J.S.I.A.M. Control
Ser. A, Vol. 1, No. 2
*Printed in U.S.A., 1963*

# MATHEMATICAL DESCRIPTION OF LINEAR DYNAMICAL SYSTEMS*

R. E. KALMAN†

**Abstract.** There are two different ways of describing dynamical systems: (i) by means of state variables and (ii) by input/output relations. The first method may be regarded as an axiomatization of Newton's laws of mechanics and is taken to be the basic definition of a system.

It is then shown (in the linear case) that the input/output relations determine only one part of a system, that which is completely observable and completely controllable. Using the theory of controllability and observability, methods are given for calculating irreducible realizations of a given impulse-response matrix. In particular, an explicit procedure is given to determine the minimal number of state variables necessary to realize a given transfer-function matrix. Difficulties arising from the use of reducible realizations are discussed briefly.

**1. Introduction and summary.** Recent developments in optimal control system theory are based on vector differential equations as models of physical systems. In the older literature on control theory, however, the same systems are modeled by transfer functions (i.e., by the Laplace transforms of the differential equations relating the inputs to the outputs). Two different languages have arisen, both of which purport to talk about the same problem. In the new approach, we talk about state variables, transition equations, etc., and make constant use of abstract linear algebra. In the old approach, the key words are frequency response, pole-zero patterns, etc., and the main mathematical tool is complex function theory.

Is there really a difference between the new and the old? Precisely what are the relations between (linear) vector differential equations and transfer-functions? In the literature, this question is surrounded by confusion [1]. This is bad. Communication between research workers and engineers is impeded. Important results of the "old theory" are not yet fully integrated into the new theory.

In the writer's view—which will be argued at length in this paper—the difficulty is due to insufficient appreciation of the concept of a *dynamical system.* Control theory is supposed to deal with physical systems, and not merely with mathematical objects such as a differential equation or a transfer function. We must therefore pay careful attention to the relationship between physical systems and their representation via differential equations, transfer functions, etc.

* Received by the editors July 7, 1962 and in revised form December 9, 1962.

Presented at the Symposium on Multivariable System Theory, SIAM, November 1, 1962 at Cambridge, Massachusetts.

This research was supported in part under U. S. Air Force Contracts AF 49(638)-382 and AF 33(616)-6952 as well as NASA Contract NASr-103.

† Research Institute for Advanced Studies (RIAS), Baltimore 12, Maryland.

To clear up these issues, we need first of all a precise, abstract definition of a (physical) dynamical system. (See sections 2–3.) The axioms which provide this definition are generalizations of the Newtonian world-view of causality. They have been used for many years in the mathematical literature of dynamical systems. Just as Newtonian mechanics evolved from differential equations, these axioms seek to abstract those properties of differential equations which agree with the "facts" of classical physics. It is hardly surprising that under special assumptions (finite-dimensional state space, continuous time) the axioms turn out to be equivalent to a system of ordinary differential equations. To avoid mathematical difficulties, we shall restrict our attention to linear differential equations.

In section 4 we formulate the central problem of the paper:

*Given an (experimentally observed) impulse response matrix, how can we identify the linear dynamical system which generated it?*

We propose to call any such system a *realization* of the given impulse response. It is an *irreducible realization* if the dimension of its state space is minimal.

Section 5 is a discussion of the "canonical structure theorem" [2, 14] which describes abstractly the coupling between the external variables (input and output) and the internal variables (state) of any linear dynamical system. As an immediate consequence of this theorem, we find that *a linear dynamical system is an irreducible realization of an impulse-response matrix if and only if the system is completely controllable and completely observable.* This important result provides a link between the present paper and earlier investigations in the theory of controllability and observability [3–5].

Explicit criteria for complete controllability and complete observability are reviewed in a convenient form in section 6.

Section 7 provides a constructive computational technique for determining the canonical structure of a constant linear dynamical system.

In section 8 we present, probably for the first time, a complete and rigorous theory of how to define the state variables of a multi-input/multi-output constant linear dynamical system described by its transfer-function matrix. Since we are interested only in irreducible realizations, there is a certain unique, well-defined number $n$ of state variables which must be used. We give a simple proof of a recent theorem of Gilbert [5] concerning the value of $n$. We give canonical forms for irreducible realizations in simple cases. We give a constructive procedure (with examples) for finding an irreducible realization in the general case.

Many errors have been committed in the literature of system theory by carelessly regarding transfer functions and systems as equivalent concepts. A list of these has been collected in section 9.

The field of research outlined in this paper is still wide open, except perhaps in the case of constant linear systems. Very little is known about irreducible realizations of nonconstant linear systems. It is not clear what additional properties—besides complete controllability and complete observability—are required to identify the stability type of a system from its impulse response. Nothing is known about nonlinear problems in this context.

Finally, the writer would like to acknowledge his indebtedness to Professor E. G. Gilbert, University of Michigan, whose work [5] predates this and whose results were instrumental in establishing the canonical structure theorem.

**2. Axiomatic definition of a dynamical system.** Macroscopic physical phenomena are commonly described in terms of cause-and-effect relationships. This is the "Principle of Causality". The idea involved here is at least as old as Newtonian mechanics. According to the latter, the motion of a system of particles is fully determined for all future time by the present positions and momenta of the particles and by the present and future forces acting on the system. How the particles actually attained their present positions and momenta is immaterial. Future forces can have no effect on what happens at present.

In modern terminology, we say that the numbers which specify the instantaneous position and momentum of each particle represent the *state* of the system. The state is to be regarded always as an abstract quantity. Intuitively speaking, the state is the minimal amount of information about the past history of the system which suffices to predict the effect of the past upon the future. Further, we say that the forces acting on the particles are the *inputs* of the system. Any variable in the system which can be directly observed is an *output*.

The preceding notions can be used to give a precise mathematical definition of a dynamical system [6]. For the present purposes it will be convenient to state this definition in somewhat more general fashion [14].

Definition 1. A dynamical system is a mathematical structure defined by the following axioms:

($D_1$) There is given a *state space* $\Sigma$ and a set of values of *time* $\Theta$ at which the behavior of the system is defined; $\Sigma$ is a topological space and $\Theta$ is an ordered topological space which is a subset of the real numbers.

($D_2$) There is given a topological space $\Omega$ of functions of time defined on $\Theta$, which are the admissible *inputs* to the system.

($D_3$) For any initial time $t_0$ in $\Theta$, any initial state $x_0$ in $\Sigma$, and any input $u$ in $\Omega$ defined for $t \geqq t_0$, the future states of the system are determined by the transition function $\varphi: \Omega \times \Theta \times \Theta \times \Sigma \rightarrow \Sigma$, which is written as $\varphi_u(t; t_0, x_0) = x_t$. This function is defined only for $t \geqq t_0$. Moreover, any $t_0 \leqq t_1 \leqq t_2$ in $\Theta$, any $x_0$ in $\Sigma$, and any fixed $u$ in $\Omega$ defined over $[t_0, t_1] \cap \Theta$, the following relations hold:

$$(\mathrm{D}_3\text{-i}) \qquad \varphi_u(t_0\,; t_0\,, x_0) = x_0\,,$$

$$(\mathrm{D}_3\text{-ii}) \qquad \varphi_u(t_2\,; t_0\,, x_0) = \varphi_u(t_2\,; t_1\,, \varphi_u(t_1\,; t_0\,, x_0)).$$

In addition, the system must be *nonanticipatory*, i.e., if $u, v \in \Omega$ and $u \equiv v$ on $[t_0, t_1] \cap \Theta$ we have

$$(\mathrm{D}_3\text{-iii}) \qquad \varphi_u(t; t_0\,, x_0) \equiv \varphi_v(t; t_0\,, x_0).$$

($\mathrm{D}_4$) Every *output* of the system is a function $\psi : \Theta \times \Sigma \rightarrow$ reals.

($\mathrm{D}_5$) The functions $\varphi$ and $\psi$ are continuous, with respect to the topologies defined for $\Sigma$, $\Theta$, and $\Omega$ and the induced product topologies.

In this paper we will study only a very special subclass of dynamical systems: those which are *real*, *finite-dimensional*, *continuous-time*, and *linear*.

"Real, finite-dimensional" means that $\Sigma = R^n$ = $n$-dimensional real linear space. "Continuous-time" means that $\Theta = R^1$ = set of real numbers. "Linear" means that $\varphi$ is linear on $\Omega \times \Sigma$ and $\psi$ is linear on $\Sigma$.

By requiring $\varphi$ and $\psi$ to be sufficiently "smooth" functions, we can deduce from the axioms a set of equations which characterize every real, finite-dimensional, continuous-time, and linear dynamical system. The proof of this fact is outside the scope of the present paper [14]. Here we shall simply assume that every such system is governed by the equations

$$(2.1) \qquad \frac{dx}{dt} = F(t)x + G(t)u(t),$$

$$(2.2) \qquad y(t) = H(t)x(t),$$

defined on the whole real line $-\infty < t < \infty$, where $x$, $u$, and $y$ are $n$, $m$, and $p$-vectors* respectively, and the matrices $F(t)$, $G(t)$, and $H(t)$ are continuous functions of the time $t$.

We call (2.1–2) the *dynamical equations* of the system.

It is instructive to check whether the axioms are satisfied. ($\mathrm{D}_1$) is obviously true; we have $\Sigma = R^n$, $\Theta = R^1$. The *state* of the system is the vector $x$. To satisfy ($\mathrm{D}_2$), we must specify the class of all inputs, that is, a subclass of all vector functions $u(t) = (u_1(t), \cdots, u_m(t))$. To define $\Omega$, we shall assume that these functions are piecewise continuous; this is sufficiently

* Vectors will be denoted by small Roman letters, matrices by Roman capitals. The components of a vector $x$ are $x_i$, components of a matrix $A$ are $a_{ij}$. On the other hand, $x^1, x^2, \cdots$, are vectors, and $F^{AA}$, $F^{AB}$ are matrices. $A'$ is the transpose of $A$.

general for most applications. We have exactly $p$ observations on the system (the components of the vector $y$) and by (2.2) they are functions of $t$, $x$. Hence ($D_4$) is satisfied. To check ($D_3$), we recall that the general solution of (2.1) is given by

$$(2.3) \qquad \varphi_u(t; t_0, x_0) \equiv x_t = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \tau)G(\tau)u(\tau)\, d\tau,$$

where $\Phi(t, \tau)$ is the transition matrix of the free differential equation defined by $F(t)$ [4, 7]†. Since (2.3) is valid for any $t \geqq t_0$ (in fact, also for $t < t_0$), $\varphi$ is well defined. Property ($D_3$-i) is obvious. ($D_3$-ii) follows from the composition property [4, 7] of the transition matrix:

$$(2.4) \qquad \Phi(t, \sigma) = \Phi(t, \tau)\Phi(\tau, \sigma),$$

which holds for every set of real numbers $t$, $\tau$, $\sigma$. Indeed, (2.4) is simply the linear version of ($D_3$-ii). ($D_3$-iii) is obvious from formula (2.3). The continuity axiom ($D_5$) is satisfied by hypothesis.

Evidently $\varphi$ given by (2.3) is linear on the cartesian product of $\Sigma$ with the linear space of vector-valued piecewise continuous functions.

We call a linear dynamical system (2.1–2) *constant*, *periodic*, or *analytic* whenever $F$, $G$, and $H$ are constant, periodic, or analytic in $t$.

It is often convenient to have a special name for the couple $(t, x) \mid \in \Theta \times \Sigma$. Giving a fixed value of $(t, x)$ is equivalent to specifying at some time $(t)$ the state $(x)$ of the system. We shall call $(t, x)$ a *phase* and $\Theta \times \Sigma$ the *phase space*. (Recall the popular phrase: "phases" of the Moon.)

To justify our claim—implicit in the above discussion—that equations (2.1–2) are a good model of physical reality, we wish to point out that these equations can be concretely simulated by a simple physical system: a general-purpose analog computer. Indeed, the numbers (or functions) constituting $F$, $G$, and $H$ may be regarded as specifying the "wiring diagram" of the analog computer which simulates the system (2.1–2) (see, for instance, [8]).

**3. Equivalent dynamical systems.** The state vector $x$ must always be regarded as an abstract quantity. By definition, it cannot be directly measured. On the other hand, the inputs and outputs of the system (2.1–2) have concrete physical meaning. Bearing this in mind, equations (2.1–2) admit two interpretations:

(a) They express relations involving the abstract linear transformations $F(t)$, $G(t)$, and $H(t)$.

