[10] ——, "Recursive least squares ladder forms for fast parameter tracking," in *Proc. 1978 IEEE Conf. Decision Contr.*, San Diego, CA, Jan. 1979, pp. 1362–1367.
[11] E. H. Satorius and J. D. Pack, "Application of least squares lattice algorithms to adaptive equalization," *IEEE Trans. Commun.*, vol. COM-29, pp. 136–142, Feb. 1981.
[12] E. H. Satorius and M. J. Shensa, "On the application of recursive least squares methods to adaptive processing," presented at the Int. Workshop Applications of Adaptive Contr., Yale Univ., New Haven, CT, Aug. 1979.
[13] D. Lee and M. Morf, "Recursive square-root ladder estimation algorithms," presented at the IEEE Conf. Acoust., Speech, Signal Processing, Denver, CO, Apr. 1980.
[14] M. J. Shensa, "A least squares lattice decision feedback equalizer," presented at the IEEE Int. Conf. Commun., Seattle, WA, June 1980.
[15] B. Friedlander, M. Morf, T. Kailath, and L. Ljung, "New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices," to be published.

Solvability, Controllability, and Observability of Continuous Descriptor Systems

ELIZABETH L. YIP AND RICHARD F. SINCOVEC

Abstract—**In this paper, we investigate the properties of the continuous descriptor system**

$$E\dot{x}(t)=Ax(t)+Bu(t), \qquad 0\leq t\leq b$$

where E, A, and B are complex and possibly singular matrices and $u(t)$ is a complex function differentiable sufficiently many times. The traditional approach to such systems is to separate the state equations from the algebraic equations. However, such algorithms usually destroy the natural, physically-based sparsity and structure of the original system. Therefore, we consider descriptor systems in their original form. Such systems possess numerous properties not shared by the well-known state variable systems. First, we relate classical theories of matrix pencils to the solvability of descriptor systems. Then we extend the concepts of reachability, controllability, and observability of state variable systems to descriptor systems, and describe the set of reachable states for descriptor systems.

I. Introduction

In this paper, we consider the *continuous* descriptor system

$$E\dot{x}(t)=Ax(t)+Bu(t), \qquad x(0)=x_0 \tag{1}$$

where E, A are $n\times n$ complex matrices, B is an $n\times k$ complex matrix, $x(t)$ is an unknown complex vector, and $u(t)$ is a function differentiable sufficiently many times (we shall say $u(t)$ is sufficiently differentiable).

Solvability of the system (1) has been discussed in Gantmacher [1] and Wilkinson [2] for B equal to the identity matrix. In the next section we shall show how their results can be generalized to arbitrary B. Luenberger [4] has considered solvability of discrete descriptor systems. We summarize these discussions in Section II and establish our notation for the remainder of this paper. We include the proofs of these well-known results in the Appendix for completeness. In Section III we define reachability for the system (1), and describe precisely the reachable set. The results in Section III are what is expected; however, the proofs of these results are far from trivial. In Section III we also extend Wonham's [5] geometrical treatment of state variable systems to descriptor systems.

Manuscript received September 6, 1979; revised June 6, 1980 and December 9, 1980. Paper recommended by A. J. Laub, Chairman of the Computational Methods and Discrete Systems Committee. This work was supported by the Department of Energy under Contract ET-78-C-01-2876. Part of the material of this paper was presented at the SIAM Spring Conference, Toronto, Canada, June 1979.

E. L. Yip is with Boeing Computer Services Company, Numerical Analysis Group, Tukwila, WA 98188.

R. F. Sincovec is with the College of Engineering and Applied Science, University of Colorado, Colorado Springs, CO 80907.

We find Wonham's approach an invaluable tool for rigorous formal proofs of mathematical theories, even though the methodology seems to be far removed from the physical problem. In Section IV we define observability and two types of controllabilities and derive results corresponding to those obtained in Section III. The results in Section IV are again what is expected since they are intuitive extensions of the results for state variable systems. The proofs of these results are directly derivable from the results concerning reachable sets in Section III. Finally, in the last section we summarize our results.

II. Solvability

It is reasonable to define solvability of the system (1) as the existence of a unique solution for any given sufficiently differentiable $u(t)$ and any given admissible initial condition corresponding to the given $u(t)$. Gantmacher's [1] analysis shows that the system (1) with $B=I$ is solvable if and only if the matrix pencil $E+\lambda A$ (or equivalently, $A-\lambda E$) is regular. Note that, since the set $\{f(t)\colon f \text{ sufficiently differentiable}\}$ contains the set $\{Bu(t)\colon u \text{ sufficiently differentiable}\}$, the following is true.

Fact 1: If the system represented by (1) with B equal to the identity matrix is solvable for any sufficiently differentiable $u(t)$, then (1), with arbitrary B, is solvable for any sufficiently differentiable $u(t)$.

For convenience, we make the following definition.

Definition 1: (A,E) is *solvable* if the matrix pencil $E+\lambda A$ is regular, i.e., $\det(E+\lambda A)\neq 0$ for all except a finite number of $\lambda\in\mathbb{C}$ where $\mathbb{C}$ is the field of complex numbers.

The following characterization of solvability is based on Gantmacher's [1] analysis of matrix pencils and Luenberger's [4] analysis of discrete descriptor systems.

