$, *then for all* $z \in T$,

$$-az^2 + bz \leqq -\frac{a}{2}z^2 + \frac{b^2}{2a}.$$

Proof.

$$-az^2 + bz \leqq -\frac{a}{2}z^2 + \frac{b^2}{2a}$$

if and only if

$$-\frac{a}{2}z^2 + \frac{b^2}{2a} - \frac{a}{2}z^2 + bz - \frac{b^2}{2a} \leqq -\frac{a}{2}z^2 + \frac{b^2}{2a}$$

if and only if

$$\left(-\frac{a}{2}z^2 + \frac{b^2}{2a}\right) - \frac{1}{2a}(az - b)^2 \leqq -\frac{a}{2}z^2 + \frac{b^2}{2a}.$$

[4] The author is indebted to J. K. Hale for this proof.

Now consider the simple chain P_c of Fig. 4.1(a). It is intuitively obvious that such an interconnection of individually stable systems will be stable because of the "weak" interconnections involved (no loading assumed, see §4). The following theorem gives a simple proof of this fact.

THEOREM 6.1. *Consider the simple chain P_c with $u = u_1 = 0$ and $u_i = x_{i-1}$ for $i = 2, \cdots, m$. If the individual transfer systems S_i, $i = 1, \cdots, m$, are in class E, then the null solution of the composite system is ASL.*

Proof. Since the individual transfer systems are in class E, the ith transfer system is modeled by (4.4) and (4.5). Moreover, $f_i(x_i, t) \in B$, and when $u_i = 0$, $\dot{x}_i = f_i(x_i, t)$ is ESL. Thus there is a positive definite function $v(x, t)$ satisfying (a), (b), and (c) of Theorem 5.1. Now the total derivative of this $v(x, t)$ with respect to (4.4) is

$$\dot{v}_i = v_{i1}(x_i, t) + \nabla v_i' D_i u_i \leqq -c_{i3} \| x_i \|^2 + c_{i4} \| x_i \| \cdot \| D_i \| \cdot \| u_i(t) \|,$$

where $\dot{v}_{i1}$ is the total derivative of $v_i(x, t)$ with respect to $\dot{x}_i = f_i(x_i, t)$. By Lemma 6.4 then,

$$\dot{v}_i \leqq -\frac{c_{i3}}{2} \| x_i \|^2 + \frac{c_{i4}^2 \| D_i \|^2}{2c_{i3}} \| u_i \|^2,$$

or

$$\dot{v}_i \leqq -\alpha_i v_i + \gamma_i \| u_i \|^2,$$

where

$$\alpha_i = \frac{c_{i3}}{2c_{i2}} > 0, \qquad \gamma_i = \frac{c_{i4}^2 \| D_i \|^2}{2c_{i3}} > 0.$$

When the interconnections are made, $u_1 = 0$ and

$$\| u_i \|^2 = \| x_{i-1} \|^2 \leqq \frac{1}{c_{i-1,1}} v_{i-1} \quad \text{for} \quad i = 2, 3, \cdots, m.$$

With $\beta_i = \gamma_i / c_{i-1,1}$, $i = 2, \cdots, m$, the resulting system of differential inequalities becomes

$$\begin{aligned} \dot{v}_1 &\leqq -\alpha_1 v_1, \\ \dot{v}_2 &\leqq -\alpha_2 v_2 + \beta_2 v_1, \\ &\cdots \\ \dot{v}_m &\leqq -\alpha_m v_m + \beta_m v_{m-1}, \end{aligned} \tag{6.5}$$

where $v_i \geqq 0$. Since $\beta_1 = 0$, the trivial solution of the system (6.5) with inequalities replaced by equalities (a system of auxiliary equations) is ASL (Lemma 6.3). Then by Lemma 6.2, the system (6.5) is also ASL, implying that each v_i goes to 0 as t goes to ∞. Since $\| x_i \|^2 \leqq v_i / c_{i1}$, the equilibrium solution $x = 0$ of the composite system is therefore ASL.

A more impressive result is obtained for the simple closed loop P_L. In this case the stability of the individual systems is not sufficient. The additional requirement is that the loop gain estimated from the chosen Lyapunov functions be less than unity—again an intuitively plausible requirement.

THEOREM 6.2. *Consider the simple closed loop P_L with $u = 0$ (no external input), $u_1 = x_m$, and $u_i = x_{i-1}$ for $i = 2, \cdots, m$. If the individual transfer systems S_i are in class E with gain estimates η_i, $i = 1, \cdots, m$, then the null solution of the composite system will be ASL if $\prod_{i=1}^{m} \eta_i < 1$.*

Proof. As in the proof of Theorem 6.1, there is again a system of differential inequalities

$$\begin{aligned} \dot{v}_1 &\leqq -\alpha_1 v_1 + \beta_1 v_m , \\ \dot{v}_2 &\leqq -\alpha_2 v_2 + \beta_2 v_1 , \\ &\cdots \\ \dot{v}_m &\leqq -\alpha_m v_m + \beta_m v_{m-1} , \end{aligned} \tag{6.6}$$

where

$$\alpha_i = \frac{c_{i3}}{2c_{i2}} > 0, \qquad i = 1, \cdots, m,$$

$$\beta_1 = \frac{c_{14}^2 \| D_1 \|^2}{2c_{13}\, c_{m1}} > 0,$$

$$\beta_i = \frac{c_{i4}^2 \| D_i \|^2}{2c_{i3}\, c_{i-1,1}} > 0, \qquad i = 2, \cdots, m,$$

where the $v_i \geqq 0$, for $i = 1, \cdots, m$. The trivial solution of the composite system will be ASL if the system of auxiliary equations [(6.6) with inequalities replaced by equalities] is ASL. But, by Lemma 6.3, this will occur if

$$\prod_{i=1}^{m} \frac{\beta_i}{\alpha_i} < 1.$$

Now

$$\frac{\beta_i}{\alpha_i} = \frac{c_{i2}\, c_{i4}^2 \| D_i \|^2}{c_{i3}^2\, c_{i-1,1}} = \frac{c_{i1}}{c_{i-1,1}}\, \eta_i^2,$$

(where $c_{i-1,1} = c_{m1}$ when $i = 1$) and so the requirement that $\prod_{i=1}^{m} \eta_i^2 < 1$, or $\prod_{i=1}^{m} \beta_i/\alpha_i < 1$, thereby gives ASL.

The main theorem now gives sufficient conditions for ASL of the null solution of a composite system with *arbitrary interconnections*, that is, a system where C has an arbitrary number of nonzero elements. It will, how-

ever, be assumed that $B_{ii} = C_{ii} = 0$ since a nonzero B_{ii} represents an internal connection within the transfer system that can be included in the transfer system model itself.

THEOREM 6.3. *Let P be a general composite system made up of m transfer systems, S_i, of order n_i, $i = 1, \cdots, m$, and modeled by (4.7) with $C_{ii} = 0$. Assume that each transfer system S_i is in class E and has a Lyapunov function $v_i(x_i, t)$ satisfying the bounds listed in Theorem 5.1 with coefficients c_{i1}, c_{i2}, c_{i3}, c_{i4} (all coefficients are positive). Consider the mth order linear system of auxiliary equations,*

$$\dot{r} = Ar, \tag{6.7}$$

where A is an $m \times m$ matrix of elements,

$$a_{ij} = \begin{cases} \dfrac{c_{i3}}{2c_{i2}} & \text{for } i = j, \\ \dfrac{c_{i4}^2 \sum_{j=1}^{m} \| C_{ij} \|^2}{c_{i3}\, c_{j1}} & \text{for } i \neq j, \end{cases}$$

with C_{ij} an element (submatrix) of C. The null solution, $x = 0$, of the unforced composite system model [(4.7) with $u = 0$] will be ASL if the null solution, $r = 0$, of the mth order linear auxiliary system (6.7) is ASL.

Proof. The ith transfer system can be modeled by (4.4) and (4.5) and due to the interconnections,

$$u_i = \sum_{j=1}^{m} B_{ij} y_j . \tag{6.8}$$

Since this ith transfer system is in class E, there is a Lyapunov function $v_i(x_i, t)$ such that, with respect to (4.4),

$$\dot{v}_i \leqq -c_{i3} \| x_i \|^2 + c_{i4} \| x_i \| \cdot \| D_i u_i \|. \tag{6.9}$$

Now a substitution of (4.5) and (6.8) into (6.9) and application of Lemma 6.4 gives

$$\dot{v}_i \leqq -\frac{1}{2} c_{i3} \| x_i \|^2 + \frac{c_{i4}^2}{2c_{i3}} \sum_{j=1}^{m} \| D_i B_{ij} H_j \|^2 \sum_{\substack{j=1 \\ j \neq i}}^{m} \| x_j \|^2.$$

A reintroduction of the inequalities

$$c_{i1} \| x_i \|^2 \leqq v_i(x_i, t) \leqq c_{i2} \| x_i \|^2, \qquad i = 1, \cdots, m,$$

and a use of the relation $C_{ij} = D_i B_{ij} H_j$ yields

$$\dot{v}_i \leqq -\frac{c_{i3}}{2c_{i2}} v_i + \frac{c_{i4}^2}{2c_{i3}} \sum_{j=1}^{m} \| C_{ij} \|^2 \sum_{\substack{j=1 \\ j \neq i}}^{m} \frac{v_j}{c_{j1}} .$$

The resulting system of inequalities is

$$\dot{v}_1 \leqq -\frac{c_{13}}{2c_{12}} v_1 + \left(\frac{c_{14}^2}{2c_{13}} \sum_{j=1}^{m} \| C_{ij} \|^2\right)\left(\sum_{j=1}^{m} \frac{v_j}{c_{j1}}\right),$$

$$\cdots$$

$$\dot{v}_m \leqq -\frac{c_{m3}}{2c_{m2}} v_m + \left(\frac{c_{m4}^2}{2c_{m3}} \sum_{j=1}^{m} \| C_{mj} \|^2\right)\left(\sum_{j=1}^{m-1} \frac{v_j}{c_{j1}}\right),$$

or

$$\dot{v} \leqq Av, \tag{6.10}$$

where $v = \text{col}\,[v_1, v_2, \cdots, v_m]$ and $\leqq$ means that the inequality holds componentwise. Now by Lemma 6.2, $v(t) \leqq r(t)$ if $v(t_0) = r(t_0)$. Thus

$$\| x_i(t; x_0, t_0) \|^2 \leqq \frac{1}{c_{i1}} v(t) \leqq \frac{1}{c_{i1}} r(t), \qquad i = 1, \cdots, m,$$

and so asymptotic stability-in-the-large of the solution $r = 0$ of (6.7) implies asymptotic stability-in-the-large of the solution $x = 0$ of the composite system.

This result does not have the intuitive appeal found in the results of Theorems 6.1 and 6.2. That is, there is nothing like a loop gain to suggest that the resulting stability criterion is reasonable. However, this is not surprising. The same difficulty is encountered even in a linear time-invariant composite system with arbitrary interconnections.

Theorems 6.1 and 6.2 are now corollaries to Theorem 6.3 when the matrix C has the special form C_c or C_L.

7. Example—a ninth-order composite system. Consider the ninth-order composite system shown in Fig. 7.1. The individual transfer systems are assumed to have the following models:

S_1 : linear, constant coefficient, third order,

$$\dot{x}_1 = A_1 x_1 + D_1 u_1(t), \quad \text{where} \quad A_1 = T\begin{bmatrix} -2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -5 \end{bmatrix} T^{-1},$$

$$y_1 = H_1 x_1,$$

and T is any nonsingular 3×3 matrix;

S_2 : linear, variable coefficient, second order,

$$\dot{x}_2 = A_2(t)x_2 + D_2 u_2(t), \quad \text{where} \quad A_2(t) = \begin{bmatrix} 0 & a(t) \\ -1 & -2a(t) \end{bmatrix},$$

$$y_2 = H_2 x_2,$$

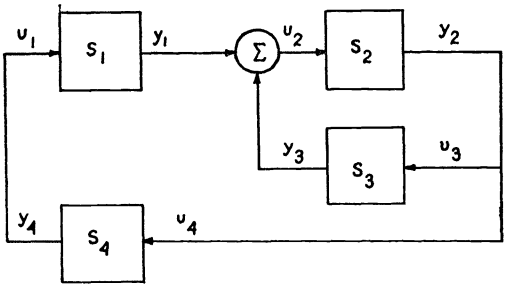

FIG. 7.1. *Composite system for example*

and $a(t)$ is a continuous real valued function such that $a^{-1}(t)$ exists for all $t \in T$ and, in addition,

$$0.5 \leqq a^{-1}(t) \leqq 1 \quad \text{and} \quad 0 \leqq \frac{da^{-1}(t)}{dt} \leqq 1;$$

S_3 : nonlinear, first order,

$$\dot{x}_3 = f_3(x_3) + D_3u_3(t), \text{ where } f_3(x_3) = -x_3 - \tfrac{1}{2}\sin 2x_3 ,$$

$$y_3 = H_3x_3 ;$$

S_4 : nonlinear, third order,

$$\dot{x}_4 = f_4(x_4) + D_4u_4(t), \text{ where } f_4(0) = 0,$$

$$y_4 = H_4x_4 .$$

The problem here is to determine a value of the positive constant k that will insure that the composite system is ASL if

$$x_4'f_4(x_4) \leqq -k\| x_4 \|^2.$$

The interconnections suggested in Fig. 7.1 indicate that

$$u_1 = y_4 ,$$

$$u_2 = y_1 + B_{23}y_3 ,$$

$$u_3 = y_2 ,$$

$$u_4 = y_2 .$$

These interconnections are described by the matrix C whose partitioned elements are $C_{ij} = D_iB_{ij}H_j$. In this case, C has the partitioned form

$$C = \begin{bmatrix} 0 & 0 & 0 & C_{14} \\ C_{21} & 0 & C_{23} & 0 \\ 0 & C_{32} & 0 & 0 \\ 0 & C_{42} & 0 & 0 \end{bmatrix},$$

and the composite system model is

$$\dot{x} = f(x, t) + Cx, \tag{7.1}$$

where $x = \text{col}\,[x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{31}, x_{41}, x_{42}, x_{43}]$ is the composite system state vector. When written out in detail, the composite system model has the form

$$\begin{aligned}
\dot{x}_{11} &= a_{11}x_{11} + a_{12}x_{12} + a_{13}x_{13} + c_{14}^{11}x_{41} + c_{14}^{12}x_{42} + c_{14}^{13}x_{43}, \\
\dot{x}_{12} &= a_{21}x_{11} + a_{22}x_{12} + a_{23}x_{13} + c_{14}^{21}x_{41} + c_{14}^{22}x_{42} + c_{14}^{23}x_{43}, \\
\dot{x}_{13} &= a_{31}x_{11} + a_{32}x_{12} + a_{33}x_{13} + c_{14}^{31}x_{41} + c_{14}^{32}x_{42} + c_{14}^{33}x_{43}, \\
\dot{x}_{21} &= a(t)x_{22} + c_{21}^{11}x_{11} + c_{21}^{12}x_{12} + c_{21}^{13}x_{13} + c_{23}^{11}x_{31}, \\
\dot{x}_{22} &= -x_{21} - 2a(t)x_{22} + c_{21}^{21}x_{11} + c_{21}^{22}x_{12} + c_{21}^{23}x_{13} + c_{23}^{21}x_{31}, \\
\dot{x}_{31} &= f_3(x_{31}) + c_{32}^{11}x_{21} + c_{32}^{12}x_{22}, \\
\dot{x}_{41} &= f_{41}(x_{41}, x_{42}, x_{43}) + c_{42}^{11}x_{21} + c_{42}^{12}x_{22}, \\
\dot{x}_{42} &= f_{42}(x_{41}, x_{42}, x_{43}) + c_{42}^{21}x_{21} + c_{42}^{22}x_{22}, \\
\dot{x}_{43} &= f_{43}(x_{41}, x_{42}, x_{43}) + c_{42}^{31}x_{21} + c_{42}^{32}x_{22},
\end{aligned} \tag{7.2}$$

where c_{ij}^{mn} is the (m, n)th element of C_{ij}, the (i, j)th submatrix in the partition of C.

A straightforward approach to this problem would involve choosing a Lyapunov function involving the 9 state variables, evaluating its derivative with respect to (7.2), and then studying this derivative to determine conditions under which it is negative definite. Clearly, this approach would involve the solution of some very difficult problems. On the other hand, the techniques developed in §6 provide a method for obtaining a solution to this problem rather quickly.

The first step is to find Lyapunov functions of the type described in Theorem 5.1 for each of the transfer systems. This will, of course, be possible if and only if these transfer systems are in class E.

S_1 : Since this is a constant-coefficient linear system, standard techniques [2] can be used to obtain the Lyapunov function $v_1(x_1) = x_1'Px_1$ with $\dot{v}_1(x_1) = -x_1'Qx_1$, where

$$P = (T^{-1})'\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}T^{-1} \quad \text{and} \quad Q = (T^{-1})'\begin{bmatrix} 4 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 30 \end{bmatrix}T^{-1}.$$

For this Lyapunov function the inequalities of Theorem 5.1 are satisfied with $c_1 = 1$, $c_2 = 3$, $c_3 = 4$, and $c_4 = 6$.

S_2 : In this case, choose $v_2(x_2, t) = x_2'P(t)x_2$ with $\dot{v}_2(x_2, t) = -x_2'Q(t)x_2$,

where

$$P(t) = \begin{bmatrix} 2 + a^{-1}(t) & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad Q(t) = \begin{bmatrix} 2 & 0 \\ 0 & 4a(t) \end{bmatrix}.$$

In this case the inequalities of Theorem 5.1 are satisfied with $c_1 = 0.5$, $c_2 = 3.5$, $c_3 = 2$, and $c_4 = 7$.

S_3 : Since this model is first order, $x_3 = x_{31}$. Choose

$$v_3(x_3) = \tfrac{1}{2}x_{31}^2,$$

and then $\dot{v}_3(x_3) = x_{31}f_3(x_{31})$. The inequalities are now satisfied with $c_1 = c_2 = \frac{1}{2}$, $c_3 = 0.56$, and $c_4 = 1$.

S_4 : In this case, choose

$$v_4(x_4) = \tfrac{1}{2}x_4'x_4,$$

and note that

$$x_4'f(x_4) \leqq -k\| x_4 \|^2.$$

Then $c_1 = c_2 = \frac{1}{2}$, $c_3 = k$, and $c_4 = 1$.

With Lyapunov functions chosen for each of the transfer systems, it is possible to apply Theorem 6.3. The system of auxiliary equations is, in this case, the fourth order system

$$\begin{aligned} \dot{r}_1 &= -\alpha_1 r_1 + \beta_1 r_4, \\ \dot{r}_2 &= -\alpha_2 r_2 + \beta_2 r_1 + \gamma_2 r_3, \\ \dot{r}_3 &= -\alpha_3 r_3 + \beta_3 r_2, \\ \dot{r}_4 &= -\alpha_4 r_4 + \beta_4 r_2. \end{aligned} \tag{7.3}$$

In (7.3)

$$\alpha_1 = \frac{2}{3}, \qquad \alpha_2 = \frac{1}{3.5}, \qquad \alpha_3 = 0.56, \qquad \alpha_4 = k,$$

and

$$\beta_1 = 18 \| C_{14} \|,$$

$$\beta_2 = \frac{49}{2}(\| C_{21} \| + \| C_{23} \|),$$

$$\beta_3 = \frac{2}{0.56} \| C_{32} \|,$$

$$\beta_4 = \frac{2}{k} \| C_{42} \|,$$

$$\gamma_2 = 49(\| C_{21} \| + \| C_{23} \|),$$

where C_{ij} are submatrices of C. The numerical values in the above constants are obtained from the bounds on the Lyapunov functions chosen for the individual transfer systems. (See Theorem 6.3.) According to Theorem 6.3, the null solution of the composite system (Fig. 7.1) will be ASL if the null solution of the system of auxiliary equations is ASL. The latter will occur if and only if the roots of the characteristic equation,

$$[-\alpha_4(k) - \lambda][-\alpha_1 - \lambda][(-\alpha_1 - \lambda)(-\alpha_3 - \lambda) - \gamma_2\beta_3] + \beta_4(k)[(-\alpha_3 - \lambda)\beta_1\beta_2] = 0,$$

all have negative real parts. [Here the dependence of α_4 and β_4 on k is indicated as $\alpha_4(k)$ and $\beta_4(k)$.] Thus the original problem has been reduced to a much simpler problem that can be solved by root locus or numerical techniques.