(*b*) At any fixed time, we take an arbitrary but fixed coördinate system

† I.e., $\Phi$ is a solution of $d\Phi/dt = F(t)\Phi$, subject to the initial condition $\Phi(\tau, \tau) = I =$ unit matrix for all $\tau$.

in the (abstract) vector space $\Sigma$. Then the symbol $x \equiv (x_1, \cdots, x_n)$ is interpreted as the numerical $n$-tuple consisting of the coördinates of the abstract state vector which is also denoted by $x$. $F$, $G$, and $H$ are interpreted as the matrix representations of the abstract linear transformations denoted by the same letters under (a).

To describe the behavior of a dynamical system in concrete terms, the second point of view must be used. Then we must also ask ourselves the question: To what extent does the description of a dynamical system depend on the arbitrary choice of the coordinate system in the state space? (No such arbitrariness occurs in the definition of the numerical vectors $u$, $y$ since the input and output variables $u_i$ and $y_j$ are concrete physical quantities.) This question gives rise to the next definition.

DEFINITION 2. Two linear dynamical systems (2.1–2), with state vectors $x$, $\bar{x}$, are *algebraically equivalent* whenever their numerical phase vectors are related for all $t$ as

$$(t, \bar{x}) = (t, T(t)x), \tag{3.1}$$

where $T(t)$ is a $n \times n$ matrix, nonsingular for all $t$ and continuously differentiable in $t$. In other words, there is a 1-1 differentiable correspondence between the phase spaces $\Theta \times \Sigma$ and $\Theta \times \bar{\Sigma}$.

*Remark:* We could generalize this definition of equivalence to $(\bar{t}, \bar{x}) = (\tau(t), T(t)x)$ where $\tau$ is an increasing function of $t$. But this involves distortion of the time scale which is not permitted in Newtonian physics.

Algebraic equivalence implies the following relations between the defining matrices of the two systems:

$$\begin{aligned} \bar{\Phi}(t, \tau) &= T(t)\Phi(t, \tau)T^{-1}(\tau), \\ \bar{F}(t) &= \dot{T}(t)T^{-1}(t) + T(t)F(t)T^{-1}(t), \\ \bar{G}(t) &= T(t)G(t), \\ \bar{H}(t) &= H(t)T^{-1}(t). \end{aligned} \tag{3.2}$$

In general, algebraic equivalence does not preserve the stability properties of a dynamical system [7, 9, 10]. For this it is necessary and sufficient to have *topological equivalence*: algebraic equivalence plus the condition

$$\| T(t) \| \leqq c_1 \quad \text{and} \quad \| T^{-1}(t) \| \leqq c_2, \tag{3.3}$$

where $c_1$ and $c_2$ are fixed constants, and $\| \quad \|$ is the euclidean norm*.

A nonconstant system may be algebraically and even topologically equivalent to a constant system. The latter case is called by Markus [11]

* Let $\Theta$, $\Sigma$, and $\bar{\Sigma}$ have the usual topologies induced by the euclidean norm. Then the product topologies induced on $\Theta \times \Sigma$ and $\Theta \times \bar{\Sigma}$ are equivalent if and only if (3.3) holds.

"kinematic similarity". Moreover, two constant systems may be algebraically and topologically equivalent without $T(t)$ being a constant. To bypass these complications, we propose

DEFINITION 3. Two constant linear dynamical systems are *strictly equivalent* whenever their numerical phase vectors are related for all $t$ as $(t, \bar{x}) = (t, Tx)$, where $T$ is a nonsingular constant matrix.

Evidently strict equivalence implies topological equivalence.

## 4. The impulse-response matrix and its realization by a linear dynamical system.

Sections 2–3 were concerned with mathematics, that is, abstract matters. If we now take the point of view of physics, then a dynamical system must be "defined" in terms of quantities which can be directly observed. For linear dynamical systems, this is usually done in the following way.

We consider a system which is at rest at time $t_0$; i.e., one whose input and outputs have been identically zero for all $t \leqq t_0$. We apply at each input in turn a very sharp and narrow pulse. Ideally, we would take $u_i^{(j)}(t) = \delta_{ij}\delta(t - t_0)$, where $\delta$ is the Dirac delta function, $\delta_{ij}$ is the Kronecker symbol, and $1 \leqq i, j \leqq m$. We then observe the effect of each vector input $u^{(j)}(t)$ on the outputs, which are denoted by $u(t; j)$. The matrix $S(t, t_0) = [s_{ij}(t, t_0)] = [y_i(t; j)]$ so obtained is called the *impulse-response matrix* of the system. Since the system was at rest prior to $t = t_0$, we must define $S(t, t_0) \equiv 0$ for $t < t_0$. We also assume, of course, that $S$ is continuous in $t$ and $t_0$ for $t > t_0$.

With these conventions, the output of a linear system originally at rest is related to its input by the well-known convolution integral:

$$y(t) = \int_{t_0}^{t} S(t, \tau)u(\tau)\, d\tau. \tag{4.1}$$

In much of the literature of system theory [12] (and also at times in physics) formula (4.1) is the basic definition of a system. The Fourier transform of $S$ is often called "the system function" [13, p. 92].

Unfortunately, this definition does not explain how to treat systems which are not "initially at rest". Hence we may ask, "To what extent, if any, are we justified in equating the physical definition (4.1) of a system with the mathematical one provided by (2.1–2)?"

Suppose that the system in question is actually (2.1–2). Then (2.3) shows that

$$\begin{aligned} S(t, \tau) &= H(t)\Phi(t, \tau)G(\tau), && t \geqq \tau, \\ &= 0, && t < \tau.\dagger \end{aligned} \tag{4.2}$$

† The right-hand side of the first equation (4.2) is defined also for $t < \tau$; then the left-hand side may be regarded as the "backward impulse response", whose physical interpretation is left to the reader.

Thus it is trivial to calculate the impulse-response matrix of a given linear dynamical system. The converse question, however, is non trivial and interesting. *When and how does the impulse-response matrix determine the dynamical equations of the system?*

This problem is commonly called the *identification* of the system from its impulse-response matrix.

Having been given an impulse-response matrix, suppose that we succeed in finding matrices $F$, $G$, and $H$ such that (4.2) holds. We have then identified a physical system that may have been the one which actually generated the observed impulse-response matrix. We shall therefore call (2.12) a *realization* of $S(t, \tau)$. This terminology is justified because the axioms given in section 2 are patterned after highly successful models of classical macroscopic physics; in fact, the system defined by (2.1–2) can be concretely realized, actually built, using standard analog-computer techniques in existence today. In short, proceeding from the impulse-response matrix to the dynamical equations we get closer to "physical reality". But we are also left with a problem: Which one of the (possibly very many) realizations of $S(t, \tau)$ is the actual system that we are dealing with?

It is conceivable that certain aspects of a dynamical system cannot ever be identified from knowledge of its impulse response, as our knowledge of the physical world gained from experimental observation must always be regarded as incomplete. Still, it seems sensible to ask how much of the physical world can be determined from a given amount of experimental data.

The first clear problem statement in this complex of ideas and the first results appear to be due to the writer [2, 14].

First of all we note

THEOREM 1. *An impulse-response matrix $S(t, \tau)$ is realizable by a finite-dimensional dynamical system* (2.1–2) *if and only if there exist continuous matrices $P(t)$ and $Q(t)$ such that*

$$(4.3) \qquad S(t, \tau) = P(t)Q(\tau) \quad \textit{for all} \quad t, \tau.$$

*Proof.* Necessity follows by writing the right-hand side of (4.2) as $H(t)\Phi(t, 0)\Phi(0, \tau)G(\tau)$, with the aid of (2.4). Sufficiency is equally obvious. We set $F(t) = 0$, $G(t) = Q(t)$, and $H(t) = P(t)$. Then $\Phi(t, \tau) \equiv I$ and the desired result follows by (4.2).

A realization (2.1–2) of $S(t, \tau)$ is *reducible* if over some interval of time there is a proper (i.e., lower-dimensional) subsystem of (2.1–2) which also realizes $S(t, \tau)$. As will be seen later, a realization of $S$ (particularly the one given in the previous paragraph) is often reducible.

An impulse-response matrix $S$ is *stationary* whenever $S(t, \tau) = S(t + \sigma, \tau + \sigma)$ for all real numbers $t$, $\tau$, and $\sigma$. $S$ is *periodic* whenever the preceding relation holds for all $t$, $\tau$, and some $\sigma$. An impulse-response matrix is *analytic* whenever $S$ is analytic in $t$ and $\tau$; if (4.3) holds, then $P$ and $Q$ must be analytic in $t$.

The main result, whose proof will be discussed later, is the following [14]:

THEOREM 2. *Hypothesis: The impulse-response matrix $S$ satisfies* (4.3) *and is either periodic* (*and continuous*) *or analytic.*

*Conclusions:* (i) *There exist irreducible realizations of $S$, all of which have the same constant dimension $n$ and are algebraically equivalent.* (ii) *If $S$ is periodic* [*analytic*] *so are its irreducible realizations.*

Topological equivalence cannot be claimed in general. It may happen that $S$ has one realization which is asymptotically stable and another which is asymptotically unstable [15]. Hence it may be impossible to identify the stability of a dynamical system from its impulse response! This surprising conclusion raises many interesting problems which are as yet unexplored [15]. If $S$ is not periodic or analytic, it may happen that the dimension $n(t)$ of an irreducible realization is constant only over finite time intervals.

In the stationary case, Theorem 2 can be improved [14].

THEOREM 3. *Every stationary impulse-response matrix $S(t, \tau) = W(t - \tau)$ satisfying* (4.3) *has constant irreducible realizations. All such realizations are strictly equivalent.*

In view of this theorem, we may talk indifferently about a stationary impulse-response matrix or the dynamical system which generates it—as has long been the practice in system theory on intuitive grounds. But note that we must require the realization to be irreducible. For nonconstant systems, such a conclusion is at present not justified. The requirement of irreducibility in Theorem 3 is essential; disregarding it can lead—and has led—to serious errors in modeling dynamical systems. (See section 9.)

In many practical cases, it is not the *weighting-function matrix* $W(t - \tau)$ (see Theorem 3) which is given, but its Laplace transform, the *transfer-function matrix* $Z(s) = \mathfrak{L}[W(t)]$. Then condition (4.3) has an interesting equivalent form, which is often used as a "working hypothesis" in engineering texts:

THEOREM 4. *A weighting-function matrix $W(t - \tau)$ satisfies* (4.3) *if and only if its elements are linear combinations of terms of the type $t^i e^{s_j t}$* ($i = 0, 1, \cdots, n - 1, j = 1, \cdots, n$). *Hence every element of the transfer-function matrix is a ratio of polynomials in $s$ such that the degree of the denominator polynomial always exceeds the degree of the numerator polynomial.*

This result is provedd in [14]. It implies that the realization of an impulse-response matrix is equivalent to expressing the elements of $F$, $G$, and $H$ as functions of the coefficients of the numerator and denominator polynomials of elements of $Z(s)$. (See section 8.)

In the remainder of the paper, we wish to investigate two main problems arising in the theory sketched above:

(i) Explicit criteria for reducibility.

(ii) Construction of irreducible realizations.

*Remark.* Elementary expositions of system theory often contain the statement that the operator $d/dt$ ($\equiv s$) is a "system." Is a it system in the same sense as that word is used here? The answer is no. To define such a system rigorously in accordance with the axioms introduced in section 2, one must proceed as follows. The output of the system, which by definition is the derivative of the input, is given by

$$y(t) = \frac{du(t)}{dt} = \psi(t, x(t)), \tag{3.4}$$

so that at any fixed $t$, $u(t)$ must be a *point* function of $(t, x(t))$. Therefore the state space $\Sigma$ must include the space $\Omega$ of functions on which the operator $d/dt$ is defined. It is simplest to let $\Sigma = \Omega$. Then $\Sigma$ is usually infinite dimensional because $\Omega$ is. Thus we define the state $x \equiv x(t)$ as the function $u(\tau)$, defined for all $\tau \leqq t$. The mapping $\varphi_u(t; t_0, x_{t_0})$ assigns to the function $x_0$ defined for $\tau \leqq t_0$ the function $x_t$, which is equal to $x_{t_0}$ on $\tau \leqq t_0$ and equal to $u$ on $t_0 < \tau \leqq t$.

