Funkcialaj Ekvacioj, **34** (1991) 355–374

Bilinear Control: Rank-One Inputs

By

Otomar HÁJEK

(Case Western Reserve University, U.S.A.)

Introduction

This paper establishes some basic results on bilinear control systems of the form

$$\dot{x} = (A + uD)x \tag{1}$$

in n-space, with scalar control functions $u(\cdot)$. A crucial subsequent specialization is that the control matrix have rank one, $D = bc^*$ with n-vectors b, c.

This last class arises quite naturally when one switches between two dynamical systems, e.g.

$$y^{(n)} = \sum_0^{n-1} \alpha_{n-k} y^{(k)} \quad \text{and} \quad y^{(n)} = \sum_0^{n-1} \beta_{n-k} y^{(k)}.$$

Here the standard phase-space description has the coefficient matrix A in companion form, with $\frac{1}{2}(\alpha_k + \beta_k)$ in the last row; and the control matrix D indeed has $D = b \cdot c^*$ for $b^* = (0, \dots, 0, 1)$ and c^* with entries $\frac{1}{2}(\beta_k - \alpha_k)$. (As Jan Willems once pointed out, all the entries are constants; however, the zeros and ones are stiff structure constants, while only the last row has "soft" parameters, to be encompassed by a rank-one control matrix.)

Section 1 presents a canonic decomposition of the state space of (1) into linear subspaces (which are controllable or observable, or not, in a suitable sense); a somewhat surprising analogue of the Kalman decomposition that applies to *linear* control systems. What makes this possible is the rather technical observation (Lemma 1) that, in the Taylor expansion of $Dx(t)$, the first nonvanishing term does not depend on the controls.

The basic result of Section 2 is that, for any initial point p and small times $t > 0$, the set $\mathscr{A}_t \cdot p$ attainable from p at time t is convex. Ultimately this establishes strict convexity of the reachable and attainable sets, and normality (uniqueness of extremal controls) for small time; and also that time-optimal controls are bang-bang and piecewise constant.

Some proofs are modifications of known reasonings applying to linear control systems. The interplay between convexity of attainable sets and bang-bang controls is classical. The more precise results involving strict convexity and uniqueness that appear in the present Theorem 6 obviously come from Sections 14 and 15 of [6]. Our Lemma 7 is almost exactly the Corollary, [7, p. 72]. Lemma 10 here is an immediate analogue of the fundamental lemma [3, Lemma 1]. Conjecture 2 would extend the result of Theorem 9.4 in [4]. None of these techniques could have been unleashed without the key result of Brockett [1, Lemma 1].

§1. One-dimensional inputs

In this section we shall treat control systems in n-space, with dynamical equation of the form

$$\dot{x} = (A + uD)x; \tag{1}$$

the data are the constant real n-square matrices A, D. Thus the control system is bilinear, homogeneous, autonomous, and single-input, with state space $\boldsymbol{R}^n$ and control space $\boldsymbol{R}^1$.

It will be useful to consider, in parallel, the associated matrix system

$$\dot{X} = (A + uD)X, \qquad X(0) = I, \tag{2}$$

where the state space is $\boldsymbol{R}^{n^2}$. To each locally integrable control $u\colon \boldsymbol{R}^1 \to \boldsymbol{R}^1$ there then corresponds a unique matrix-valued solution $X = X(u)$ of (2), with values denoted as in $X_t = X_t(u)$. Returning to (1), the solution $t \mapsto x_t$ of (1) with initial value $x_0 = p \in \boldsymbol{R}^n$ is then $x_t = X_t \cdot p$.

In the sequel the control functions u are often $u\colon \boldsymbol{R}^1 \to [-1, 1]$ measurable (to be called *admissible*, or relaxed); sometimes we allow $u\colon \boldsymbol{R}^1 \to \boldsymbol{R}^1$ to be merely locally integrable (*unrestricted* or unbounded controls), or $u\colon \boldsymbol{R}^1 \to \{-1, 1\}$ (the *bang-bang* controls).

In (1) and (2) the control matrix D will have some rank r, $0 \leq r \leq n$. If we ignore the trivial case $D = 0$, it may be decomposed as in

$$D = B \cdot C \tag{3}$$

with B of type (n, r) and rank r, and C of type (r, n) and rank r again (and also $D = BT \cdot T^{-1}C$ for any nonsingular r-square matrix T). In a purely formal manner we then associate with (1) the observed linear control system

$$\dot{x} = Ax + Bv, \qquad y = Cx \tag{4}$$

with r-dimensional inputs; we will refer to (4) as the system (C, A, B).

Using (3) in (1) we see that the control term is $u(t)Dx(t) = u(t)B \cdot Cx(t)$. One may recognize here an open loop control term $u(t)B$ and a linear feedback term $Cx(t)$. Since these are multiplied (rather than added), obviously near the subspace $Cx = 0$ the control effect is largely neutralized, and x_t evolves approximately as in the dynamical system $\dot{x} = Ax$ without controls. (More precisely, if $C \cdot p = 0$, then $\frac{d}{dt} X_t(u) \cdot p = Ap$ at $t = 0$.) An elaboration of this yields our first result.

Lemma 1. *For each point $p \in \boldsymbol{R}^n$ we have an alternative: Either $CA^k p = 0$ for all $k = 0, 1, 2, \ldots$, and then*

$$X_t(u)p = e^{At}p, \qquad CX_t(u)p \equiv 0 \tag{5}$$

for each (unrestricted) control $u(\cdot)$ and all t. Or $CA^k p \neq 0$ for some first integer $k \geq 0$ (necessarily $k \leq n - 1$), and then

$$CX_t(u)p = CA^k p \cdot \frac{t^k}{k!} + O(t^{k+1}) \tag{6}$$

as $t \to 0$, uniformly for all admissible controls $u(\cdot)$; in particular, $t \mapsto CX_t(u)p$ has only isolated zeros, each of multiplicity $n - 1$ at most.

Proof. First assume that (the initial point) p is such that

$$CA^k p = 0 \quad \text{for} \quad k = 0, 1, \ldots. \tag{7}$$

Consider the Picard iterates $\Phi_k(\cdot)$ to (2), for an arbitrarily chosen but fixed unrestricted control $u(\cdot)$:

$$\Phi_0(t) = I, \quad \Phi_{k+1}(t) = I + \int_0^t (A + u(s)BC)\Phi_k(s)\,ds.$$

Picard's theorem yields, in our linear case, that $\Phi_k(t) \to X_t(u)$ as $k \to \infty$.