Theorem 1: The following statements are equivalent.

a) (A,E) is solvable.

b) If X_0 is the null space of A (denoted by $\operatorname{Ker} A$) and

$$X_i=\{x\colon Ax\in EX_{i-1}\} \text{ then } \operatorname{Ker} E\cap X_i=0 \text{ for } i=0,1,2,3,\cdots.$$

c) If $Y_0=\operatorname{Ker} A^T$ and $Y_i=\{x\colon A^Tx\in E^TY_{i-1}\}$ then $\operatorname{Ker} E^T\cap Y_i=0$ for $i=0,1,2,3,\cdots$.

d) The matrix

$$G(n)=\left.\begin{bmatrix} E & 0 & \cdots & 0\\ A & E & \cdots & 0\\ 0 & A & \cdots & 0\\ & & & E\\ 0 & & \cdots & A\end{bmatrix}\right\} n+1$$

has full column rank for $n=1,2,\cdots$.

e) The matrix

$$F(n)=\underbrace{\begin{bmatrix} E & A & 0 & \cdots & 0\\ 0 & E & A & \cdots & 0\\ \vdots & & & & \vdots\\ 0 & & \cdots & E & A\end{bmatrix}}_{n+1}$$

has full row rank for $n=1,2,\cdots$.

f) There exist nonsingular matrices P and Q such that (1) is decomposed into possibly two subsystems: a subsystem with only state variables, and an algebraic-like subsystem, i.e., $PEQQ^{-1}\dot{x}(t)=PAQQ^{-1}x(t)+PBu(t)$ has one of the following forms.

i)
$$\begin{cases}\dot{y}_1=E_1y_1+B_1u\\ E_2\dot{y}_2=y_2+B_2u, \quad E_2^m=0, \quad E_2^{m-1}\neq 0.\end{cases}$$

(In this case, both E and A are singular, or A is nonsingular and E is singular but not nilpotent, i.e., $E^m\neq 0$ for all positive integer m.)

ii)
$$\dot{y}_1=E_1y_1+B_1u.$$

(In this case, E is nonsingular.)

iii) $E_2\dot{y}_2 = y_2 + B_2u, \quad E_2^m = 0, \quad E_2^{m-1} \neq 0.$

(In this case, A is nonsingular and E is nilpotent.)

In all cases

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Q^{-1}x, \quad \text{and} \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = PB$$

and the exact solution is

$$y_1(t) = e^{tE_1}y_{10} + \int_0^t e^{(t-s)E_1}B_1u(s)\,ds$$

$$y_2(t) = -\sum_{i=0}^{m-1} E_2^iB_2u^{(i)}(t) \tag{2}$$

where y_{10} is the transformed initial condition, i.e.,

$$\begin{bmatrix} y_{10} \\ y_{20} \end{bmatrix} = Q^{-1}x_0.$$

The proof of this theorem is given in the Appendix.

Note that statements b) and d) are equivalent to the absence of the row minimum indices (see Gantmacher [1] for the definition of minimum indices) and that statements c) and e) are equivalent to the absence of the column minimum indices. Since the matrices E and A are square, the absence of the row minimum indices is equivalent to the absence of the column minimum indices. Also, in the special case that $B = I$, Gantmacher [1] has shown that the presence of the column minimum indices implies constraints on $u(t)$, and so the system is not solvable for all sufficiently differentiable $u(t)$, while the presence of row minimum indices yields infinitely many solutions for any sufficiently differentiable $u(t)$. Thus, when E and A are square matrices and B is the identity (or more generally, B is nonsingular), existence of a solution for any sufficiently differentiable $u(t)$ for the system (1) implies uniqueness of the solution for any sufficiently differentiable $u(t)$. Concerning statements d) and e), we note that Luenberger [4] defined the system (1) as "conditionable" if $G(n)$ is of full column rank for all $n = 1,2,\cdots$, and the system (1) as "solvable" if $F(n)$ is of full row rank for all $n = 1,2\cdots$.

In the remainder of this paper, we shall assume (A, E) to be solvable and that the system (1) is of the form

$$\dot{x}_1(t) = E_1x_1(t) + B_1u(t)$$

$$E_2\dot{x}_2(t) = x_2(t) + B_2u(t) \tag{3}$$

where

E_1 is an $n_1 \times n_1$ complex matrix,
E_2 is an $n_2 \times n_2$ complex matrix, and $E_2^m = 0$, $E_2^{m-1} \neq 0$,
B_1 is an $n_1 \times k$ complex matrix,
B_2 is an $n_2 \times k$ complex matrix,
$x_1(t)$ is an unknown vector in $\mathbb{C}^{n_1}$, complex vectors of length n_1,
$x_2(t)$ is an unknown vector in $\mathbb{C}^{n_2}$, complex vectors of length n_2,
$u(t)$ is a function differentiable at least $m-1$ times.

We call (3) the *standard canonical form* of the descriptor system and m the *degree of nilpotency* of (3). Note that if $E_2 = 0$ then (3) has its degree of nilpotency $m = 1$. If the second equation of (3) is missing, we say the degree of nilpotency of (3) is zero. Sometimes we refer to the first and second equations of (3) as the state variable equation and the algebraic equation, respectively.

Note that in the case of the state variable equation, every vector in the vector space is an admissible initial condition. This is not the case with the descriptor system.

From (2), $y_1(0) = y_{01} \in \mathbb{C}^{n_1}$, and $y_2(0) = -\sum_{i=0}^{m-1} E_2^iB_2u^{(i)}(0)$. Thus, the following statement is true.

Let U be the set of functions $u(t) \in \mathbb{C}^k$, complex vectors of length k, such that $u(t)$ is differentiable at least $m-1$ times. The set of admissible initial conditions for the system (3) is

$$I = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : x_1 \in \mathbb{C}^{n_1}, x_2 = -\sum_{i=0}^{m-1} E_2^iB_2u^{(i)}(0), u \in U \right\}.$$

Note that the zero vector 0 is in I because there exists $u \in U$ such that $u^i(0) = 0$ for $i = 1,2,\cdots$, and the set I is a subspace.

III. Reachability

For (3), we say a state x_1 is reachable from a state x_0 if and only if there exists $u \in U$ such that the solution

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

has $x(0) = x_0$ and $x(t_1) = x_1$ for some $t_1 > 0$. Let $R(x)$ be the set of reachable states from $x \in I$. In this section we precisely describe $R(x)$. First, we restrict our attention to $x_0 = 0$ and describe $R(0)$. To accomplish this, we define the notation $\langle \cdot | \cdot \rangle$ for an arbitrary matrix pair (E, B), where E is a square matrix and the product EB is well defined:

$$\langle E | B \rangle = \beta + E\beta + E^2\beta + \cdots + E^{n-1}\beta$$

where n is the order of E and $\beta = \text{Image } B = \text{Im } B = \{y: y = Bx, \text{ all possible } x\}$.