In this paper, the finite dimensionality of $\Sigma$ is used in an essential way, which rules out consideration of the "system" $d/dt$ in all but trivial cases.

**5. Canonical structure of linear dynamical systems.** The concept of irreducibility can be understood most readily with the help of the writer's "canonical structure theorem" for linear dynamical systems [2, 14].

Before presenting and illustrating this central result, it is necessary to recall some definitions and facts concerning the *controllability* and *observability* of linear dynamical systems.

DEFINITION 4. A linear dynamical system (2.1-2) is *completely controllable* at time $t_0$ if it is not algebraically equivalent, for all $t \geqq t_0$, to a system of the type

$$\begin{aligned} &\text{(a)} \quad dx^1/dt = F^{11}(t)x^1 + F^{12}(t)x^2 + G^1(t)u(t) \\ (5.1)\quad &\text{(b)} \quad dx^2/dt = F^{22}(t)x^2 \\ &\text{(c)} \quad y(t) = H^1(t)x^1(t) + H^2(t)x^2(t). \end{aligned}$$

(In (5.1), $x^1$ and $x^2$ are vectors of $n_1$ and $n_2 = n - n_1$ components respectively.)

In other words, it is *not* possible to find a coördinate system in which the state variables $x_i$ are separated into two groups, $x^1 = (x_1, \cdots, x_{n_1})$ and $x^2 = (x_{n_1+1}, \cdots, x_n)$, such that the second group is not affected either by the first group or by the inputs to the system. If one could find such a

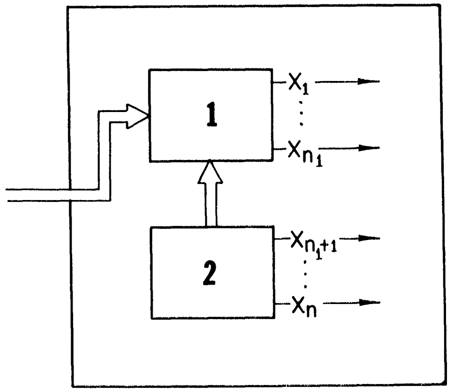

FIGURE 1.

coördinate system, we would have the state of affairs depicted schematically in Fig. 1.

Clearly, controllability is a system property which is completely independent of the way in which the outputs of the system are formed. It is a property of the couple $\{F(t), G(t)\}$.

The "dual" of controllability is observability, which depends only on the outputs but not on the inputs.

DEFINITION 5. A linear dynamical system (2.1–2) is *completely observable* at time $t_0$ if it is not algebraically equivalent, for all $t \leqq t_0$, to any system of the type

$$\begin{aligned}
&\text{(a)} && dx^1/dt = F^{11}(t)x^1(t) + G^1(t)u(t)\\
(5.2)\quad &\text{(b)} && dx^2/dt = F^{21}(t)x^1(t) + F^{22}(t)x^2 + G^2(t)u(t)\\
&\text{(c)} && y(t) = H^1(t)x^1(t).
\end{aligned}$$

(Again, $x^1$ is an $n_1$-vector and $x^2$ is an $(n - n_1)$-vector.)

In other words, it is not possible to find a coördinate system in which the state variables $x_i$ are separated into two groups, such that the second group does not affect either the first group or the outputs of the system. If such a coördinate system could be found, we would have the state of affairs depicted in Fig. 2.

The above definitions show that controllability and observability are preserved under algebraic equivalence. These properties are coördinate-free, i.e., independent of the particular choice of basis in the state space.

The equivalence of the present definitions with other more abstract

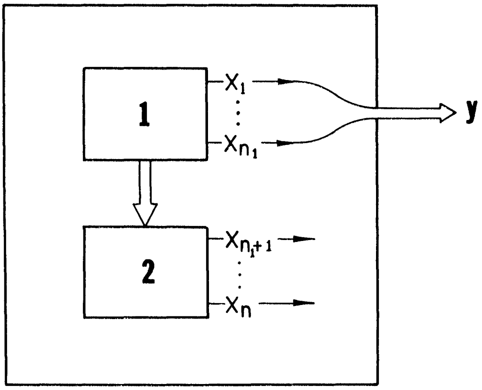

FIGURE 2.

definitions of controllability may be found in [4]. As to observability, we note that the *duality relations*

$$
(5.3) \qquad
\begin{array}{ll}
\text{(a)} & t - t_0 = t_0 - t', \\
\text{(b)} & F(t - t_0) \Leftrightarrow F'(t_0 - t'), \\
\text{(c)} & G(t - t_0) \Leftrightarrow H'(t_0 - t'), \\
\text{(d)} & H(t - t_0) \Leftrightarrow G'(t_0 - t'),
\end{array}
$$

transform the system (5.2) into (5.1). Hence all theorems on controllability can be "dualized" to yield analogous results on observability.

It can be shown that in applying definitions 4–5 to constant systems it is immaterial whether we require algebraic or strict equivalence [14]. Hence—as one would of course expect—for constant systems the notions of complete controllability and complete observability do not depend on the choice of $t_0$.

EXAMPLE 1. A simple, well-known, and interesting case of a physical system which is neither completely controllable nor completely observable is the so-called constant-resistance network shown in Fig. 3.

Let $x_1$ be the magnetic flux in the inductor and $x_2$ the electric charge on the capacitor in Fig. 3, while $u_1(t)$ is a voltage source (zero short-circuit resistance) and $y_1(t)$ is the current into the network. The inductor and capacitor in the network may be time-varying, but we assume—this is the constant-resistance condition—that $L(t)$ and $C(t)$ are related by:

$$L(t)/C(t) = R^2 = 1 \qquad (L(t), C(t) > 0).$$

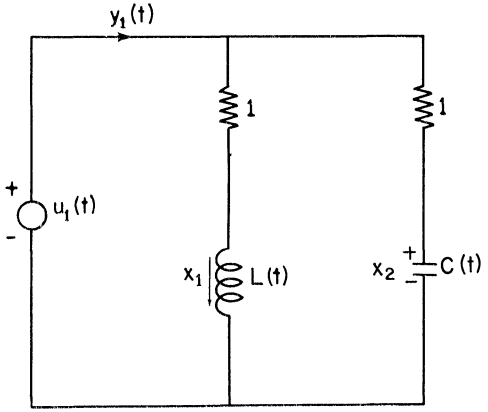

FIGURE 3.

The differential equations of the network are

$$\begin{aligned} dx_1/dt &= -[1/L(t)]x_1 + u_1(t), \\ dx_2/dt &= -[1/C(t)]x_2 + u_1(t), \\ y_1(t) &= [1/L(t)]x_1 - [1/C(t)]x_2 + u_1(t). \end{aligned}$$

If we let

$$\begin{aligned} \bar{x}_1 &= (x_1 + x_2)/2, \\ \bar{x}_2 &= (x_1 - x_2)/2, \end{aligned}$$

the dynamical equations become

$$\begin{aligned} d\bar{x}_1/dt &= -[1/L(t)]\bar{x}_1 + u_1(t), \\ d\bar{x}_2/dt &= -[1/L(t)]\bar{x}_2\,, \\ y_1(t) &= 2[1/L(t)]\bar{x}_2 + u_1(t).^* \end{aligned} \tag{5.4}$$

Here the state variable $\bar{x}_1$ is controllable but not observable, while $\bar{x}_2$ is observable but not controllable.

For obvious reasons, the subsystem (b) of (5.1) may be regarded as (completely) uncontrollable, while subsystem (b) of (5.2) is (completely) unobservable. In view of linearity, it is intuitively clear that it must be possible to arrange the components of the state vector—referred to a

* Note that this equation does not correspond to (2.2) but to $y(t) = H(t)x(t) + J(t)u(t)$. This is a minor point. In fact, Axiom ($D_4$) may be generalized to: "($D_4$): Every output is a function of $t$, $x(t)$, and $u(t)$." This entails only minor modifications as far as the results and arguments of the present paper are concerned.

suitable (possibly time-varying) coördinate system—into four mutally exclusive parts, as follows:

Part (A): Completely controllable but unobservable.

Part (B): Completely controllable and completely observable.

Part (C): Uncontrollable and unobservable.

Part (D): Uncontrollable but completely observable.

The precise statement of this idea is [2, 14]:

THEOREM 5 (*Canonical Structure Theorem*). *Consider a fixed linear dynamical system* (2.1–2).

(i) *At every fixed instant t of time, there is a coördinate system in the state space relative to which the components of the state vector can be decomposed into four mutually exlusive parts*

$$x = (x^A, x^B, x^C, x^D),$$

*which correspond to the scheme outlined above.*

(ii) *This decomposition can be achieved in many ways, but the number of state variables* $n_A(t), \cdots, n_D(t)$ *in each part is the same for any such decomposition.*

(iii) *Relative to such a choice of coördinates, the system matrices have the canonical form*

$$F(t) = \begin{bmatrix} F^{AA}(t) & F^{AB}(t) & F^{AC}(t) & F^{AD}(t) \\ 0 & F^{BB}(t) & 0 & F^{BD}(t) \\ 0 & 0 & F^{CC}(t) & F^{CD}(t) \\ 0 & 0 & 0 & F^{DD}(t) \end{bmatrix},$$

$$G(t) = \begin{bmatrix} G^A(t) \\ G^B(t) \\ 0 \\ 0 \end{bmatrix},$$

*and*

$$H(t) = [0 \quad H^B(t) \quad 0 \quad H^D(t)].$$

In view of this theorem, we shall talk, somewhat loosely, about "Parts $(A), \cdots, (D)$ of the system." Thus the system (5.4) consists of Parts $(A)$ and $(D)$.

The canonical form of $F$, $G$, and $H$ can be easily remembered by reference to the causal diagram shown on Fig. 4.

It is intuitively clear (and can be easily proved) that algebraically equivalent systems have the same canonical structure.

Unfortunately, the coördinate system necessary to display the canonical form of $F$, $G$, and $H$ will not be continuous in time unless $n_A(t), \cdots, n_D(t)$ are constants. If these dimension numbers vary, we cannot call the various

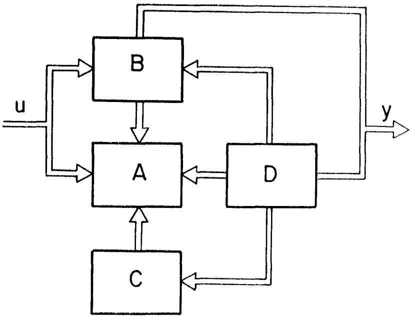

FIGURE 4.

parts of the canonical structure "subsystems." For constant systems this difficulty does not arise. More generally, we have:

THEOREM 6. *For a periodic or analytic linear dynamical system* (2.1–2) *the dimension numbers $n_A, \cdots, n_D$ are constants, and the canonical decomposition is continuous with respect to t.*

An illustration of the canonical structure theorem is provided by

EXAMPLE 2. Consider the constant system defined by

$$F = \begin{bmatrix} -3 & -3 & 0 & 1 \\ 26 & 36 & -3 & -25 \\ 30 & 39 & -2 & -27 \\ 30 & 43 & -3 & -32 \end{bmatrix},$$

$$G = \begin{bmatrix} 3 & 3 \\ -2 & -1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$H = [-5 \quad -8 \quad 1 \quad 5].$$

We introduce new coördinates by letting $\bar{x} = Tx$, where

$$T = \begin{bmatrix} 2 & 3 & 0 & -2 \\ 1 & 1 & 0 & -1 \\ -2 & -3 & 0 & 3 \\ -6 & -9 & 1 & 6 \end{bmatrix},$$

and

$$T^{-1} = \begin{bmatrix} 0 & 3 & 1 & 0 \\ 1 & -2 & 0 & 0 \\ 3 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

With respect to these new coördinates the system matrices assume the canonical form:

$$\bar{F} = TFT^{-1} = \begin{bmatrix} 2 & 4 & 1 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\bar{G} = TG = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$\bar{H} = HT^{-1} = [0 \quad 1 \quad 0 \quad 1].$$

On the other hand, if we define the new coördinates by

$$T = \begin{bmatrix} 3 & 4 & 0 & -3 \\ 1 & 1 & 0 & -1 \\ -5 & -7.5 & 0.5 & 6 \\ -6 & 9 & 1 & 6 \end{bmatrix},$$

$$T^{-1} = \begin{bmatrix} 0 & 3 & 1 & -0.5 \\ 1 & -3 & 0 & 0 \\ 3 & -3 & 0 & 1 \\ 1 & -1 & 1 & -0.5 \end{bmatrix},$$

then the system matrices become

$$\bar{F} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\bar{G} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$\bar{H} = [0 \quad 1 \quad 0 \quad 1].$$

The numerical values of these two canonical forms are different, yet Theorem 5 is verified in both cases. In the second case the connections from Part (D) to Parts (A) and (C) are missing. This is not a contradiction since Theorem 5 does not require that all the indicated casual connections in Fig. 4 be actually present.