8. Conclusions. In the study of complex physical systems frequently it is found that the difficulties encountered are strongly order dependent. Stability study through LSM is a typical example of a technique that is theoretically applicable to problems of arbitrary order but actually encounters formidable practical limitations that grow rapidly with the order of the system being treated. Thus the application of LSM to low order (third, fourth and sometimes fifth) problems has received a considerable amount of attention while high order problems are still inaccessible.

This paper describes an attempt to circumvent these order dependent limitations of LSM by noting that many high order systems are actually, or effectively, an interconnection of lower order systems. The recognition of this interconnection structure has led to the definition of a composite system as an interconnection of simpler subsystems, termed transfer systems, and a study of the properties of these transfer systems with the aid of LSM and the associated concept of the auxiliary equation. For a composite system made up of transfer systems belonging to a reasonably broad class, the auxiliary equations for the individual transfer systems can be interconnected, following the interconnections existing in the given composite system, to obtain a system of auxiliary equations whose solutions have the same stability properties as those of the given composite system. Moreover, in the construction of this system of auxiliary equations it has been necessary to apply LSM only to the lower order transfer systems. The difficult problem of constructing a single Lyapunov function for the high order composite system has actually been avoided.

As might be suspected, this simplification of the original problem is not obtained without some sacrifice. The main disadvantage of this approach, a fault common to most attempts to obtain general sufficient conditions for stability, is that it may sometimes be overly restrictive (overly suffi-

cient) due to a failure to make the best use of available information about the fine structure of the system under analysis. This fine structure is "washed out" at points where matrix norms or absolute values are used. Hopefully, this is compensated by the fact that the introduction of transfer systems and interconnection information into stability analysis has increased the class of problems to which LSM has practical application.

Appendix. The following proof of Lemma 6.3 uses the root locus concept described in [10].

Proof of Lemma 6.3. The characteristic equation for this system is

$$\prod_{i=1}^{n} (\lambda + a_i) - \prod_{i=1}^{n} b_i = 0.$$

Now replace b_1 with μb_1 and consider the root locus obtained for $\mu \geqq 0$. If $\mu = 0$ there are n negative real roots $\lambda = -a_i$, $i = 1, \cdots, n$. As μ increases the most positive root moves to the right along the real axis reaching the origin when

$$\prod_{i=1}^{n} a_i - \mu \prod_{i=1}^{n} b_i = 0. \tag{A-1}$$

It is now only necessary to ascertain that no root has crossed the imaginary axis for a smaller value of μ. However, this cannot occur since at all points λ *on the root locus* it is easily shown that

$$\mu = \prod_{i=1}^{n} \frac{\lambda + a_i}{b_i}.$$

Thus the point where the locus crosses the imaginary axis with the smallest value of μ is at the origin and the root moving along the real axis must be the first to cross. Now if $\mu = 1$ the necessary and sufficient condition (A-1) becomes

$$\prod_{i=1}^{n} \frac{b_i}{a_i} < 1,$$

completing the proof.

Acknowledgment. The author would like to thank K. B. Irani, A. W. Naylor, and B. F. Barton of the Cooley Electronics Laboratory, Department of Electrical Engineering, The University of Michigan, for their encouragement and support throughout the course of this research.

REFERENCES

[1] G. KRON, *Diakoptics*, Macdonald, London, 1963.

[2] W. HAHN, *Theory and Application of Lyapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[3] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Academic Press, New York, 1963.
[4] A. M. Lyapunov, *Problème général de la stabilité du mouvement*, Annals of Mathematics Studies, vol. 17, Princeton University Press, Princeton, 1947.
[5] R. E. Kalman and J. E. Bertram, *Control system analysis and synthesis via the second method of Lyapunov, I. Continuous time systems*, Trans. ASME Ser. D. J. Basic Engrg., 82D (1960), pp. 371–393.
[6] A. I. Lur'e, *Some Nonlinear Problems in the Theory of Automatic Control*, Her Majesty's Stationery Office, London, 1957.
[7] T. Yoshizawa, *Lyapunov's functions and boundedness of solutions*, Funkcial. Ekvac., 2 (1959), pp. 95–142.
[8] R. Conti, *Sulla prolingabilità della soluzioni di un sistema di equazioni differenziali ordinarie*, Boll. Un. Mat. Ital., 11 (1956), pp. 510–514.
[9] F. Brauer, *Global behavior of solutions of ordinary differential equations*, J. Math. Anal. Appl., 2 (1961), pp. 145–158.
[10] J. J. D'Azzo and C. H. Houpis, *Feedback Control System Analysis and Synthesis*, McGraw-Hill, New York, 1960.
[11] N. N. Krasovskii, *Stability of Motion*, Stanford University Press, Stanford, California, 1963.
[12] J. L. Massera and J. J. Schaffer, *Linear differential equations and functional analysis, I.* Ann. of Math., 67 (1958), pp. 517–573.
[13] F. N. Bailey, *Stability of interconnected systems*, Tech. Rpt. 152, Cooley Electronics Laboratory, The University of Michigan, Ann Arbor, 1964.
[14] E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin, 1961.

J. SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

ON THE EXISTENCE OF OPTIMAL STOCHASTIC CONTROLS*

HAROLD J. KUSHNER†

1. Introduction. We will prove several theorems which state, under their prescribed conditions, that if there exists one stochastic control which accomplishes a given task, then there is an optimal stochastic control. Up to Theorem 3, the systems of concern are governed by the stochastic vector differential equations,

$$(1)\quad dx(\omega, t) = f(x(\omega, t), u(\omega, t))\, dt + \sigma(x(\omega, t), u(x(\omega, t), t))\, dz(\omega, t),$$

or

$$(2)\qquad dx(\omega, t) = f(x(\omega, t), u(x(\omega, t), t))\, dt + dz(\omega, t),$$

where $x(\omega, t)$ is an r-dimensional vector with components $x_0(\omega, t), \cdots, x_{r-1}(\omega, t)$; $u(x(\omega, t), t)$ is a vector control; $\sigma(x, u)$ is an $r \times r$ matrix, and $z(\cdot, \cdot)$ is a vector stochastic process. For both forms there is the restriction

$$dx_0(\omega, t) = f_0(x(\omega, t), u(x(\omega, t), t))\, dt.$$

In the form (1), $z(\cdot, \cdot)$ is assumed to be Brownian motion; in the form (2), $z(\cdot, \cdot)$ is a more general process to be described later. In Theorem 3, $u(x(\omega, t), t)$ is replaced by the more general form $u(\omega, t)$. Many stochastic systems may be put into the form of (1) or (2), but this will not be pursued here. The problem will be investigated with two types of tasks, or terminal conditions. The first is that $x(\omega, t)$ satisfies (*with probability one*)

$$(3)\qquad g(x(\omega, T_\omega)) = 0,$$

where $g(\cdot)$ is continuous, and T_ω is a random stopping time (see [6, p. 578]). The second is a terminal condition on the expectation

$$(4)\qquad g(Ex(\omega, T)) = 0,$$

where T is a nonrandom terminal time. (See [1] for examples where the latter case is of importance.) In both cases, the risk to be minimized by the optimal control is the expectation

$$(5)\qquad R(u) = E \int_0^{T_\omega} f_0(x(\omega, t), u(x(\omega, t), t))\, dt = Ex_0(\omega, T_\omega),$$

* Received by the editors June 20, 1964, and in final revised form August 9, 1965.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island. This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research, Office of Aerospace Research, under Contract No. AF 49(638)-1206 and in part by the National Aeronautics and Space Administration under Contract No. NASw-845.

where T_ω is either the first (random) time that (3) is satisfied or, for the second problem, $T_\omega = T$, any time that (4) is satisfied.

For the deterministic problem, which has already received substantial attention [2], [3], [4], the question of existence (assuming that there is one control which accomplishes the desired task) is equivalent to the question of the closure of a set of attainable states $x(t)$, over all possible controls $u(\cdot)$. In the stochastic case, the question also reduces to that of closure of an appropriate set of attainable sample states or of expectations.

2. Existence theorems. Define the norm of any vector v with components v_i as $\| v \| = \sum_i | v_i |$. Let K and K_i be any positive, finite and nonrandom numbers. The abbreviations a.a.ω and w.p.1 are used for "almost all ω" and "with probability one", respectively. Ω is the space of points ω. $\Sigma(t)$ is the minimal σ-field over which $z(\cdot, \tau)$, $0 \leqq \tau \leqq t$, is measurable. Define[1] $\tilde{\Sigma}(T) = \Sigma(T) \times \mathfrak{I}(T)$, where $\mathfrak{I}(T)$ is the Borel field over the interval $[0, T]$. Since $z(\cdot, \cdot)$ is assumed to be measurable in the pair (ω, t) (see (A4)), it is measurable with respect to $\tilde{\Sigma}(T)$, for $t \leqq T$. The measure on the sets of $\Sigma(t)$ is $m(d\omega)$ for all t, and the measure on the sets of $\tilde{\Sigma}(T)$ is $\mu(d\omega \times dt)$ for all T. Let E^r denote the Euclidean r-dimensional vector space.

The following conditions will be used.

(A1) $| f_i(x, u) | \leqq K(1 + \| x \| + \| u \|)$.

(A2) $| f_i(x + \delta x, u + \delta u) - f_i(x, u) | \leqq K(\| \delta x \| + \| \delta u \|)$.

(A3) $E \max_{t \leqq T} \| z(\omega, t) \| < \infty$, $Ez(\omega, t) = 0$.

(A4) $z(\cdot, \cdot)$ is measurable in the pair (ω, t).

(A5) $| \sigma_{ij}(x, u) | \leqq K(1 + \| x \| + \| u \|)$.

(A6) $| \sigma_{ij}(x + \delta x, u + \delta u) - \sigma_{ij}(x, u) | \leqq K(\| \delta x \| + \| \delta u \|)$.

(A7) $g(\cdot)$ is uniformly continuous.

LEMMA 1. *Assume form* (2) *and* (A1) *to* (A4) *and* (A7), *and that almost all sample functions of* $z(\cdot, \cdot)$ *are continuous. Let* K_1 *be a given finite number and let the family of admissible controls be of the form* $u(x, t)$, *where* $u(\cdot, \cdot)$ *satisfies the uniform Lipschitz condition, for fixed* K_1,

$$\begin{aligned} \| u(x + \epsilon, t + \delta) - u(x, t) \| &\leqq K_1(\| \epsilon \| + | \delta |), \\ \| u(0, 0) \| &\leqq K_1 . \end{aligned} \tag{6}$$

Let the initial condition $x(\omega, 0)$ *be a fixed constant. Then there is a unique continuous solution to* (2) *in the* t *interval* $[0, \infty)$, *for a.a.*ω. *Also there is a nonrandom* $K_2(T) < \infty$, *not depending on* $u(\cdot, \cdot)$, *such that, for any* $u(\cdot, \cdot)$

[1] $\tilde{\Sigma}(T)$ is defined as the minimal σ-field over the product sets in $\Sigma(T) \times \mathfrak{I}(T)$.

satisfying (6), *the corresponding* $x(\cdot, \cdot)$ *satisfies*

$$E \max_{t \leqq T < \infty} \| x(\omega, t) \| \leqq K_2(T) < \infty. \tag{7}$$

Assume that, for some control $u(\cdot, \cdot)$ *satisfying* (6),

$$T_\omega = \inf \{t : g(x(\omega, t)) = 0\} \tag{8}$$

*is defined and finite w.p.*1. *Then* T_ω *and* $x(\omega, T_\omega)$ *are random variables, and*

$$g(x(\omega, T_\omega)) = 0 \tag{9}$$

*w.p.*1.

Proof. The proof of the existence, uniqueness, and continuity of $x(\cdot, \cdot)$ w.p.1 and of (7) is essentially a paraphrase of the proof for the vector Brownian motion case given by Ito. (See [5] for the proof in one dimension.) This will not be proved here, but is easy to see since, by virtue of (6), $f(\cdot, \cdot)$ satisfies a uniform Lipschitz condition, and, since $dz(\omega, t)$ in (2) is not multiplied by a function of x, no stochastic integrals are involved. (See, e.g., the remark on p. 286 of [5].) Inequality (7) holds for all $u(\cdot, \cdot)$ satisfying (6), since $K_2(T)$ depends only upon K, K_1, T, and $x(\omega, 0)$.

By the continuity of $x(\omega, \cdot)$ w.p.1 and uniform continuity of $g(\cdot)$, and $T_\omega < \infty$ w.p.1, we have $g(x(\omega, T_\omega)) = 0$ w.p. 1. Define the random variable T_ω' by

$$C_t = \{\omega : T_\omega' \leqq t\} = \bigcap_{n=1}^{\infty} \bigcup_{0 \leqq r \leqq t} \left\{\omega : | g(x(\omega, r))| < \frac{1}{n}\right\},$$

where $\{r\}$ are the rational numbers. Then, if $\omega \in C_t$,

$$\inf_{s \leqq t} | g(x(\omega, s)) | = 0,$$

and, since $g(\cdot)$ is uniformly continuous, $| g(x(\omega, T_\omega')) | = 0$. Fix ω; then if $T_\omega' > t$, there is no $s \leqq t$ such that $| g(x(\omega, s)) | = 0$ (unless ω is in some null set which does not have to depend on t). Thus, $T_\omega = T_\omega'$ w.p.1 and T_ω can be defined (by changing its value on some null set) to be a random variable.

Since $x(\cdot, \omega)$ is continuous w.p.1 and $T_\omega < \infty$ w.p.1 and is a random time, $x(\omega, T_\omega)$ is a random variable (by arguments in [6, pp. 578–579]).

THEOREM 1. *Assume the conditions of Lemma* 1 *and let* $f_0 \geqq 0$. *Restrict the family of admissible controls to those which satisfy* (6), *for fixed* K_1, *and for which the corresponding quantity* (8) *is defined and finite w.p.*1, *and for which* $R(u) < \infty$, *and*

$$ET_\omega \leqq \hat{T} < \infty, \tag{10}$$

where $\hat{T}$ *is a given constant not depending upon the particular* $u(\cdot, \cdot)$. *Then,*

if there is one admissible control, there is an (optimal) admissible control, which absolutely minimizes (5).

Proof. Define $\alpha = \inf R(u)$, where the infimum is over all admissible $u(\cdot, \cdot)$. Thus, there are an infinite sequence of controls $u^n(\cdot, \cdot)$ and corresponding $T_\omega{}^n$ (perhaps the same) such that $R(u^n) = \alpha_n \to \alpha$ monotonically. To prove the theorem, it must be shown that there are an admissible control $\bar{u}(\cdot, \cdot)$ and corresponding T_ω such that $R(\bar{u}) = \alpha$, $ET_\omega \leqq \hat{T}$, and $g(\bar{x}(\omega, T_\omega)) = 0$ w.p.1.

By (6), the family $\{u^n(\cdot, \cdot)\}$ is equicontinuous. Therefore, by Ascoli's theorem, for any compact set D in $E^r \times [0, \infty)$, there is a function $\bar{u}(\cdot, \cdot)$ satisfying (6), and there is a subsequence (indexed by n) such that $u^n(x, t) \to \bar{u}(x, t)$ uniformly in D. By the diagonal process, we can find a further subsequence (indexed by n) and a function $\bar{u}(\cdot, \cdot)$ satisfying (6) such that

$$u^n(x, t) \to \bar{u}(x, t) \tag{11}$$

uniformly on all compact sets in $E^r \times [0, \infty)$. We will show that $\bar{u}(\cdot, \cdot)$ is the desired optimal control.

For any control $u(\cdot, \cdot)$ satisfying (6), (6) and (7) yield (r is the dimension of x)

$$E \max_{t \leqq T < \infty} \| u(x(\omega, t), t) \| \leqq K_1[1 + K_2(T) + T] < \infty. \tag{12}$$

By (A2), and $x^n(\omega, 0) = \bar{x}(\omega, 0)$,

$$\begin{aligned} \delta x^n(\omega, t) &\equiv x^n(\omega, t) - \bar{x}(\omega, t) \\ &= \int_0^t [f(x^n(\omega, s), u^n(x^n(\omega, s), s)) - f(\bar{x}(\omega, s), \bar{u}(\bar{x}(\omega, s), s))]\, ds, \end{aligned} \tag{13}$$

$$\| \delta x^n(\omega, t) \| \leqq \int_0^t Kr[\| \delta x^n(\omega, s) \| + \| u^n(x^n(\omega, s), s) - \bar{u}(\bar{x}(\omega, s), s) \|]\, ds.$$

Substituting

$$\| u^n(x^n, s) - \bar{u}(\bar{x}, s) \| \leqq \| u^n(x^n, s) - u^n(\bar{x}, s) \| + \| u^n(\bar{x}, s) - \bar{u}(\bar{x}, s) \|$$

and

$$\| u^n(x^n, s) - u^n(\bar{x}, s) \| \leqq K_1 \| \delta x^n \|$$

into (13) yields

$$\| \delta x^n(\omega, t) \| \leqq (Kr)(K_1 + 1) \int_0^t [\| \delta x^n(\omega, s) \| + \delta u^n(\omega, s)]\, ds,$$

where

$$\delta u^n(\omega, s) \equiv \| u^n(\bar{x}(\omega, s), s) - \bar{u}(\bar{x}(\omega, s), s) \|,$$

from which follows, via an application of Gronwall's lemma, with $K_3 < \infty$,

$$\max_{t \leqq T} \| \delta x^n(\omega, t) \| \leqq K_3 \int_0^T \delta u^n(\omega, s)\, ds, \tag{14}$$

$$E \max_{t \leqq T} \| \delta x^n(\omega, t) \| \leqq K_3 \int_0^T \delta u^n(\omega, s)\, \mu(d\omega \times ds). \tag{15}$$

By (6),

$$\delta u^n(\omega, s) \leqq 2K_1[1 + \| \bar{x}(\omega, s \| + | s |],$$

which, by Lemma 1, has a finite integral over $\Omega \times [0, T]$. Since $u^n(x, s) \to \bar{u}(x, s)$ pointwise on $E^r \times [0, T]$, and since $\max_{s \leqq T} \| \bar{x}(\omega, s) \| < \infty$ w.p.1 by (7), we have $\| \delta u^n(\omega, s) \| \to 0$ w.p.1 almost everywhere on $\Omega \times [0, T]$. Thus, the dominated convergence theorem is applicable and yields that the right side of (15) goes to zero as $n \to \infty$; hence

$$E \max_{t \leqq T} \| \delta x^n(\omega, t) \| \to 0. \tag{16}$$

There is also a further subsequence (also indexed by n) so that

$$\max_{t \leqq T} \| \delta x^n(\omega, t) \| \to 0 \tag{17}$$

w.p.1.

Next, it will be shown that

$$\liminf_n \| x^n(\omega, T_\omega^n) - \bar{x}(\omega, \bar{T}_\omega) \| = 0,$$

where $\bar{T}_\omega$ is a random time with $E\bar{T}_\omega \leqq \hat{T}$.