We shall prove that

$$\Phi_k(t)p = \sum_0^k \frac{1}{j!} A^j t^j p,$$

by induction on k (the case $k = 0$ is trivial). For the inductive step assume the above, and observe that then $C\Phi_k(t)p \equiv 0$ by (7). We have

$$\begin{aligned}\Phi_{k+1}(t)p &= p + \int_0^t (A + u(s)BC)\sum_0^k \frac{1}{j!} A^j s^j p\,ds \\ &= p + \int_0^t \Big(\sum_0^k \frac{1}{j!} A^{j+1} s^j p + 0\Big)\,ds = p + \sum_0^k \frac{1}{(j+1)!} A^{j+1} t^{j+1} p.\end{aligned}$$

i.e., the $(k + 1)$-st assertion. Having established this,

$$X_t(u)p = \lim_k \Phi_k(t)p = e^{At}p,$$

$$CX_t(u)p = Ce^{At}p = \sum_0^\infty CA^k p \cdot \frac{t^k}{k!} = 0;$$

these are the relations is (5).

Second, assume that $CA^k p \neq 0$ for some first $k \geq 0$. We write the Picard iteration in the Neumann series form,

$$X_t(u) = I + \sum_1^\infty \Psi_j(t),$$

(8) $$\Psi_j(t) = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{j-1}} (A + u(s_1)BC) \cdot \cdots \cdot (A + u(s_j)BC)\, ds_j \cdots ds_1$$

and note that $I + \sum_1^k \Psi_j(t) = \Phi_k(t)$ (for the iterates Φ_k as above).

Our assumption yields $CA^j p = 0$ for $0 \leq j \leq k-1$, so that

$$C(I + \sum_1^{k-1} \Psi_j(t))p = C\Phi_{k-1}(t)p \equiv 0,$$

and

$$C\Psi_k(t)p = \frac{1}{k!} CA^k p t^k \neq 0$$

(on considering all the remaining terms in the k-fold product in (8)). Finally, for all $\ell > k$,

$$|C\Psi_\ell(t)p| \leq \frac{t^\ell}{\ell!}(|A| + |B| \cdot |C|)^\ell \quad \text{if} \quad |u(s)| \leq 1.$$

This yields (6), and shows that $CX_t(u)p$ has at most an isolated zero of order $\leq n-1$ at $t = 0$.

For any other t we write

$$CX_s(u)p = CX_{s-t}(v) \cdot X_t(u)p$$

(with admissible control v a t-shift of u) and apply the preceding to the initial point $p_1 = X_t(u)p$. Note that again $CA^\ell p_1 \neq 0$ for some ℓ; indeed, otherwise we would also have all $CA^k p = 0$. This concludes the proof.

According to (5), the linear subspace L of $\boldsymbol{R}^n$,

(9) $$L = \{p \colon CA^k p = 0 \text{ for } k = 0, 1, 2, \ldots\}$$

is significant. Its points p, and L itself, may be called completely uncontrollable; and if $L = 0$, the system (1) may be termed controllable. From (9), L is the unobservable subspace of the linear system (C, A, B) in (4).

Theorem 2. *If coordinates are chosen in* $\boldsymbol{R}^n$ *in such a way that* $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ *is in the completely uncontrollable subspace precisely when* $x_1 = 0$, *then* (1) *decomposes, with dynamical equations of the form*

$$
\begin{aligned}
\dot{x}_1 &= (A_{11} + uB_1C_1)x_1 \\
\dot{x}_2 &= (A_{21} + uB_2C_1)x_1 + A_{22}x_2,
\end{aligned}
\tag{10}
$$

and the system (10) *corresponding to* $x_2 = 0$ *is controllable. Thus in L the system* (1) *reduces to* $\dot{x}_2 = A_{22}x_2$ *without controls.*

Proof. First partition (1) conformably, obtaining matrices A_{ij}, B_i, C_j for $i, j = 1, 2$. Now choose any point x_2 and any constant control u. Then the point $p = \begin{pmatrix} 0 \\ x_2 \end{pmatrix} \in L$, so that also $X_t(u)p \in L$ (see (5)), and the x_1-coordinate vanishes. Hence the first equation reads

$$0 = (A_{11} + uB_1C_1)\cdot 0 + (A_{12} + uB_1C_2)x_2(t).$$

On taking $t = 0$ we find that $0 = (A_{12} + uB_1C_2)x_2$ for all u, x_2, and therefore $A_{12} = 0$, $B_1C_2 = 0$. Thus the first equation has the form (10).

Next, again for the completely uncontrollable initial point $p = \begin{pmatrix} 0 \\ x_2 \end{pmatrix}$, the second equation reads

$$\dot{x}_2(t) = (A_{21} + uB_2C_1)0 + (A_{22} + uB_2C_2)x_2(t).$$

Now, $X_t(u)p$ is independent of u by (5); hence so are the components $x_2(t)$, $\dot{x}_2(t)$, and therefore $B_2C_2 = 0$ since x_2 was arbitrary. This concludes the proof.

Remarks. For linear control systems $\dot{x} = Ax + Bv$ we have a unique (largest) controllable subspace $\langle A; B\rangle$, namely the span of the columns of the controllability matrix

$$(B, AB, A^2B, \ldots, A^{n-1}B);$$

and every complementary subspace is uncontrollable (better: not completely controllable). The situation is quite different for bilinear systems (1): we have the unique completely uncontrollable subspace L of (9), and every complementary subspace is a "controllable" (better: the only uncontrollable point is the origin).

Another consequence is that the uncontrollable subspace L is strongly invariant, in both time directions: no point of L can be reached from, nor steered to, any point outside L even by unrestricted controls.

If, as in Lemma 1, $k \geq 0$ is the first integer with $CA^k p \neq 0$, it seems appropriate to call $n - k$ the degree of controllability of p. Then points p with $Cp \neq 0$ have highest degree of controllabililty; and completely uncontrollable points might be said to have zero degree of controllability.

Naturally, Lemma 1 has a dual version. The proof is omitted; it is analogous to that of Lemma 1, or may be obtained from it by passing to adjoint equations

$$\dot{y} = -(A^* + uC^* B^*)y$$

(the adjoint to (2) has $(X_t^{-1})^*$ as solution).