The transfer-function matrix of the system is easily found from the canonical representation. The coördinate transformations affect only the internal (state) variables, but not the external (input and output) variables; consequently the impulse response matrix is invariant under such transformations. We get by inspection:

$$Z(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{1}{s+1} \end{bmatrix}.$$

It would be rather laborious to determine these transfer functions directly from the signal-flow graph [16] corresponding to $F$, $G$, and $H$.

Example 3. A far less trivial illustration of the canonical decomposition theorem is provided by the following dynamical system, which occurs in the solution of a problem in the theory of statistical filtering [17]. Let $A$ be an arbitrary positive function of $t$ and define

$$F = \begin{bmatrix} -t^4/4A & 1 & 0 \\ -t^3/2A & 0 & 1 \\ -t^2/2A & 0 & 0 \end{bmatrix},$$

$$G = \begin{bmatrix} t^4/4A \\ t^3/2A \\ t^2/2A \end{bmatrix},$$

and

$$H = [0 \quad 1 \quad 0].$$

We introduce new state variables

$$\bar{x}(t) = T(t)x(t),$$

where

$$T(t) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & -t & 0 \\ 0 & 1 & -t \end{bmatrix},$$

$$T^{-1}(t) = \begin{bmatrix} t^2/2 & 1/2 & t/2 \\ t & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Then

$$\bar{F}(t) = T(t)F(t)T^{-1}(t) - \dot{T}(t)T^{-1}(t) = \left[\begin{array}{c|c|c} -t^4/4A & -t^3/4A & -t^2/4A \\ \hline 0 & 0 & 1 \\ \hline 0 & 0 & 0 \end{array}\right],$$

$$\bar{G}(t) = T(t)G(t) = \left[\begin{array}{c} t^2/2A \\ \hline 0 \\ \hline 0 \end{array}\right],$$

and

$$H(t) = H(t)T^{-1}(t) = [t \mid 0 \mid 1].$$

Hence the system consists of Parts $(B - D)$, with $n_B = nc = n_D = 1$. It is interesting that the canonical decomposition is of constant dimension, even though the system may be neither periodic nor analytic.

The preceding examples illustrate special cases of a noteworthy general relationship which exists between the canonical structure of a dynamical system and irreducible realizations of an impulse-response matrix. The main facts here are the following:

THEOREM 7. (i) *The impulse-response matrix of a linear dynamical system* (2.1–2) *depends solely on Part* (*B*) *of the system and is given explicitly by*

$$(5.5) \qquad S(t, \tau) = H^B(t)\Phi^{BB}(t, \tau)G^B(\tau),$$

*where* $\Phi^{BB}$ *is the transition matrix corresponding to* $F^{BB}$.

(ii) *Any two completely controllable and completely observable realizations of S are algebraically equivalent.*

(iii) *A realization of S is irreducible if and only if at all times it consists of Part* (*B*) *alone*; *thus every irreducible realization of S is completely controllable and completely observable.*

*Proof.* The first statement can be read off by inspection from Fig. 4. The second statement is proved in [14]. The necessity of the third statement follows from Theorem 5, while the sufficiency is implied by (ii).

It is clear that Theorem 2 is a consequence of Theorems 5–7.

We can now answer the question posed in section 4 in a definite way:

THEOREM 8 (*Main Result*). *Knowledge of the impulse-response matrix* $S(t, \tau)$ *identifies the completely controllable and completely observable part, and this part alone, of the dynamical system which generated it. This part* (*"B"* in Theorem 5) *is itself a dynamical system and has the smallest dimension among all realizations of S. Moreover, this part is identified by S uniquely up to algebraic equivalence.*

Using different words, we may say that an impulse-response matrix is a *faithful representation* of a dynamical system (2.1–2) if and only if the latter is completely controllable and completely observable.

*Remark.* It is very interesting to compare this result with Theorem 4 of E. F. Moore, in one of the early papers on finite automata [26]:

"*The class of all machines which are indistinguishable from a given strongly connected machine S by any single experiment has a unique* (*up to isomorphism*) *member with a minimal number of states. This unique machine, called the reduced form of S, is strongly connected and has the property that any two of its states are distinguishable.*"

"Indistinguishable machines" in Moore's terminology correspond in ours to alternate realizations of the same input/output relation. "Strongly connected" in his terminology means completely controllable in ours. "Indistinguishable states" in our terminology corresponds to states whose difference, not zero, is an unobservable state in the sense of [3].

Evidently the two theorems are concerned with the same abstract facts, each being stated in a different mathematical framework.

**6. Explicit criteria for complete controllability and observability.** The canonical structure theorem is so far merely an abstract result, since we have not yet given a constructive procedure for obtaining the coördinate transformation which exhibits the system matrices in canonical form. We shall do this in section 7. The method rests on the possibility of finding explicit criteria for complete controllability and complete observability. The following lemmas, proved in [4], play a central role:

LEMMA 1. $n_A(t_0) + n_B(t_0) = \text{rank}\, W(t_0, t_1)$ *for* $t_1 > t_0$ *sufficiently large, where*

$$(6.1) \qquad W(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, \tau)G(\tau)G'(\tau)\Phi'(t_0, \tau)\, d\tau$$

*or*

$$(6.2) \qquad dW/dt_0 = F(t_0)W + WF'(t_0) - G(t_0)G'(t_0),\ W(t_1) = 0.$$

LEMMA 2. $n_c(t_0) + n_D(t_0) = \text{rank}\, M(t_0, t_{-1})$ *for* $t_{-1} < t_0$ *sufficiently small, where*

$$(6.3) \qquad M(t_0, t_{-1}) = \int_{t_{-1}}^{t_0} \Phi'(\tau, t_0)H'(\tau)H(\tau)\Phi(\tau, t_0)\, d\tau$$

*or*

$$(6.4) \quad -dM/dt_0 = F'(t_0)M + MF(t_0) - H'(t_0)H(t_0),\ M(t_{-1}) = 0.$$

For constant systems, the preceding lemmas can be considerably improved [4]:

LEMMA 3. *For a constant system,*

$$(6.5) \qquad n_A + n_B = \text{rank}\, [G, FG, \cdots, F^{n-1}G].$$

LEMMA 4. *For a constant system,*

$$(6.6) \qquad nc + n_D = \text{rank}\, [H', F'H', \cdots, (F')^{n-1}H'].$$

EXAMPLE 4. For $F$ and $G$ defined in Example 2, the matrix (6.5) is

$$(6.7) \qquad \begin{bmatrix} 3 & 3 & -3 & -3 & 3 & 3 & -3 & -3 \\ -2 & -1 & 6 & 8 & 2 & 6 & 14 & 22 \\ 0 & 3 & 12 & 18 & 12 & 24 & 36 & 60 \\ 0 & 1 & 4 & 6 & 4 & 8 & 12 & 20 \end{bmatrix}.$$

The rank of this matrix is 2, which checks with the fact that $n_A = 1$ and $n_B = 1$ in Example 2.

The determination of the rank of (6.7), while elementary, is laborious. For practical purposes it might be better to compute $W$; for instance, by solving the differential equation (6.2).

In the constant case, there is another criterion of complete controllability which is particularly useful in theoretical investigations. The most general form of this theorem (which may be found in [14]) is complicated; we state here a simplified version which is adequate for the present purposes:

LEMMA 5. *Hypothesis: The matrix $F$ is similar to a diagonal matrix. In other words, there is a nonsingular coördinate transformation $\bar{x} = Tx$ with the property that in the new coördinate system $F$ has the form*

$$\bar{F} = TFT^{-1} = \begin{bmatrix} \lambda_1 I_{q_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_r I_{q_r} \end{bmatrix},$$

*where $I_{q_i}$ is a $q_i \times q_i$ unit matrix,*

$$\sum_{i=1}^{r} q_i = n,$$

*and the matrix $G$ has the form*

$$\bar{G} = TG = \begin{bmatrix} \bar{G}^{(1)} \\ \hdashline \vdots \\ \hdashline \bar{G}^{(r)} \end{bmatrix} \begin{matrix} \} \; q_1 \text{ rows} \\ \vdots \\ \} \; q_r \text{ rows.} \end{matrix}$$

*Conclusion: The system is completely controllable if and only if*

$$\text{rank } \bar{G}^{(1)} = q_1, \cdots, \text{rank } \bar{G}^{(r)} = q_r. \tag{6.8}$$

We leave it to the reader to dualize this result to complete observability.

EXAMPLE 5. Consider the special case $q_1 = \cdots = q_r = 1$ of Lemma 5. The eigenvalues of $F$ are then distinct. If condition (6.8) is satisfied, every element of the one-column matrix $\bar{G}$ is nonzero; by a trivial transformation, all of these elements can be made equal to 1, without affecting $\bar{F}$. Thus we can choose a coordinate system in which $F$, $G$ have the representation:

$$\bar{F} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} (\lambda_i = \lambda_j \Rightarrow i = j), \bar{G} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6.9}$$

This is the canonical form of Lur'e [18]. It is closely related to the partial-fraction expansion of transfer functions. To illustrate this, consider the $1 \times 1$ transfer-function matrix:

$$z_{11}(s) = \frac{s+2}{(s+1)(s+3)(s+4)} = \frac{1/6}{s+1} + \frac{1/2}{s+3} - \frac{2/3}{s+4}.$$

This transfer function is realized by the system:

$$F = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -4 \end{bmatrix}, \tag{6.10}$$

$$G = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{6.11}$$

and

$$H = [\tfrac{1}{6} \quad \tfrac{1}{2} \quad -\tfrac{2}{3}] \tag{6.12}$$

which is in the canonical form of Lur'e.

By Lemma 5, (6.10–11) is completely controllable; by the dual of Lemma 5, (6.10–12) is completely observable.

We can double-check these facts by means of Lemmas 3–4. For (6.9) the matrix (6.5) is

$$\begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{bmatrix}, \tag{6.13}$$

where the $\lambda_j$ are the diagonal elements (= eigenvalues) of $F$ in (6.9). But the determinant of (6.13) is the well-known Vandermonde determinant. The latter is nonzero if and only if all the $\lambda_i$ are distinct, which is what we have assumed.

**7. Computation of the canonical structure.** We show now how to determine explicitly the change of coördinates which reduces $F$, $G$, $H$ to the canonical form. We consider only the constant case of (2.1–2). The computations are elementary; it is not necessary to diagonalize the matrix $F$ or even to determine its eigenvalues.

The procedure is as follows:

(a) We compute the controllability matrix $W = W(0, 1)$* given by

* It can be shown [4, Theorem 10] that in the constant case one may choose any $t_1 > t_0$ in Lemma 1.

(6.1); for instance, by solving the differential equation (6.2). Then we find a nonsingular matrix $T$ such that

$$T'WT = E = \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} \tag{7.1}$$

where $I_{n_1}$ is the $n_1 \times n_1$, $0 \leqq n_1 \leqq n$, unit matrix and the 0's are zero matrices of appropriate size. Clearly $n_1 = n_A + n_B$ is the number of controllable state variables.

The matrix $T$ defines the change of coördinates

$$x = T\tilde{x}; \tag{7.2}$$

in terms of the new coördinates, the system matrices are

$$\bar{F} = T^{-1}FT, \qquad \bar{G} = T^{-1}G, \qquad \bar{H} = HT, \qquad \bar{W} = E. \tag{7.3}$$

$$\bar{x} = \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \end{bmatrix}, F = \begin{bmatrix} \bar{F}^{11} & \bar{F}^{12} \\ 0 & \bar{F}^{22} \end{bmatrix}, \bar{G} = \begin{bmatrix} \bar{G}^1 \\ 0 \end{bmatrix}, \text{ and } \quad H = [\bar{H}^1 \quad \bar{H}^2]. \tag{7.4}$$

This decomposition is trivial (and therefore omitted) if $n_1 = n$, i.e., when the system is completely controllable.