Theorem 2: $R(0) = \langle E_1 | B_1 \rangle \oplus \langle E_2 | B_2 \rangle$. E_1, E_2, B_1, B_2 are as defined in (3).

The proof of Theorem 2 is constructive and requires the following two lemmas.

Lemma 1: For any polynomial $f(s) \in \mathbb{C}$ not identically zero and for any $z \in \mathbb{C}^{n_1}$, define $W(f,t)$: $\mathbb{C}^{n_1} \to \mathbb{C}^{n_1}$ by

$$W(f,t)z = \int_0^t \left(f(s)e^{sE_1}B_1B_1^Te^{sE_1^T}f(s) \right) z\,ds.$$

If $\text{Im } W(f,t)$ is the range (or image space) of $W(f,t)$, then $\text{Im } W(f,t) = \langle E_1 | B_1 \rangle$.

Lemma 2: For arbitrary vectors, $x_i, y_i \in \mathbb{C}^k$, $i = 0,1,2,\cdots,m-1$, and $t > 0$, there exists a vector polynomial $u(t) \in \mathbb{C}^k$ of degree $2m-1$ such that $u^{(i)}(0) = x_i$ and $u^{(i)}(t) = y_i$.

Proof of Theorem 2: We shall first prove that $R(0) \subseteq \langle E_1 | B_1 \rangle \oplus \langle E_2 | B_2 \rangle$. Suppose

$$x = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \in R(0);$$

then, by definition of $R(0)$, there exists $u \in U$ such that

$$x_1(t) = \int_0^t e^{(t-s)E_1}B_1u(s)\,ds = \sum_{i=0}^{n_1-1} \int_0^t \psi_i(t-s)E_1^iB_1u(s)\,ds$$

for some polynomial function $\psi_i(s)$, $i = 0,1,2,\cdots,n_1$ over $\mathbb{C}$ because

$$e^{(t-s)E_1} = I + (t-s)E_1 + \frac{1}{2}(t-s)^2E_1^2 + \cdots + \frac{1}{(n_1-1)!}(t-s)^{n_1-1}E_1^{n_1-1} + \cdots$$

and $E_1^{n_1} = P(E_1)$, where $P(x)$ is a polynomial of degree less than n_1 and

$$x_2(t) = -\sum_{i=0}^{m-1} E_2^iB_2u^{(i)}(t).$$

Thus, $x_1 \in \langle E_1 | B_1 \rangle$ and $x_2 \in \langle E_2 | B_2 \rangle$.

Now we shall show $R(0) \supseteq \langle E_1 | B_1 \rangle \oplus \langle E_2 | B_2 \rangle$. Assume

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

is such that $x_1 \in \langle E_1 | B_1 \rangle$, $x_2 \in \langle E_2 | B_2 \rangle$, and $x_1 \neq 0$, $x_2 \neq 0$. We have to find $u \in U$ such that X is a solution of (3) at some $t > 0$. We choose $u(t)$ to

be of the form $u(t)=u_1(t)+u_2(t)$. Thus

$$x_1(t)=\int_0^t e^{(t-s)E_1}B_1u_1(s)\,ds+\int_0^t e^{(t-s)E_1}B_1u_2(s)\,ds,$$

$$x_2(t)=-\sum_{i=0}^{m_1-1}E_2^iB_2u_1^{(i)}(t)-\sum_{i=0}^{m-1}E_2^iB_2u_2^{(i)}(t).$$

We choose $\tilde{t}>0$ arbitrarily and choose $u_1(t)$ to be of the form

$$u_1(t)=t^m(t-\tilde{t})^m y(t) \qquad \text{for some } y(t)\in\mathbb{C}^k. \tag{4}$$

Thus, $u_1^{(i)}(t)=u_1^{(i)}(0)=0$ if $i<m$ and $u_1(t)$ contributes nothing to $x_2(t)$ at $t=0$ and $t=\tilde{t}$. We choose y to satisfy the following equation.

$$\int_0^{\tilde{t}}e^{(\tilde{t}-s)E_1}B_1s^m(s-\tilde{t})^m y(s)\,ds=x_1-\int_0^{\tilde{t}}e^{(\tilde{t}-s)E_1}B_1u_2(s)\,ds\equiv\tilde{x}_1. \tag{5}$$

How we choose a y that satisfies (5) will be described later. We call the expression on the right-hand side of (5), $\tilde{x}_1$. For now, let us consider the choice of $u_2(t)$. Note that $x_2\in\langle E_2|B_2\rangle$ implies

$$x_2=-\sum_{j=0}^{m-1}x_{2,j}$$

where $x_{2,j}\in E_2^j\beta_2$ and β_2 is defined as the image (or range space) of B_2. Note that $x_{2,j}\in E_2^j\beta_2$ implies that there exists $y_j\in\mathbb{C}^k$ such that $x_{2,j}=E_2^jB_2y_j$. Now, by Lemma 2, there exists a polynomial $h(s)\in\mathbb{C}^k$ of degree $2m-1$ such that $h^{(j)}(0)=0$ and $h^{(j)}(\tilde{t})=y_j$ for $j=0,1,2,\cdots,m-1$. Let $u_2(t)=h(t)$. Thus, $x_2(t)=-\sum_{i=0}^{m-1}E_2^iB_2u_2^{(i)}(t)=x_2$, $x_2(0)=0$.