Lemma 3. *For each point* $q \in \boldsymbol{R}^n$ *we have an alternative: Either* $q^* A^k B = 0$ *for all* $k = 0, 1, \ldots,$ *and then*

$$q^* X_t^{-1}(u) = q^* e^{-At}, \qquad q^* X_t^{-1}(u) B \equiv 0$$

for each (unrestricted) control $u(\cdot)$ *and all* t. *Or* $q^* A^k B \neq 0$ *for some first integer* $k \geq 0$ $(k \leq n-1)$, *and then*

$$q^* X_t^{-1}(u) B = q^* A^k B \frac{t^k}{k!} + O(t^{k+1})$$

as $t \to 0$, *uniformly for all admissible controls* $u(\cdot)$; *in particular,* $t \mapsto q^* X_t^{-1}(u) B$ *has only isolated zeros, of multiplicity* $n-1$ *at most.*

Remarks. As an illustration, consider a single input linear control system $\dot{x} = Ax + bu$ in $\boldsymbol{R}^n$. Professor Brockett suggested an interpretation of this as a bilinear system in $\boldsymbol{R}^{n+1}$, namely

$$\begin{pmatrix} x \\ \xi \end{pmatrix}^{\cdot} = \left(\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} + u \begin{pmatrix} b \\ 0 \end{pmatrix} (0, 1) \right) \begin{pmatrix} x \\ \xi \end{pmatrix},$$

with the original system corresponding to the hyperplane $\xi = 1$. The completely controllable subspace is the hyperplane $\xi = 0$ (in verifying this note that $\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}^k$ is $\begin{pmatrix} A^k & 0 \\ 0 & 0 \end{pmatrix}$ for $k \geq 1$, but $\begin{pmatrix} I & 0 \\ 0 & 1 \end{pmatrix}$ for $k = 0$). As concerns the dual observation vectors $\begin{pmatrix} q \\ \eta \end{pmatrix}$, we have complete unobservability precisely when $q^* A^k b = 0$ for all $k \geq 0$, i.e., when q is perpendicular to the controllability subspace $\langle A; b \rangle$.

There is also a dual version of the canonic decomposition in Theorem 2. Actually we may combine these as follows. Choose coordinates in $\boldsymbol{R}^n$ in such a way that a point

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

has all $CA^k x = 0$ precisely when $x_1 = 0$, $x_2 = 0$, and has all $x^* A^k B = 0$ precisely when $x_1 = 0$, $x_3 = 0$. Then the matrices A, B, C decompose conformably into A_{ij}, B_i, C_j $(i, j = 1, 2, 3, 4)$; and the zero entries are as indicated in

$$A = \begin{bmatrix} \cdot & \cdot & 0 & 0 \\ 0 & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & \cdot \end{bmatrix} \qquad B = \begin{bmatrix} \cdot \\ 0 \\ \cdot \\ 0 \end{bmatrix} \qquad C = (\cdot \ \cdot \ 0 \ 0).$$

There is an interesting re-interpretation of Lemma 3, again in terms of the controllability subspace $\langle A; B \rangle$ of the linear system (C, A, B). Take any vector q perpendicular to $\langle A; B \rangle$. Then $q^* A^k B = 0$ for all $k \geq 0$, so that (Lemma 3)

$$q^* X_t^{-1}(u) B = 0$$

for all t, and even unrestricted controls $u(\cdot)$. In other words, $q^* x = 0$ whenever x can be steered to some point p in the span of columns of B, at some time t via some control $u(\cdot)$. Since q was an arbitrary point in $\langle A; B \rangle^\perp$, we have: the union of reachable sets to points in span B is contained in the controllability space of (C, A, B).

It is then natural to ask whether these reachable sets span the controllability space. A partial answer is provided in Lemma 10 to follow.

Corollary 4 *In* (1) *let $D = b \cdot c^*$ have rank* 1, *and assume that* (c^*, A, b) *is controllable and observable (i.e., minimal). Then, for any nonzero points $p, q \in \boldsymbol{R}^n$, and each admissible control $u(\cdot)$, the real-valued function*

$$t \longmapsto q^* X_t^{-1}(u) b \cdot c^* X_t(u) p \tag{11}$$

has a continuous derivative, and only isolated zeros.

Proof. Almost everywhere the derivative of (11) is

$$\begin{aligned} -\, q^* X_t^{-1}(A + u_t bc^*) bc^* X_t p + q^* X_t^{-1} bc^*(A + u_t bc^*) X_t p \\ = q^* X_t^{-1} [bc^*, A] X_t p, \end{aligned}$$

and this is (locally absolutely) continuous. The rest follows from Lemmas 1, 3 applied to the factors $q^* X_t^{-1} b$, $c^* X_t p$ of (11).

§ 2. Rank one inputs

Here we shall consider the special case of single input systems $\dot{x} = (A + uD)x$, for which the control matrix D has rank 1. Thus $D = bc^*$ with b, c nonzero,

$$\dot{x} = (A + ubc^*)x \tag{1}$$

$$\dot{X} = (A + ubc^*)X, \qquad X_0 = I. \tag{2}$$

The following construction is due to Brockett [1].

Lemma 5. *Let $x(\cdot)$, $y(\cdot)$ be solutions of* (1) *corresponding to controls $u(\cdot)$, $v(\cdot)$; assume that*

$$c^*(x_t + y_t) \neq 0 \qquad a.e.\ in\ [0, T].$$

Then the control

$$w_t := u_t \cdot \frac{c^* x_t}{c^*(x_t + y_t)} + v_t \cdot \frac{c^* y_t}{c^*(x_t + y_t)} \tag{3}$$

steers the initial point $\frac{1}{2}(x_o + y_o)$ along the midway solution $\frac{1}{2}(x_t + y_t)$.

If, furthermore, $c^* x_t$ and $c^* y_t$ are both ≥ 0 or both ≤ 0 a.e. in $[0, T]$, and $u(\cdot)$, $v(\cdot)$ are admissible, then $w(\cdot)$ is also admissible.

Proof. By direct substitution,

$$\begin{aligned}
&\frac{d}{dt}\frac{1}{2}(x_t + y_t) - (A + w_t bc^*)\frac{1}{2}(x_t + y_t) \\
&\quad = \frac{1}{2}(Ax_t + u_t bc^* x_t) + \frac{1}{2}(Ay_t + v_t bc^* y_t) \\
&\qquad - \left(A\frac{1}{2}(x_t + y_t) + w_t bc^* \frac{1}{2}(x_t + y_t)\right) \\
&\quad = \frac{1}{2} b(u_t c^* x_t + v_t c^* y_t - w_t c^*(x_t + y_t)) = 0,
\end{aligned}$$

having used (3).

With the added sign assumptions we have that both

$$\frac{c^* x_t}{c^*(x_t + y_t)}, \qquad \frac{c^* y_t}{c^*(x_t + y_t)}$$

are in $[0, 1]$ a.e.; hence if both u_t, v_t are in $[-1, 1]$, so is their convex combination w_t: admissibility. This concludes the verification.