Write

$$S^n = \| x^n(\omega, T_\omega^n) - \bar{x}(\omega, \bar{T}_\omega) \| \leqq S_1^n + S_2^n,$$

$$S_1^n = \| x^n(\omega, T_\omega^n) - \bar{x}(\omega, T_\omega^n) \|,$$

$$S_2^n = \| \bar{x}(\omega, T_\omega^n) - \bar{x}(\omega, \bar{T}_\omega) \|.$$

Let $T_i \to \infty$ be a sequence of real numbers. From (16), there is a sequence n_i so that

$$E \max_{t \leqq T_i} \| \delta x^{n_i}(\omega, t) \| < 2^{-i}.$$

Thus, there is a subsequence such that $\max_{t \leqq T_i} \delta x^{n_i}(\omega, t) \to 0$ w.p.1 as $i \to \infty$. This implies that $S_1^n \to 0$ (along the subsequence n_i) for a.a.ω such that $T_\omega^{n_i} > T_i$ only finitely often. Replace n_i by i, and define T_i so that $m\{\omega: T_\omega^i > T_i\} < 2^{-i}$. Since

$$\sum_i m\{\omega: T_\omega^i > T_i\} < \infty,$$

the Borel-Cantelli lemma implies that $T_\omega^i > T_i$ only finitely often w.p.1 (or, equivalently, $m\{\omega: T_\omega^i < T_i, \text{ all } i > I\} \to 1$ as $I \to \infty$). Thus, as $i \to \infty$,

$$S_1^i = \| x^i(\omega, T_\omega^i) - \bar{x}(\omega, T_\omega^i) \| \to 0 \text{ w.p.1.} \tag{18}$$

Define the random variable

$$\bar{T}_\omega \equiv \liminf_i T_\omega^i.$$

By Fatou's lemma and (10),

$$E\bar{T}_\omega \leqq \liminf_i ET_\omega^i \leqq \hat{T}. \tag{19}$$

Now, since $\bar{x}(\omega, \cdot)$ is continuous w.p.1. in $[0, \infty)$ and $\bar{T}_\omega < \infty$ w.p.1,

$$\liminf_i \| \bar{x}(\omega, T_\omega^i) - \bar{x}(\omega, \bar{T}_\omega) \| = 0.$$

Thus, $\liminf_i S_2^i = 0$. Since

$$\liminf_i (S_1^i + S_2^i) = \lim_i S_1^i + \liminf_i S_2^i,$$

$\liminf_i S^i = 0$. Since $g(\cdot)$ is uniformly continuous, this implies that

$$g(x^i(\omega, T_\omega^i)) - g(\bar{x}(\omega, \bar{T}_\omega)) = 0.$$

Since $\bar{x}_0(\omega, \cdot)$ is continuous w.p.1 and nondecreasing as t increases,

$$\liminf_i \bar{x}_0(\omega, T_\omega^i) - \bar{x}_0(\omega, \bar{T}_\omega) = 0 \tag{20}$$

w.p.1. Now, by (20), and since (18) converges to zero w.p.1,

$$\begin{aligned}
&\liminf_i [x_0^i(\omega, T_\omega^i) - \bar{x}_0(\omega, \bar{T}_\omega)] \\
&\quad = \liminf_i [(x_0^i(\omega, T_\omega^i) - \bar{x}_0(\omega, T_\omega^i)) + (\bar{x}_0(\omega, T_\omega^i) - \bar{x}_0(\omega, \bar{T}_\omega))] \\
&\quad = \lim_i [x_0^i(\omega, T_\omega^i) - \bar{x}_0(\omega, T_\omega^i)] + \liminf_i [\bar{x}_0(\omega, T_\omega^i) - \bar{x}_0(\omega, \bar{T}_\omega)] \\
&\quad = 0 \text{ w.p.1.}
\end{aligned}$$

Fatou's lemma now yields

$$E \liminf_i x_0^i(\omega, T_\omega^i) = E\bar{x}_0(\omega, \bar{T}_\omega) \leqq \liminf_i Ex_0^i(\omega, T_\omega^i) = \alpha,$$

proving that $\bar{u}(\cdot, \cdot)$ is the optimal control.

Let T_ω be the first time $g(\bar{x}(\omega, t)) = 0$; we have $T_\omega \leqq \bar{T}_\omega$. If $T_\omega < \bar{T}_\omega$ on some ω set, then the loss obtained by stopping at T_ω will also be α, since $x_0(\omega, \cdot)$ is nondecreasing in t. This completes the proof.

An easy consequence of the proof is a corresponding result when the

terminal constraint is the set of expectations $g(Ex(\omega, T)) = 0$, where T is nonrandom and finite. The proof is based on the fact that $Ex^n(\omega, T) \to E\bar{x}(\omega, \bar{T})$ for all finite T. The proof requires neither $f_0 \geqq 0$ nor the continuity of $x(\omega, \cdot)$, and, consequently, the continuity of $z(\omega, \cdot)$ may be dropped.

COROLLARY. *Assume* (A1) *to* (A4) *and* (A7). *Let all admissible controls satisfy* (6), *and restrict the terminal times to be nonrandom and bounded by some arbitrary but finite number. Let the target set be the set of expectations such that* (4) *is satisfied, where T is nonrandom and finite. Then if there is one admissible control such that* (4) *is satisfied, there is an admissible optimal control.*

We state the following theorem for form (1). The proof, although differing in detail, is essentially the same as the proof of Theorem 1, and will not be given. $z(\cdot, \cdot)$ is confined to Brownian motion to assure that the various stochastic integrals are defined and have suitable properties.

THEOREM 2. *Assume all the conditions of Theorem* 1, *except let* (1) *replace* (2), *and let $z(\cdot, \cdot)$ be vector Brownian motion. Assume* (A5) *and* (A6). *Then the conclusions of Theorem* 1 *hold.*

The existence and uniqueness of solutions to (1) and (2) has not yet been proved under much more general conditions on $u(x, t)$ than those of Theorem 1. In this sense, Theorems 1 and 2 represent about the best currently attainable result with the use of the control form $u(x, t)$, depending explicitly upon x.

In order to present existence results with other control forms, a different approach is taken in the sequel. Assume that the information upon which the values of the control are to depend are observations on $z(\cdot, \cdot)$, and that almost all sample functions of these observations are Borel measurable (as a function of t). We now write the control as an explicit function of ω and t, namely, $u(\omega, t)$. Without specifying the type of observations further, let there be a sub σ-field $\tilde{\Sigma}_c(T) \subset \tilde{\Sigma}(T)$ with respect to which the observations, as functions of ω and t ($t \leqq T$), are measurable. Then the admissible controls $u(\cdot, \cdot)$ will be measurable over $\tilde{\Sigma}_c(T)$. Let $\Sigma_c(t)$ be the fixed t section of $\tilde{\Sigma}_c(T)$, for $t \leqq T$; then we also require $\Sigma_c(t) \subset \Sigma(t)$, i.e., $u(\cdot, t)$ does not depend on the future.[2] For any function $u(\cdot, \cdot)$ measurable over $\tilde{\Sigma}_c(T)$, the section $u(\cdot, t)$ is measurable over the section $\Sigma_c(t)$, for each $t \leqq T$. Under (A9), all sections of $u(\cdot, \cdot)$ are integrable (over the respective sections of $\tilde{\Sigma}_c(T)$).

We will use the following.

(A8) All admissible $u(\cdot, \cdot)$ are measurable with respect to $\tilde{\Sigma}_c(T)$, for

[2] As one example, choose $\Sigma_c(t) \subset \Sigma(t)$ with $\Sigma_c(s) \subset \Sigma_c(t)$, $s < t$, and define $\tilde{\Sigma}_c(T)$ as the minimal σ-field over the union $\bigcup_{t \leqq T} [\Sigma_c(t) \times \mathfrak{I}(t, T)]$, where $\mathfrak{I}(t, T)$ is the Borel field over $[t, T]$.

any finite T. Also, almost all $u(\omega, \cdot)$ are Lebesgue measurable, and $\Sigma_s(t) \subset \Sigma(t)$ for each t.

We will also require the following.

(A9) Let $u(\omega, t)$ take values only in the convex compact set U.

(A10) Let $k(\cdot)$ be a continuous function and $k(U)$ a convex set.

The following theorems are also of interest for the reason that *if* $\tilde{\Sigma}_c(T) = \tilde{\Sigma}(T)$ *and* $\Sigma_c(t) = \Sigma(t)$, *then the optimal control in this family of controls yields at least as small a risk as the optimal control of any other family.* There are a number of important cases where the control may be chosen based on observations on $z(\cdot, \cdot)$ only; e.g., if $dz(\cdot, \cdot) = 0$, except at random times determined by a Poisson or other distribution, when it takes an impulsive form; or when a stochastic process, say $z_0(\cdot, \cdot)$, correlated with the $z(\cdot, \cdot)$ which drives the dynamical system, is the only function whose values are observed.

LEMMA 2. *Let B be a space of points σ, $\mathfrak{B}$ a σ-field over B. Let $\gamma(\sigma)$ be measurable with respect to $\mathfrak{B}$. Let U be a compact set in E^q, and $k(\cdot)$ a continuous function mapping U into E^m. Assume there is a not necessarily measurable function $u'(\sigma)$ mapping B into U such that*

$$\gamma(\sigma) = k(u'(\sigma)) = k(u_1'(\sigma), \cdots, u_q'(\sigma)). \tag{21}$$

Then there is a measurable function $u(\cdot)$ taking values in U such that

$$\gamma(\sigma) = k(u(\sigma)).$$

Remark. With replacement of the real line by a general measurable space, the proof is actually equivalent to the proof of a similar lemma given by Datco [7] (if his $K_t = K$ or, equivalently, U does not depend on σ—since under this condition, the metric space notions of [7] are not necessary).

Proof. Since U is compact, the range of $k(\cdot)$ is bounded and, by (21), so is the range of $\gamma(\cdot)$. Owing to this, we may define a sequence (indexed by n) of finite and measurable partitions $\{A_i^n\}$ of B such that $\bigcup_i A_i^n = B$, $A_i^n \cap A_j^n = \varnothing$, the empty set, and such that the oscillation of $\| \gamma(\sigma) \|$ on A_i^n is less than 2^{-n}.

Let $u = \{u_1, \cdots, u_q\}$ and assume that $u_1(\cdot), \cdots, u_{p-1}(\cdot)$ are measurable and that

$$\gamma(\sigma) = k(u_1(\sigma), \cdots, u_{p-1}(\sigma), u_p'(\sigma), \cdots, u_q'(\sigma)).$$

Equation (21) implies that there are numbers $\alpha_p^n(i), \cdots, \alpha_q^n(i)$ such that, *for each* $\sigma \in A_i^n$,

$$\tilde{u}^n(\sigma) \equiv \{u_1(\sigma), \cdots, u_{p-1}(\sigma), \alpha_p^n(i), \cdots, \alpha_q^n(i)\}$$

is in U and

$$\| \gamma(\sigma) - k(\tilde{u}^n(\sigma)) \| \leqq 2^{-n}.$$

The vector valued function $\hat{u}^n(\cdot)$, whose first $p - 1$ components are $u_1(\cdot)$,

$\cdots, u_{p-1}(\cdot)$, and whose last $q - p + 1$ components, $\tilde{u}_p{}^n(\cdot), \cdots, \tilde{u}_q{}^n(\cdot)$, take values $\alpha_p{}^n(i), \cdots, \alpha_q{}^n(i)$ in $A_i{}^n$, is measurable. Also,

$$\lim_n \| \gamma(\sigma) - k(u^n(\sigma)) \| \to 0$$

uniformly in B.

Let

$$u_p(\sigma) = \lim_n \inf \hat{u}_p{}^n(\sigma).$$

Then $u_p(\cdot)$ is measurable. *Fix* σ. Let n index a subsequence so that $\hat{u}_p{}^n(\sigma) \to u_p(\sigma)$ and $u_i{}^n(\sigma) \to \beta_i(\sigma)$, $i > p$. Let

$$v^n(\sigma) = \{u_1(\sigma), \cdots, u_p(\sigma), \beta_{p+1}^n(\sigma), \cdots\},$$

and $v(\sigma) = \lim_n v^n(\sigma)$. $v(\sigma)$ is in U and

$$\lim_n \| \gamma(\sigma) - k(v^n(\sigma)) \| \to 0.$$

By continuity, $\gamma(\sigma) = k(v(\sigma))$. By induction, the lemma is proved.

THEOREM 3. *Assume* (A3), (A4) *and* (A7) *to* (A10). *Let*

$$x(\omega, t) = x(\omega, 0) + \int_0^t A(s)\, x(\omega, s)\, ds + \int_0^t k(u(\omega, s))\, ds + z(\omega, t) - z(\omega, 0), \tag{22}$$

where $z_0(\omega, t) \equiv 0$, *and* $A(s)$ *is a matrix with bounded and Borel measurable components. Let* (4) *be the terminal condition, where* T *is nonrandom. Let* $\hat{T} < \infty$ *be given. Define an admissible control* $u(\cdot, \cdot)$ *as one satisfying* (A8) *and* (A9) *and for which there is a nonrandom* $T \leqq \hat{T}$ *such that* $g(Ex(\omega, T)) = 0$ *and* $R(u) < \infty$. *Assume there exists one admissible control. Then, there is an optimal admissible control (minimizing* $R(u)$*).*

Proof. Let $u(\cdot, \cdot)$ satisfy (A8) and (A9). Under the conditions on $u(\cdot, \cdot)$, $k(\cdot)$, $A(\cdot)$, and $z(\cdot, \cdot)$, the existence and uniqueness w.p.1 of solutions to (22) in $[0, \infty)$ is a special case of Lemma 1. Also, for $T < \infty$,

$$\max_{t \leqq T} E\| x(\omega, t) \| \leqq K_2(T) < \infty, \tag{23}$$

where $K_2(T)$ does not depend on the particular admissible control used.

Let α be the infimum of $R(u)$ over the class of admissible controls. Then there is a sequence of admissible controls $u^n(\cdot, \cdot)$ with corresponding terminal times $T^n \leqq \hat{T}$, and with $R(u^n) = \alpha^n$ decreasing monotonically to α. We must show that there are a $\bar{u}(\cdot, \cdot)$ and a corresponding $T \leqq \hat{T}$ such that $g(Ex(\omega, T)) = 0$.

Define

$$T = \lim_n \inf T^n \leqq \hat{T} < \infty.$$

By the conditions on $k(\cdot)$ and on each $u^n(\cdot, \cdot)$,

$$E \int_0^T \| k(u^n(\omega, t)) \| \, dt = \int_0^T E \| k(u^n(\omega, t)) \| \, dt \leqq K_1 < \infty, \tag{24}$$

where K_1 does not depend on n. The sequence $\| k(u^n(\cdot, \cdot)) \|$ is uniformly bounded (in n, ω, and t) since $\| u^n(\cdot, \cdot) \|$ is uniformly bounded and $k(\cdot)$ is continuous. Thus

$$\int_A k(u^n(\omega, t)) \, \mu(d\omega \times dt) \to 0$$

uniformly in n as $\mu(A) \to 0$. Owing to this and to (24) we have, by [8, Theorem IV.8.9], that $\{k(u^n(\cdot, \cdot)\}$ is weakly sequentially compact; there is a subsequence, also indexed by n, and there is a $\tilde{\Sigma}_c(T)$ measurable function $\gamma(\cdot, \cdot)$ so that, for all sets A in $\tilde{\Sigma}_c(T)$,

$$\int_A k(u^n(\omega, t)) \, \mu(d\omega \times dt) \to \int_A \gamma(w, t) \, \mu(d\omega \times dt) \tag{25}$$

and $T^n \to T$ *monotonically*.

The rest of the proof follows a method of Roxin [4]. For any A in $\tilde{\Sigma}_c(T)$ and *any* constant vector y,

$$\begin{aligned}
&\int_A \operatorname*{lub}_{v \in U} y'k(v) \, \mu(d\omega \times dt) \\
&\qquad \geqq \lim_n \int_A y'k(u^n(\omega, t)) \, \mu(d\omega \times dt) \\
&\qquad = \int_A y'\gamma(\omega, t) \, \mu(d\omega \times dt) \\
&\qquad = \lim_n \int_A y'k(u^n(\omega, t)) \, \mu(d\omega \times dt) \\
&\qquad \geqq \int_A \operatorname*{glb}_{v \in U} y'k(v) \, \mu(d\omega \times dt).
\end{aligned}$$

(Note that the arguments of the first and last integrals are constants.) Thus, except for a set of measure zero, all (ω, t) in $\Omega \times [0, T]$ satisfy

$$\operatorname*{lub}_{v \in U} y'k(v) \geqq y'\gamma(\omega, t) \geqq \operatorname*{glb}_{v \in U} y'k(v). \tag{26}$$

Redefine $\gamma(\omega, t)$ on the remaining set of μ measure zero so that (26)

holds everywhere. Since U, the range of v, is a closed convex set, (26) implies that for each (ω, t) there is a number $u(\omega, t)$ in U so that

$$\gamma(\omega, t) = k(u(\omega, t)). \tag{27}$$

By Lemma 2, the $u(\cdot, \cdot)$ in (27) can be defined to be measurable with respect to $\tilde{\Sigma}_c(T)$. Denote this measurable $u(\cdot, \cdot)$ by $\bar{u}(\cdot, \cdot)$. It will be shown that $\bar{u}(\cdot, \cdot)$ and T are the optimal control and stopping time, respectively.

Let $\bar{x}(\cdot, \cdot)$ correspond to $\bar{u}(\cdot, \cdot)$. Since (22) holds for all admissible controls, and $k(\cdot)$ is uniformly bounded on U, all future interchanges of the order of integration are justified. We have

$$Ex^n(\omega, T^n) - Ex^n(\omega, T) = \int_T^{T^n} Ex^n(\omega, t)\, dt + \int_T^{T^n} Ek(u^n(\omega, t))\, dt \to 0 \tag{28}$$

as $n \to \infty$, since both integrands are finite, and $T^n \to T$ monotonically. Also

$$Ex^n(\omega, t) - E\bar{x}(\omega, t) = \int_0^t A(s)[Ex^n(\omega, s) - E\bar{x}(\omega, s)]\, ds + \delta_n(t),$$

where

$$\delta_n(t) = \int_0^t [Ek(u^n(\omega, s)) - Ek(\bar{u}(\omega, s))]\, ds$$

(the $Ez(\omega, t)$ terms cancel). By the weak convergence, $\delta_n(t) \to 0$ for each $t \leqq T$ as $n \to \infty$. This and the boundedness of $A(\cdot)$ in $[0, T]$ imply, by Gronwall's lemma, that

$$\| Ex^n(\omega, T) - E\bar{x}(\omega, T) \| \to 0. \tag{29}$$

Combining (29) with (28) yields $Ex^n(\omega, T^n) \to E\bar{x}(\omega, T)$ and, hence, $R(u^n) \to R(\bar{u})$. By the continuity of $g(\cdot)$, $g(E\bar{x}(\omega, T)) = 0$, and the proof is concluded.

Remark. A stronger type of convergence than the weak convergence argument used here appears to be necessary to establish convergence of the sample functions in general (or even of their expectations in the nonlinear case) as was done in Theorem 1. This is the reason for restricting the target to a set of expectations, rather than sample functions, and the linear assumption (22). It would be useful to prove whether or not sample function convergence is necessary, in general, in order to satisfy terminal constraints on $x(\omega, t)$.

Acknowledgment. It is a pleasure to acknowledge the anonymous assistance of the referee, through whose painstaking reading several ambiguities and errors were removed.

REFERENCES

[1] H. J. KUSHNER, *On the stochastic maximum principle with "average" constraints*, J. Math. Anal. Appl., 12 (1965), pp. 13–26.

[2] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] E. B. LEE AND L. MARCUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[4] E. ROXIN, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[5] S. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[6] M. LOEVE, *Probability Theory*, 3d ed., Van Nostrand, Princeton, 1963.

[7] R. DATCO, *An implicit function theorem with an application to control theory*, Michigan Math. J., 11 (1964), pp. 345–351.

[8] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Part 1, Interscience, New York, 1958.

J.SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

EXISTENCE THEOREMS FOR OPTIMAL SOLUTIONS IN PONTRYAGIN AND LAGRANGE PROBLEMS*

LAMBERTO CESARI†

In §2 we state an existence theorem (Theorem 1) for Pontryagin's problem which represents a useful generalization of Filippov's existence theorem. In §3 we consider the more general Lagrange problem with unilateral constraints written in terms of a Pontryagin problem with variable control space $U(t, x)$ which is closed but not necessarily compact. If $U(t, x)$ is compact for every (t, x), then the problem reduces to a Pontryagin problem of optimal control; if U is fixed and coincides with the whole space then the problem is of a generality comparable to the usual Lagrange problem. With a further particularization the problem reduces to a free problem of the calculus of variations. For the general problem thus formulated where $U(t, x)$ is any closed variable set, we state existence theorems for optimal solutions (Theorems 3 and 4), which contain as particular cases the Filippov statement, Theorem 1, existence statements for the Lagrange problem of the calculus of variations, and the Nagumo-Tonelli theorem for free problems. Also, we formulate existence statements (Theorems 5 and 6) for the case where the differential equations are linear in the control variables. Finally, we state the analogous existence theorems (Theorems 7 and 8) for weak solutions (in the sense of Gamkrelidze) of the same general Lagrange problems with unilateral constraints ($U(t, x)$ closed, not necessarily compact).