Now we describe how we choose y in (4) to satisfy (5). Note that $\tilde{x}_1$, which is the expression on the right-hand side of (5), is in $\langle E_1|B_1\rangle$, which, by Lemma 1, is the image (or range space) of $W(f,t)$ for any polynomial $f\in\mathbb{C}$. In other words, Lemma 1 implies there exists $z\in\mathbb{C}^{n_1}$ such that $W(f,\tilde{t})z=\tilde{x}_1$. Let $f(s)=s^m(s-\tilde{t})^m$ and

$$y(s)=f(s)B_1^Te^{(\tilde{t}-s)E_1^T}z.$$

Then, by (4),

$$u_1(s)=\big(f(s)\big)^2B_1^Te^{(\tilde{t}-s)E_1^T}z.$$

Thus,

$$\begin{aligned}\int_0^{\tilde{t}}e^{(\tilde{t}-s)E_1}B_1u_1(s)\,ds&=\int_0^{\tilde{t}}f(s)e^{(\tilde{t}-s)E_1}B_1B_1^Te^{(\tilde{t}-s)E_1^T}f(s)z\,ds\\&=W(f,\tilde{t})z\\&=x_1.\end{aligned}$$

Thus, (5) is satisfied and we have accomplished our purpose. We have shown how to choose $u\in U$ such that $u(t)=u_1(t)+u_2(t)$ with

$$\int_0^t e^{(t-s)E_1}B_1u(s)\,ds=x_1$$

and

$$-\sum_{i=0}^{m-1}E_2^iB_2u^{(i)}(t)=x_2.$$

To complete the proof of Theorem 2, we need to prove Lemmas 1 and 2.

Proof of Lemma 1: (Lemma 1 is actually a generalization of equation (4) on p. 36 of Wonham [5].)

To show $\mathrm{Im}(W(f,t))=\langle E_1|B_1\rangle$ is equivalent to showing that

$$\operatorname{Ker}W(f,t)=\bigcap_{i=0}^{n_1-1}\operatorname{Ker}B_1^T\big(E_1^T\big)^i. \tag{6}$$

We first show $\operatorname{Ker}W(f,t)\subseteq\cap_{i=0}^{n_1-1}\operatorname{Ker}B_1^T(E_1^T)^i$.

If $x\in\operatorname{Ker}W(f,t)$ and $x\neq0$, then

$$0=x^TW(f,t)x=\int_0^t\|B_1^Te^{sE_1^T}f(s)x\|^2\,ds$$

and thus,

$$0=B_1^Te^{sE_1^T}f(s)x \qquad \text{for } 0\leqslant s\leqslant t.$$

Since $f(s)$ can have only finitely many zeros in the interval $0\leqslant s\leqslant t$, it follows that

$$0=B_1^Te^{sE_1^T}x \qquad \text{for } 0\leqslant s\leqslant t. \tag{7}$$

Repeated differentiation of (7) gives the result

$$0=B_1^T\big(E_1^T\big)^ie^{sE_1^T}x \qquad \text{for } i=0,1,2,\cdots,n_1-1 \tag{8}$$

and $0\leqslant s\leqslant t$.

It follows from (8) that

$$0=B_1^T\big(E_1^T\big)^ix \qquad \text{for } i=0,1,2,\cdots,n_1-1 \tag{9}$$

and, therefore,

$$x\in\bigcap_{i=0}^{n_1-1}\operatorname{Ker}B_1^T\big(E_1^T\big)^i.$$

To show $\operatorname{Ker}W(f,t)\supseteq\cap_{i=0}^{n_1-1}\operatorname{Ker}B_1^T(E_1^T)^i$, note that if

$$x\in\bigcap_{i=0}^{n_1-1}\operatorname{Ker}B_1^T\big(E_1^T\big)^i$$

then

$$B_1^Te^{sE_1^T}x=\sum_{i=0}^{n_1-1}\psi_i(s)B_1^T\big(E_1^T\big)^ix=0$$

for some polynomial ψ_i; therefore, $x\in\operatorname{Ker}W(f,t)$. Thus, (6) is true and the proof is complete.

Proof of Lemma 2: Let $x_i=(x_{i1},x_{i2},\cdots,x_{ik})^T$ and $y_i=(y_{i1},y_{i2},\cdots,y_{ik})^T$ for $i=0,1,\cdots,m-1$. For each $j=1,2,\cdots,k$, the classical Hermite interpolation [3] gives a polynomial $h_j(t)$ of degree $2m-1$ such that

$$h_j^{(i)}(0)=x_{ij},\quad h_j^{(i)}(t)=y_{ij},\qquad i=0,1,2,\cdots,m-1.$$

Thus, the equations in Lemma 2 are satisfied by $u(t)$ defined as follows:

$$u(t)=\big(h_1(t),h_2(t),\cdots,h_k(t)\big)^T.$$

The following theorem, which we use in a later section, generalizes Theorem 2.

Theorem 3: Let I_0 be the set of admissible initial conditions for (3) such that the components corresponding to the state equations are zero, i.e.,

$$I_0=\left\{\begin{bmatrix}x_1\\x_2\end{bmatrix}:x_1=0,\ x_2=-\sum_{i=0}^{m-1}E_2^iB_2u^{(i)}(0),\ u\in U\right\}.$$

Then, for $x_0\in I_0$, $R(x_0)=\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle$.

The complete set of admissible initial conditions for (3) is

$$I=\mathbb{C}^{n_1}\oplus\langle E_2|B_2\rangle.$$

If $\hat{x}\in I$ then $R(\hat{x})=\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle+H(x)$ where

$$H(x)=\left\{\begin{bmatrix}z_1\\z_2\end{bmatrix}:z_1=e^{tE_1}x_1,\ t>0,\ z_2=0\right\}.$$

Thus, the complete set of reachable states is

$$R=\bigcup_{x\in I} R(x)=\mathbb{C}^{n_1}\oplus\langle E_2|B_2\rangle$$

and, hence, $I=R$.