We recall the concepts of attainable and reachable sets: referring to the matrix system (2), the attainable set $\mathscr{A}_t$ (at time t, from initial value I) consists of all $X_t(u)$ as $u(\cdot)$ varies over all admissible controls; and $\mathscr{A}_t^{-1} = \{X^{-1}: X \in \mathscr{A}_t\}$ is the reachable set (of initial values from which I can be reached at t by using admissible controls). Then for the system (1) in $\boldsymbol{R}^n$, $\mathscr{A}_t \cdot p$ is the set attainable from p, and $\mathscr{A}_t^{-1} \cdot p$ the reachable set, both at time t.

Theorem 6 (Local convexity of attainable sets). *Consider system* (1). *For every initial point p there exists T, $0 < T \leq +\infty$, such that the attainable and reachable sets*

$$\mathscr{A}_t \cdot p \textit{ and } \mathscr{A}_t^{-1} \cdot p \textit{ for } 0 < t < T$$

are convex and compact.

Furthermore, in the generic case that the linear system (c^*, A, b) *is controllable and observable, and $p \neq 0$, these attainable and reachable sets are strictly convex* (*in particular* $\operatorname{int} \mathscr{A}_t \cdot p \neq \emptyset \neq \operatorname{int} \mathscr{A}_t^{-1} p$); *all extremal controls are bang-bang and piecewise constant, and are determined uniquely a.e. by the initial point p, terminal point x, and terminal time t.*

Proof. In (2) the vectograms $\{A + ubc^*: |u| \leq 1\}$ are convex and compact. By Filippov's theorem, $\mathscr{A}_t$ is compact; hence so are $\mathscr{A}_t \cdot p$ and $\mathscr{A}_t^{-1} \cdot p$.

Consider any initial point $p \in \boldsymbol{R}^n$. If the first alternative of Lemma 1 applies, then $\mathscr{A}_t \cdot p = \{e^{At}p\}$, and singleton sets are, of course, convex (in this case we may take $T = +\infty$).

Assume, then, that $c^* A^k p \neq 0$ for a first $k \geq 0$. Since we are restricting the controls to admissible ones, (6) in Lemma 1 yields that there exists $T > 0$ such that $c^* X_t(u) p$ has constant sign in $(0, T)$ (namely, $\operatorname{sgn} c^* A^k p$) for all admissible controls $u(\cdot)$.

We assert that this is the desired T: that each $\mathscr{A}_t \cdot p$ with $t \in (0, T)$ is convex. Since $\mathscr{A}_t \cdot p$ is compact, hence closed, one need only verify that

$$x, y \in \mathscr{A}_t \cdot p \quad \text{implies} \quad \frac{1}{2}(x + y) \in \mathscr{A}_t \cdot p.$$

For this we use Brockett's lemma. The points x, y are values, at time t, of solutions $x(\cdot)$, $y(\cdot)$ of (1), issuing from the same initial point p, and corresponding to some admissible controls $u(\cdot)$, $v(\cdot)$. Since

$$c^* x_t = c^* X_t(u) p, \quad c^* y_t = c^* X_t(v) p$$

have the same sign in $(0, T)$ (namely sgn $c^* A^k p$ again), we have the situation described in Lemma 5. Hence the control $w(\cdot)$ from (3) is also admissible, and steers the initial point $\frac{1}{2}(x_o + y_o) = p$ to the terminal point

$$\frac{1}{2}(x_t + y_t) = \frac{1}{2}(x + y).$$

Thus indeed $\frac{1}{2}(x + y)$ is in $\mathscr{A}_t \cdot p$. (The assertion on the reachable sets $\mathscr{A}_t^{-1} \cdot p$ is obtained similarly, using initial points x, y and terminal point p.)

Next, assume that (c^*, A, b) is minimal (i.e., controllable and observable). Take any admissible control $u(\cdot)$ on $[0, t]$ steering p to a point x on the boundary of the convex set $\mathscr{A}_t \cdot p$. (We are not asserting yet that int $\mathscr{A}_t \cdot p \neq \emptyset$). Choose a non-zero exterior normal q to $\mathscr{A}_t \cdot p$ at x.

We wish to show that $u(\cdot)$ is necessarily bang-bang. Begin by choosing arbitrarily numbers $s < s + h$ in $[0, t]$, and a constant value $v \in [-1, 1]$; then consider the admissible control which has value v in $[s, s + h]$, but coincides with $u(\cdot)$ outside this subinterval. The corresponding matrix solution of (2) is (we abbreviate $X_s(u) = X_s$)

$$X_t X_{s+h}^{-1} e^{(A + vbc^*)h} \cdot X_s$$

at t, and then the exterior normal condition yields

$$q^* X_t X_{s+h}^{-1} e^{(A + vbc^*)h} X_s p \leq q^* X_t p.$$

Subtract the right side, divide by $h > 0$, and take $h \to 0$. There results (by differentiation at s, for almost all $s \in [0, t)$),

$$q^* X_t X_s^{-1} (v - u_s) bc^* X_s p \leq 0.$$

For the moment denote $q^* X_t$ by q_1^*, and consider the scalar function

$$\alpha(s) = q_1^* X_s^{-1} bc^* X_s p$$

as in Corollary 4. We have shown that

$$v\alpha(s) \leq u(s)\alpha(s)$$

for almost all $s \in [0, t)$. In principle, the exceptional set of times s could depend on the choice of v. However, we can take $v = \pm 1$, and then the union of the two null sets. This proves that

$$u(s)\alpha(s) = \max_{|v| \leq 1} v \cdot \alpha(s) = |\alpha(s)| \qquad a.e.$$

Since $\alpha(\cdot)$ has only isolated zeros (Corollary 4), this specifies $u(\cdot)$ completely a.e., and it must be bang-bang and piecewise constant.

Next we show that $u(\cdot)$ is uniquely determined by the terminal data x, t (above we showed that it is uniquely determined by the exterior normal q; but there might be several of these at x). Thus, let $u(\cdot)$, $v(\cdot)$ be admissible controls, both steering p to $x \in \partial \mathscr{A}_t \cdot p$ over $[0, t]$; denote by $x(\cdot)$, $y(\cdot)$ the corresponding solutions; and invoke the admissible control $w(\cdot)$ from Lemma 5, which steers p along the midway solution to the point $\frac{1}{2}(x_t + y_t) = \frac{1}{2}(x + x) = x$. Since $x \in \partial \mathscr{A}_t \cdot p$, all of $u(\cdot)$, $v(\cdot)$, $w(\cdot)$ must be bang-bang and piecewise constant, as we have just shown. If it were not true that $u = v$ a.e., then $u \neq v$ on a set of positive measure. Now, $w(\cdot)$ is both bang-bang and pointwise a convex combination of $u(\cdot)$, $v(\cdot)$. Thus, on a further subset of positive measure, w coincides with u or with v, e.g., the latter. But then from (3),

$$(u(s) - v(s)) \cdot c^* x(s) = 0$$

on a set of positive measure, on which also $u \neq v$. Thus $c^* x(\cdot) = 0$ on a set of positive measure. On the other hand, from minimality and Lemma 1, $c^* x(\cdot)$ has isolated zeros only. This contradiction now yields that indeed $u = v$ a.e.