Theorems 2–8 are merely stated here; their proofs will appear elsewhere. Theorem 1 is a corollary of both Theorems 3 and 4, but it admits also of a direct proof which is very simple and similar to Filippov's proof. In the Appendix we give—for the convenience of the reader—both the deduction of Theorem 1 from Theorems 3 and 4, and its direct proof.

In other papers we shall discuss analogous problems for multidimensional Lagrange problems with unilateral constraints involving partial differential equations.

1. Notations for Lagrange problems with unilateral constraints.

1a. Let A be a closed subset of the tx-space $E_1 \times E_n$, $t \in E_1$, $x = (x^1, \cdots, x^n) \in E_n$; and for each $(t, x) \in A$, let $U(t, x)$ be a closed subset of the u-space E_m, $u = (u^1, \cdots, u^m)$. We do not exclude that A

* Received by the editors May 17, 1965, and in revised form August 23, 1965.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan. This research was supported by the National Science Foundation under Grant GP-3920.

coincides with the whole tx-space and that U coincides with the whole u-space. Let M denote the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let $\tilde{f}(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function from M into E_{n+1}. Let B be a closed subset of points (t_1, x_1, t_2, x_2) of E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$. We shall consider the class of all pairs $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, of vector functions $x(t)$, $u(t)$ satisfying the following conditions:

(a) $x(t)$ is absolutely continuous (AC) in $[t_1, t_2]$;
(b) $u(t)$ is measurable in $[t_1, t_2]$;
(c) $(t, x(t)) \in A$ for every $t \in [t_1, t_2]$;
(d) $u(t) \in U(t, x(t))$ almost everywhere (a.e.) in $[t_1, t_2]$;
(e) $f_0(t, x(t), u(t))$ is L-integrable in $[t_1, t_2]$;
(f) $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$;
(g) $(t_1, x(t_1), t_2, x(t_2)) \in B$.

By (f) we mean that the n ordinary differential equations

$$\frac{dx^i}{dt} = f_i(t, x(t), u(t)), \qquad i = 1, \cdots, n, \tag{1}$$

are satisfied a.e. in $[t_1, t_2]$. Since $x(t)$ is AC, that is, each component $x^i(t)$ of $x(t)$ is AC, we conclude that all $f_i(t, x(t), u(t))$, $i = 1, \cdots, n$, are L-integrable in $[t_1, t_2]$ as f_0.

A pair $x(t)$, $u(t)$ satisfying (a b c d e f g) is said to be *admissible*, $x(t)$ is called a *trajectory*, and $u(t)$ a *strategy*, or *control*, or *steering function*. As usual, $U(t, x)$ is said to be the *control space* at the time t and space point x.

The functional

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt \tag{2}$$

is called the *cost functional*, and we seek the minimum of $I[x, u]$ in the total class Ω of admissible pairs $x(t)$, $u(t)$, or in some well defined subclass Ω.

In the particular case where $U(t, x)$ is a compact subset of E_m for every $(t, x) \in A$, the problem of the minimum of $I[x, u]$ is called a *Pontryagin problem* of optimal control theory. The general case above, where $U(t, x)$ is a closed subset of E_m for every $(t, x) \in A$, not necessarily compact, will be denoted as a *Lagrange problem with unilateral constraints*. The classical Lagrange problem corresponds essentially to the case where $U = E_m$ is the whole u-space, with the side conditions being here differential equations in normal form.

There is a particular case of the Lagrange problem which shall be taken into consideration, namely, $m = n$, $U = E_m$, and the vector function $f(t, x, u)$ given by $f(t, x, u) = u$, or $f_i(t, x, u) = u^i$, $i = 1, \cdots, m = n$, and hence $\tilde{f}(t, x, u) = (f_0, u)$. Then the differential system (1)

reduces to $dx^i/dt = u^i$, $i = 1, \cdots, n$, and the cost functional becomes

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), x'(t))\, dt. \tag{3}$$

This problem is called a *free problem.*

1b. If we denote by X the space of all continuous vector functions $x(t) = (x^1, \cdots, x^n)$, $a \leqq t \leqq b$, from arbitrary finite intervals $[a, b]$ to E_n, it is convenient to define a *distance* function $\rho(x, y)$ for elements $x(t)$, $a \leqq t \leqq b$, and $y(t)$, $c \leqq t \leqq d$, of X, so as to make X a metric space. For this purpose we extend $x(t)$ in all $(-\infty, +\infty)$ by defining it equal to $x(a)$ for $t \leqq a$ and equal to $x(b)$ for $t \geqq b$, and, analogously, for $y(t)$. We then define

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

where max is taken for all t, $-\infty < t < +\infty$. Then ρ is a distance function and X is a metric space.

Every element $x(t)$ of an admissible pair $[x(t), u(t)]$ is an element of X but, of course, the converse is not true.

A class Ω of admissible pairs is said to be *complete* provided it satisfies the following property: If $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, and $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, are all admissible pairs, if $x_k(t) \to x(t)$ in the ρ-metric, and all pairs $x_k(t)$, $u_k(t)$, $k = 1, 2, \cdots$, belong to Ω, then $x(t)$, $u(t)$ also belongs to Ω. The classes usually taken into consideration in applications are complete. The class of all admissible pairs (satisfying (a b c d e f g)) is certainly complete.

Given any point $(t_0, x_0) \in A$ and $\delta > 0$, we denote by $N_\delta(t_0, x_0)$ the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from (t_0, x_0). The set $U(t, x)$ is said to be an *upper semicontinuous function of* (t, x) *in* A provided for every $(t_0, x_0) \in A$ there is a $\delta > 0$ such that

$$U(t, x) \subset [U(t_0, x_0)]_\epsilon$$

for every $(t, x) \in N_\delta(t_0, x_0)$, where U_ϵ denotes the closed ϵ-neighborhood of U in E_m.

1c. In Filippov's existence theorem [2a] the most typical requirement is that for every $(t, x) \in A$ the set

$$\tilde{Q}(t, x) = \tilde{f}(t, x, U(t, x)) = \{\tilde{z} = (z^0, z) \mid \tilde{z} = \tilde{f}(t, x, u), u \in U(t, x)\}$$

$$= \{\tilde{z} = (z^0, z) \mid z^0 = f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+1}$$

be a convex subset of E_{n+1}. We prove here that a much weaker condition

suffices; namely, we shall require that for every $(t, x) \in A$ the set

$$\tilde{\tilde{Q}}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+1}$$

be a convex set of E_{n+1}. Using the set $\tilde{\tilde{Q}}$ we have obtained results for optimal control problems comparable in generality with those obtained by Tonelli for free problems in the calculus of variations. Moreover, the new theorems can be extended to the case where $U(t, x)$ is only closed (in particular, $U = E_m$), that is, for the Lagrange problem with unilateral constraints. In addition, the corresponding existence theorems for the Lagrange problem contain as a particular case both the Tonelli-Nagumo theorem for free problems, and the Filippov existence theorem for problems of optimal control.

It may be noted that for free problems the sets $\tilde{Q}$, $\tilde{\tilde{Q}}$ —thought of as subsets of the z^0u-space E_{n+1} —are the sets

$$\tilde{Q}(t, x) = \{\tilde{z} = (z^0, u) \mid z^0 = f_0(t, x, u), u \in E_n\} \subset E_{n+1},$$

$$\tilde{\tilde{Q}}(t, x) = \{\tilde{z} = (z^0, u) \mid z^0 \geqq f_0(t, x, u), u \in E_n\} \subset E_{n+1},$$

and thus the convexity of $\tilde{\tilde{Q}}$ reduces to the usual convexity condition of $f_0(t, x, u)$ as a function of u in $U(t, x)$—a condition which is familiar in the calculus of variations—while the set $\tilde{Q}(t, x)$ is convex if and only if f_0 is linear in u. This particular case of the free problems already shows the amount of generality introduced by the consideration of the sets $\tilde{\tilde{Q}}$ instead of $\tilde{Q}$.

We mention here that a function $\phi(u)$, $u \in E_m$, is said to be convex in u, provided $u, v \in E_m$, $0 \leqq \alpha \leqq 1$, implies $\phi(\alpha u + (1 - \alpha)v) \leqq \alpha\phi(u) + (1 - \alpha)\phi(v)$.

2. Existence theorems for Pontryagin problems.

2a. THEOREM 1. (Existence theorem for Pontryagin problems). *Let A be any compact subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ let $U(t, x)$ be a compact subset of the u-space E_m. Let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$, and let $f(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function on M. Let $U(t, x)$ be an upper semicontinuous function of (t, x) in A, and for every $(t, x) \in A$ let*

$$\tilde{\tilde{Q}}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq f(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+1}$$

be a convex subset of E_{n+1}. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Then the cost functional $I[x, u]$ has an absolute minimum in any nonempty complete class Ω of admissible pairs $u(t)$, $x(t)$.

If A is not compact, but closed and contained in a slab $\{t_0 \leqq t \leqq T, -\infty < x^i < +\infty, i = 1, \cdots, n, t_0, T$ finite$\}$, then Theorem 1 still holds

under the additional hypotheses that
(a)* $x^1 f_1 + \cdots + x^n f_n \leqq N[(x^1)^2 + \cdots + (x^n)^2 + 1]$ for all $(t, x, u) \in M$ and some constant $N > 0$, and
(b) each trajectory $x(t)$ of the class Ω contains at least one point $(t^*, x(t^*))$ on a compact subset P of A (for instance, the initial point $(t_1, x(t_1))$ is fixed, or the final point is fixed).

If A is not compact, nor contained in a slab as above, but A is closed, then Theorem 1 still holds under the additional hypotheses (a), (b) and (c):
(c′) $f_0(t, x, u) \geqq -G$ for all $(t, x, u) \in M$ and some constant $G \geqq 0$;
(c″) $f_0(t, x, u) \geqq \mu > 0$ for all $(t, x, u) \in M$ with $| t | \geqq N_1$ and some constants $\mu > 0$, $N_1 \geqq 0$. Also, condition (a) can be replaced by the following condition (d):
(d′) $f_0(t, x, u) \geqq -G$ for all $(t, x, u) \in M$ and some constant $G \geqq 0$;
(d″) $f_0(t, x, u) \geqq E \mid f(t, x, u) \mid$ for all $(t, x, u) \in M$ with $| x | \geqq F$ for some constants $E > 0$ and $F \geqq 0$.

Condition (b) is certainly verified if for instance the projection B_1 of B on the t_1x_1-space is compact, in particular, if the initial points $(t_1, x(t_1))$ of the trajectories $x(t)$ of Ω belong to a compact subset of the same space. The same holds if the projection B_2 of B on the t_2x_2-space is compact, in particular if the endpoints $(t_2, x(t_2))$ of the trajectories $x(t)$ of Ω belong to a compact subset of the same space.

The existence Theorem 1 contains as a particular case Filippov's theorem [2a]. Indeed, in the latter it is requested that the set $\tilde{Q}(t, x)$ is convex, and the convexity of $\tilde{Q}$ certainly implies the convexity of $\tilde{\tilde{Q}}$. Therefore, Theorem 1 contains as a particular case the analogous existence theorem proved by Markus and Lee in [5] for $\tilde{f}$ linear in u. Also, an extension analogous to the one of Roxin [10] (see also [2b]) can be duplicated in the present more general situation, by replacing the continuity requirement for f_0 and f by measurability and the hypothesis that $| f_i(t, x, u) | \leqq \varphi(t)[A + B \mid x \mid]$ for all $(t, x, u) \in M$ where $A, B \geqq 0$ are constants, and $\varphi(t)$ is a fixed function of t which is L-integrable on every finite interval.

2b. As an example of an application of Theorem 1 we may consider the Pontryagin problem with $m = n = 2$,

$$I = \int_{t_1}^{t_2} (x^2 + y^2 + u^2 + v^2 + 1)\, dt = \text{minimum},$$

* The following weaker assumption from differential equation theory would suffice. There exist a (Lyapunov-like) positive, continuously differentiable function $V(x, t)$ and a positive constant c such that $| \operatorname{grad}_x V(x, t) \cdot f(x, t, u) + \partial V/\partial t | \leqq cV(x, t)$ for all $(t, x, u) \in M$, and the set $\{x \mid V(x, t) \leqq \alpha, (t, x) \in A\}$ is compact for every α. I wish to thank the referee for having suggested this remark.

$$U(x, y) = \{-1 \leqq u \leqq 1,\ -1 \leqq v \leqq 1\},$$

$$\frac{dx}{dt} = u, \qquad \frac{dy}{dt} = v,$$

$$t_1 = 0, \qquad x(0) = y(0) = 0, \qquad x(t_2) = 1, \qquad t_2\,, y(t_2) \text{ undetermined.}$$

Here A is the halfspace (t, x, y) with $t \geqq 0$, $(x, y) \in E_2$. Also, B is the closed set $B = \{(t_1\,, x_1\,, y_1\,, t_2\,, x_2\,, y_2) \mid t_1 = 0, x_1 = 0, y_1 = 0, x_2 = 1\}$, that is, a 2-plane in E_6, and B_1 is the single point $(0, 0, 0)$ of E_3, certainly a compact set, $P = B_1$, and condition (b) is satisfied. The set

$$\tilde{Q}(x, y)$$

$$= \{(z, u, v) \mid z = x^2 + y^2 + u^2 + v^2 + 1,\ -1 \leqq u \leqq 1,\ -1 \leqq v \leqq 1\}$$

is not convex, and Filippov's theorem does not apply. Instead

$$\tilde{\tilde{Q}}(x, y)$$

$$= \{(z, u, v) \mid z \geqq x^2 + y^2 + u^2 + v^2 + 1,\ -1 \leqq u \leqq 1,\ -1 \leqq v \leqq 1\}$$

is certainly convex. Finally, by the usual relation $|x| \leqq 2^{-1}(1 + x^2)$, x scalar, we deduce

$$xu + yv \leqq |x| + |y| \leqq 1 + x^2 + y^2,$$

and condition (a) is satisfied. Condition (c) is satisfied with $G = \mu = 1$. By Theorem 1 the problem above has an optimal solution.

2c. Instead of the assumption that the sets $\tilde{Q}(t, x)$ be convex as in Filippov's theorem, or of the much less demanding assumption that the sets $\tilde{\tilde{Q}}(t, x)$ be convex, we may ask whether an alternate assumption would suffice for the existence of the minimum. The question has been proposed as to whether the joint assumption that the sets $Q(t, x) = f[t, x, U(t, x)] \subset E_n$ be convex and that $f_0(t, x, u)$ be convex in u for every (t, x) would suffice. The answer is negative as it was proved by an example in [1c; §5] and by an easier example in [1d, I, §8]. We proved in [1b] and [1c] that the hypothesis $Q(t, x)$ convex and $f_0(t, x, u)$ convex in u still guarantees existence provided an additional hypothesis is satisfied. We expressed this additional hypothesis by saying that the "curvature" of f is small with respect to the "convexity" of f_0. We state below the theorem (Theorem 2) in its precise form in order to point out its connection with Theorem 1 (see Remark after Theorem 2).

For every $(t, x) \in A$ we shall denote by $U^*(t, x)$ the closed convex hull of $U(t, x)$, or $U^*(t, x) = \text{cl co } U(t, x) \subset E_m$, and by M^* the set of all (t, x, u) with $(t, x) \in A$, $u \in U^*(t, x)$. We shall assume here that $\tilde{f}(t, x, u)$ admits of a continuous extension on the set M^*, though we assume that $U(t, x)$ is still the control space, that is, that any admissible strategy $u(t)$

has its values only in $U(t, x)$, that is, $u(t) \in U(t, x(t))$. The following hypotheses (or definitions) (α), (β) and (γ) will be needed.

(α) *Hypothesis of convexity of* f_0. There is a nonnegative bounded and Borel measurable function $C = C(t, x, u)$, $(t, x, u) \in M^*$, with the following property: for each $\epsilon > 0$ and $(t_0, x_0, u_0) \in M^*$, there are a $\delta = \delta(t_0, x_0, u_0, \epsilon) > 0$ and a linear scalar function $z(u) = r + b \cdot u$ (also depending on t_0, x_0, u_0, ϵ) such that, for every $(t, x) \in A$ at a distance $\leqq \delta$ from (t_0, x_0), we have

$$f_0(t, x, u) \geqq z(u) + C(t_0, x_0, u_0, \epsilon) \mid u - u_0 \mid^2 \quad \text{for all} \quad u \in U^*(t, x),$$

$$f_0(t, x, u) \leqq z(u) + \epsilon \quad \text{for all} \quad u \in U^*(t, x) \quad \text{with} \quad \mid u - u_0 \mid \leqq \delta.$$

Condition (α) is certainly satisfied if, for each $(t, x) \in A$, the function f_0 is of class C^2 in $u = (u^1, \cdots, u^m)$, if the second partial derivatives f_{0hk} are continuous in $A \times U^*$, and if the quadratic form

$$F(\xi) = \sum a_{hk}\xi_h\xi_k, \qquad a_{hk} = a_{kh} = f_{0hk}(t, x, u), \qquad A = [a_{hk}],$$

where $h, k = 1, \cdots, m$, and $\xi = (\xi_1, \cdots, \xi_m)$ is a real vector, is positive semidefinite (positive definite if we want $C > 0$). Then $F(\xi) \geqq \lambda_0(\xi_1^2 + \cdots + \xi_m^2)$, where $\lambda_0 \geqq 0$ is the smallest root of the equation $\det(A - \lambda I) = 0$, I is the unit matrix, and we can take $C(t, x, u) = \lambda_0/2$. Condition (α) as given above is only a generalized form of this familiar condition, which does not require second order partial derivatives. This generalized form of stating convexity is often used in the calculus of variations.

(β) *Hypothesis of boundedness of the curvature of* f. There exists a nonnegative bounded, Borel measurable function $D = D(t, x, u)$, $(t, x, u) \in M^*$, with the following property: for each $\epsilon > 0$ and $(t_0, x_0, u_0) \in M^*$, there are a $\delta = \delta(t_0, x_0, u_0, \epsilon) > 0$ and a linear vector function $Z(u) = R + Bu$, where R is an n-vector, B is an $n \times m$ matrix, and δ, R, B depend on $(t_0, x_0, u_0, \epsilon)$, such that for each $(t, x) \in A$ at a distance $\leqq \sigma$ from (t_0, x_0) we have

$$\mid f(t, x, u) - Z(u) \mid \leqq \epsilon + D(t_0, x_0, u_0, \epsilon) \mid u - u_0 \mid^2 \quad \text{for all } u \in U^*(t, x).$$

Condition (β) is certainly satisfied if, for each $(t, x) \in A$, $f(t, x, u)$ is of class C^2 in $u = (u^1, \cdots, u^m)$ with second order partial derivatives continuous in M^*, and

$$\sum_{i=1}^{n} \mid \sum_{hk} a_{ihk}\xi_h\xi_k \mid \leqq 2D(\xi_1^2 + \cdots + \xi_m^2),$$

$$a_{ihk} = a_{ikh} = f_{ihk} = \frac{\partial^2 f_i}{\partial u_h \, \partial u_k},$$

and $\sum_{hk}$ ranges over all $h, k = 1, \cdots, m$.

Finally, we shall require certain Lipschitz-type conditions on both the scalar function f_0 and the vector function $f = (f_1, \cdots, f_n)$.