(b) Next we consider the two subsystems defined by

$$\begin{aligned} &\bar{F}^{11}, \bar{G}^1, \quad \text{and} \quad \bar{H}^1; \\ &\bar{F}^{22}, 0, \quad \text{and} \quad \bar{H}^2. \end{aligned} \tag{7.5}$$

We compute the observability matrices $\bar{M}^1 = \bar{M}^1(0, 1)$ and $\bar{M}^2 = \bar{M}^2(0, 1)$ given by (6.3) for both of these subsystems. Then we determine two nonsingular matrices $\bar{U}^1$, $\bar{U}^2$ such that

$$(\bar{U}^1)'\bar{M}^1\bar{U}^1 = \bar{E}^1 = \begin{bmatrix} 0 & 0 \\ 0 & I_{n_B} \end{bmatrix}, \tag{7.6}$$

and

$$(\bar{U}^2)'\bar{M}^2\bar{U}^2 = \bar{E}^2 = \begin{bmatrix} 0 & 0 \\ 0 & I_{n_d} \end{bmatrix}. \tag{7.7}$$

These results define another change of coördinates

$$\bar{x} = \begin{bmatrix} \bar{x}^1 \\ \bar{x}^2 \end{bmatrix} = \bar{U}\tilde{x} = \begin{bmatrix} \bar{U}^1 & 0 \\ 0 & \bar{U}^2 \end{bmatrix} \cdot \begin{bmatrix} \tilde{x}^1 \\ \tilde{x}^2 \end{bmatrix}.$$

One or the other of these transformations is superfluous if $n_B = n_1$ or $n_d = n - n_1$.

After the coördinate changes (7.2) and (7.8), we obtain the following matrices

$$(7.9)\qquad \tilde{x} = \begin{bmatrix} x^A \\ x^B \\ \hline x^c \\ x^d \end{bmatrix}, \qquad \tilde{F} = \bar{U}^{-1}\bar{F}\bar{U} = \left[\begin{array}{cc|cc} F^{AA} & F^{AB} & F^{Ac} & F^{Ad} \\ 0 & F^{BB} & \tilde{F}^{Bc} & F^{Bd} \\ \hline 0 & 0 & \tilde{F}^{Bc} & F^{cd} \\ 0 & 0 & 0 & F^{dd} \end{array}\right],$$

$$\bar{U}^{-1}G = \tilde{G} = \begin{bmatrix} G^A \\ G^B \\ \hline 0 \\ 0 \end{bmatrix}, \qquad \tilde{H} = \bar{H}\bar{U} = [0 \quad H^B \quad 0 \quad H^d],$$

$$\tilde{M}^1 = \bar{E}^1, \qquad \tilde{M}^2 = \bar{E}^2.$$

Clearly, $n_B$ is the number of state variables which are both controllable and observable. But, in general, $n_d < n_D$ and $n_c > n_C$ .

(c) It remains to transform the element $\tilde{F}^{Bc}$ into 0, if this is not already the case. (If $\tilde{F}^{Bc} = 0$, then $n_c = n_C$ , $n_d = n_D$ and (7.9) has the desired canonical structure.)

We consider the subsystem

$$(7.10)\qquad \tilde{F}^* = \left[\begin{array}{c|c} F^{BB} & \tilde{F}^{Bc} \\ \hline 0 & \tilde{F}^{cc} \end{array}\right], \quad \tilde{G}^* = \begin{bmatrix} G^B \\ \hline 0 \end{bmatrix}, \quad \text{and} \quad \tilde{H}^* = [H^B \mid 0].$$

The corresponding observability matrix given by (6.3) is

$$\tilde{M}^*(0, 1) = \tilde{M}^* = \left[\begin{array}{c|c} I_{n_B} & A \\ \hline A' & Q \end{array}\right], \quad (Q = Q' \text{ nonnegative definite.})$$

(The upper left element of $\tilde{M}^*$ is $I_{n_B}$ in view of (7.9); all we know about the other elements is their symmetry properties.) Letting

$$\tilde{V}^* = \left[\begin{array}{c|c} I_{n_B} & -A \\ \hline 0 & I_{n_c} \end{array}\right],$$

we find that

$$(\tilde{V}^*)'\tilde{M}^*\tilde{V}^* = \tilde{M}^* = \left[\begin{array}{c|c} I_{n_B} & 0 \\ \hline 0 & R \end{array}\right],$$

where $R = Q - A'A$ is a symmetric, nonnegative-definite matrix.

Now let $\tilde{V}^{**}$ be a nonsingular matrix such that

$$(\tilde{V}^{**})'\tilde{M}^{**}\tilde{V}^{**} = \left[\begin{array}{c|cc} I_{n_B} & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & I_{n_e} \end{array}\right],$$

where $n_e = \operatorname{rank} R$. Let $\tilde{V} = \tilde{V}^*\tilde{V}^{**}$. Since $\tilde{V}^*$ and $\tilde{V}^{**}$ are upper triangular relative to the partitioning in (7.10), so is $\tilde{V}$, which will take $\tilde{F}^*$ into the upper triangular form

$$\tilde{V}^{-1}\tilde{F}^*\tilde{V} = \left[\begin{array}{c|cc} F^{BB} & F^{BC} & F^{Be} \\ \hline 0 & F^{CC} & F^{Ce} \\ 0 & F^{eC} & F^{ee} \end{array}\right].$$

where $n_C = n_c - n_e$. But these transformations decompose $\tilde{F}^*$ into a completely observable and an unobservable part. Hence $F^{BC} = F^{eC} = 0$. Moreover,

$$\tilde{H}^*\tilde{V} = [H^B \;\vdots\; 0]\tilde{V} = [H^B \;|\; 0 \quad H^e]$$

Theorem 9. *The explicit transformation which takes the constant matrices $F$, $G$, and $H$ into the canonical form required by Theorem* (5-iii) *is given by $x \to \tilde{V}^{-1}\tilde{U}^{-1}\tilde{T}^{-1}x$. We partition*

$$F^{Ac} = [F^{AC} \quad F^{Ae}],$$

*and partition*

$$F^{cd} = \begin{bmatrix} F^{Cd} \\ F^{ed} \end{bmatrix}.$$

*Then we define $n_D = n_d + n_e$ and find*

$$F^{AD} = [F^{Ae} \qquad F^{Ad}],$$

$$F^{BD} = [F^{Be} \qquad F^{Bd}],$$

$$F^{CD} = [F^{Ce} \qquad F^{Cd}],$$

$$F^{DD} = \begin{bmatrix} F^{ee} & F^{ed} \\ 0 & F^{dd} \end{bmatrix},$$

$$H^D = [H^e \quad H^d].$$

## 8. Construction of irreducible realizations.

Now we give an explicit procedure for the construction of an irreducible realization of a weighting-function matrix $W(t - \tau)$. In view of Theorem 7, part (iii), we can do this in two stages:

(I) We construct a realization of $W$, then

(II-A) we prove, using Lemmas 1–5, that the resultant system is completely controllable and completely observable, hence irreducible; or

(II-B) we carry out explicitly the canonical decomposition and remove all parts other than (B).

Instead of the weighting-function matrix $W$, it is usually more convenient to deal with its Laplace transform $Z$.

Let us consider the problem with Method A in order of increasing difficulty.

*Case 1.* $m = p = 1$. This is equivalent to the problem of simulating a single transfer function on an analog computer. There are several well-known solutions. They may be found in textbooks on classical servomechanism theory or analog computation.

Without loss of generality (see Theorem 4) we may consider transfer functions of the form

$$z_{11}(s) = \frac{a_n s^{n-1} + \cdots + a_1}{s^n + b_n s^{n-1} + \cdots + b_1} = \frac{N(s)}{D(s)} \tag{8.1}$$

where the $a_n, \cdots, a_1$; $b_n, \cdots, b_1$ are real numbers. Of course, at least one of the $a_i$ must be different from zero. We assume also that the numerator $N(s)$ and denominator $D(s)$ of $z_{11}(s)$ have no common roots.

There are two basic realizations of (8.1). See Figs. 5–6, where the standard signal-flow-graph notation [16] is used. In either case, one verifies almost by inspection that the transfer functions relating $y_1$ to $u_1$ are indeed given by $z_{11}$.

In Fig. 5, the system matrices are

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -b_1 & -b_2 & -b_3 & \cdots & -b_{n-1} & -b_n \end{bmatrix}, \tag{8.2}$$

$$G = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{8.3}$$

and

$$H = [a_1 \quad a_2 \quad \cdots \quad a_{n-1} \quad a_n]. \tag{8.4}$$

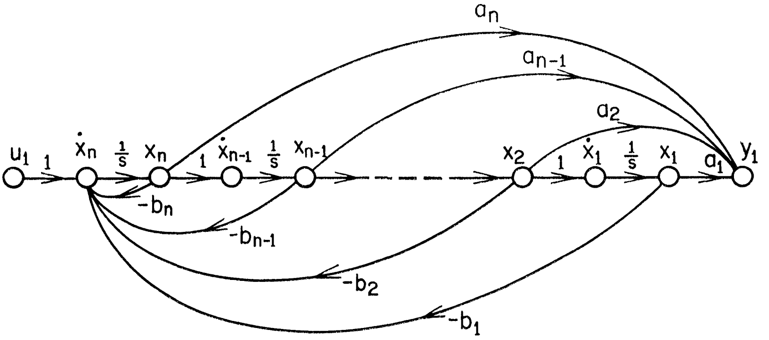

FIGURE 5.

In Fig. 6, the system matrices are

$$(8.5) \qquad F = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -b_1 \\ 1 & 0 & 0 & \cdots & 0 & -b_2 \\ 0 & 1 & 0 & \cdots & 0 & -b_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -b_n \end{bmatrix},$$

$$(8.6) \qquad G = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix},$$

and

$$(8.7) \qquad H = [0 \quad 0 \quad 0 \quad \cdots \quad 1].$$

It is very easy to check by means of (6.5) and (6.6) that the system (8.2, 3) is completely controllable and (8.5, 7) is completely observable.

However, if we attempt to check the controllability of (8.5, 6) by means of (6.5) we get a matrix whose elements are complicated products of the coefficients of $N(s)$ and $D(s)$. To prove that the determinant of this matrix does not vanish, we have only one fact at our disposal: the assumption that $N(s)$ and $D(s)$ have no common roots. Guided by this observation, we find that the following is true:

LEMMA 7. *Suppose F has the form* (8.5) *and G has the form* (8.6). *Then* (i) *we have the relation*

$$(8.8) \qquad K(F, G) = [G \quad FG \quad \cdots — \quad F^{n-1}G] = N(F),$$

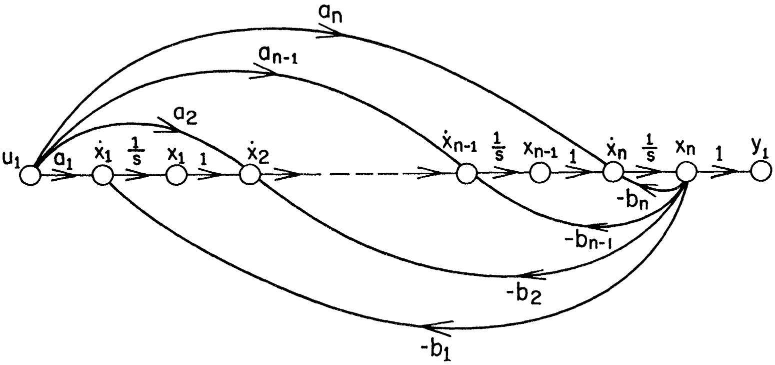

FIGURE 6.

*and (ii) the polynomials $N(s)$ and $D(s)$ have no root in common if and only if* $\det K(F, G) \neq 0$.

The main fact to be proved is (ii), for then the complete controllability of (8.5, 6) follows by Lemma 3. A straightforward way of establishing (ii) is to transform the standard Euler-Sylvester determinantal criterion [19, p. 84] for the nonexistence of common roots of $N(s)$ and $D(s)$ (the so-called *resolvent* of $N(s)$ and $D(s)$) into the form (8.8). This can be easily done, but the details are not very transparent. Therefore we prefer to give another

*Proof.* Let $e_i$, $i = 1, \cdots, n$, be the set of $n$-vectors in which the $j$-th component of $e_i$ is $\delta_{ij}$. Since $F$ is given by (8.5), we see that $e_{i+1} = Fe_i$, $1 \leqq n - 1$, and $K(F, e_1) = [e_1, e_2, \cdots, e_n] = I$. Hence $K(F, e_i) = K(F, F^{i-1}e_1) = F^{i-1}K(F, e_1) = F^{i-1}$, when $1 \leqq i \leqq n$. Then (8.8) follows by linearity.