It remains to show that $\mathscr{A}_t \cdot p$ is strictly convex. For this take distinct points $x \neq y$ in $\mathscr{A}_t \cdot p$, and assume that their mid-point is not in the interior; i.e., that $\frac{1}{2}(x + y) \in \partial \mathscr{A}_t \cdot p$. Since $\mathscr{A}_t \cdot p$ is convex, if either of x, y were interior points, so would their midpoint; thus both are also on $\partial \mathscr{A}_t \cdot p$. Now proceed as above, choosing controls $u(\cdot)$, $v(\cdot)$ steering p to x, y respectively, and then $w(\cdot)$ steering to $\frac{1}{2}(x + y)$. Since all these endpoints are on the boundary, all the controls are bang-bang. Again, $u = v$ a.e., since the converse would lead to a contradiction. But then $x = y$; and this contradiction with the assumption yields strict convexity. (Similarly for $\mathscr{A}_t^{-1} \cdot p$.)

Remarks. The preceding result concerns the vector system (1). It would be a trivial consequence if we could prove the analogous assertion for the matrix system (1), on applying the linear mapping $X \mapsto X \cdot p$ taking $\boldsymbol{R}^{n^2}$ to $\boldsymbol{R}^n$. This, however, is not available. Indeed, if $\mathscr{A}_t$ were convex for all times t in some interval $(0, T)$, then each $\mathscr{A}_t \cdot p$ would be convex, with a time-interval $(0, T)$ common to all initial points p. However, it seems that the time extent T that appears in Theorem 6 depends strongly on p, and there is no common positive lower bound.

This result is necessarily local, in the time sense. Indeed, consider the bilinear system which switches ($u = \pm 1$) between the dynamical systems

$$\begin{cases} \dot{x} = y \\ \dot{y} = -x \end{cases} \quad \text{and} \quad \begin{cases} \dot{x} = y \\ \dot{y} = 0 \end{cases}.$$

For the initial point $p = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and time $t = \pi$, the trajectory of the first dynamical system reaches the point $-p = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, while the second remains constantly at p. Thus both $\pm p$ are in $\mathscr{A}_\pi(p)$; however, their midpoint is 0, and this cannot be reached from any point outside the origin. Hence $\mathscr{A}_\pi(p)$ is not convex. (A more detailed examination yields that the last time T that $\mathscr{A}_T(p)$ is convex is $T = \pi/2$.) A consequence not confined to small time intervals appears in Theorem 9.

The rank-one assumption seems essential for local convexity; this does not extend to single-input systems. E.g., in the planar system

$$\dot{x} = ux, \qquad \dot{y} = -uy$$

we have $\dot{x}/x + \dot{y}/y = 0$, so that the product xy remains a constant independent of the control. Thus each attainable set is an arc of a hyperbola $xy = c$, and hence is not convex for small times.

Local convexity does generalize partly to a larger class of bilinear systems, namely to

$$\dot{x} = (A + \Sigma u_k \cdot b_k c_k^*)x$$

with independent controls $u_k(\cdot)$: each initial point p in the open set described by $c_k^* p \neq 0$ for all k, has convex attainable set for small time. This is the class of systems introduced in [1].

In Theorem 6 is it natural to inquire into the "position" of the convex sets $\mathscr{A}_t \cdot p$; more precisely, to specify the affine span of $\mathscr{A}_t \cdot p$ (in the nontrivial case that $\operatorname{int} \mathscr{A}_t \cdot p = \emptyset$). The point is, of course, that each convex set has nonvoid interior in its affine span. The main result of [5] is a direct description of the affine span of $\mathscr{A}_t$, namely

$$\operatorname{af span} \mathscr{A}_t = e^{At}(I + \mathscr{V})$$

where $\mathscr{V}$ is the linear space generated by finite products of the matrices

$$e^{-As} D e^{As} \qquad (s \geq 0),$$

or also by finite products of matrices $[D, A]_k$, where

$$[D, A]_0 = D \qquad [D, A]_{k+1} := [[D, A]_k, A]$$

(and, as usual, $[M, N] = MN - NM$).

Theorem 6 concerns extremal controls, steering the initial point p to the boundary of the attainable set $\mathscr{A}_t \cdot p$. We are more interested in time-optimal controls, i.e. those admissible controls u on $[0, 1]$ such that the point $x = X_t(u) \cdot p$ has t as minimal time. It seems obvious that each optimal control is extremal (and we shall prove this explicitly in Corollary 8); however, the converse may well fail. Indeed consider the schematic figure 1. Here p is steered to q time-optimally; it is also steered to q on the boundary of $\mathscr{A}_1 \cdot p$, but $t = 1$ is obviously not the minimal time for q. Of course, it may happen that all extremal controls are optimal, if the initial point p is locally controllable ($\mathscr{A}_s \cdot p \subset \operatorname{int} \mathscr{A}_t \cdot p$ whenever $0 \leq s < t$).

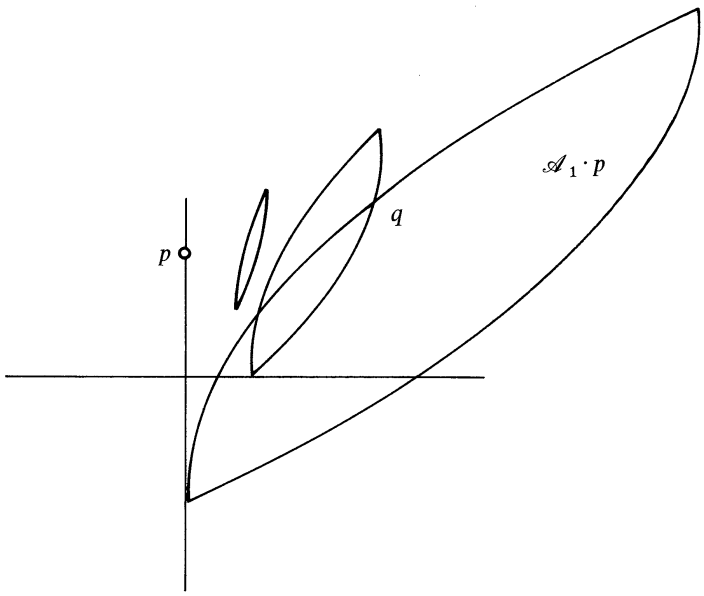

Fig. 1 Point q is on the boundary of the attainable set $\mathscr{A}_1 \cdot p$ but the minimal time is less than 1.