(γ) *Lipschitz-type conditions for f_0 and f.* There are two functions $L(t, x, u)$, $\Lambda(t, x, u)$, $(t, x, u) \in M^*$, both nonnegative and Borel measurable, with points of infinity (if any) whose coordinates t lie in a subset of measure zero of E_0, with the following properties:

(γ_1) $| f_0(t, x, u) - f_0(t, x, u_0) | \leqq L(t, x, u_0) | u - u_0 |$ for all $(t, x) \in A$ and any two points $u, u_0 \in U^*(t, x)$;

(γ_2) if, for any $(t, x) \in A$, for any n-vector $z_0 = f(t, x, u_0)$, $u_0 \in U^*(t, x)$, and for any other n-vector $z \in f(t, x, U^*(t, x))$, we take $u \in U^*(t, x)$ in such a way that $z = f(t, x, u)$, and $| u - u_0 |$ = minimum, then we have

$$| u - u_0 | \leqq \Lambda(t, x, u_0) | z - z_0 |.$$

Condition (γ_2) obviously does not require the monotonicity of f in the vector u, a condition which would be impossible to verify if, for instance, $n < m$. Nevertheless, if $n \geqq m$, then condition (γ_2) is certainly verified if f is monotone in u for every $(t, x) \in A$, and if

$$| u - u_0 | \leqq \Lambda(t, x, u_0) | f(t, x, u) - f(t, x, u_0) |$$

for every $(t, x) \in A$ and any two $u, u_0 \in U^*(t, x)$. We say that $f(t, x, u)$ is *monotone in u* if $u, u_0 \in U^*(t, x)$, $u \neq u_0$, implies $f(t, x, u) \neq f(t, x, u_0)$.

2d. THEOREM 2. (Existence theorem for Pontryagin problems). *Let A be any compact subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ let $U(t, x)$ be a compact subset of the u-space E_m. For every $(t, x) \in A$ let $U^*(t, x)$ be the closed convex hull of $U(t, x)$, and let $\tilde{f}(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function on M^*. Let $U(t, x)$ be an upper semicontinuous function on M, and for every $(t, x) \in A$ let*

$$Q(t, x) = \{z \mid z = f(t, x, u),\ u \in U(t, x)\} \subset E_n$$

be a convex subset of E_n. Let hypotheses (α), (β), (γ) be satisfied and assume that

$$\Lambda(t, x, u) L(t, x, u) D(t, x, u) \leqq C(t, x, u), \qquad (t, x, u) \in M^*,$$

where equality holds at most on a set of points $(t, x, u) \in M^*$ whose coordinates t lie in a subset of measure zero of E_1. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Then $I[x, u]$ has an absolute minimum in any nonempty complete class Ω of admissible pairs $u(t)$, $x(t)$.

If A is not compact, but closed and contained in a slab $\{t_0 \leqq t \leqq T$, $-\infty < x^i < +\infty$, $i = 1, \cdots, n$, t_0, T finite$\}$, then Theorem 2 holds under the additional hypotheses (a), (b) at the end of Theorem 1. If A is not compact, nor contained in any slab as above, but A is closed, then Theorem 2 still holds under the additional hypotheses (a), (b), and (c) at the end of Theorem 1. Finally, condition (a) can be replaced by condition (d) at the end of Theorem 1.

A series of examples concerning the use of Theorem 2 has been given in previous papers [1b], [1c]. For a direct proof of Theorem 2 we refer to [1b].

Remark. Whenever the condition $\Lambda LD < C$ of Theorem 2 implies that the set $\tilde{\tilde{Q}}(t, x)$ of Theorem 1 be convex, then Theorem 2 becomes a particular case of Theorem 1. For instance, assume that $m = n$ and that the transformation $z = f(t, x, u)$ is one-to-one and admits a continuous inverse $u = f^{-1}(t, x, z)$. Then $\tilde{\tilde{Q}}$ can be defined as the set $\tilde{\tilde{Q}}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq F(t, x, z), z \in Q(t, x)\}$, where $F(t, x, z) = f(t, x, f^{-1}(t, x, z))$. In this situation the requirement of the convexity of $\tilde{\tilde{Q}}(t, x)$ reduces to the requirement of the convexity of the function F in z. If $m = n = 1$, if for any (t_0, x_0, z_0) we have, in a neighborhood of u_0,

$$f_0(t_0, x_0, u) = z_0^0 + l(u - u_0) + 2^{-1}c(u - u_0)^2 + \cdots,$$

$$z = f(t_0, x_0, u) = z_0 + \lambda^{-1}(u - u_0) + 2^{-1}d(u - u_0)^2 + \cdots,$$

with $c > 0$, $\lambda l\, d < c$, $\lambda \neq 0$, and the second relation can be inverted, then in a neighborhood of z_0 we have

$$u = u_0 + \lambda(z - z_0) - 2^{-1}\lambda^3 d(z - z_0)^2 + \cdots,$$

$$F(t_0, x_0, z) = z_0^0 + \lambda l(z - z_0) + 2^{-1}\lambda^2(c - \lambda l\, d)(z - z_0)^2 + \cdots.$$

Thus, the condition that $\lambda l d < c$ implies that F is convex in z at z_0, and hence everywhere, since z_0 is arbitrary, and $\tilde{\tilde{Q}}(t_0, x_0)$ is a convex set.

3. Existence theorems for Lagrange problems with unilateral constraints.

3a. We shall use the same notations as in §§1, 2, but we shall now assume that the sets $U(t, x)$ and $Q(t, x)$ are not compact but only closed. The usual condition of upper semicontinuity need be replaced by slightly more general conditions, namely condition (U) for the sets $U(t, x)$ and property (Q) for the sets $Q(t, x)$. Given a set E, we shall denote by cl E, co E, bd $E = \partial E$, int E the closure, the convex hull, the boundary, and the subset of the interior points, respectively, of E. Thus cl co E denotes the closure of the convex hull of E.

As usual, let A be a closed subset of the tx-space $E_1 \times E_n$. For any $(t_0, x_0) \in A$ and $\delta > 0$ let $N_\delta(t_0, x_0)$ denote the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from (t_0, x_0). For any point $(t, x) \in A$ let $U(t, x)$ be a closed nonempty set of the u-space E_m. For any $(t_0, x_0) \in A$ let $U(t_0, x_0, \delta)$ denote the set $U(t_0, x_0, \delta) = \bigcup U(t, x)$, where the union is taken for all $(t, x) \in N_\delta(t_0, x_0)$. We shall say that $U(t, x)$ satisfies property (U) at the point (t_0, x_0) of A if

$$U(t_0, x_0) = \bigcap_{\delta > 0} \text{cl } U(t_0, x_0, \delta).$$

We shall say that $U(t, x)$ satisfies property (U) in A if $U(t, x)$ satisfies property (U) at every point (t_0, x_0) of A. We shall say that $U(t, x)$ sat-

isfies property (Q) at the point (t_0, x_0) of A if

$$U(t_0, x_0) = \bigcap_{\delta>0} \operatorname{cl} \operatorname{co} U(t_0, x_0, \delta).$$

We shall say that $U(t, x)$ satisfies property (Q) in A if $U(t, x)$ satisfies property (Q) at every point $(t_0, x_0) \in A$.

Sets $U(t, x)$ satisfying property (U) are necessarily closed; sets $Q(t, x)$ satisfying property (Q) are necessarily closed and convex. Both properties (U) and (Q) are generalizations of the usual upper semicontinuity in the following sense. If the sets $U(t, x)$ are closed and upper semicontinuous, then they satisfy property (U); if the sets $U(t, x)$ are closed, convex, and upper semicontinuous, then they satisfy property (Q). On the other hand, there are closed sets $U(t, x)$ satisfying property (U) which are not upper semicontinuous, and there are sets $U(t, x)$ which are closed, convex, and satisfy property (Q) which are not semicontinuous. An example of this last occurrence is given by $U(t, x) = \{u = (u^1, u^2) \mid 0 \leqq u^2 \leqq tu^1, 0 \leqq u^1 < +\infty\}$, where $0 \leqq t \leqq 1, 0 \leqq x \leqq 1$.

3b. THEOREM 3. (Existence theorem for Lagrange problems with or without unilateral constraints). *Let A be any compact subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ let $U(t, x)$ be a closed subset of the u-space E_m. Let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$, and let $\tilde{f}(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function on M. For every $(t, x) \in A$ let*

$$\tilde{\tilde{Q}}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+1}$$

be a closed convex subset of E_{n+1}. Let us assume that $U(t, x)$ satisfies condition (U) in A, and that $\tilde{\tilde{Q}}(t, x)$ satisfies condition (Q) in A. Let $\varphi(z)$, $0 \leqq z < +\infty$, be a given continuous function of z satisfying the relation $\varphi(z)/z \to +\infty$ as $z \to +\infty$, and assume that $f_0(t, x, u) \geqq \varphi(|u|)$ for all $(t, x, u) \in M$. Also, let C, D be constants and assume that $f(t, x, u) \leqq C + D|u|$ for all $(t, x, u) \in M$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Then the cost functional $I[x, u]$ has an absolute minimum in any nonempty complete class Ω of admissible pairs $x(t)$, $u(t)$.

If A is not compact, but closed and contained in a slab $\{t_0 \leqq t \leqq T, -\infty < x^i < +\infty, i = 1, \cdots, m, t_0, T_0$ finite$\}$, then Theorem 3 holds under the additional hypotheses (a) and (b) at the end of Theorem 1. If A is not compact and not contained in any slab as above, but A is closed, then Theorem 3 still holds under the additional hypotheses (a), (b), (c) at the end of Theorem 1.

Also, condition (a) can be replaced in either case by condition (d) at the end of Theorem 1. Finally, for A not compact but closed, the conditions $f_0 \geqq \varphi(|u|), |f| \leqq C + D|u|$ above can be replaced by the following set

of conditions:
(e) $f_0(t, x, u) \geqq -L$ for all $(t, x, u) \in M$ and some constant L;
(f) $f_0(t, x, u) \geqq \mu > 0$ for all $(t, x, u) \in M$ with $|t| \geqq R$, and some constants $\mu > 0$, $R \geqq 0$;
(g) for every compact subset A_0 of A there are functions φ_0 as above and constants $C_0 \geqq 0$, $D_0 \geqq 0$ (all may depend on A_0) such that $f_0 \geqq \varphi_0(|u|)$, $|f| \leqq C_0 + D_0|u|$ for all $(t, x, u) \in M$ with $(t, x) \in A_0$.

For a proof of Theorem 3 we refer to [1d].

For free problems, that is, $m = n, f(t, x, u) = u$, $U = E_n$, the condition $|f| \leqq C + D|u|$ is trivially satisfied with $C = 0$, $D = 1$, the set $\tilde{Q}(t, x)$ reduces to

$$\tilde{Q}(t, x) = \{\tilde{z} = (z^0, u) \mid z^0 \geqq f_0(t, x, u), u \in E_n\} \subset E_{n+1},$$

and the hypothesis of convexity of $\tilde{Q}$ reduces to the usual convexity of $f_0(t, x, u)$ with respect to u in E_n. In addition, this convexity and the growth condition $f_0(t, x, u) \geqq \varphi(|u|)$ together assure that $\tilde{Q}(t, x)$ satisfies property (Q) in A (see the proof in [1d]). Thus, Theorem 3 contains as a particular case the Nagumo-Tonelli existence theorem for free problems [6], [11].

3c. THEOREM 4. (Existence theorem for Lagrange problems with unilateral constraints). *Let A be any compact subset of the tx-space $E_1 \times E_u$, and for every $(t, x) \in A$ let $U(t, x)$ be a closed subset of the u-space E_m. Let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$, and let $\tilde{f}(t, x, u) = (f_0, f) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function on M. For every $(t, x) \in A$ let*

$$\tilde{Q}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+1}$$

be a closed convex subset of E_{n+1}. Let us assume that $U(t, x)$ satisfies condition (U) in A and that $\tilde{Q}(t, x)$ satisfies condition (Q) in A. Let $\varphi(t)$ be a given function, which is L-integrable in any finite interval, such that $f_0(t, x, u) \geqq \varphi(t)$ for all $(t, x, u) \in M$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Let Ω be a nonempty complete class of admissible pairs $x(t)$, $u(t)$ such that

$$\int_{t_1}^{t_2} \left|\frac{dx^i}{dt}\right|^p dt \leqq N_i, \qquad i = 1, \cdots, n, \tag{4}$$

for some constants $p > 1$, $N_i \geqq 0$. Then the cost functional $I[x, u]$ has an absolute minimum in Ω.

If A is not compact, but closed and contained in a slab $\{t_0 \leqq t \leqq T, -\infty < x^i < +\infty, i = 1, \cdots, n, t_0, T_0 \text{ finite}\}$, then Theorem 4 holds under the additional hypothesis (b) at the end of Theorem 1. If A is not compact, nor contained in any slab as above, but A is closed, then Theorem 4 still holds under the additional hypotheses (b) and (c*): $f_0(t, x, u) \geqq \varphi(t)$ for all

$(t, x, u) \in M$, where $\varphi(t)$ is a given function which is L-integrable in any finite interval and

$$\int_0^{+\infty} \varphi(t)\, dt = +\infty, \quad \int_{-\infty}^{0} \varphi(t)\, dt = +\infty.$$

Finally, if for some $i = 1, \cdots, n$, and any $N > 0$, there is some $N_i > 0$ such that $(x, u) \in \Omega$, $I[x, u] \leqq N$ implies

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^p dt \leqq N_i,$$

then the corresponding requirement (4) can be disregarded.

For a proof of this theorem we refer to [1d].

3d. Existence theorems for Lagrange problems with f linear in u. We state here two corollaries of Theorems 3 and 4 for Lagrange problems with f linear in u and no unilateral constraints. For the sake of simplicity we take into consideration only a particular type of boundary conditions.

THEOREM 5. *Let us consider the Lagrange problem*

$$I[x, u] = \int_{t_1}^{t_2} [g(t, x)\Phi(u) + g_0(t, x)]\, dt = \text{minimum}, \tag{5}$$

$$\frac{dx^i}{dt} = \sum_{j=1}^{m} g_{ij}(t, x)u_j + g_i(t, x), \qquad i = 1, \cdots, n, \tag{6}$$

where $x = (x^1, \cdots, x^n) \in E_n$, $u = (u^1, \cdots, u^m) \in E_m$, *and* $\Phi(u)$ *denotes a continuous nonnegative convex function of* $u = (u^1, \cdots, u^m)$. *Assume that there is some continuous function* $\varphi(z)$, $0 \leqq z < +\infty$, *such that* $\varphi(z)/z \to +\infty$ *as* $z \to +\infty$ *and* $\Phi(u) \geqq \varphi(|\, u\,|)$ *for every* $u \in E_m$. *Assume that all* $g(t, x)$, $g_0(t, x)$, $g_{ij}(t, x)$, $g_i(t, x)$ *are continuous functions of* (t, x) *in* $E_1 \times E_n$ *such that*

$$g, g_0 \geqq \mu > 0, \qquad \sum_{ij} |\, g_{ij}\,| \leqq Cg, \qquad \sum_{ij} |\, g_{ij}\,| + \sum_{i} |\, g_i\,| \leqq Cg_0,$$

for given constants $\mu > 0$, $C > 0$ *and all* $(t, x) \in E_1 \times E_n$. *Let* Ω *be the class of all pairs* $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, $x(t)$ *absolutely continuous,* $u(t)$ *measurable, satisfying* (6), *and such that the graph* $(t, x(t))$ *joins the fixed point* $(t_1 = 0, x(t_1) = (0, \cdots, 0))$ *to a given closed subset* B *of the halfspace* $t \geqq 0$, $x \in E_n$ *of* $E_1 \times E_n$. *Then the Lagrange problem above has an optimal solution in* Ω.

The function $\Phi(u) = \varphi(|\, u\,|) = |\, u\,|^p$, $u \in E_m$, $p > 1$, certainly satisfies the requirements above for Φ. The requirement $g_0 \geqq \mu > 0$ can be disregarded if B is contained in a slab $\{0 \leqq t \leqq T, x \in E_n, T \text{ finite}\}$.

For a detailed proof of this statement as a corollary of Theorem 3, see [1d]. In the course of the proof it is shown that the subset $\tilde{\tilde{Q}}(t, x)$ of E_{n+1} relative to the problem above is convex for every $(t, x) \in A$ and satisfies condition (Q).

THEOREM 6. *Let us consider the Lagrange problem*

$$I[x, u] = \int_{t_1}^{t_2} [g(t, x)\Phi(u) + g_0(t, x)]\, dt = \text{minimum}, \tag{5}$$

$$\frac{dx^i}{dt} = \sum_{j=1}^{m} g_{ij}(t, x)u^j + g_i(t, x), \qquad i = 1, \cdots, n, \tag{6}$$

or $dx/dt = H(t, x)u + h$, *where* $x = (x^1, \cdots, x^n)$, $u = (u^1, \cdots, u^m)$, $H = (g_{ij})$, $h = (g_i)$, *and* $\Phi(u)$ *is a continuous nonnegative convex function of* u *in* E_m. *Assume that the (convex) set* $\tilde{Q}(t, x) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq g\Phi(u) + g_0,\ z = Hu + h,\ u \in E_m\}$ *satisfies condition* (Q). *Assume that all* $g(t, x)$, $g_0(t, x)$, $g_{ij}(t, x)$, $g_i(t, x)$ *are continuous functions of* (t, x) *in* $E_1 \times E_n$ *such that*

$$g(t, x) \geqq 0, \qquad g_0(t, x) \geqq -G_0 \quad \textit{for all} \quad (t, x) \in E_1 \times E_n,$$

$$g_0(t, x) \geqq \mu > 0 \quad \textit{for all} \quad (t, x) \in E_1 \times E_n \quad \textit{with} \quad |t| \geqq D_0,$$

for some constants $\mu > 0$, $G_0 \geqq 0$, $D_0 \geqq 0$. *Let* Ω *be the class of all pairs* $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, $x(t)$ *absolutely continuous,* $u(t)$ *measurable, such that the graph* $(t, x(t))$ *joins the fixed point* $(t_1 = 0, x(t_1) = (0, \cdots, 0))$ *to a given closed subset* B *of the halfspace* $t \geqq 0$, $x \in E_n$ *of* $E_1 \times E_n$, *and such that*

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^p dt \leqq N_i, \qquad i = 1, \cdots, n, \tag{7}$$

for some constants $p > 1$, $N_i \geqq 0$. *If* Ω *is not empty then the Lagrange problem above has an optimal solution in* Ω.

The requirement $g_0 \geqq \mu > 0$ can be disregarded if B is contained in a slab $\{0 \leqq t \leqq T, x \in E_n, T \text{ finite}\}$.

If $g(t, x) \geqq \mu > 0$, where μ is a constant, and if $\Phi(u) \geqq \varphi(|u|)$, where $\varphi(z)$, $0 \leqq z < +\infty$, is a nonnegative function of z with $\varphi(z) \to +\infty$ as $z \to +\infty$, then $\tilde{Q}$ certainly satisfies condition (Q).

Also, any of the n requirements (7) which is a consequence of a relation of the form

$$\int_{t_1}^{t_2} [g\Phi + g_0]\, dt \leqq N,$$

can be disregarded. For a detailed proof of Theorem 6 as a corollary of Theorem 4, see [1d]. We shall denote by $r = r(t, x)$ the rank of the matrix $H(t, x)$.

Example 1. Let us consider the (free) problem

$$I[x] = \int_{t_1}^{t_2} (1 + |x'|^2)\, dt = \text{minimum},$$

with $x = (x^1, \cdots, x^n)$ in the class Ω of all absolutely continuous functions $x(t) = (x^1, \cdots, x^n)$, $0 \leqq t \leqq t_2$, whose graph $(t, x(t))$ joins the point

$(t_1 = 0, x(t_1) = (0, \cdots, 0))$ to a nonempty closed subset B of the half-space $t_2 \geqq 0$, $x \in E_n$. This problem can be written as a Lagrange problem with

$$J[x, u] = \int_{t_1}^{t_2} (1 + |u(t)|^2)\, dt = \text{minimum},$$

$$\frac{dx^i}{dt} = u^i, \qquad i = 1, \cdots, n,$$

where $x(t) = (x^1, \cdots, x^n)$, $u(t) = (u^1, \cdots, u^n)$, $m = n$, $f_0 = 1 + |u|^2$, $f_i = u^i$, $i = 1, \cdots, n$, and the control space $U(t, x)$ is fixed and coincides with the whole space E_n. Here $\tilde{\tilde{Q}}(t, x) = \{(z, u) \mid z \geqq 1 + |u|^2, u \in E_n\}$ is a fixed and convex subset of E_{n+1}. The conditions of Theorem 5 are satisfied with $g = 1$, $g_0 = 1$, $\Phi(u) = \varphi(|u|) = |u|^2$, or $\varphi(z) = z^2$, $0 \leqq z < +\infty$, $A = \{(t, x) \mid t \geqq 0, x \in E_n\} \subset E_{n+1}$, $g_{ii} = 1$, $g_{ij} = 0$ for $i \neq j$, $g_i = 0$. Thus, the problem above has an optimal solution.