Let $\lambda_i[A]$, $i = 1, \cdots, n$ denote the eigenvalues (not necessarily distinct) of a square matrix $A$. Then

$$\det K(F, G) = \prod_{i=1}^{n} \lambda_i[N(F)] = \prod_{i=1} N(\lambda_i[F]),$$

where the second equality follows from (8.8) by a well-known identity in matrix theory. Thus $\det K(F, G) = 0$ if and only if $N(\lambda_i[F]) = 0$ for some $i$; that is, when an eigenvalue of $F$ is a root of $N(\lambda)$. Since the eigenvalues of $F$ are roots of $D(\lambda)$, this proves (ii).*

It is interesting that (8.8) provides a new representation for the resolvent, which is preferable in some respects to the Euler-Sylvester determinant. The latter is a $2n \times 2n$ determinant, whereas $\det K(F, G)$ is $n \times n$.

The complete observability of (8.2, 4) is proved similarly.

The systems given by (8.2–4) and (8.5–7) are duals of one another in

* The present proof of Lemma 6 was suggested by Drs. John C. Stuelpnagel and W. M. Wonham of RIAS.

the sense defined by (5.3). Fig. 6 is a reflection of Fig. 5 about the vertical axis, with all arrows reversed.

A third type of realization in common use is obtained from the partial-fraction expansion of $z_{11}(s)$ (see Example 5). Note, however, that this requires factorization of the denominator of $z_{11}(s)$, whereas the preceding realizations can be written down by inspection, using only the coefficients of $z_{11}(s)$.

These considerations may be summarized as the following result, which is a highly useful fact in control theory:

THEOREM 10. *Consider a linear constant dynamical system with* $m = p = 1$, *which is completely controllable and completely observable. Then one may always choose a basis in the state space so that* $F$, $G$, $H$ *have the form* (8.2–4) *or* (*with respect to a different basis*) (8.5–7).

*Proof.* Let (8.1) be the transfer-function matrix of the given dynamical system. By Theorem 8, the given system is an irreducible realization of (8.1). So are the systems specified by (8.2–4) and (8.5–7). By Theorem (7-ii), all three systems are algebraically equivalent and by constancy (Theorem 3) they are even strictly equivalent.

Extensions of this theorem may be found in [14]. For an interesting application to the construction of Lyapunov functions, see [25].

The procedure described here may be generalized to the non-constant case. Assuming the factorization (4.3) of $S(t, \tau)$ is known (with $m = p = 1$), Batkov [20] shows how to determine the coefficients of the differential equation

$$\begin{aligned} d^n y_1/dt^n + b_n(t)^{n-1}y_1/dt^{n-1} + \cdots + b_1(t)y_1 & \\ = a_n(t)d^{n-1}u_1/dt^{n-1} + \cdots + a_1(t)u_1 . & \end{aligned} \tag{8.9}$$

Laning and Battin [21, p. 191–2] show how one converts (8.9) into a system of first-order differential equations (2.1) with variable coefficients. We shall leave to the reader the proof of the irreducibility of the realization so obtained.

*Case 2-a.* $m = 1$, $p > 1$. We have a single-input/multi-output system. We can realize $Z(s)$, without factoring the denominators of its transfer functions, by the following generalization of the procedure given by Fig. 5 and (8.2–4).

First, we find the smallest common denominator of the elements of $Z(s)$. (This can be done, of course, without factorization.) $Z(s)$ assumes the form

$$Z(s) = \frac{1}{s^n + b_n s^{n-1} + \cdots + b_1} \begin{bmatrix} a_{1n} s^{n-1} + \cdots + a_{11} \\ \cdots \\ a_{pn} s^{n-1} + \cdots + a_{p1} \end{bmatrix}.$$

Then the following dynamical system provides an irreducible realization of $Z(s)$: $F$ and $G$ are as in (8.2–3), while $H$ given by (8.4) is generalized to

$$H = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pn} \end{bmatrix}.$$

Complete controllability is trivial; complete observability is established by a straightforward generalization of Lemma 6.

In this case we form $p$ linear functions of the state, rather than merely one as in Fig. 5.

*Case 2-b.* $m > 1$, $p = 1$. We can realize this multi-input/single-output system analogously to Case 2-a by generalizing the procedure given by Fig. 6 and (8.5–8.7). Let us write the elements of $Z(s)$ in terms of their smallest common denominator:

$$Z(s) = \left[\frac{a_{n1}\, s^{n-1} + \cdots a_{11}}{s^n + b_n\, s^{n-1} + \cdots + b_1} \cdots \frac{a_{nm}\, s^{n-1} + \cdots + a_{1m}}{s^n + b_n\, s^{n-1} + \cdots + b_1}\right].$$

Then the desired irreducible realization consists of $F$ and $H$ as defined by (8.5–6), while

$$G = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}.$$

This case is the dual of Case 2-a.

Even in Case 2, it is impractical to give a general formula which expresses the coefficients of $F$, $G$, and $H$ in terms of the coefficients of the transfer functions in $Z(s)$ if the denominators are not all the same. When we pass to the general case, determination of $F$, $G$, and $H$ often requires extensive numerical computation.

*Case 3. m, p arbitrary.* Here Method (A) is very complicated if any transfer function in $Z(s)$ has multiple poles [14]. In most practical applications, however, such complications are of no interest. Ruling them out, E. G. Gilbert gave an elegant and relatively simple solution [5].

Let $s_1, \cdots, s_q$ be distinct complex numbers corresponding to the poles of all the elements of $Z(s)$. Assume that all poles are simple. Then

$$R(k) = \lim_{s \to s_k} (s - s_k) Z(s), \; k = 1, \cdots, q$$

is the $k$-th residue matrix of $Z(s)$. If $s_\ell = \bar{s}_k$, then $R(s_\ell) = \bar{R}(s_k)$, where the bar denotes the complex conjugate. In terms of the residue matrices, the weighting-function matrix $W(t)$ corresponding to $Z(s)$ has the explicit form

$$W(t) = \mathfrak{L}^{-1}[Z(s)] = \sum_{k=1}^{q} R(k) e^{s_k t}.$$

We have then:

THEOREM 11. (*Gilbert*). *Hypotheses: No element of the transfer-function* *matrix $Z(s)$ has multiple poles. $Z(s)$ has a total of $q$ distinct poles $s_1, \cdots, s_q$, with corresponding residue matrices $R(1), \cdots, R(q)$.*

*Conclusions:* (i) *The dimension of irreducible realizations of $Z(s)$ is*

$$(8.11) \qquad n = \sum_{k=1}^{q} r_k, \text{ where } r_k = \operatorname{rank} R(k).$$

(ii) *Write*

$$(8.12) \qquad R(k) = H(k)G(k), \qquad k = 1, \cdots, q,$$

*where $H(k)$ is a $p \times r_k$ matrix and $G(k)$ is an $r_k \times m$ matrix, both of rank $r_k$. Then $Z(s)$ has the irreducible realization*

$$(8.13) \qquad F = \begin{bmatrix} s_1 I_{r_k} & & 0 \\ & \ddots & \\ 0 & & s_q I_{r_q} \end{bmatrix}, \qquad (I_r = r \times r \text{ unit matrix}),$$

$$(8.14) \qquad G = \begin{bmatrix} G(1) \\ \vdots \\ G(q) \end{bmatrix},$$

*and*

$$(8.15) \qquad H = [H(1) \quad \cdots \quad H(q)].$$

*Proof.* This is one of the main results in [5]. With the aid of machinery developed here, we can give a shorter (though more abstract) demonstration. The factorization (8.12) is well known in linear algebra. We give in the Appendix various explicit formulae (which are easily machine-computable) for $G(k)$ and $H(k)$. Applying Lemma 5 shows that the dynamical system defined by (8.13–15) is completely controllable and completely observable. Hence it is irreducible, which implies formula (8.11). By elementary changes of variables, (8.13–15) can be transformed into matrices which have only real elements.

A serious disadvantage of Method (A), as expressed by Theorem 11, is that the denominators of the transfer functions in $Z(s)$ must be factored in order to determine the poles. This is not easily done numerically. Moreover, the residue matrices $R(k)$ corresponding to complex poles are complex, which makes the factorization (8.11) more complicated (see Appendix).

Now we turn to Method (B). This method does not require computation of eigenvalues, and it is not bothered by multiple poles. This is a decided advantage in numerical calculations. On the other hand, the method is not convenient for simple illustrative examples. Nor is it possible to display the elements of $F$, $G$, and $H$ as simple functions of the coefficients in $z_{ij}(s)$.

An easy way of realizing $Z(s)$ (without guaranteeing irreducibility) is the following. Let $\alpha_i$ be the number of distinct poles (counting each pole with its maximum multiplicity) in the $i$-th row of $Z(s)$, and let $\beta_i$ be the number of poles in the $i$-th column. Then the maximum number $n_0$ of state variables required to realize $Z(s)$ by repeatedly using the scheme given under Case 2-a or 2-b is

$$n_0 = \min\left\{ \sum_{i=1}^{p} \alpha_i , \quad \sum_{j=1}^{m} \beta_j \right\}.$$

As before, we can determine the $\alpha_i$ and $\beta_j$ without factoring the transfer functions of $Z(s)$. There is in general no simple way in this method to determine the dimension $n \leqq n_0$ of irreducible realizations without performing the computations outlined in Section 7.

The two methods are best compared via an example. This example must be of fairly high order, since we wish to provide accurate numerical checks.

Example 6. Consider the transfer-function matrix

$$Z(s) = \begin{bmatrix} \dfrac{3(s+3)(s+5)}{(s+1)(s+2)(s+4)} & \dfrac{6(s+1)}{(s+2)(s+4)} & \dfrac{2s+7}{(s+3)(s+4)} & \dfrac{2s+5}{(s+2)(s+3)} \\[2ex] \dfrac{2}{(s+3)(s+5)} & \dfrac{1}{(s+3)} & \dfrac{2(s-5)}{(s+1)(s+2)(s+3)} & \dfrac{8(s+2)}{(s+1)(s+3)(s+5)} \\[2ex] \dfrac{2(s^2+7s+18)}{(s+1)(s+3)(s+5)} & -\dfrac{2s}{(s+1)(s+3)} & \dfrac{1}{(s+3)} & \dfrac{2(5s^2+27s+34)}{(s+1)(s+3)(s+5)} \end{bmatrix}.$$

Applying Method (A) first, we find that the residue matrices are:

$$R(1) = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 3 & 1 & 0 & 3 \end{bmatrix}; \qquad r_1 = 3.$$

$$R(2) = \begin{bmatrix} -4.5 & -3 & 0 & 1 \\ 0 & 0 & -6 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}; \qquad r_2 = 2.$$

$$R(3) = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ -3 & -3 & 1 & 1 \end{bmatrix}; \qquad r_3 = 2.$$

$$R(4) = \begin{bmatrix} -0.5 & 9 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}; \qquad r_4 = 1.$$

$$R(5) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -3 \\ 2 & 0 & 0 & 6 \end{bmatrix}; \qquad r_5 = 1.$$

Thus $n = 9$.