Lemma 7. *Assume that $\mathscr{A}_t \cdot p$ is convex for all $t \in [0, T)$. If $x \in \operatorname{int} \mathscr{A}_t \cdot p$ then also $x \in \operatorname{int} \mathscr{A}_s \cdot p$ for all s sufficiently close to t.*

Proof. Choose a ball U centered at x and entirely within $\operatorname{int} \mathscr{A}_t \cdot p$. In U take a simplex E with center x again, and with vertices $v_1, \ldots, v_{n+1} \in U$. Since each $v_k \in \mathscr{A}_t \cdot p$, there exist admissible controls $u_k(\cdot)$ such that $v_k = X_t(u_k)p$. If s is sufficiently close to t, we will still have x in the interior of the simplex E_s with vertices $X_s(u_k)p$. Since $\mathscr{A}_s \cdot p$ is convex (for small $|s - t|$),

$$x \in \operatorname{int} E_s \subset \mathscr{A}_s \cdot p,$$

as we wished to prove.

Corollary 8. *If $u(\cdot)$ is a time-optimal control for initial point p, then $u(\cdot)$ is* *extremal on* $[0, t]$ *for small* $t > 0$ (*i.e.*, $X_t(u)p \in \partial \mathscr{A}_t \cdot p$).

Proof. Choose T as described in Theorem 6. Assume the assertion fails, and that a point $x := X_t(u)p \notin \partial \mathscr{A}_t \cdot p$ for some $t \in [0, T)$. Since $x \in \mathscr{A}_t \cdot p$, necessarily $x \in \operatorname{int} \mathscr{A}_t \cdot p$; by Lemma 7,

$$x \in \operatorname{int} \mathscr{A}_s \cdot p \subset \mathscr{A}_s \cdot p$$

for all times $s < t$ close to t. But this contradicts the assumption that t was the least time to reach x from p.

Theorem 9. *For the system* (1) *with* (c^*, A, b) *controllable and observable, every time-optimal control is bang-bang and piecewise constant.*

Proof. Let $u(\cdot)$ be an admissible control on $[0, \theta]$ which is time-optimal for steering between two given points; let $x(\cdot)$ be the corresponding solution of (1). We invoke Lemma 1: to each point $x(t)$ there is a $\delta_t > 0$ such that

$$c^* x_s \neq 0 \qquad \text{in } (t, t+\delta_t] \text{ and in } [t-\delta_t, t)$$

(see (6) in Section 1). The cover $(t-\delta_t, t+\delta_t)$ of $[0, \theta]$ has a finite subcover. From the principle of optimality, $u(\cdot)$ is also time-optimal on each subinterval, and, in particular, on

$$[t-\delta_t, t] \quad \text{and} \quad [t, t+\delta_t].$$

By Corollary 8, $u(\cdot)$ is extreme on $[t, t+\delta_t]$ for the provisional initial point $x(t)$, and so it is bang-bang and piecewise constant there (Theorem 6). Similarly for $[t-\delta_t, t]$ (one may wish to change time orientation). Thus $u(\cdot)$ is indeed bang-bang on $[0, \theta]$, and piecewise constant over each member of the finite cover of $[0, \theta]$.

§ 3. Synthesis of extreme controls

The setting will again be that of a bilinear system in $\boldsymbol{R}^n$ with rank one inputs,

$$(1) \qquad \dot{x} = (A + ubc^*)x.$$

We begin with a technical lemma on which all else will be based. Our aim is to present the synthesis of extreme controls.

Lemma 10. *If the linear system* (c^*, A, b) *is controllable, then there exists* ε, $0 < \varepsilon \leq +\infty$, *such that, for any* n *numbers* $t_0, \ldots, t_{n-1}$ *subject to* $t_k > 0$, $\Sigma t_k < \varepsilon$, *the* n *vectors*

$$e^{A_0 t_0} b, \quad e^{A_1 t_1} e^{A_0 t_0} b, \quad e^{A_2 t_2} e^{A_1 t_1} e^{A_0 t_0} b, \ldots, e^{A_{n-1} t_{n-1}} \ldots e^{A_0 t_0} b$$

are linearly independent; here we have denoted $A_k := A + (-1)^k bc^*$.

Proof. From continuity it suffices to treat the case of $t_0 = 0$; we then wish to show that

$$D(t_1,\ldots,t_{n-1}) := \det(b,\ e^{A_1t_1}b,\ e^{A_2t_2}e^{A_1t_1}b,\ldots,e^{A_{n-1}t_{n-1}}\ldots e^{A_1t_1}b)$$

is nonzero for small $t_k > 0$. It is probably obvious that $D(\cdot)$ is an entire function of its variables, so that it can be expressed as a power series in the t_k. Evidently $D(\cdot)$ vanishes when some $t_k = 0$, so that it has t_k (or some higher power) as root factor. We shall examine this further.

By subtracting the first column from the second we obtain $(e^{A_1t_1} - I)b$, and hence

$$\lim_{t_1\to 0}\frac{1}{t_1}D(t_1,\ldots,t_{n-1}) = \det(b,\ A_1b,\ e^{A_2t_2}b,\ldots,e^{A_{n-1}t_{n-1}}\ldots e^{A_2t_2}b).$$

In the second column

$$A_1b = (A - bc^*)b = Ab - b\cdot(c^*b),$$

so that another column operation, using b, yields

$$\det(\cdots) = \det(b,\ Ab,\ e^{A_2t^2}b,\ldots,e^{A_{n-1}t_{n-1}}\cdot\ldots\cdot e^{A_2t_2}b).$$

In processing the third column we shall subtract $(I + A_2t_2)b$, which is a linear combination of the first two columns:

$$\det(\cdots) = \det(b,\ Ab,\ (e^{A_2t_2} - I - A_2t_2)b,\ e^{A_3t_3}e^{A_2t_2}b,\ldots).$$

Therefore

$$\lim_{t_1,t_2\to 0}\frac{D(t_1,\ldots,t_{n-1})}{t_1t_2^2} = \det\left(b,\ Ab,\ \frac{1}{2}A_2^2b,\ e^{A_3t_3}b,\ldots,e^{A_{n-1}t_{n-1}}\cdot\ldots\cdot e^{A_3t_3}b\right).$$