Proof of Theorem 3: As defined in the definition of I_0, x_2 is in $\langle E_2|B_2\rangle$ and $x_1=0$. Thus, by Theorem 2, $R(x_0)=\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle$. Since any vector in $\mathbb{C}^{n_1}$ is an admissible initial condition for the state equations, all the admissible initial conditions for the algebraic equation have to be of the form x_2. Hence, the complete set of admissible initial conditions is $I=\mathbb{C}^{n_1}\oplus\langle E_2|B_2\rangle$. Thus, if $x\in I$, i.e., $x(0)=\hat{x}_1$, then

$$x_1(t)=e^{tE_1}x_1-\int_0^t e^{(t-s)E_1}B_1u(s)\,ds$$

$$x_2(t)=\sum_{i=0}^{m-1}E_2^iB_2u^i(t).$$

Therefore, $R(x)=\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle+H(x)$.

Since $H(x)\in\mathbb{C}^{n_1}\oplus\{0\}$, where 0 indicates the vector 0 in $\mathbb{C}^{n_2}$, the reachable set is

$$R=\bigcup_{x\in I} R(x)=\mathbb{C}^{n_1}\oplus\langle E_2|B_2\rangle.$$

We have described precisely the complete set of reachable states for descriptor systems in terms of their standard canonical forms.

We do not have corresponding statements regarding sets of reachable states and admissible initial conditions for descriptor systems in their original forms. However, Theorem 3 has significant impact in our subsequent work on the numerical solution of descriptor systems [10] and on the controllability and observability of descriptor systems. In each case, we are able to arrive at significant results without referring to the standard canonical forms. We shall discuss this further in Section IV.

IV. Controllability and Observability

The theory developed in Section III is essential for the extension of the concept of controllability and observability from state variable systems to descriptor systems. We note that Paige [7] and Cline [9] have defined controllability for discrete descriptor systems:

$$Ex_n=Ax_{n-1}+Bu_{n-1}.$$

The conventional definition of controllability for state variable systems is as follows.

Definition C1: A system is completely controllable (*C*-controllable) if one can reach any state from any initial state.

An obvious extension of the concept of controllability to continuous descriptor systems is as follows.

Definition C2: The system (1) is controllable within the set of reachable states (*R*-controllable) if one can reach any state in the set of reachable states from any admissible initial state.

Note that, in the case of the state variable systems, *C*-controllable and *R*-controllable are equivalent. This is not so with descriptor systems, as indicated by Theorems 4 and 5 below.

Theorem 4—Regarding C-Controllability: The descriptor system in standard canonical form (3) is *C*-controllable if and only if

$$\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle=\mathbb{C}^{n_1+n_2}.$$

Proof of Theorem 4: If $\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle$ is the entire vector space then any admissible initial condition is in $\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle$. Thus, by definition, (3) is controllable.

Theorem 5—Regarding R-Controllability:

a) The descriptor system (3) is *R*-controllable if and only if

$$\langle E_1|B_1\rangle=\mathbb{C}^{n_1}.$$

b) The descriptor system (3) is *R*-controllable if and only if the subsystem described by the first equation of (3) (i.e., the subsystem that is described by the state variable equation) is controllable.

Proof of Theorem 5: Let $x^T=(x_1^T, x_2^T)$ be an admissible initial condition. Let $x_0^T=(0, x_2^T)$. Then x_0 is in I_0, where I_0 is defined in Theorem 3. By Theorem 3, the set of reachable states from x_0 is

$$R(x_0)=\langle E_1|B_1\rangle\oplus\langle E_2|B_2\rangle.$$

Any state in $R(x_0)$ is reachable from x if and only if $e^{tE_1}x_1$ is in $\langle E_1|B_1\rangle$. Thus, the system in (3) is *R*-controllable if and only if $\langle E_1|B_1\rangle=\mathbb{C}^{n_1}$, and we have proved statement a). From classical control theories, $\langle E_1|B_1\rangle=\mathbb{C}^{n_1}$ is equivalent to the first equation of (3) being controllable. Therefore, statement b) is true.

Corollary 1:

a) The system in (3) is *C*-controllable if and only if the augmented matrices

$$S_1=\left[B_1|E_1B_1|\cdots|E_1^{n_1-1}B_1\right]$$

and

$$S_2=\left[B_2|E_2B_2|\cdots|E_2^{n_1-1}B_2\right]$$

have ranks n_1 and n_2, respectively.

b) The system in (3) is *R*-controllable if and only if the augmented matrix S_1 defined above has rank n_1.

Proof of Corollary 1: The subspace $\langle E_1|B_1\rangle$ is spanned by the columns of S_1 and the subspace $\langle E_2|B_2\rangle$ is spanned by the columns of S_2. Thus, $\langle E_1|B_1\rangle=\mathbb{C}^{n_1}$ if and only if S_1 is of rank n_1, and $\langle E_2|B_2\rangle=\mathbb{C}^{n_2}$ if and only if S_2 is of rank n_2. Thus, the corollary is proved.

Note that, if B_2 of (3) is a *vector* (a matrix with one column) with no zero entries, then the subsystem described by the second equation of (3) [i.e., the algebraic part of the system (3)] is *C*-controllable if and only if the nilpotency of the system is n_2-1. For many practical models, the matrix E in (1) is rank deficient by more than 1, and so the nilpotency of the system is less than n_2-1. Thus, many practical models with simple input will not be *C*-controllable. This property indicates the "restrictiveness" of the property of *C*-controllability. In the case of the state variable system, controllability is a "dual" of observability. However, *C*-controllability will not be a "dual" of the obvious extension of observability to descriptor systems.

Observability deals with the following system:

$$E\dot{x}(t)=Ax(t)+Bu(t)$$
$$y(t)=Cx(t) \tag{10}$$

where E, A, $x(t)$, $u(t)$, and B are as defined for (1), C is an $n\times r$ complex matrix, and $y(t)$ is a complex vector in $\mathbb{C}^r$.