Employing the procedure given in the Appendix, we find the following factors for matrices $R(k)$ (the products are accurate up to four places beyond the decimal point):

$$H(1) = \begin{bmatrix} 8.0000 & 0.0000 & 0.0000 \\ 0.0000 & 4.1231 & 0.0000 \\ 3.0000 & 0.7276 & 3.0774 \end{bmatrix},$$

$$G(1) = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.9701 & 0.2425 \\ 0.0000 & 0.3249 & -0.2294 & 0.9175 \end{bmatrix};$$

$$H(2) = \begin{bmatrix} 5.5000 & 0.0000 \\ 0.0000 & 6.0000 \\ 0.0000 & 0.0000 \end{bmatrix},$$

$$G(2) = \begin{bmatrix} -0.8182 & -0.5455 & 0.0000 & 0.1818 \\ 0.0000 & 0.0000 & -1.0000 & 0.0000 \end{bmatrix};$$

$$H(3) = \begin{bmatrix} 1.3416 & 0.4472 \\ 3.1305 & -0.4472 \\ 0.0000 & 4.4721 \end{bmatrix},$$

$$G(3) = \begin{bmatrix} 0.2236 & 0.2236 & 0.6708 & 0.6708 \\ -0.6708 & -0.6708 & 0.2236 & 0.2236 \end{bmatrix};$$

$$H(4) = \begin{bmatrix} 9.0692 \\ 0.0000 \\ 0.0000 \end{bmatrix}, \qquad G(4) = [-0.0551 \quad 0.9924 \quad 0.1103 \quad 0.0000];$$

$$H(5) = \begin{bmatrix} 0.0000 \\ -3.1623 \\ 6.3246 \end{bmatrix}, \qquad G(5) = [0.3162 \quad 0.0000 \quad 0.0000 \quad 0.9487].$$

Using these numerical results, we find that the dynamical equations of the irreducible realization are given by

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix},$$

$$G = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.9701 & 0.2425 \\ 0.0000 & 0.3249 & -0.2294 & 0.9175 \\ -0.8182 & -0.5455 & 0.0000 & 0.1818 \\ 0.0000 & 0.0000 & -1.0000 & 0.0000 \\ 0.2236 & 0.2236 & 0.6708 & 0.6708 \\ -0.6708 & 0.6708 & 0.2236 & 0.2236 \\ -0.0551 & 0.9924 & 0.1103 & 0.0000 \\ 0.3162 & 0.0000 & 0.0000 & 0.9487 \end{bmatrix},$$

$$H = \begin{bmatrix} 8.0000 & 0.0000 & 0.0000 & 5.5000 & 0.0000 & 1.3416 & 0.4472 & 9.0692 & 0.0000 \\ 0.0000 & 4.1231 & 0.0000 & 0.0000 & 6.0000 & 3.1305 & -0.4472 & 0.0000 & -3.1623 \\ 3.0000 & 0.7276 & 3.0774 & 0.0000 & 0.0000 & 0.0000 & 4.4721 & 0.0000 & 6.3246 \end{bmatrix}.$$

Now we apply Method (B). First of all we note that $\alpha_1 = \alpha_2 = 4$, $\alpha_3 = 3$, while $\beta_1 = 5$, $\beta_2 = \beta_3 = 4$ (see p. 181). Hence it is best to choose for the preliminary realization three structures of the type discussed under Case 2-b. This will require $n_0 = p(\alpha_1 + \alpha_2 + \alpha_3) = 11$ dimensions.

Next, we find the least common denominator of the rows of $Z(s)$. See Fig. 7.

$$Z(s) = \begin{bmatrix} \dfrac{3(s^3 + 11s^2 + 39s + 45)}{s^4 + 10s^3 + 35s^2 + 50s + 24} & \dfrac{6(s^3 + 5s^2 + 7s + 3)}{\cdots} & \dfrac{2s^3 + 13s^2 + 25s + 14}{\cdots} & \dfrac{2s^3 + 15s^2 + 33s + 20}{\cdots} \\ \dfrac{2(s^2 + 3s + 2)}{s^4 + 11s^3 + 41s^2 + 61s + 30} & \dfrac{s^3 + 8s^2 + 17s + 10}{\cdots} & \dfrac{2(s^2 + 10s + 25)}{\cdots} & \dfrac{8(s^2 + 4s + 4)}{\cdots} \\ \dfrac{2(s^2 + 7s + 18)}{s^3 + 9s^2 + 23s + 15} & \dfrac{-2(s^2 + 5s)}{\cdots} & \dfrac{s^2 + 6s + 5}{\cdots} & \dfrac{2(5s^2 + 27s + 34)}{\cdots} \end{bmatrix}$$

FIGURE 7.

The desired realization of $Z(s)$ can be read off by inspection from Fig. 7, using (8.5) and (8.6):

$$F = \left[\begin{array}{cccc|cccc|ccc} 0 & 0 & 0 & -24 & & & & & & & \\ 1 & 0 & 0 & -50 & & & 0 & & & 0 & \\ 0 & 1 & 0 & -35 & & & & & & & \\ 0 & 0 & 1 & -10 & & & & & & & \\ \hline & & & & 0 & 0 & 0 & -30 & & & \\ & & 0 & & 1 & 0 & 0 & -61 & & 0 & \\ & & & & 0 & 1 & 0 & -41 & & & \\ & & & & 0 & 0 & 1 & -11 & & & \\ \hline & & & & & & & & 0 & 0 & -15 \\ & & 0 & & & & 0 & & 1 & 0 & -23 \\ & & & & & & & & 0 & 1 & -9 \end{array}\right],$$

$$G = \left[\begin{array}{rrrr} 135 & 18 & 14 & 20 \\ 117 & 42 & 25 & 33 \\ 33 & 30 & 13 & 15 \\ 3 & 6 & 2 & 2 \\ \hline 4 & 10 & 50 & 32 \\ 6 & 17 & 20 & 32 \\ 2 & 8 & 2 & 8 \\ 0 & 1 & 0 & 0 \\ \hline 36 & 0 & 5 & 68 \\ 14 & -10 & 6 & 54 \\ 2 & -2 & 1 & 10 \end{array}\right],$$

and

$$H = \left[\begin{array}{ccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

By virtue of its construction, this system is completely observable but we cannot tell by inspection whether or not it is completely controllable. (From the results obtained above with Method (A), we know that the system is not completely controllable since $11 = n_0 > n = 9$.) Therefore the canonical decomposition may contain Parts (B) and (D).

To see what the dimensions of these parts are, we compute numerically the decomposition of the system into completely controllable and uncontrollable parts according to the method described in Section 8. These calculations involve only the matrices $F$ and $G$, but the resulting transformations must be applied also to the matrix $H$.

$$\hat{F} = \left[\begin{array}{rrrrrrrrr|rr} -0.3346 & 0.1182 & 0.0139 & -0.0299 & 0.0097 & -0.0001 & -0.0663 & -0.0113 & 0.0000 & 0.0000 & 0.8334 \\ 0.2455 & -0.2025 & -0.0115 & 0.0268 & -0.0101 & 0.0000 & 0.0734 & 0.0120 & 0.0001 & 0.0000 & 0.7189 \\ -0.8333 & -0.3850 & -0.1023 & 1.0535 & -0.2230 & -0.0005 & -0.9998 & 0.0237 & -0.0120 & 21.5463 & -9.9631 \\ -0.2943 & 0.2032 & 0.0361 & 0.0022 & -0.0610 & -0.0044 & 0.1773 & -0.0569 & 0.0154 & -1.6475 & 0.8726 \\ -0.8896 & 0.8321 & 0.1838 & 0.4999 & -0.4287 & -0.0275 & 1.1199 & 0.0024 & 0.0089 & -1.1882 & 2.5562 \\ 0.2477 & 1.3097 & 2.0439 & -0.3777 & -0.9685 & -0.4965 & -0.4046 & 0.3657 & 0.0199 & -462.2221 & -73.0068 \\ -0.1358 & 0.0429 & 0.0009 & -0.0321 & 0.0315 & 0.0016 & -0.3252 & -0.0152 & -0.0010 & 0.0000 & 0.4698 \\ 1.1634 & -0.0290 & 0.1114 & 1.1645 & -0.4233 & -0.0196 & 0.4057 & -0.0480 & -0.0040 & -25.8717 & -2.1911 \\ -0.2787 & -0.6854 & 2.4604 & 2.4604 & -1.5802 & -0.2801 & 0.8992 & 0.8863 & -0.2649 & -710.3771 & -32.0706 \\ \hline 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.5\text{[illegible]}2 & -0.0003 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0401 & -0.2993 \end{array}\right] \times 10$$

FIGURE 8.

$$\hat{G} = \left[\begin{array}{cccc} 1.2012 & -0.7886 & -0.3459 & -0.5320 \\ 1.6822 & 0.6039 & 0.3594 & 0.4745 \\ -1.7962 & 0.8640 & 2.5387 & -1.7927 \\ -0.1683 & -0.2420 & 2.2981 & 0.4039 \\ -0.0663 & 0.9596 & 1.4137 & 2.7344 \\ 1.0412 & 1.2734 & 1.5800 & -2.9382 \\ -0.1117 & -0.3946 & -0.0977 & 1.5554 \\ 1.3130 & -2.8589 & 1.9252 & -0.4989 \\ -2.8756 & -1.7637 & 0.9281 & -2.5137 \\ \hdashline 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{array}\right] \times 10$$

$$\hat{H} = \left[\begin{array}{ccccccccc:cc} -0.2928 & 0.4076 & 0.0160 & -0.0373 & 0.0141 & 0.0000 & -0.0981 & -0.0167 & -0.0001 & 0.0000 & 1.0000 \\ -0.0286 & 0.0023 & -0.0116 & -0.0169 & 0.0373 & 0.0022 & -0.0830 & -0.0033 & -0.0007 & 1.0000 & 0.0000 \\ -0.0599 & 0.2173 & 0.0161 & 0.0354 & -0.0603 & -0.0039 & 0.6736 & 0.0178 & 0.0024 & 0.0000 & 0.0000 \end{array}\right]$$

FIGURE 9.

The final results may be seen in Figs. 8–9, which give the matrices $\hat{F}$, $\hat{G}$, and $\hat{H}$. Elements in the lower left-hand corner of $\hat{F}$ should be exactly zero. In fact, they are zero to at least the number of digits indicated in Fig. 8.

To check the accuracy of these two irreducible realizations of the transfer function matrix on p. 181, we have computed the corresponding weighting-function matrices $W^{(1)}(t)$ and $W^{(2)}(t)$. The equality $W^{(1)}(t) = W^{(2)}(t)$ was found to be correct to at least four significant digits.

**9. Other applications to system theory.** The literature of system theory contains many instances of errors, incomplete or misleading solutions of problems, etc., which can be traced to a lack of understanding of the issues discussed in this paper. This section presents some cases of this known to the writer; other examples may be found in the paper of Gilbert [5].

*Analog computers.* According to Theorem 8, a linear dynamical system (2.1–2) is a "faithful" realization of an impulse-response matrix if and only if it is irreducible. Suppose the dynamical equations (2.1–2) are programmed on an analog computer. (See [8].) Then it is clear from Theorem 8 that *the computer will simulate the impulse-response matrix correctly if and only if a minimal number of integrators are used.* Otherwise the system programmed on the analog computer will have, besides Part (B), at least one of the Parts (A), (C), or (D). Since the impulse-response matrix determines Part (B), and that alone, the nature of the redundant parts will depend not on the impulse-response matrix but on the particular method used to obtain the dynamical equations. It should be borne in mind that the canonical decomposition is an abstract thing; usually it is not possible to identify the redundant integrators without a change of variables.

The writer is not aware of any book or paper on analog computation where this is explicitly pointed out. But the facts of life seem to be well known (intuitively) to practitioners of the analog art.

That redundancy in the number of integrators used *can* cause positive harm is quite clear from the canonical structure theorem.

EXAMPLE 7. Let the simulated system consist of Parts (A) and (B) and suppose that Part (A) is unstable. Because of noise in the computer, Part (A) will be subject to perturbations; they will be magnified more and more, because of the instability. As long as assumptions of linearity hold exactly, the unstable (A) component of the state vector will not be noticed, but soon the computer will cease to function because its linear range will be exceeded.

*Lur'e canonical form.* In his book on the Lur'e problem, Letov implies [18; equation (2.4) and (2.23)] that every vector system

$$(9.1) \qquad dx/dt = Fx + g\cdot\sigma \qquad (\sigma = \text{scalar})$$

can be reduced to the canonical form

$$(9.2) \qquad dx_i/dt = \lambda_i x_i + \sigma, \qquad i = 1, \cdots, n$$

whenever the eigenvalues $\lambda_i$ of $F$ are distinct. Since (9.2) is completely controllable, this assertion, if true, would imply that (9.1) is also completely controllable, which is false. In fact, the system defined by

$$(9.3) \qquad F = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \qquad g = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is obviously not equivalent to

$$(9.4) \qquad F = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \qquad g = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

whenever $\lambda \neq \mu$.

In examining the derivation originally given by Lur'e for his canonical form [27; Chapter 1, §2–3], it is clear that the last step before equation (3.5) is valid if and only if det $[\mathrm{H}_k(\lambda_p)] \neq 0$ (in the notation of Lur'e [27].) It is easy to show that this condition is equivalent to complete controllability, whenever the eigenvalues of $F$ are distinct. Unfortunately, the condition det $[H_k(\lambda_p)] \neq 0$ was not emphasized explicitly by Lur'e [28] in the original publications.