Again we may operate with the first two columns to replace A_2^2b by A^2b; indeed,

$$A_2^2b = (A + bc^*)^2b = A^2b + Ab\cdot(c^*b) + b\cdot(c^*Ab + (c^*b)^2).$$

In point of fact all we need for these operations is the observation that

$$A_k^kb \in A^kb + \operatorname{span}(b,\ Ab,\ldots,A^{k-1}b).$$

After $n - 1$ such steps we obtain

$$\lim_{t_k\to 0}\frac{D(t_1,\ldots,t_{n-1})}{t_1t_2^2\cdots t_{n-1}^{n-1}} = \frac{1}{1!\,2!\cdots(n-1)!}\det(b,\ Ab,\ldots,A^{n-1}b).$$

By controllability the last determinant is $\neq 0$. We have thus obtained that

$$D(t_1, \ldots, t_{n-1}) = t_1 t_2^2 \cdots t_{n-1}^{n-1} \cdot D_0(t_1, \ldots, t_{n-1})$$

where $D_0(\cdot)$ is again entire, and $D_0(0, \ldots, 0) \neq 0$. Thus indeed $D(t_1, \ldots, t_{n-1}) \neq 0$ for small $t_k > 0$.

Theorem 11 (Synthesis of exteme controls). *Consider system* (1), *assuming* (c^*, A, b) *is controllable and observable; and any initial point* $p \neq 0$. *Then there exists* ε, $0 < \varepsilon \leq +\infty$, *such that an admissible control* $u(\cdot)$ *on* $[0, t] \subset [0, \varepsilon)$ *is an extreme control for steering* p *if, and only if,* $u(\cdot)$ *is a bang-bang piecewise constant control with at most* $n - 1$ *switches.*

Proof. We shall first select ε; then show that every extreme control is of the described type; and last, prove that every such switching control is extremal.

From controllability of (c^*, A, b), each $p \neq 0$ is completely controllable. Thus by Lemma 1, there exists $\varepsilon > 0$ such that

$$c^* X_s(u) p \neq 0 \qquad \text{on } (0, \varepsilon) \tag{2}$$

for all admissible controls $u(\cdot)$. By decreasing $\varepsilon > 0$ further we ensure that all attainable sets $\mathscr{A}_t \cdot p$ are strictly convex for $0 < t < \varepsilon$ (Theorem 6). A further decrease of $\varepsilon > 0$ provides the conclusion of Lemma 10.

Consider now any extreme control $u(\cdot)$ on $[0, t] \subset [0, \varepsilon)$. By Theorem 6, $u(\cdot)$ is bang-bang and piecewise constant; we wish to check the number of switches. By the proof of Theorem 6, $u(s) = \operatorname{sgn} \alpha(s)$, where

$$\alpha(s) = (q^* X_t \cdot X_s^{-1} b) \cdot (c^* X_s p) \tag{3}$$

(and $X_s = X_s(u)$). The second factor in (3) does not vanish in $(0, \varepsilon)$, by (2). Assume now that $u(\cdot)$ has at least n switches, at points s_k,

$$0 < s_1 < s_2 < \cdots < s_n < \varepsilon.$$

Then $\alpha(s)$ must vanish at the s_k, and hence also the first factor in (3). In other terms, the nonzero vector $q_1 = X_t^* q$ is perpendicular to the n vectors $X_{s_k}^{-1}(u) \cdot b$. This is impossible, since these are independent: Lemma 10 with $t_o = s_1$, $t_1 = s_2 - s_1, \ldots, t_{n-1} = s_n - s_{n-1}$ (and $-A$ in place of A). This contradiction shows that indeed an extreme control on $[0, \varepsilon]$ has at most $n - 1$ switches.

Finally, consider any $t \in (0, \varepsilon)$; let E be the collection of all extreme controls on $[0, t]$, steering p to the various points of $\partial \mathscr{A}_t \cdot p$; let F be the collection of all bang-bang admissible controls with at most $n - 1$ switches on $[0, t]$. We have already shown that

$$E \subset F, \tag{4}$$

and now we wish to prove equality.

This will be a little roundabout. From Theorem 6, $\mathscr{A}_t \cdot p$ is a strictly convex (and compact nonvoid) subset of $\boldsymbol{R}^n$; thus its boundary is homeomorphic to the $n-1$ sphere, $\partial \mathscr{A}_t \cdot p \approx S^{n-1}$. Next, the natural mapping $u(\cdot) \mapsto X_t(u)p$ maps E onto $\partial \mathscr{A}_t \cdot p$. If we use the weak topology (or the L_1-topology), the mapping is continuous. Uniqueness in Theorem 6 then yields $E \approx \partial \mathscr{A}_t \cdot p$, so that $E \approx S^{n-1}$. With the same topology, $F \approx S^{n-1}$ (e.g. use a linear control system in $\boldsymbol{R}^n$, controllable and single-input (hence, normal) and the natural correspondence between F and the boundary of a reachable set R_t, $\partial R_t \approx S^{n-1}$; the details are in [3]).

Thus in (4) we have $E \approx S^{n-1} \approx F$. Now use the Preservation of Domain Theorem, e.g. in the formulation of [2, Thm. 3.9, p. 303]: if two subsets of an m-manifold M are homeomorphic, and one is open in M, then so is the other. Indeed, $F \approx S^{n-1}$ is an $(n-1)$-manifold, open in itself; then $E \subset F$ is homeomorphic to $S^{n-1} \approx F$, so E is also open in F. In addition, E is compact and non-void, and F is connected. Thus necessarily $E = F$.

This concludes the proof.

The synthesis of extreme controls from Theorem 11 provides a construction of attainable and reachable sets. Given (1) and $p \neq 0$, one finds (at least in principle) $\varepsilon > 0$ as described in the statement. For each $t \in (0, \varepsilon)$ and $k = 0, 1, \ldots, n-1$ let $E_k^+(t)$ be the collection of all bang-bang control functions $[0, t]$ with precisely k switches, and with value 1 on the last interval; and similarly for $E_k^-(t)$, and last value -1. ($E_k^+(t)$ is a k-manifold, and $\bigcup_{k=0}^{n-1} E_k^+(t) \cup E_k^-(t)$ is a disjoint decomposition of S^{n-1}.)

Next, let

$$M_k^+(t) = \{X_t(u)p : u \in E_k^+(t)\}, \tag{5}$$

and analogously for $M_k^-(t)$. Then we obtain that

$$\partial \mathscr{A}_t \cdot p = \bigcup_{k=0}^{n-1} M_k^+(t) \cup M_k^-(t),$$

a disjoint composition into k-manifolds.