Definition: We say the system (10) is observable if and only if, for $t\geq 0$, $x(t)$ can be computed from E, A, B, C, $y(\tilde{t})$ and $u(\tilde{t})$ for any $\tilde{t}\in[0, b]$.

Consider observability for a descriptor system in standard canonical form:

$$\dot{x}_1(t)=E_1x_1(t)+B_1u(t)$$
$$E_2\dot{x}_2(t)=x_2(t)+B_2u(t)$$
$$y(t)=[C_1, C_2]\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}. \tag{11}$$

Note that $x_2(t)$ can always be computed since $x_2(t)=-\sum_{i=0}^{n_1-1}E_2^iB_2u^{(i)}(t)$ for a given B_2 and $u(t)$.

The system (11) is observable if and only if we can compute $x_1(t)$ from E_1, B_1, C_1, $y(t)-C_2x_2(t)$, and $u(t)$. Thus, we conclude the following.

Theorem 6: The system (11) is observable if and only if the system described by the state variable equation is observable, i.e., the system (11) is observable if and only if the following system is observable:

$$\dot{x}_1(t)=E_1x(t)+B_1u(t)$$
$$y(t)=C_1x(t).$$

Corollary 2: The system (11) is observable if and only if the the augmented matrix

$$T=\left[C_1^T|E_1^TC_1^T|\cdots|(E_1^T)^{n_1-1}C_1^T\right]$$

is of rank n_1.

Corollary 2 is the classical result of observability for state variable systems.

Theorem 6 indicates that observability is a "dual" of R-controllability.

Results in Rosenbrock [11] can be applied to Theorems 4, 5, and 6 to express controllability and observability of the descriptor system (10). We apply the following result from Rosenbrock to obtain Theorem 7.

Let A and B be $n\times n$ and $n\times r$ matrices, respectively. The following are equivalent.

a) $(sI-A|B)$ is of full row rank for all finite s. (I is the identity matrix.)

b) The augmented matrix $[B|AB|A^2B|\cdots|A^nB]$ is of full row rank.

Consider $E=I\oplus E_2$, $A=E_1\oplus I$ (where $\oplus$ denotes direct sums), and $B^T=(B_1^T, B_2^T)$. Then $(sE-A|B)$ is of full row rank for all finite s if and only if the augmented matrix $(sI-E_1|B_1)$ is of rank n_1, since $(sE_2-I|B)$ is of rank n_2 for all finite s. To see that $(sE_2-I|B)$ is of rank n_2 for all finite s, note that E_2 is nilpotent and, therefore, it is similar to an upper triangular matrix with zeros on the diagonal. Thus, (sE_2-I) is similar to an upper triangular matrix with -1's on the diagonal and, therefore, it is of full rank. Also, $(E|B)$ is of full rank if and only if the augmented matrix $(E_2|B_2)$ is of rank n_2, since $(I|B_1)$ is always of rank n_1. Thus, applying Rosenbrock's result above, we conclude the following.

1) $(sE-A|B)$ is of full row rank if and only if the augmented matrix S_1 defined in Corollary 1 has full row rank.

2) $(E|B)$ is of full row rank if and only if the augmented matrix S_2 in Corollary 1 has full row rank.

Applying the above two statements and Corollary 1, we can conclude that system (3) is R-controllable if and only if $(sE-A|B)$ is of full row rank and that system (3) is C-controllable if and only if $(sE-A|B)$ and $(E|B)$ are of full row rank. Also, it is observable if and only if $((sE-A)^T|C^T)$ is of full row rank. Since controllability and observability are invariant under a transformation of basis or variables, we have proved the following.

Theorem 7:

a) The descriptor system (1) is R-controllable if and only if the augmented matrix $(sE-A|B)$ is of full rank.

b) The descriptor system (1) is C-controllable if and only if the augmented matrices $(sE-A|B)$ and $(E|B)$ are of full rank.

c) The descriptor system (1) is observable if and only if the augmented matrix $((sE-A)^T|C^T)$ is of full rank.

We speculate that there could be other suitable extensions of controllability to continuous descriptor systems. The definition should be made in the light of solutions of various optimal control problems involving continuous descriptor systems. Verghese and Kailath [12] and Verghese, Kailath, and Van Dooren [13] defined strong controllability, which appears to be intermediate between C- and R-controllability. By their definition, the descriptor system (10) is strongly controllable if $\langle E_1, B_1\rangle = \mathbb{C}^{n_1}$ and $\langle E_2|B_2\rangle \supset$ range of E_2.

V. Conclusion

We have characterized solvability of descriptor systems in standard canonical form (3). Transforming a descriptor system to standard canonical form is computationally unacceptable and often not even feasible. The results stated in Section II have led to the development of numerical methods for descriptor systems that are stable and preserve sparsity. (See Sincovec, Yip, and Epton [6].)

We have characterized reachability for descriptor systems in standard canonical forms. The theory developed in Section III is essential for the extension of the concept of controllability and observability from the state variable systems to descriptor systems, as described in Section IV. We have defined observability and two types of controllabilities and proved corresponding necessary and sufficient conditions. We note the restrictiveness of some of our definitions, and speculate that there could be more suitable definitions for controllability and observability with respect to the solutions of various optimal control problems.

Appendix

Proof of Theorem 1: We show a) is equivalent to b).