Example 2. The (free) problem

$$I[x] = \int_0^1 tx'^2\, dt = \text{minimum}, \qquad x(0) = 1, \qquad x(1) = 0,$$

is known to have no solution [11c, vol. 3, p. 91]. The same problem written as a Lagrange problem

$$J_1[x, u] = \int_0^1 tu^2\, dt = \text{minimum}, \qquad x(0) = 1, \qquad x(1) = 0,$$

$$\frac{dx}{dt} = u, \qquad u \in U = E_1,$$

gives rise to sets $\tilde{\tilde{Q}}$:

$$\tilde{\tilde{Q}}(t) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq tu^2, z = u, u \in E_1\} \subset E_2, \qquad 0 \leqq t \leqq 1,$$

which are convex and do satisfy property (Q) ($g = t$, $g_0 = 0$, $g_{11} = 1$, $g_1 = 0$, constant rank $r = 1$). The same problem written as a Lagrange problem

$$J_2[x, u] = \int_0^1 t^3u^2\, dt = \text{minimum}, \quad x(0) = 1, \quad x(1) = 1,$$

$$\frac{dx}{dt} = tu, \quad u \in U = E_1,$$

gives rise to sets $\tilde{\tilde{Q}}$:

$$\tilde{\tilde{Q}}(t) = \{\tilde{z} = (z^0, z) \mid z^0 \geqq t^3u^2, z = tu, u \in E_1\} \subset E_2, \qquad 0 \leqq t \leqq 1,$$

which are still convex but do not satisfy property (Q). Indeed, for $t = 0$

we have $\tilde{Q}(0) = \{(z^0, z) \mid z^0 \geqq 0, z = 0\}$, that is, the positive half z^0-axis, while each set $\tilde{Q}(t)$ for $t > 0$ is the set of all points $(z^0, z) \in E_2$ above or on the line $z^0 = tz^2$, and then $\bigcap_\delta \operatorname{cl} \operatorname{co} \tilde{Q}(t, \delta) = P$ is the entire positive halfplane $z^0 \geqq 0$, that is, $\tilde{Q}(0) \neq P$. (Note that $g = t^3$, $g_0 = 0$, $g_{11} = t$, $g_1 = 0$, hence, $r = 0$ for $t = 0$ and $r = 1$ for $0 < t \leqq 1$. Thus, r is not constant in any neighborhood of $t = 0$, and $g = 0$ at $t = 0$.) In either case J_1, or J_2, the growth condition $f_0 \geqq \varphi(|u|)$ with $\varphi(z)/z \to +\infty$ as $z \to +\infty$ is not satisfied at $t = 0$. The problem

$$I[x] = \int_0^1 tx'^2\, dt = \text{minimum}, \qquad x(0) = 1, \qquad x(1) = 0,$$

$$\int_0^1 x'^2\, dt \leqq N_0,$$

for any given constant $N_0 \geqq 1$, has an optimal solution. Indeed as a Lagrange problem ($J_1 = \text{minimum}$, $dx/dt = u$, $u \in E_1$), it satisfies the conditions of Theorem 6. Of course, the optimal solution and the value of the minimum may depend on N_0. The class Ω is not empty for $N_0 \geqq 1$ since, for $x(t) = 1 - t$, we have

$$\int_0^1 x'^2\, dt = 1.$$

3e. Existence theorems for weak solutions. Instead of considering the usual cost functional, differential equations and constraints,

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt,$$

$$\frac{dx}{dt} = f(t, x(t), u(t)), \qquad f = (f_1, \cdots, f_n),$$

$$(t, x(t)) \in A, \qquad u(t) \in U(t, x(t)), \qquad (t_1, x(t_1), t_2, x(t_2)) \in B,$$

we have to consider the new cost functional, differential equations and constraints,

$$J[x, p, v] = \int_{t_1}^{t_2} g_0(t, x(t), p(t), v(t))\, dt,$$

$$\frac{dx}{dt} = g(t, x(t), p(t), v(t)), \qquad (t_1, x(t_1), t_2, x(t_2)) \in B,$$

$$(t, x(t)) \in A, \qquad (p(t), v(t)) \in \Gamma \times V(t, x(t)).$$

Here $x = (x^1, \cdots, x^n)$, $p = (p_j,\ j = 1, \cdots, \nu)$, $\nu \geqq n + 1$, $v = (u^{(j)},\ j = 1, \cdots, \nu)$, $\Gamma \equiv \{p \mid p_j \geqq 0,\ p_1 + \cdots + p_\nu = 1\}$,

$u^{(j)} \in U(t, x)$, and therefore the new control variable (p, v) takes on its values in the set $\Gamma \times V(t, x)$, where $V = [U]^\nu$ is the logic product of U by itself taken ν times. In other words, $v(t) = (u^{(1)}, \cdots, u^{(\nu)})$ represents a finite system of $\nu \geqq n + 1$ ordinary strategies $u^{(1)}, \cdots, u^{(\nu)}$, each $u^{(j)}$ having its values in U, that is, $u^{(j)}(t) \in U(t, x(t)) \subset E_m$, $j = 1, \cdots, \nu$. Thus, $v = (u^{(1)}, \cdots, u^{(\nu)})$ is a vector variable whose ν components $u^{(1)}, \cdots, u^{(\nu)}$ are themselves vectors with values in $U(t, x) \subset E_m$, hence $v(t) \in V(t, x(t)) \subset E_{m\nu}$. Above, $p(t) = (p_1, \cdots, p_\nu)$ represents a probability distribution; hence p is an element of the simplex $\Gamma \subset E_\nu$ defined above, and for the new control variable we have $(p(t), v(t)) \in \Gamma \times V(t, x(t)) \subset E_{\nu+m\nu}$. As usual, we denote by g and $\tilde{g}$ the two vectors $g = (g_1, \cdots, g_n)$, $\tilde{g} = (g_0, g_1, \cdots, g_n) = (g_0, g)$, and then

$$g_i(t, x, p, v) = \sum_{j=0}^{n} p_j f_i(t, x, u^{(j)}), \qquad i = 0, 1, \cdots, n.$$

As usual, we require that all functions $p_j(t)$, $u^{(j)}(t)$ are measurable, and that $x(t)$ is absolutely continuous.

We say that $[p(t), v(t)]$ is a generalized strategy, that $p(t) = (p_1, \cdots, p_\nu)$ is a *probability distribution*, and that $x(t)$ is a *generalized trajectory*. We shall also say for the sake of brevity that $[x(t), p(t), v(t)]$ is a *weak solution*.

If we introduce, as usual, the auxiliary variable x^0 with initial values $x^0(t_1) = 0$, and the vector $\tilde{x} = (x^0, x) = (x^0, x^1, \cdots, x^n)$, then instead of the system $d\tilde{x}/dt = \tilde{f}$, we shall consider the system

$$\frac{d\tilde{x}}{dt} = \tilde{g}(t, x(t), p(t), v(t)), \qquad \tilde{g} = (g_0, g) = (g_0, g_1, \cdots, g_\nu),$$

and as usual we have

$$J[x, p, v] = x^0(t_2).$$

Instead of the usual sets $Q(t, x) = f[t, x, U(t, x)] \subset E_n$ and

$$\tilde{Q}(t, x) = \tilde{f}(t, x, U(t, x)) = \{\tilde{z} = (z^0, z) \mid \tilde{z} = \tilde{f}(t, x, u), u \in U(t, x)\} \subset E_{n+1},$$

we shall now consider the sets

$$R(t, x) = g[t, x, \Gamma \times V(t, x)] = \{z \mid z = g(t, x, p, v), (p, v) \in \Gamma \times V(t, x)\} \subset E_n,$$

$$\tilde{R}(t, x) = \tilde{g}[t, x, \Gamma \times V(t, x)] = \{\tilde{z} = (z^0, z) \mid \tilde{z} = \tilde{g}(t, x, p, v), (p, v) \in \Gamma \times V(t, x)\} \subset E_{n+1}.$$

Since

$$\tilde{R}(t, x) = \Big\{\tilde{z} = (z^0, z) \mid \tilde{z} = \sum_{j=1}^{\nu} p_j \tilde{f}(t, x, u^{(j)}),$$

$$p \in \Gamma, u^{(j)} \in U(t, x), j = 1, \cdots, \nu\Big\},$$

with $\nu \geqq n + 1$, we see that $\tilde{R}(t, x)$ is the convex hull of the set $\tilde{Q}(t, x)$ in E_{n+1}, and hence $\tilde{R}(t, x)$ is always convex. For weak solutions there is no reason to consider sets analogous to the sets $\tilde{\tilde{Q}}(t, x)$. Any usual admissible pair $[x(t), u(t)]$ can be thought of as a generalized element $[y(t), p(t), v(t)]$ by taking $p_j(t) = 1/\nu$, $j = 1, \cdots, \nu$, $u^{(j)}(t) = u(t)$, and then $y = x$, $y^0 = x^0$, $J = I$.

Let Ω be the class of all admissible pairs, say $[\tilde{x} = (x^0, x), u(t)]$, satisfying differential equations, constraints, and boundary conditions, and let Ω^* be the class of all generalized elements $[\tilde{y}(t) = (y^0, y), p(t), v(t)]$ satisfying the corresponding differential equations and constraints, and the same boundary conditions. As mentioned above we have $\Omega \subset \Omega^*$. If

$$i = \inf_{\Omega} I[x, u], \qquad j = \inf_{\Omega^*} J[y, p, v],$$

then $\Omega \subset \Omega^*$ implies $i \geqq j$.

It is a fairly general phenomenon that generalized trajectories and corresponding values of J can be approached by means of usual trajectories and corresponding values of I, so that $i = j$. We shall say that *property* (P) holds whenever $i = j$.

Under the hypotheses that $A = E_1 \times E_n$, that $U(t, x)$ depends on t only, that $U(t)$ is compact for every t, that $U(t)$ is an upper semicontinuous function of t, that $\tilde{f}$ satisfies a Lipschitz condition, and that Ω is the class of all admissible pairs $x(t)$, $u(t)$ (satisfying the differential equations, the constraints, and the given boundary conditions), Gamkrelidze [3] has proved that property (P) is always valid.

In the present more general situation—A any closed set, $U(t, x)$ depending both on t and x, $U(t, x)$ closed, $U(t, x)$ satisfying condition (U), Ω any given class of admissible pairs $x(t)$, $u(t)$—a proof of property (P), that is, that $i = j$, is much more difficult. Particularly the condition that $U(t, x)$ depends on both t and x gives rise to a number of difficulties. Nevertheless, we were able to prove property (P) under a set of simple requirements which are satisfied in most cases and which are easier to verify than property (P). We refer to [1d] for the requirements and for the corresponding statement, which contains the remark of Gamkrelidze as a particular case.

Analogously, in Existence Theorem 8 we shall denote by Ω_0 the class of all usual admissible pairs $[\tilde{x}(t) = (x^0, x), u(t)]$ of Ω satisfying the given inequalities

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^p dt \leqq N_i , \qquad i = 1, \cdots, n,$$

for certain constants $N_i \geqq 0$, $p > 1$, and we shall denote by Ω_0^* the class of all generalized elements $[\tilde{y}(t) = (y^0, y), p(t), v(t)]$ of Ω^* satisfying the same inequalities. Then $\Omega_0 \subset \Omega_0^*$. If i_0 and j_0 denote the infimum of I in Ω_0 and the infimum of J in Ω_0, then again we have $i_0 \geqq j_0$. We shall say that property (P_0) holds whenever $i_0 = j_0$. We have proved that simple requirements analogous to the ones for (P) assure that also property (P_0) holds (see [1d]).

For definitions of generalized or weak solutions slightly different from the one of Gamkrelidze used above, we refer to Young [14], McShane [4], Warga [12], and Wazewski [13]. Property (P) is proved by Warga and Wazewski under assumptions different from Gamkrelidze's.

THEOREM 7. (Existence theorem for weak solutions). *Let A be a compact subset of the tx-space $E_1 \times E_n$, let $U(t, x)$ be a closed subset of E_m for every $(t, x) \in A$, and let $\tilde{f}(t, x, u) = (f_0, \cdots, f_n)$ be a continuous vector function on the set M of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let us assume that there is some continuous scalar function $\varphi(z)$, $0 \leqq z < +\infty$, with $\varphi(z)/z \to +\infty$ as $z \to +\infty$, such that $f_0(t, x, u) \geqq \varphi(|u|)$ for all $(t, x, u) \in M$ and that there are constants C, $D \geqq 0$ such that $|f(t, x, u)| \leqq C + D|u|$ for all $(t, x, u) \in M$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Let us assume that $U(t, x)$ satisfies property* (U) *in A, that $\tilde{R}(t, x)$ satisfies property* (Q) *in A, that property* (P) *holds, and that Ω is not empty. Then the infimum i of $I[x, u]$ in Ω is attained by a weak solution (that is, i is attained by $J[x, p, v]$ in the class Ω^*).*

When A is not compact, but closed, then the theorem still holds under the additional hypotheses stated at the end of Existence Theorem 1.

THEOREM 8. (Existence theorem for weak solutions). *Let A be a compact subset of the tx-space $E_1 \times E_n$, and for every $(t, x) \in A$ let $U(t, x)$ be a closed subset of the u-space E_m. Let $\tilde{f}(t, x, u) = (f_0, f_1, \cdots, f_n)$ be a continuous vector function on the set M of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$. Let B be a closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}. Let us assume that the set $U(t, x)$ satisfies property* (U) *in A, and that the set $\tilde{R}(t, x)$ satisfies property* (Q) *in A. Let us assume that $f_0(t, x, u) \geqq -G_0$ for some constant $G_0 \geqq 0$ and all $(t, x, u) \in M$. Let Ω_0 be the class of all admissible pairs $[\tilde{x}(t) = (x^0, x), u(t)]$ of Ω satisfying the inequalities*

$$\int_{t_1}^{t_2} \left| \frac{dx^i}{dt} \right|^p dt \leqq N_i , \qquad i = 1, \cdots, n, \tag{7}$$

for some constants $N_i \geqq 0$, $p > 1$, *and let* Ω_0^* *be the analogous subclass of all generalized elements* $[\tilde{y}(t) = (y^0, y), p(t), v(t)]$ *of* Ω^* *satisfying the same inequalities. Assume that* Ω_0 *is not empty, and property* (P_0) *holds. Then the infimum* i *of* $I[x, u]$ *in* Ω_0 *is attained by a weak solution (that is,* i *is attained by* $J[y, p, v]$ *in* Ω_0^**).*

When A is not compact, but A is closed, then Theorem 8 still holds under the additional hypotheses stated at the end of Theorem 4. As for Theorem 4, every inequality (7) which is a consequence of an inequality of the form $J \leqq N$ can be disregarded.

4. Appendix.

4a. Deduction of Theorem 1 from Theorems 3 and 4. First let us consider the case of A compact. Under the conditions of Theorem 1 and A compact, M is also a compact set, and hence $|f_0(t, x, u)| \leqq N$, $|f(t, x, u)| \leqq N$ for some constant N and all $(t, x, u) \in M$. On the other hand we know that $U(t, x)$ satisfies condition (U) (see §3a). Also, $\tilde{Q}(t, x) = \tilde{f}(t, x, U(t, x))$ is necessarily compact and upper semicontinuous, and so is the set $\tilde{\tilde{Q}}_N(t, x) = \tilde{\tilde{Q}}(t, x) \cap \{z^0 \leqq N\}$. If $Q(t, x)$ is convex, then $\tilde{\tilde{Q}}_N(t, x)$ is convex, compact, and upper semicontinuous, hence $\tilde{\tilde{Q}}_N(t, x)$ satisfies property (Q) (§3a), and so obviously $\tilde{\tilde{Q}}(t, x)$ does also. Finally, because of the boundedness of f_0 and f, the growth conditions of Theorem 3, say, $f_0 \geqq \varphi(|u|)$, $|f| \leqq C + D|u|$, are trivially satisfied. Thus, for A compact, Theorem 1 is a corollary of Theorem 3. Analogously, $dx^i/dt = f_i$, $i = 1, \cdots, n$, almost everywhere, implies $|dx^i/dt| \leqq N$, and relations (4) are trivially satisfied. Thus, for A compact, Theorem 1 is a corollary of Theorem 4 also. For A closed and hypotheses (a), (b), or (a), (b), (c), we can reduce A to a convenient compact subset A_0 of A by well-known arguments which are given at the end of §4b below. Thus, Theorem 1 is a particular case of Theorem 3 as well as of Theorem 4.

4b. Direct proof of Theorem 1. Since A is compact and $U(t, x)$ is compact for every (t, x) and an upper semicontinuous function of (t, x) in A, we deduce that M is compact; hence $\tilde{f}(t, x, u)$ is bounded on M, say, $|\tilde{f}(t, x, u)| \leqq N$ for all $(t, x, u) \in M$.

Let us consider the auxiliary variables u^0 and x^0, and let $\tilde{u}$ be the $(m + 1)$-vector $\tilde{u} = (u^0, u) = (u^0, u^1, \cdots, u^m)$, and $\tilde{x}$ be the $(n + 1)$-vector $\tilde{x} = (x^0, x) = (x^0, x^1, \cdots, x^n)$. For every $(t, x) \in A$ let $\tilde{U}(t, x)$ be the compact set

$$\tilde{U}(t, x) = \{\tilde{u} = (u^0, u) \mid N \geqq u^0 \geqq f_0(t, x, u),\ u \in U(t, x)\} \subset E_{m+1},$$

let $\tilde{f}_0(t, x, \tilde{u})$ be defined by $\tilde{f}_0(t, x, \tilde{u}) = u^0$, and let $\tilde{\tilde{f}}(t, x, \tilde{u})$ be the $(n + 1)$-vector function $\tilde{\tilde{f}}(t, x, \tilde{u}) = (\tilde{f}_0, f) = (\tilde{f}_0, f_1, \cdots, f_n)$, which is continuous

for all $(t, x, \tilde{u})$ with $(t, x) \in A$, $\tilde{u} = (u^0, u) \in \tilde{U}(t, x)$, and

$$\tilde{\tilde{f}}_0(t, x, \tilde{u}) = \tilde{f}_0(t, x, \tilde{u}) = u^0, \qquad \tilde{\tilde{f}}_i(t, x, \tilde{u}) = f_i(t, x, u), \quad i = 1, \cdots, n.$$

Let $J[x, \tilde{u}]$ be the auxiliary cost functional

$$J[x, \tilde{u}] = \int_{t_1}^{t_2} u^0 \, dt.$$

We introduce for x^0 the auxiliary differential equation and initial condition

$$x^0(t_1) = 0, \qquad \frac{dx^0}{dt} = \tilde{f}_0(t, x, \tilde{u}) = u^0.$$

Now the set $\tilde{Q}(t, x)$ (precisely the part of $\tilde{Q}$ with $u^0 \leqq N$) has the following simple interpretation

$$\begin{aligned} \tilde{Q}(t, x) &= \{\tilde{z} = (z^0, z) \mid N \geqq z^0 \geqq f_0(t, x, u), z = f(t, x, u), \\ &\qquad u \in U(t, x)\} \\ &= \{\tilde{z} = (z^0, z) \mid z^0 = u^0, z = f(t, x, u), \\ &\qquad N \geqq u^0 \geqq f_0(t, x, u), u \in U(t, x)\} \\ &= \{\tilde{z} = (z^0, z) \mid \tilde{z} = \tilde{\tilde{f}}(t, x, \tilde{u}), \tilde{u} = (u^0, u) \in \tilde{U}(t, x)\} \\ &= \tilde{\tilde{f}}(t, x, \tilde{U}(t, x)), \end{aligned} \tag{8}$$

in other words, $\tilde{\tilde{Q}}$ is the image of $\tilde{U}$ with respect to $\tilde{\tilde{f}}$.