In practice one might carry this through for trial values of terminal time $t > 0$ without first checking that $t < \varepsilon$. This is particularly simple in the case of $n = 2$: with $t > 0$ fixed, the extreme controls are

$$u(r) = \begin{cases} -1 & \text{for } r \in [0, s] \\ 1 & \text{for } r \in (s, t] \end{cases}$$

(or $-u(\cdot)$), with a single switch parameter $s \in [0, t]$. The resulting trajectory end-points trace two parametric curves with common end-points; these

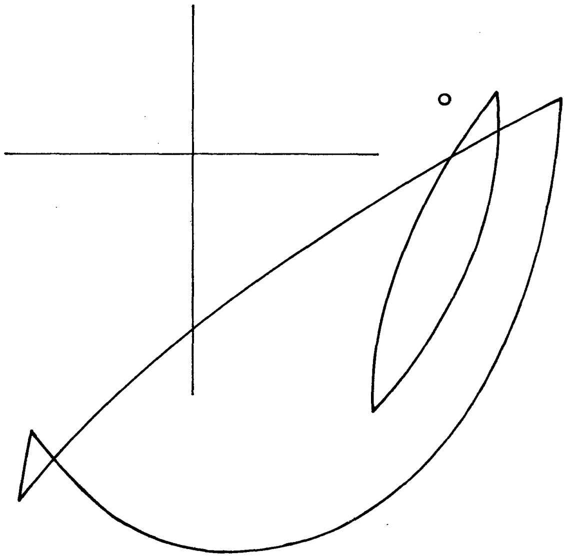

Fig. 2 Control of $\dot{x} = y$, $\dot{y} = (1 - u)x$, $-1 \leq u \leq 1$; attainable sets at times 0.7 and 1.7 from initial point (2, 0.5).

constitute the boundary of an attainable set $\mathscr{A}_t \cdot p$. Obvious misbehavior of these boundary curves signals that the time length t has been taken too large; see Fig. 2.

Implicit in the above is a construction of feedback controls: still given p and ε, define $\psi(\cdot)$ thus: for $t \in (0, \varepsilon)$ let

$$\psi(x, t) = \begin{cases} 1 & \text{if } x \in \bigcup_{k=1}^{n-1} M_k^+(t) \\ -1 & \text{if } x \in \bigcup_{k=1}^{n-1} M_k^-(t) \end{cases}$$

(see (5)); then the extreme solutions $x(\cdot)$ issuing from p satisfy

$$\dot{x}(t) = (A + \psi(x(t), t)bc^*)x(t). \tag{6}$$

Here the domain of ψ consists of all (x, t) subject to

$$x \in \partial\mathscr{A}_t \cdot p, \qquad 0 < t < \varepsilon$$

(from Corollary 8, $\cup\, \partial\mathscr{A}_t \cdot p = \cup\, \mathscr{A}_t \cdot p$).

Uniqueness provides a state-dependent feedback for optimal controls. Indeed, consider any $x \in \cup \{\mathscr{A}_t \cdot p : 0 \leq t < \varepsilon\}$; let $\theta = T(x)$ be the minimal time,

$$T(x) = \min \{t \geq 0 : x \in \mathscr{A}_t \cdot p\}.$$

Then $x \in \partial\mathscr{A}_\theta \cdot p$ (Corollary 8); thus x is attained by an admissible control $u(\cdot)$ on $[0, \theta]$ uniquely determined by x, and $u(\cdot)$ is bang-bang (Theorem 6). One may then define

$$\varphi(x) := \lim_{s \to \theta -} u(s),$$

and time-optimal solutions $x(\cdot)$ issuing from p satisfy

$$\dot{x} = (A + \varphi(x) bc^*)x \tag{7}$$

a.e. Here the domain of φ is $\cup\{\mathscr{A}_t \cdot p\colon\ 0 \leq t < \varepsilon\}$.

§ 4. Concluding remarks

In Section 3, a complete description was given for the class of exteme controls (over short time intervals). The time-optimal controls form a subset of these (Corollary 8); it is natural to ask how can one distinguish them.

Conjecture 1. For (1) with (c^*, A, b) minimal, let $u(\cdot)$ be extremal, and $q \neq 0$ an exterior normal to $\mathscr{A}_t \cdot p$ at $x = X_t(a)p$. Then $u(\cdot)$ is time-optimal if

$$q^* X_s(u)p \leq q^* X_t(u)p$$

for $s \to t-$.

This is suggested by Fig 1, which also shows that the condition is not necessary (e.g. see the corner points).

In constructing feedback controls, it is rather superficial to present equations such as (6) and (7), observing that extreme and optimal solutions also satisfy these feedback equations. This is often quite irrelevant: it is the converse that is of some practical interest (whether solutions of the feedback equation are necessarily the optimal solutions to the original control problem).

Conjecture 2. For (1) with (c^*, A, b) minimal and $p \neq 0$, the time-optimal feedback equation (7) has stability with respect to measurement (in the sense of [4]); in particular, every generalized solution $x(\cdot)$ of (7) with initial value $x(0) = p$ is a time-optimal solution of (1).

Added in proof: An obvious modification of the proof of Theorem 9 extends the conclusion to all controls that are extremal (cf. Corollary 8).

References

[1] Brockett, R., On the reachable set for bilinear systems, *Variable Structure Systems*, Lecture Notes in Economics and Mathematical Systems 111, 55–63, Springer-Verlag, 1971.

[2] Eilenberg, S. and Steenrod, N., *Foundations of Algebraic Topology*, Princeton University Press, 1952.

[3] Hájek, O., Terminal manifolds and switching locus, Math. Systems Theory, **6** (1973) 298–301.

[4] Hájek, O., Discontinuous differential equations I, II, J. Diff. equations, **32** (1979) 149–185.
[5] Hájek, O. and Loparo, K. A., Bilinear control: geometric properties of reachable sets, Advances in Optimization and Control (eds., M.A. Eiselt, G. Pederzoli), Lecture Notes in Economics and Mathematical Systems **302**, 252–273, Springer Verlag, 1988.
[6] Hermes, H. and LaSalle, J. P., *Functional Analysis and Time-Optimal Control*, Academic Press, 1969.
[7] Lee, E. B. and Markus, L., *Foundations of Optimal Control Theory*, Wiley, 1967.

nuna adreso:
Department of Mathematics and Statistics
Case Western Reserve University
Cleveland, Okio 44106
U.S.A.

(Ricevita la 4-an de aŭgusto, 1989)