We first show that the negation of a) implies the negation of b): (A, E) not solvable is equivalent to the the existence of a nonzero vector x such that $(E+\lambda A)x=0$ for all $\lambda\in\mathbb{C}$. We can express x as a minimum degree polynomial in λ:

$$x=x_0+\lambda x_1+\lambda^2x_2+\cdots+\lambda^kx_k, \qquad x_0\neq 0,\ x_k\neq 0. \tag{A.1}$$

Substituting the right-hand side of (A.1) into $(E+\lambda A)x=0$, we obtain the following system of equations:

$$\begin{aligned} Ex_0&=0\\ Ex_{i-1}&=-Ax_i, \qquad 1\leq i\leq k-1\\ Ax_k&=0. \end{aligned} \tag{A.2}$$

Note that $x_k\in X_0, x_{k-1}\in X_1, x_{k-2}\in X_2, \cdots, x_0\in X_k$, and also $x_0\in \operatorname{Ker} E$ and $x_0\neq 0$. Thus, b) implies a). Now we show that the negation of b) implies the negation of a): Suppose x_k is the first such subspace with $\operatorname{Ker} E\cap X_n$ nonzero. Take x_0 nonzero in this intersection and, with the definition of x_k, $Ax_0=-Ex_1$ where $x_1\in X_{k-1}$ is uniquely defined. Similarly, $x_2\in X_{k-2}$, $x_k\in X_0$ are uniquely defined, so (A.2) holds, which negates a). Therefore, if x is as defined in equation (A.1), then x is nonzero, and $(E+\lambda A)x=0$ for all $\lambda\in\mathbb{C}$. Thus, a) implies b).

$\operatorname{Det}(E^T-\lambda A^T)\equiv 0$ is the same as $\det(E-\lambda A)\equiv 0$, so, by analogy with b), a) is equivalent to c).

The equivalence of e) and a) can be shown in a similar manner.

To prove the equivalence a) and f), we say a matrix D is a direct sum of two matrices D_1 and D_2 if D is of the form

$$D=\begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix}$$

and we write $D=D_1\oplus D_2$ and call D_1, D_2 the direct summands of D. Gantmacher [1] has shown that a) is equivalent to the existence of nonsingular matrices P and Q such that

$$PEQ+PAQ=(I+\lambda E_1)\oplus(E_2+\lambda I)$$

where E_2 is nilpotent and the first and second direct summands are contributed by the finite and infinite elementary divisors, respectively. Thus, a) is equivalent to f).

Thus, we complete the proof of Theorem 1.

Acknowledgment

The authors would like to thank Dr. A. Erisman for introducing them to the problem. They would also like to thank Dr. B. Dembart, Dr. M. A. Epton, and J. Manke, all coinvestigators of this project, for the many fruitful discussions, especially Dr. Dembart for pointing out the inadequacy of their original definition of controllability, namely, C-controllability, in practical applications, and for suggesting R-controllability as a possible alternative. The authors also wish to thank the referees for pointing out certain technical difficulties in their first draft, and for directing them to the work of Rosenbrock, Verghese, Kailath, and Van Dooren, which instigated the proof of Theorem 7.

References

[1] F. R. Gantmacher, *The Theory of Matrices*, vol. 2. New York: Chelsea, 1974.
[2] J. H. Wilkinson, "Linear differential equations and Kronecker's canonical form," in *Recent Advances in Numerical Analysis*, C. de Boor and G. H. Golub, Eds. New York: Academic, 1978.
[3] G. Dahlquist and A. Bjorck, *Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1974.
[4] D. G. Luenberger, "Time-invariant descriptor systems," *Automatica*, vol. 14, pp. 473–480, 1978.

[5] W. M. Wonham, *Linear Multivariable Control, A Geometric Approach*. Berlin: Springer-Verlag, 1974.
[6] R. F. Sincovec, E. L. Yip, and M. A. Epton, "Numerical algorithms for solving descriptor systems," Dep. Energy, Contract ET-78-C-01-2876, Task 3 Rep., 1979.
[7] C. C. Paige, "Controllability of general discrete linear time-invariant systems," to be published.
[8] M. Athans and P. Falb, *Optimal Control*. New York: McGraw-Hill, 1966.
[9] T. B. Cline, D. G. Luenberger, and D. N. Stengel, "Descriptor variable representation of large-scale deterministic systems," Dep. Energy, Contract EX-76-C-0102090, Tech. Memo. 5186-6, 1977.
[10] R. F. Sincovec, A. M. Erisman, E. L. Yip, and M. A. Epton, "Analysis of descriptor systems using numerical algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 139–147, Feb. 1981.
[11] H. H. Rosenbrock, *State Space and Multivariable Theory*. New York: Wiley, 1970.
[12] G. Verghese and T. Kailath, "Impulsive behavior in dynamical systems, structure and significance," in *Proc. 4th Int. Symp. Math. Theory, Networks Syst.*, Delft, The Netherlands, July 1979.
[13] G. Verghese, T. Kailath, and P. Van Dooren, "Properties of the system matrices of a generalized state-space system," *Int. J. Contr.*, vol. 30, pp. 235–243, 1979.

On the Existence of a Negative Semidefinite, Antistabilizing Solution to the Discrete-Time Algebraic Riccati Equation

EDMOND JONCKHEERE, MEMBER, IEEE

***Abstract*—In the problem of infimizing a not necessarily positive semidefinite quadratic cost subject to a linear dynamical constraint, it is usually expected that the existence of a lower bound to the cost is equivalent to the existence of a negative semidefinite, antistabilizing solution to the algebraic Riccati equation. By a counterexample, it is shown that this equivalence breaks down in the discrete-time case. This phenomenon, as well as the whole question of the existence of the appropriate solution to the algebraic Riccati equation, are investigated in detail.**

I. INTRODUCTION

Consider the discrete-time finite-dimensional linear system

$$x(k+1) = Ax(k) + Bu(k); \qquad k = i, \cdots, t; \quad x(i) = \xi \tag{1}$$

where $x(k) \in \mathbf{R}^n$ and $u(k) \in \mathbf{R}^r$; A and B are time-invariant matrices of compatible sizes; the pair (A, B) is reachable; the matrix A is asymptotically stable; by feedback invariance, this last condition is not restrictive [7], [14], [15]. Together with (1), define the quadratic cost

$$J[\xi, u(i,t)] = \sum_{k=i}^{t-1} w[x(k), u(k)], \tag{2}$$

$$w(x,u) = x'Qx + 2x'Su + u'Ru \quad (x \in \mathbf{R}^n, u \in \mathbf{R}^r)$$

where $u(i,t) = [u'(i) \cdots u'(t-1)]'$. The overall weighting matrix $W = \begin{pmatrix} Q & S \\ S' & R \end{pmatrix}$ is symmetric, but not necessarily positive semidefinite.