For each pair $x(t)$, $u(t)$ which is admissible for $I[x, u]$, we let correspond the pair $\tilde{x}(t) = (x^0, x)$, $\tilde{u}(t) = (u^0, u)$ with

$$\frac{dx^0}{dt} = u^0(t) = f_0(t, x(t), u(t)), \qquad \frac{dx}{dt} = f(t, x(t), u(t)),$$

$$(t, x(t)) \in A, u(t) \in U(t, x(t)),$$

and hence $\tilde{u}(t) \in \tilde{U}(t, x(t))$, that is, the pair $\tilde{x}(t)$, $\tilde{u}(t)$ is admissible for $J[x, \tilde{u}]$ with $J[x, \tilde{u}] = I[x, u]$. Conversely, if $\tilde{x}(t) = (x^0, x)$, $\tilde{u}(t) = (u^0, u)$ is admissible for $J[\tilde{x}, \tilde{u}]$, then

$$\frac{dx^0}{dt} = u^0(t) \geqq f_0(t, x(t), u(t)), \qquad \frac{dx}{dt} = f(t, x(t), u(t)),$$

with $(t, x(t)) \in A$, $u(t) \in U(t, x(t))$, and $-N \leqq f_0(t, x(t), u(t)) \leqq u^0(t) \leqq N$. Thus $f_0(t, x(t), u(t))$ is L-integrable, and $(x(t), u(t))$ is an admissible pair for $I[x, u]$ with $J[x, \tilde{u}] \geqq I[x, u]$.

Now let i be the infimum of $I[x, u]$ in Ω. Then i is finite, and we may take a (minimizing) sequence $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, of ad-

missible pairs (for I) with $I[x_k, u_k] \to i$. If we consider the corresponding pairs $\tilde{x}_k(t) = (x_k^0, x_k)$, $\tilde{u}_k(t) = (u_k^0, u_k)$ relatively to J we have

$$u_k^0(t) = f_0(t, x_k(t), u_k(t)), \qquad u_k(t) \in U(t, x_k(t)),$$

$$\frac{dx_k^0}{dt} = u_k^0(t) = f_0(t, x_k(t), u_k(t)), \qquad \frac{dx_k}{dt} = f(t, x_k(t), u_k(t)).$$

Since $|f_0(t, x_k(t), u_k(t))| \leqq N$, $|f(t, x_k(t), u_k(t))| \leqq N$, $(t, x(t)) \in A$, $|x^0(t)| \leqq ND$, where D is the diameter of A, we conclude that the sequence $\tilde{x}_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, is equibounded and equicontinuous, and by Ascoli's theorem there exists a subsequence, say still $x_k(t)$, which is convergent in the ρ-metric toward a continuous vector function $\tilde{x}(t) = (\tilde{x}_0, x)$, $t_1 \leqq t \leqq t_2$, and $\tilde{x}(t)$ is also Lipschitzian. The usual Filippov's argument [2a] applied to J assures now—in view of (8)—that $\tilde{x}(t)$ is a trajectory for J, that is, there is a measurable vector $\tilde{u}(t) = (u^0, u)$ with

$$\frac{dx^0}{dt} = u^0(t) \geqq f_0(t, x(t), u(t)), \qquad \frac{dx}{dt} = f(t, x(t), u(t)),$$

$$(t, x(t)) \in A, \qquad u(t) \in U(t, x(t)).$$

Thus $x(t)$, $u(t)$ is admissible for I and even belongs to Ω since Ω is complete. Thus

$$i = J[\tilde{x}, \tilde{u}] = \int_{t_1}^{t_2} u^0(t)\, dt \geqq \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt = I[x, u] \geqq i,$$

where i is finite. Thus, we must have $u^0(t) = f_0(t, x(t), u(t))$ almost everywhere in $[t_1, t_2]$, and

$$i = J[\tilde{x}, \tilde{u}] = \int_{t_1}^{t_2} u^0(t)\, dt = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt = I[x, u] = i,$$

that is, i is attained by $I[x, u]$ in Ω. This proves Theorem 1 in the case A is compact. For A closed and contained in the slab $\{t_0 \leqq t \leqq T, -\infty < x_i < +\infty, i = 1, \cdots, n, t_0, T$ finite$\}$ we deduce from hypotheses (a) and (b) that

$$0 \leqq |x(t)|^2 + 1 \leqq [|x(t^*)|^2 + 1] \exp C(T - t_0),$$

and since $(t^*, x(t^*)) \in P$, where P is a given compact subset of A, we conclude that $|x(t)| \leqq D_0$ for a convenient constant D_0. Then we may restrict A to the subset A_0 of all $(t, x) \in A$ with $|x| \leqq D_0$, and A_0 is compact.

Let us consider now the last case, where A is any closed subset of $E_1 \times E_n$, but the additional conditions (a), (b), and (c) hold. Let $\bar{x}(t)$, $\bar{u}(t)$ be an admissible pair contained in Ω, and let j denote the corresponding value of the cost functional. Let D be a number large enough so that $D \geqq N_1$

and the interval $[-D, D]$ contains in its interior the projection P_0 of P on the t-axis. Let $L = \mu^{-1}[2\,DG + |\,j\,| + 1]$, and $D_0 = D + L$. If any admissible pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, of Ω possesses a point $(t_0, x(t_0))$ with $|\,t_0\,| > D_0$, then $x(t)$ possesses also a point $(t^*, x(t^*)) \in P$ with $t_1 \leqq t^* \leqq t_2$, $|\,t^*\,| \leqq D$. Thus, there is a subarc Γ of $(t, x(t))$, $t' \leqq t \leqq t''$, along which $|\,t\,| \leqq D$ and $f_0 \geqq -G$, while on the remaining part, say E, of $(t, x(t))$ we have $|\,t\,| \geqq N_1$, $f_0 \geqq \mu > 0$, and E contains at least one interval of length L. Then

$$I[x, u] = \int_{t_1}^{t_2} f_0\,dt = \left(\int_{t'}^{t''} + \int_E\right) f_0\,dt \geqq -2DG + \mu L$$

$$= |\,j\,| + 1 \geqq j + 1.$$

Since we are seeking the minimum of $I[x, u]$, there is no reason to consider those pairs $x(t)$, $u(t)$ of Ω for which $I[x, u] \geqq j + 1$. Thus, we may restrict A to the closed subset A_0 of all $(t, x) \in A$ with $|\,t\,| \leqq D_0$, and we are reduced to the previous case.

Let us prove now that condition (a) can be replaced by condition (d). There are numbers $E > 0$, $F \geqq 0$ such that $f_0(t, x, u) \geqq E\,|\,f(t, x, u)\,|$ for all $(t, x, u) \in M$ with $|\,x\,| \geqq F$. Let us assume first that A is closed and contained in a slab $\{t_0 \leqq t \leqq T, x \in E_n\}$ as above, and that condition (b) holds. Let us take the number $F \geqq 0$ so large that the projection P^* of P on the x-space E_n is completely in the interior of the solid sphere $|\,x\,| \leqq F$, and also so large that $F \geqq T - t_0$. Let $\bar{x}(t)$, $\bar{u}(t)$ be an admissible pair contained in Ω, and let j denote the corresponding value of the cost functional. Let $L = E^{-1}(FG + |\,j\,| + 1)$, and let us take $F_0 = F + L$. If any admissible pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, of Ω possesses a point $(t_0, x(t_0))$ with $|\,x(t_0)\,| \geqq F_0$, then $x(t)$ possesses also a point $(t^*, x(t^*)) \in P$, with $t_1 \leqq t^* \leqq t_2$, $|\,x(t^*)\,| \leqq F$. Thus, there is at least one subarc $\Gamma : x = x(t)$, $t' \leqq t \leqq t''$, of the graph $(t, x(t))$ along which $|\,x(t)\,| \geqq F$ and $|\,x(t)\,|$ passes from the value F to the value $F_0 = F + L$. Such an arc Γ has a length $\geqq L$. If E denotes the part of $[t_1, t_2]$ not covered by $[t', t'']$, then

$$I[x, u] = \int_{t_1}^{t_2} f_0\,dt = \left(\int_E + \int_{t'}^{t''}\right) f_0\,dt \geqq -FG + \int_{t'}^{t''} E\,|\,f\,|\,dt$$

$$= -FG + E\int_{t'}^{t''} \left|\frac{dx}{dt}\right| dt \geqq -FG + EL \geqq |\,j\,| + 1.$$

Since we are seeking for the minimum of $I[x, u]$ in Ω there is no reason to consider those pairs $x(t)$, $u(t)$ of Ω for which $I[x, u] \geqq j + 1$. Thus, we may restrict A to the compact subset A_0 of all $(t, x) \in A$ with $|\,x\,| \leqq F_0 = F + L$. We shall now consider the case in which A is not compact, nor is

A contained in any slab as above, but A is closed and conditions (b), (c), (d) hold. The previous argument proves that we can limit ourselves to the part A_1 of all $(t, x) \in A$ with $-L_1 \leqq t \leqq L_1$ for some L_1 sufficiently large, and then the argument above applies again.

REFERENCES

[1a] L. CESARI, *Semicontinuità e convessità nel calcolo delle variazioni*, Ann. Scuola Norm. Sup. Pisa, 18 (1964), pp. 389–423.

[1b] ———, *Un teorema di esistenza in problemi di controlli ottimi*, Ibid., 19 (1965), pp. 35–78.

[1c] ———, *An existence theorem in problems of optimal control*, this Journal, 3 (1965), pp. 7–22.

[1d] ———, *Existence theorem for weak and usual optimal solutions in Lagrange problems with unilateral constraints: I, Existence theorems; II, Existence theorems for weak solutions*, to appear.

[2a] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astronom., 2 (1959), pp. 25–32; English transl., this Journal, 1 (1962), pp. 76–84.

[2b] ———, *Differential equations with many-valued discontinuous right-hand side*, Doklady Akad. Nauk SSSR, 151 (1963), pp. 65–68; English transl., Soviet Math. Dokl., 4 (1963), pp. 941–945.

[3] R. V. GAMKRELIDZE, *Optimal sliding states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245; English transl., Soviet Math. Dokl., 3 (1962), pp. 559–561.

[4a] E. J. MCSHANE, *Curve-space topologies associated with variational problems*, Ann. Scuola Norm. Sup. Pisa, 9 (1940), pp. 45–60.

[4b] ———, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.

[4c] ———, *Necessary conditions in generalized-curve problems of the calculus of variations*, Ibid., 7 (1940), pp. 1–27.

[4d] ———, *Existence theorems for Bolza problems in the calculus of variations*, Ibid., 7 (1940), pp. 28–61.

[4e] ———, *A metric in the space of generalized curves*, Ann. of Math., 52 (1950), pp. 328–349.

[5] L. MARKUS AND E. B. LEE, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[6] M. NAGUMO, *Über die gleichmässige Summierbarkeit und ihre Anwendung auf ein Variationsproblem*, Japan J. Math., 6 (1929), pp. 173–182.

[7] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[8] L. S. PONTRYAGIN, *Optimal control processes*, Uspehi Mat. Nauk, 14 (85) (1959), pp. 3–20; English transl., Automation Express, 1 (1959), pp. 15–17; 2 (1959), pp. 26–30.

[9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962, or Pergamon, London, 1964.

[10] E. O. ROXIN, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[11a] L. TONELLI, *Sugli integrali del calcolo delle variazioni in forma ordinaria*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 401–450; Opere Scelte, Cremonese, Roma, vol. 3, 1962, pp. 192–254.

[11b] ———, *Fondamenti di Calcolo delle Variazioni*, Zanichelli, Bologna, 1921-1923.
[11c] ———, *Opere Scelte*, Cremonese, Roma, 1962.
[12] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.
[13a] T. WAZEWSKI, *Sur une généralization de la notion des solutions d'une équation au contingent*, Bull. Acad. Polon. Sci. Ser. Sci. Astronom. Phys., 10 (1962), pp. 11–15.
[13b] ———, *Sur les systèmes de commande non linéaires dont le contredomaine de commande n'est pas forcement convexe*, Ibid., 10 (1962), pp. 17–21.
[14] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, Compt. Rend. Soc. Sci. Lettr. Varsovie Cl III, 30 (1937), pp. 212–234.

J. SIAM CONTROL
Ser. A, Vol. 3, No. 3
Printed in U.S.A., 1966

SUFFICIENT CONDITIONS FOR THE OPTIMALITY OF A STOCHASTIC CONTROL*

HAROLD J. KUSHNER†

1. Introduction. This paper is concerned with the optimal control of diffusion processes governed by a vector stochastic differential (Ito) equation, where both the control and the equation coefficients are subject to local Lipschitz conditions, and the terminal time is the (random) time of entrance into a given set. In §2, theorems giving a sufficient condition for optimality are proved. The theorems are based on some results of Dynkin [1] and are more general than the corresponding theorems in [2], in that compactness of the state space is not required. The sufficient condition is a stochastic analog of the Hamilton-Jacobi equation approach for deterministic systems.

In §3, we obtain the optimal control corresponding to a stochastic version of a minimum average time problem in [3]. Section 1 contains some introductory material, and the problem formulation. We note that some of these results have been obtained (by the author and others) by a formal application of dynamic programming, but the work here is believed to be the first (excepting results in [2]) nonformal approach.

The problem formulation. The control system is governed by the vector stochastic differential equation

$$dx = f(x, u)\,dt + \sigma(x, u)\,dz, \tag{1.1}$$

where z is a vector of independent Brownian motion processes, $\sigma(x, u)$ is a matrix and u is the control. u will be a function of x only, but the functional dependence will usually be suppressed for notational simplicity. Define the matrix $S(x, u)$ with components $S_{ij}(x, u)$ by

$$\{S_{ij}(x, u)\} = S(x, u) = \sigma(x, u)\sigma'(x, u),$$

where the prime $'$ represents the transpose. Let x_t be the value of x at time t, and x_i and x_{it}, $i \geqq 1$, the ith component of x and its value at t, respectively. Define $\| x \| = (x'x)^{1/2}$ and $\| \sigma \| = (\sum_{i,j} \sigma_{ij}^2)^{1/2}$. The range of x_t is in a Euclidean space E.

If $f(x, u)$, $\sigma(x, u)$, and $u(x)$ satisfy a uniform Lipschitz condition in all

* Received by the editors July 19, 1965.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island. This research was supported in part by the National Aeronautics and Space Administration under Grant No. NGR-40-002-015 and in part by the United States Air Force through the Office of Scientific Research under Grant No. AF-AFOSR-693-65.

their arguments, then it is well-known that, w.p.1 (with probability one), there is a unique solution to (1.1) which is a Markov process and has continuous paths in the time interval $[0, \infty)$.

A random variable τ is said to be a random time (or Markov time) for a Markov process x_t if the event $\{\tau < t\}$ for any t can be verified by observing x_s, $s \leqq t$ ($\{\tau < t\}$ is in the σ-field generated by x_s, $s \leqq t$). The constant t is a random time. If τ_1 and τ_2 are random times, then so is

$$\tau = \min(\tau_1, \tau_2) \equiv \tau_1 \cap \tau_2 .$$

Let G be an open set with compact closure in E. Then the τ_1 of (1.2) is a random time ((1.2) has the interpretation that, if $x_t \in G$ for all $t < \infty$, then $\tau_1 = \infty$; otherwise τ_1 is the first exit time from G).

$$\tau_1 = \inf_t \{t : x_t \notin G\}. \tag{1.2}$$

Now, let $f(x, u)$, $u(x)$, and $\sigma(x, u)$ satisfy a local Lipschitz condition; i.e., for x and $x + \delta$ in any compact set, there is a $K < \infty$ (depending on the set) such that

$$\| u(x + \delta) - u(x) \| \leqq K \| \delta \| ;$$

for x, $x + \delta$, u, $u + \delta$ in any compact set, there is a $K < \infty$ (depending on the set) so that

$$\| f(x + \delta, u + \eta) - f(x, u) \| \leqq K(\| \delta \| + \| \eta \|),$$

$$\| \sigma(x + \delta, u + \eta) - \sigma(x, u) \| \leqq K(\| \delta \| + \| \eta \|.$$

Thus $u(x), f(x, u(x))$, and $\sigma(x, u(x))$ satisfy a uniform Lipschitz condition in x in the set G. Let τ_1 be given by (1.2). The solution process x_t is defined, unique and continuous (w.p.1) for $t < \tau_1$. In fact, the process x_t for $t \leqq \tau_1$ does not depend on the values of $f(x, u)$, $\sigma(x, u)$, or $u(x)$ for x outside G.

Associated with the control u and resulting process x_t is the partial differential operator (the differential generator)

$$L^u = \sum_i f_i(x, u) \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i,j} S_{ij}(x, u) \frac{\partial^2}{\partial x_i \, \partial x_j} . \tag{1.3}$$

For purposes of this paper, the domain of L^u is the family of nonnegative functions with continuous second partial derivatives in E. Let E_x^u be the expectation operator given that u is the control and $x_0 = x$. Then, if τ is a random time with $E_x^u \tau < \infty$ and $V(x)$ is in the domain of L^u and has *compact support*, then a formula of Dynkin,

$$V(x) - E_x^u V(x_\tau) = -E_x^u \int_0^\tau (L^u V(x_t))\, dt, \tag{1.4}$$

holds. Formula (1.4) may be loosely interpreted to mean that $V(x)$ is the average of the integral of its "stochastic derivative" $L^u V(x)$. Note that the integral

$$\int_0^\tau k(x_t, u(x_t))\, dt$$

will be written

$$\int_0^\tau k(x, u)\, dt.$$

Notes and references. For an analysis of (1.1) on the scalar (no control) case with uniform Lipschitz conditions on f and σ, see [4, pp. 277–286] or [5, §5]. The vector case is discussed in [1, Chap. 11]. Random times, also called Markov times, are discussed in [5, pp. 56–59] and in [1, Chap. 3, §3, and Chap. 4]. Chapter 11 of [1] discusses diffusion processes which are stopped at random times. For diffusion processes with control u, the operator L^u is the differential generator. On a domain including twice continuously differentiable functions with compact support, it is a restriction of the infinitesimal operator of the process (see [1, Theorem 11.5]).

Since our analysis depends on the properties of the trajectories until the first exit time from open sets G with compact closure, local Lipschitz conditions on f, σ, and u are sufficient to insure that the solutions are well defined and continuous (w.p.1) for all t less than the first exit time from G. The process x_t, with global Lipschitz conditions on u, f, and σ, is Markovian. It is also a strong Markov process ([1, Chap. 3, §3] or [5, §2]); i.e., the Markov property holds for all finite random times: for any measurable set Γ in E, and finite random time τ,

$$P[x_{t+\tau} \in \Gamma \mid x_s, s \leqq \tau] = P[x_{t+\tau} \in \Gamma \mid x_\tau].$$

For times less than the first exit time from G, the process with local Lipschitz conditions on u, f, and σ is also strongly Markovian.

If $E_x^u \tau < \infty$ and $V(x)$ has continuous second derivatives and compact support, then Dynkin's formula ([5, p. 73] or [1, Theorems 5.1 and Corollary, 5.5 and 11.5], etc.) reduces to (1.4). τ may have to be less than the first exit time from some open set with compact closure, when the Lipschitz conditions are only local (and limits taken subsequently).

The control problem. The target set is denoted by S, and its boundary by ∂S. In the theorems S is assumed to be compact, but useful extensions are possible. The control u is termed admissible if $u(x)$ satisfies a local Lipschitz condition. To each admissible control u' (and corresponding

trajectory x_t' with $x_0' = x$), there are random time τ' and a cost $C^{u'}(x)$:

$$\tau' = \inf \{t\colon x_t' \in \partial S\}, \quad \text{control} = u',$$

$$C^{u'}(x) = E_x^{u'} \int_0^{\tau'} k(x', u')\, dt, \tag{1.5}$$

where $k(x', u')$ is nonnegative and continuous in both arguments.