We may thus conclude that *when $F$ has distinct eigenvalues and there is a* *single control variable, the Lur'e-Letov canonical form exists if and only if the pair $\{F, g\}$ is completely controllable.*

It is interesting to note that (9.3) can be transformed into (9.4) when $\lambda = \mu$; in other words, when the eigenvalues are not distinct the Lur'e canonical form may exist even if the system is not completely controllable.

*Cancellations in the transfer-function.* When a mathematical model is derived from physical principles, the equations of the system are in or near the form (2.1–2). Regrettably, it has become widespread practice in system engineering to dispense with differential equations and to replace them by transfer functions $Z(s)$. Later, $Z(s)$ must be converted back into the form (2.1–2) for purposes of analog computation. In the process of algebraic manipulations, some transfer functions may have (exactly or very nearly) common factors in the numerator and denominator, which are then *canceled*. This is an indication that a part of the dynamics of the system is not represented by the transfer function.

Such cancellations are the basic idea of some elementary design methods in control theory. These methods do not bring the system under better control but merely "decouple" some of the undesirable dynamics. But then the closed-loop transfer function is no longer a faithful representation of the (closed-loop) dynamics. Stability difficulties may arise. Similar criticisms may be leveled against the large, but superficial, literature on "noninteracting" control system design.

EXAMPLE 8. Consider the system defined by the matrices

$$(9.5)\qquad F = \begin{bmatrix} 0 & 1 & 0 \\ 5 & 0 & 2 \\ -2 & 0 & -2 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix}, \qquad H = [-2 \quad 1 \quad 0].$$

The transfer function relating $y_1$ to $u_1$ is the sum of two terms:

$$\begin{aligned} \frac{y_1(s)}{u_1(s)} &= -2\frac{x_1(s)}{u_1(s)} + \frac{x_2(s)}{u_1(s)} \\ (9.6)\qquad &= -\frac{2}{s^3 + 2s^2 - 5s - 6} + \frac{s}{s^3 + 2s^2 - 5s - 6} \\ &= \frac{(s-2)}{(s+1)(s-2)(s+3)} = \frac{1}{(s+1)(s+3)}. \end{aligned}$$

Thus, by cancellation, the transfer function is reduced from the third to the second order. The system has an unstable "natural mode" (corresponding to $s_3 = 2$) about which the transfer functions gives no information.

Using (6.5) we see that the system (9.5) is completely controllable. By Theorem 5, the system cannot be completely observable: $n_B = 2$ from (9.6) and Case 1, section 8. The canonical structure consists of Parts (A) and (B). In canonical coördinates the system matrices can be taken as

$$\bar{F} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \bar{G} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \bar{H} = [0.5 \quad -0.5 \quad 0].$$

We can easily calculate the change of coördinates

$$\bar{x} = Tx$$

by the method of partial fractions discussed in [8]. First we find $T^{-1}$, then $T$. The results are

$$T^{-1} = \frac{1}{30}\begin{bmatrix} -5 & 3 & 2 \\ 5 & -9 & 4 \\ 10 & 6 & -1 \end{bmatrix}, \qquad T = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 2 \\ 8 & 4 & 2 \end{bmatrix}.$$

*Loss of controllability and observability due to sampling.* Consider a single-input/single-output constant linear system. Suppose the output is observed only at the instants $t = kT$ ($k$ = integer, $T > 0$), and that the input is constant over the intervals $kT \leqq t < (k + 1)T$. This situation is commonly called "sampling"; it arises when a digital computer is used in control or data processing. $T$ is the *sampling period.* We can regard such a setup as a discrete-time dynamical system. We define here $\Theta$ (Axiom ($D_1$)) as the set of integers and replace (2.1) by a difference equation. All theorems carry over to this situation with small modifications.

The analysis of discrete-time systems by conventional techniques requires the computation of the so-called $z$-transform of $Z(s)$ [22]. The analysis using $z$-transforms then proceeds in close analogy with analysis based on Laplace transforms.

A constant linear system which is completely controllable and completely observable will retain these properties even after the introduction of sampling if and only if [4]

$$\text{(9.7)} \qquad \operatorname{Re} s_i = \operatorname{Re} s_j \quad \text{implies} \quad \operatorname{Im}(s_i - s_j) \neq q\pi/T$$

where $i, j = 1, \cdots, n$ and $q$ = positive integer.

If this condition is violated (the sampling process "resonates" with the system dynamics) then cancellations will take place in the $z$-transform. The $z$-transform will then no longer afford a faithful representation of the system, so that if (9.7) *is violated, results based on formal manipulation of z-transforms may be invalid.*

This point is not at all clear in the literature. True, Barker [23] has drawn attention to a related phenomenon and called it "hidden oscillation." The textbooks, however, dismiss the problem without providing real insight [22, §5-3; 24, §2.13].

A practical difficulty arises from the fact that near the "resonance" point given by (9.7) it is hard to identify the dynamical equations accurately from the $z$-transform. Small numerical errors in the computation of the $z$-transform may have a large effect on the parameters of the dynamical equations.

## REFERENCES

[1] R. E. Kalman, *Discussion of paper by I. Flügge-Lotz*, Proc. 1st International Conference on Automatic Control, Moscow, 1960; Butterworths, London, 1961, Vol. 1, pp. 396–7.

[2] R. E. Kalman, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. USA, 48 (1962), pp. 596–600.

[3] R. E. Kalman, *On the general theory of control systems*, Proc. 1st International Congress on Automatic Control, Moscow, 1960; Butterworths, London, 1961, Vol. 1, pp. 481–492.

[4] R. E. Kalman, Y. C. Ho, and K. S. Narendra, *Controllability of linear dynamical systems*, (to appear in Contributions to Differential Equations, Vol. 1, John Wiley, New York.)

[5] E. G. Gilbert, *Controllability and observability in multivariable control systems*, J. Soc. Indust. Appl. Math. Ser. A: On Control, Vol. 1, No. 2 (1963), pp. 128–151.

[6] V. V. Nemitskii and V. V. Stepanov, *Qualitative Theory Of Differential Equations*, Princeton Univ. Press, Princeton, 1960.

[7] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the 'second method' of Lyapunov*, J. Basic Engr. (Trans. A.S.M.E.), 82 D (1960), pp. 371–393.

[8] R. E. Kalman, *Analysis and design principles of second and higher-order saturating servomechanisms*, Trans. Amer. Inst. Elect. Engrs., 74, II (1955), pp. 294–310.

[9] W. Hahn, *Theorie und Anwendung der direkten Methode von Ljapunov*, Springer, Berlin, 1959.

[10] J. P. LaSalle and S. Lefschetz, *Stability By Lyapunov's Direct Method*, Academic Press, New York, 1961.

[11] L. Markus, *Continuous matrices and the stability of differential systems*, Math. Z., 62 (1955), pp. 310–319.

[12] L. A. Zadeh, *A general theory of linear signal transmission systems*, J. Franklin Inst., 253 (1952), pp. 293–312.

[13] D. Middleton, *An Introduction To Statistical Communication Theory*, McGraw-Hill, New York, 1960.

[14] R. E. Kalman, *On controllability, observability, and identifiability of linear dynamical systems*, (to appear).

[15] R. E. Kalman, *On the stability of time-varying linear systems*, Trans. I.R.E. Prof. Gr. Circuit Theory, (CT-9 (1962), pp. 420–422.).

[16] S. J. Mason, *Feedback theory: some properties of signal flow graphs*, Proc. I.R.E., 41 (1953), pp. 1144–56; *Further properties of signal flow graphs*, ibid., 44 (1956), pp. 920–926.

[17] R. E. Kalman, *New results in filtering and prediction theory*, RIAS Report 61-1, Research Institute for Advanced Studies (RIAS), Baltimore, 1961.

[18] A. M. Letov, *Stability In Nonlinear Control Systems*, Princeton Univ. Press, Princeton, 1961.

[19] B. L. van der Waerden, *Modern Algebra*, Vol. 1, 2nd Ed., Ungar, New York, 1949.

[20] A. M. Batkov, *On the problem of synthesis of linear dynamic systems with two parameters*, Avtomat. i Telemeh., 19 (1958), pp. 49–54.

[21] J. H. Laning, Jr. and R. H. Battin, *Random Processes In Automatic Control*, McGraw-Hill, New York, 1956.

[22] J. R. Ragazzini and G. F. Franklin, *Sampled-Data Control Systems*, McGraw-Hill, New York, 1958.

[23] R. H. Barker, *The pulse transfer function and its application to sampling servo systems*, Proc. Inst. Elec. Engrs. 99 IV (1952), pp. 302–317.

[24] E. I. Jury, *Sampled-Data Control Systems*, John Wiley, New York, 1957.

[25] R. E. Kalman, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. USA, 49, (1963), pp. 201–205.

[26] E. F. Moore, *Gedanken-experiments on sequential machines*, *Automata Studies*, Princeton Univ. Press, Princeton, 1956.

[27] A. I. Lur'e, *Certain Nonlinear Problems in the Theory of Automatic Control.* (in Russian), Gostekhizdat, Moscow, 1951; German translation Akademie-Verlág, Berlin, 1957.

[28] Private communication, Academician A. I. Lur'e.

## APPENDIX

**Factorization of rectangular matrices.** Given an arbitrary, real, $p \times m$ matrix $R$ of rank $q \leqq \min(m, p)$. We wish to find a $p \times q$ matrix $H$ and a $q \times m$ matrix $G$, both of rank $q$, such that $R = HG$. The existence of $H$ and $G$ follows almost immediately from the definition of rank. We describe below a constructive procedure for determining $H$ and $G$ numerically from numerical values of $R$.

Let $p \leqq m$. Form the $p \times p$ matrix $S = RR'$.

As is well known, there exists a nonsingular matrix $T$ such that

$$\text{(A-1)} \qquad TRR'T' = TST' = E,$$

where precisely $q$ diagonal elements of $E$ are 1, all other elements are 0. $T$ can be calculated by steps similar to the gaussian elimination procedure.

Compute the generalized inverse $R^{\#}$ (in the sense of Penrose [4]) of $R$. $R^{\#}$ is an $m \times p$ matrix.

Using the properties of $R^{\#}$ ([4]) we obtain

$$\text{(A-2)} \quad R = RR^{\#}R = RR'R^{\#\prime} = SR^{\#\prime} = T^{-1}ET^{-1\prime}R^{\#\prime} = (T^{-1}E)(T^{-1}E)'R^{\#\prime}.$$

Now $T^{-1}E$ is a matrix which contains precisely $p - q$ zero columns. Deleting these columns, we obtain a $p \times q$ matrix $(T^{-1}E)^0 = H$. Similarly, deleting $p - q$ zero rows from $(T^{-1}E)'R^{\#\prime} = (R^{\#}T^{-1}E)'$ we obtain a $m \times q$ matrix $G' = (R^{\#}T^{-1}E)^0$. Evidently $R = HG$. Since the ranks of $H$ and $G$ are obviously less than or equal to $q$, both ranks must be exactly $q$ for otherwise *rank* $R < q$, contrary to hypothesis.

Alternately, let $T$, $U$ be nonsingular matrices such that

$$TRU = E;$$

then

$$R = (T^{-1}E)^0(EU^{-1})^0 \tag{A-3}$$

is the desired decomposition. However, the computation of (A-3) may require more steps than that of (A-2).

Suppose now that $R$ is complex. Then $S = R\bar{R}' = RR^* = A + iB$ is complex hermitian; it corresponds to the $2n \times 2n$ nonnegative matrix

$$\Sigma = \begin{bmatrix} A & B \\ -B & A \end{bmatrix} \tag{A-4}$$

where $A = A'$ and $B = -B'$. In fact, if $z = x + iy$, the hermitian form $z^*RR^*z$ (which is real-valued) is equal to the quadratic form

$$[x \quad y]\begin{bmatrix} A & B \\ -B & A \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

As is well known, there exists a nonsingular *complex* matrix $T$ such that $TST^* = E$. If $T = U + iV$, it follows further that

$$\begin{bmatrix} U & V \\ -V & U \end{bmatrix}\begin{bmatrix} A & B \\ -B & A \end{bmatrix}\begin{bmatrix} U' & -V' \\ V' & U' \end{bmatrix} = \begin{bmatrix} E & 0 \\ 0 & E \end{bmatrix}.$$

Hence the determination of the complex $n \times n$ matrix has been reduced to the determination of a real $2p \times 2p$ matrix. Similar remarks apply to the calculation of $R^\#$. Thus the problem of factoring complex $p \times m$ matrices can be embedded in the problem of factoring real $2p \times 2m$ matrices.