The lack of positive semidefiniteness of W directly leads to the question of whether or not the cost is bounded from below. It is well known [1]–[10] that this question involves a rather intricate string of time-domain and frequency-domain conditions. More precisely, let $l^2_{\mathbf{R}^r}(-\infty, t)$ be the Hilbert space of square summable control sequences $\{u(k) \in \mathbf{R}^r : k = \cdots, t-2, t-1\}$. For all $u(-\infty, t) \in l^2_{\mathbf{R}^r}(-\infty, t)$, define $x_{0,u(-\infty,t)}(t) = \sum_{k=-\infty}^{t-1} A^{t-1-k}Bu(k)$. This allows the precise definition of $J[0, u(-\infty, t)] \triangleq \sum_{k=-\infty}^{t-1} w[x_{0,u(-\infty,k)}(k), u(k)]$. From the linear-quadratic nature of the problem, it is clear that

$$J[0, u(-\infty, t)] = u'(-\infty, t)R(-\infty, t)u(-\infty, t), \qquad \forall u(-\infty, t) \in l^2_{\mathbf{R}^r}(-\infty, t) \tag{3}$$

where $R(-\infty, t)$ is a symmetric, semi-infinite matrix representing a bounded, self-adjoint Hilbert space operator, whose structure has been investigated in detail in [20] and [21]. The problem of the existence of a lower bound to the cost, and its connection with the related time-domain and frequency-domain conditions, can now be made precise.

Theorem 1: Consider the problem (1), (2) with (A, B) reachable and A asymptotically stable. The following statements are equivalent.

a) For all $t \geq i$, there exists a bounded symmetric matrix $N(t-i) \in \mathbf{R}^{n \times n}$ such that $J[\xi, u(i,t)] \geq \xi' N(t-i)\xi$ for all ξ and all $u(i,t)$.

b) The *Riccati equation*

$$\Pi(k-1) = A'\Pi(k)A + Q - [S + A'\Pi(k)B] \cdot [R + B'\Pi(k)B]^{+}[S' + B'\Pi(k)A], \tag{4a}$$

$$R + B'\Pi(k)B \geq 0, \tag{4b}$$

$$\operatorname{Ker}[R + B'\Pi(k)B] \subseteq \operatorname{Ker}[S + A'\Pi(k)B], \tag{4c}$$

$$\Pi(t) = 0 \tag{4d}$$

has a global solution $\{\Pi(k) : k = \cdots, t-1, t\}$.

c) $R(-\infty, t) \geq 0$.

d) $J[0, u(-\infty, t)] \geq 0$ for all $u(-\infty, t) \in l^2_{\mathbf{R}^r}(-\infty, t)$.

e) The *backwards infimization problem*

$$J^*_{-}(\eta) = \inf\{J[0, u(-\infty, t)] : u(-\infty, t) \in l^2_{\mathbf{R}^r}(-\infty, t) \text{ and } x_{0,u(-\infty,t)}(t) = \eta\} \tag{5}$$

has a solution of the form $J^*_{-}(\eta) = -\eta'\Pi_{-}\eta$ with $\Pi_{-} = \Pi'_{-} \leq 0$.

f) The *linear matrix inequality*

$$\Lambda(\Pi) = \begin{bmatrix} A'\Pi A - \Pi + Q & S + A'\Pi B \\ S' + B'\Pi A & R + B'\Pi B \end{bmatrix} \geq 0 \tag{6}$$

has a solution $\Pi = \Pi' \leq 0$.

Moreover, should any of these statements hold, then $\Lambda(\Pi_{-}) \geq 0$, and any solution $\Pi = \Pi'$ of $\Lambda(\Pi) \geq 0$ is such that $\Pi_{-} \leq \Pi$.

Proof: This result is proved for the single-input case in [20, Theorem 2]. The generalization to the multi-input case is straightforward.

The algebraic Riccati equation is now introduced.

Theorem 2: Consider the problem (1), (2) with (A, B) reachable and A asymptotically stable. If the *algebraic Riccati equation*

$$K(\Pi) \equiv \Pi - A'\Pi A - Q + (S + A'\Pi B) \cdot (R + B'\Pi B)^{+}(S' + B'\Pi A) = 0 \tag{7a}$$

$$R + B'\Pi B \geq 0 \tag{7b}$$

$$\operatorname{Ker}(R + B'\Pi B) \subseteq \operatorname{Ker}(S + A'\Pi B) \tag{7c}$$

has a solution $\Pi = \Pi' \leq 0$ satisfying the additional conditions $|\lambda_k[A - B(R + B'\Pi B)^{+}(S' + B'\Pi A)]| \geq 1$, $k = 1, \cdots, n$, then any of the statements of Theorem 1 is verified.

Proof: Let Π be such a solution. It is easily seen that $J[\xi, u(i,t)] = \xi'\Pi\xi + \sum_{k=i}^{t-1}[x'(k)u'(k)]\,\Lambda(\Pi)[x'(k)u'(k)]' - x'(t)\Pi x(t)$. Since Π is a solution of the algebraic Riccati equation, $\Lambda(\Pi) \geq 0$. This and $\Pi \leq 0$ yield statement a), and hence the other statements of Theorem 1.

Manuscript received September 19, 1979; revised January 19, 1981. Paper recommended by A. Z. Manitius, Past Chairman of the Optimal Systems Committee. This work was supported in part by AFOSR Grant 80-0013.

The author was with Phillips Research Laboratory, Brussels, Belgium. He is now with the Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, CA 90007.

0018-9286/81/0600-0707$00.75