An optimal control u (and corresponding time τ, defined by (1.5)) is sought such that $x'_{t\cap\tau} \to \partial S$ w.p.1 as $t \to \infty$, and

$$C^u(x) \leqq C^{u'}(x)$$

for other admissible controls u' for which $x'_{t\cap\tau'} \to \partial S$ w.p.1 as $t \to \infty$. The exact class of admissible comparison controls is defined by Theorems 1, 2, and 3.

2. Theorems. It is convenient to prove Theorem 1 with the hypotheses (2.2), (2.3) and (2.5). Theorem 2 gives a general and easily checkable condition for (2.2) and Theorems 2 and 3 together give a general and easily checkable condition for (2.3) and (2.5). The results are, essentially, a proof of the formal result that the optimal control u and cost $V(x)$ are given by the solution of

$$\min_{u'} [L^{u'} V(x) + k(x, u')] = 0.$$

THEOREM 1. *Let u be admissible, and let $V(x)$ be a scalar valued function with continuous second partial derivatives, nonnegative in $E - S$, $V(x) = 0$ for $x \in \partial S$ and $V(x) \to \infty$ as $\| x \| \to \infty$. Let $f(x, u)$, $u(x)$, and $\sigma(x, u)$ satisfy a local Lipschitz condition in all arguments. Let S be compact and*

$$L^u V(x) = -k(x, u) \leqq 0 \tag{2.1}$$

in $E - S$. Let x_t correspond to u. Define (recall (1.2))

$$Q_m = \{x\colon 0 \leqq V(x) \leqq m\} - S + \partial S,$$

$$\tau = \inf \{t\colon x_t \in \partial S\},$$

$$\tau(m) = \inf \{t\colon x_t \in \partial Q_m\}.$$

Let $x_0 = x \in Q_m$, and, for any $m > 0$, let[1]

$$x_{t\cap\tau(m)} \to \partial Q_m \quad \text{w.p.1 as } t \to \infty. \tag{2.2}$$

Then $x_{t\cap\tau} \to \partial S$ w.p.1 as $t \to \infty$. Let[1]

$$E_x^u V(x_{\tau(m)}) \to 0 \quad \text{as } m \to \infty. \tag{2.3}$$

[1] See remark above Theorem 1, and also Theorems 2 and 3.

Let, for an admissible control u',

$$L^{u'}V(x) \geqq -k(x, u'), \qquad x \in E - S. \tag{2.4}$$

With control u', *define* τ' *and* $\tau'(m)$ *analogously to* τ *and* $\tau(m)$. *Let* x_t' *correspond to* u'. *Let* $x'_{t\cap\tau'} \to \partial S$ *w.p.*1 *as* $t \to \infty$. *If*[1]

$$E_x^{u'}V(x'_{\tau'(m)}) \to 0, \tag{2.5}$$

then, if $x_0' = x$,

$$C^u(x) \leqq C^{u'}(x), \tag{2.6}$$

i.e., u is optimal with respect to the class of controls satisfying (2.4) *and* (2.5).

Proof. Let $\tau_i(m) \uparrow \tau(m)$ be a sequence of random times with $E_x^u\tau_i(m) < \infty$. Since $V(x)$ is continuous and nonnegative and $V(x) \to \infty$ as $\| x \| \to \infty$, the set Q_m is compact. Let $V_m(x)$ be a nondecreasing sequence of functions with compact support, continuous second partial derivatives, and equal to $V(x)$ in Q_m. In $Q_m - \partial Q_m$, $L^uV_m(x) = L^uV(x) = -k(x, u) \leqq 0$. Since f, u, and σ satisfy a local Lipschitz condition, the process x_t, $t < \tau(m)$, is a continuous (w.p.1) strong Markov process, and $V_m(x)$ is in the domain of its differential generator L^u. Thus, by the formula of Dynkin (1.4) and (2.1),

$$V_m(x) - E_x^uV_m(x_{\tau_i(m)}) = E_x^u\int_0^{\tau_i(m)} k(x, u)\, dt \geqq 0. \tag{2.7}$$

Since $\tau_i(m) \uparrow \tau(m)$ and $k(x, u) \geqq 0$, the monotone convergence theorem yields

$$E_x^u\int_0^{\tau_i(m)} k(x, u)\, dt \uparrow E_x^u\int_0^{\tau(m)} k(x, u)\, dt.$$

By hypothesis, $x_{t\cap\tau(m)} \to x_{\tau(m)}$ w.p.1. Since $V_m(x)$ is continuous and has compact support, and $x_{t\cap\tau(m)}$ is continuous (w.p.1) in t,

$$V_m(x_{\tau_i(m)}) \to V_m(x_{\tau(m)}) \quad \text{w.p.1.} \tag{2.8}$$

$V_m(x)$ is uniformly bounded. Thus, (2.8) and the dominated convergence theorem imply

$$E_x^uV_m(x_{\tau_i(m)}) \to E_x^uV_m(x_{\tau(m)}).$$

Since, by hypothesis, $x_{\tau(m)} \in \partial Q_m$ w.p.1, $V_m(x_{\tau(m)}) = 0$ or m w.p.1. Thus

$$E_x^uV_m(x_{\tau(m)}) = mP(\sup_{\tau \geqq t \geqq 0} V(x_t) \geqq m). \tag{2.9}$$

Relations (2.9) and (2.7) imply

$$P(\sup_{\tau(m+1) \geqq t \geqq 0} V(x_t) \geqq m) = P(\sup_{\tau \geqq t \geqq 0} V(x_t) \geqq m) \leqq V(x)/m. \tag{2.10}$$

Condition (2.10), together with the hypothesis that $x_{t\cap\tau(m)} \to \partial Q_m$ w.p.1, for *arbitrary* m, implies that

$$P(V(x_{t\cap\tau}) \nrightarrow 0) = P(\sup_{\tau \geqq t \geqq 0} V(x_t) > m, \text{ all finite } m)$$
$$\leqq \lim_m V(x)/m = 0,$$

and, hence, $x_{t\cap\tau} \to \partial S$ w.p.1. We have also proved that $\sup_{\tau \geqq t \geqq 0} \| x_t \| < \infty$ w.p.1, and that x_t is well defined and continuous up to the first contact with ∂S (w.p.1).

Now, by the preceding limit arguments,

$$V_m(x) - E_x^u V_m(x_{\tau(m)}) = E_x^u \int_0^{\tau(m)} k(x, u)\, dt. \tag{2.11}$$

By the monotone convergence theorem, the right side goes to

$$E_x^u \int_0^{\tau} k(x, u)\, dt$$

as $m \to \infty$. Also $V_m(x) = V(x)$ for fixed x and sufficiently large m. Hence (2.3) implies that (2.9) goes to zero as $m \to \infty$. Thus,

$$V(x) = -E_x^u \int_0^{\tau} L^u V(x)\, dt = E_x^u \int_0^{\tau} k(x, u)\, dt = C^u(x). \tag{2.12}$$

Now, let u' be admissible. With the use of (2.4) and (2.5), and arguments analogous to those leading to (2.12), we have (with $x_0' = x$)

$$C^u(x) = V(x) \leqq E_x^{u'} \int_0^{\tau'} k(x', u')\, dt = C^{u'}(x).$$

Thus u is optimal relative to all u' satisfying the conditions of the theorem.

THEOREM 2. *Let $k(x, u)$ be continuous and positive in $E - S$. Assume the conditions of Theorem 1 on $V(x)$. Then, with control u, $x_{t\cap\tau(m)} \to \partial Q_m$ w.p.1.*

Proof. Let $\epsilon > 0$ and let N_ϵ be neighborhoods of ∂Q_m with the properties: $N_\epsilon \cap Q_m \to \partial Q_m$ as $\epsilon \to 0$, and $k(x, u) \geqq \epsilon > 0$ in $Q_m - N_\epsilon$. Since $k(x, u)$ is positive and continuous in $Q_m - \partial Q_m$, there is such a sequence N_ϵ.

Note that $0 \leqq V(x_{t\cap\tau(m)}) \leqq m$. Hence, the left side of (2.7) as well as its limit is bounded by m. Thus

$$E_x^u \int_0^{\tau(m)} k(x, u)\, dt = \lim_i E_x^u \int_0^{\tau_i(m)} k(x, u)\, dt \leqq m$$

for all $x \in Q_m$. Then, since $k(x, u) \geqq \epsilon$ in $Q_m - N_\epsilon$, the average time that $x_{t\cap\tau(m)}$ spends in $Q_m - N_\epsilon$ is finite (less than m/ϵ). Thus the time spent in $Q_m - N_\epsilon$ is finite w.p.1. Since $\epsilon > 0$ is arbitrary and $N_\epsilon \cap Q_m \to \partial Q_m$, and $x_{t\cap\tau(m)}$ is stochastically continuous, we have $x_{t\cap\tau(m)} \to \partial Q_m$ w.p.1 as $t \to \infty$.

Remark. There are weaker conditions (than $k > 0$ in $E - S$) under which $x_{t\cap\tau(m)} \to \partial Q_m$ w.p.1, or $x_{t\cap\tau} \to \partial S$ w.p.1. These are results in stochastic stability. If k is nonnegative, then its form together with the system equations (1.1) may imply these facts. See, e.g., [6, Example 5], where k is only semidefinite.

Theorem 3 gives a criterion for condition (2.3) or (2.5). From (2.9), we need only prove (2.13).

THEOREM 3. *Let $F(\lambda)$ be a nonnegative, scalar valued function which is strictly*[2] *monotone increasing and has continuous second derivatives. Let $F(0) = 0$ and $F(\lambda)/\lambda \to \infty$ as $\lambda \to \infty$. If $L^u V(x) \leqq 0$ and*

$$L^u F(V(x)) \leqq 0$$

in $E - S$, then

$$mP(\sup_{\tau \geqq t \geqq 0} V(x_t) \geqq m) \to 0 \tag{2.13}$$

as $m \to \infty$.

Proof. We have

$$\begin{aligned} L^u F(V(x)) &= \frac{\partial F(V)}{\partial V} L^u V(x) + \frac{1}{2} \frac{\partial^2 F(V)}{\partial V^2} \sum_{i,j} \left(\frac{\partial V}{\partial x_i}\right)\left(\frac{\partial V}{\partial x_j}\right) S_{ij} \\ &= -\varphi(x, u) \leqq 0. \end{aligned} \tag{2.14}$$

Let $\tau_i(m)$, $\tau(m)$, and τ be as in Theorem 1. (We do *not* assume here that $x_{t\cap\tau(m)} \to \partial Q_m$.) Then, by Dynkin's formula (1.4),

$$F(V_m(x)) - E_x{}^u F(V_m(x_{\tau_i(m)})) = E_x{}^u \int_0^{\tau_i(m)} \varphi(x, u)\, dt \geqq 0. \tag{2.15}$$

Let

$$\sup_{\tau \geqq t \geqq 0} V(x_t) = \lim_{T \to \infty} [\sup_{0 \leqq t \leqq T \cap \tau} V(x_t)] \geqq m + 1.$$

Then $\tau(m) < \infty$. Also, if for some sample function, $\sup_{\tau \geqq t \geqq 0} V(x_t) \geqq m + 1$, then there is an i such that

$$\sup_{\tau_i(m) \geqq t \geqq 0} V(x_t) \geqq m$$

for that sample function. Thus, by an application of the monotone convergence theorem,

$$\lim_i P(\sup_{\tau_i(m) \geqq t \geqq 0} V(x_t) \geqq m) \geqq P(\sup_{\tau \geqq t \geqq 0} V(x_t) \geqq m + 1). \tag{2.16}$$

[2] The strict monotonicity is not restrictive, since $F(\lambda)$ must be at least linear in λ.

Also

$$\lim_i V(x_{\tau_i(m)}) = \lim_t V(x_{t\cap\tau(m)}) = m.$$

In any case,

$$E_x^u F(V(x_{\tau_i(m)})) \geqq F(m)P(\sup_{\tau_i(m)\geqq t\geqq 0} V(x_t) \geqq m). \tag{2.17}$$

The fact that $f(\lambda)$ has a unique inverse for all $\lambda \geqq 0$ is used in (2.17). Together, (2.15), (2.16), and (2.17) imply that

$$F(V(x)) \geqq F(m)P(\sup_{\tau\geqq t\geqq 0} V(x_t) \geqq m + 1). \tag{2.18}$$

Multiplying both sides of (2.18) by $(m + 1)/F(m)$ and using $[(m + 1)/F(m)] \to 0$ as $m \to \infty$ yields the theorem.

Remark. The criterion of Theorem 3 is a simple "rate of growth" condition and is quite usable. In cases where the matrix $\{S_{ij}\}$ has full rank (then L^u is elliptic), the condition for $E_x^u V(x_{\tau(m)}) \to 0$ reduces to a uniqueness condition for the solution of an exterior Dirichlet problem. Our results are valid whether or not L^u is elliptic.

In the sequel we use

$$F(V) = V \log (A + V), \tag{2.18}$$

for some large A. Then,

$$\begin{aligned} L^u F(V(x)) = &\left(\log (A + V) + \frac{V}{V + A}\right) L^u V(x) \\ &+ \frac{1}{2}\left(\frac{2}{A + V} - \frac{V}{(A + V)^2}\right) \sum_{i,j} \left(\frac{\partial V}{\partial x_i}\right)\left(\frac{\partial V}{\partial x_j}\right) S_{ij}. \end{aligned} \tag{2.19}$$

If $S_{ij}(x)$ is uniformly bounded and $|\partial V/\partial x_i|^2/(A + V)$ is uniformly bounded for large A, and $L^u V(x) < 0$ in $E - S$, then $L^u F(V(x)) \leqq 0$ and Theorem 3 may be applied. In addition, if S_{ij} does not depend on the control, then the u of Theorem 1 is optimal with respect to all u' such that $L^{u'} V(x) < 0$ in $E - S$.

Remark. In practice, certain components of x may not be observed, and it may be desired that the control be a function only of the instantaneous values of the observed components. Let d be the dimension of x, and let u and u' be functions of $x_1, \cdots, x_s$ only. Now, if (assuming the other conditions of Theorem 1 satisfied)

$$L^u V(x) = -k(x, u)$$

and

$$L^{u'} V(x) \geqq -k(x, u')$$

for *all* $x_{s+1}, \cdots, x_d$, then

$$C^u(x) \leqq C^{u'}(x).$$

3. A solution to a norm invariant problem. The stochastic version of the so-called *norm invariant problem* considered in [3] is the following. Let $S = \{x: \| x \| \leqq r\}$ and

$$dx = f(x)\,dt + u\,dt + \sigma I\,dz,$$

where σ is a constant and I is the identity matrix[3] and[4] $x'f(x) = 0$ (the noiseless uncontrolled system is conservative). The control values are constrained by

$$u'u \leqq \rho^2. \tag{3.1}$$

We seek the control which minimizes the *average* time required to transfer $x_0 = x$ to ∂S; $k(x, u) = 1$. Analogs of the minimum fuel and energy problems in [3] may be analyzed similarly. That the optimal stochastic control is identical to the optimum deterministic control is no surprise. The problem is interesting since few stochastic solutions are available and "conservative" systems are of some current importance.

Let u and $V(x)$ satisfy the conditions of Theorem 1 for this problem. Then for $\| x \| > r$, $L^u V(x) = -1$, $V(x) > 0$, and $V(x) = 0$ for $\| x \| = r$. It will be shown that the admissible control (3.2) is an optimal control.

$$u = -\frac{\rho x}{\| x \|}. \tag{3.2}$$

A preliminary analysis indicates that $V(x) = g(\| x \|)$ for some monotonically increasing function $g(\lambda)$. We will first find $V(x) = g(\| x \|)$ such that $L^u V(x) = -1$, then check that $V(x)$ satisfies the conditions of Theorems 1, 2, and 3, and finally determine the family of comparison admissible controls.

Let $w = \| x \|$. Since

$$\frac{\partial^2 g(w)}{\partial x_i^2} = \frac{\partial g(w)}{\partial w}\left(\frac{1}{w} - \frac{x_i^2}{w^3}\right) + \frac{\partial^2 g(w)}{\partial w^2}\left(\frac{x_i^2}{w^2}\right),$$

we have, with $V(x) = g(w)$,

$$\begin{aligned} L^u V(x) &= \frac{\partial g(w)}{\partial w}\left(\frac{x'f(x)}{w} + \frac{u'x}{w}\right) \\ &\quad + \frac{\sigma^2}{2}\sum_i \left[\frac{\partial g(w)}{\partial w}\left(\frac{1}{w} - \frac{x_i^2}{w^3}\right) + \frac{\partial^2 g(w)}{\partial w^2}\frac{x_i^2}{w^2}\right] \\ &= \frac{\partial g(w)}{\partial w}\left(-\rho + \frac{(d-1)\sigma^2}{2w}\right) + \frac{\partial^2 g(w)}{\partial w^2}\frac{\sigma^2}{2} = -1, \end{aligned} \tag{3.3}$$

where d is the dimension of the x vector. Equation (3.3) admits of a solu-

[3] Solutions for more general $\sigma(x, u)$ have not been found. In certain applications, such as the tumbling satellite, it seems reasonable that the noise would be symmetric.

[4] In this section x' is the transpose of x.

tion of the form

$$g(w) = A_0 + A_1 w + A_2 \log w + \sum_1^\infty \frac{B_i}{w^i}. \tag{3.4}$$

Substituting (3.4) into (3.3) and equating coefficients of like terms in w yields the A_i and B_i :

$$A_1 = 1/\rho, \qquad A_2 = (d-1)\frac{\sigma^2}{2\rho^2},$$

$$B_1 = \frac{-(d-2)(d-1)}{\rho}\left(\frac{\sigma^2}{2\rho}\right)^2 \leqq 0, \tag{3.5}$$

$$B_{n+1} = \frac{\sigma^2 n\, B_n(d-n-2)}{2\rho(n+1)} \leqq 0.$$

There are only d nonzero terms in (3.5). Thus,

$$g(w) = A_0 + A_1 w + A_2 \log w + \sum_1^{d-2} \frac{B_i}{w^i},$$

where the empty sum $\sum_1^0 B_i/w = 0$. Since we require $g(r) = 0$, A_0 is obtained from

$$0 = A_0 + A_1 r + A_2 \log r + \sum_1^{d-2} \frac{B_i}{r^i}.$$

Since $A_i \geqq 0$ and $B_i \leqq 0$, $g(w) > 0$ for $w > r$. (The negative terms in $g(w)$ decrease, and the positive terms increase, as w increases.)

The hypotheses of the theorems will now be verified. Define $V(x) = g(\| x \|)$. By construction, $L^u V(x) = -1$ in $E - S$. Also, $V(x) \to \infty$ as $\| x \| \to \infty$, and $V(x) = 0$ for $x \in \partial S$. Hence, by Theorem 2, (2.2) is satisfied. Since $L^u V(x) = -1$ in $E - S$ and $| \partial V/\partial x_i |^2$ and $S_{ij}(x) = \sigma^2 \delta_{ij}$ are uniformly bounded, Theorems 2 and 3 and the use of (2.19) imply that (2.3) is satisfied. Let $u' \neq u$ be admissible. Since $u = -\rho x/\| x \|$ absolutely minimizes $L^u V(x) + 1$, we have $L^{u'} V(x) \geqq -1 = -k(x, u')$. Let $L^{u'} V(x) < 0$ in $E - S$. Then, as with u, (2.5) is satisfied. Then u is optimal relative to (at least) all u' such that $L^{u'} V(x) < 0$ in $E - S$.

REFERENCES

[1] E. B. Dynkin, *Markov Processes*, Springer-Verlag, Berlin, 1965.

[2] H. J. Kushner, *Stochastic stability and the design of feedback controls*, Proceedings of the Polytechnic Institute of Brooklyn Symposium on System Theory, 1965.

[3] M. Athans, P. Falb, and R. T. Lacoss, *Time, fuel and energy optimal control of nonlinear norm invariant systems*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 196-201.

[4] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[5] K. Ito, *Lectures on Stochastic Processes*, Tata Institute, Bombay, 1961.

[6] H. J. Kushner, *On the theory of stochastic stability*, Advances in Control Systems, vol. 4, Academic Press, New York, 1966, to appear.