KTH Mathematics

AN INTRODUCTION TO MATHEMATICAL SYSTEMS THEORY

Lecture notes by

Anders Lindquist and Janne Sand

revised by Xiaoming Hu

Optimization and Systems Theory
Royal institute of technology
SE-100 44 Stockholm, Sweden

Contents

Preface

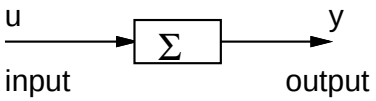
This is a set of lecture notes for the course "Mathematical Systems Theory" given at Kungliga Tekniska Högskolan, Stockholm. The compendium is initially written by Anders Lindquist and Janne Sand, and is partially revised by Xiaoming Hu in 2010. Some minor updates are made constantly.

CHAPTER 1

Linear Control Systems

1.1. Introduction

"System" as defined in the Webster is "a regularly interacting or interdependent group of items forming a unified whole". In this course the class of systems we study typically has some input signals and output signals as shown below.



When an input u is applied to the system Σ, we assume that a *unique* output y is produced. Note that the uniqueness of the output is an essential property of the systems we will study. A system is called a single-input single-output (SISO) system if both the input and the output are scalars. Otherwise it is called a multi-input multi-output (MIMO) system.

Now let us assume that there is a given set T of times which is either continuous ($T \subset \mathbb{R}$ and the system is called a *continuous-time system* accordingly) or discrete ($T \subset \mathbb{Z}$ and the system is called a *discrete-time system* accordingly), u and y are vector-valued functions of time belonging to vector spaces $\mathcal{U}$ and $\mathcal{Y}$, and the box that defines the system is a function $f_\Sigma : \mathcal{U} \to \mathcal{Y}$.

A system is called *memoryless* if its output signal at time t depends only on the input signal applied at time t. However, most practical engineering systems do have memory. Thus if an input u is applied to the system over $[t, \infty]$, the output y is in general not determinable unless we know the input applied before t. As we will see later, the key properties of the system will be the same (provided it is a linear system!) no matter what past input was applied to the system. Thus in the input-output modeling of a system, we assume that the system is *relaxed* before any input is applied, namely $f_\Sigma(0) = 0$.

The system is called a *linear* system if

$$f_\Sigma(\alpha u_1 + \beta u_2) = \alpha f_\Sigma(u_1) + \beta f_\Sigma(u_2)$$

for all $u_1, u_2 \in \mathcal{U}$ and $\alpha, \beta \in \mathbb{R}$ (superposition), and *causal* if

$$u_1(t) = u_2(t) \quad \text{for } t < t_1 \quad \Longrightarrow \quad f_\Sigma(u_1) = f_\Sigma(u_2) \quad \text{for } t < t_1,$$

namely the current output depends only on the past and current input, but not the future input.

1.2. Input-output Description of a System

A suitable model for continuous-time linear systems is

$$y(t) = \int_{t_0}^{t} G(t,s)u(s)ds + D(t)u(t) \tag{1}$$

where the functions G and D take values in the space $\mathbb{R}^{m\times k}$ of (real) $m \times k$ matrices. The set T of times could be one of the intervals $[t_0, t_1]$, $[t_0, \infty)$, or, setting $t_0 = -\infty$, $(-\infty, t_1]$, or $(-\infty, \infty)$.

Since $G(t,s)$ is the output response to the impulse input, it is called the *impulse response*. Obviously being causal implies that $G(t,s) = 0,\ \forall t < s$ and being memoryless implies that $G(t,s) \equiv 0$.

A linear system Σ is called *time invariant* if

$$y = f_\Sigma(u) \quad \Longrightarrow \quad y_T = f_\Sigma(u_T)$$

where w_T denotes the shift function: $w_T(t) = w(t-T),\ t \geq t_0 + T$ and $w_T(t) = 0,\ t < t_0 + T$. In this case,

$$\begin{aligned} f_\Sigma(u_T) &= \int_{t_0}^{t} G(t,s)u_T(s)ds + D(t)u(t-T) \\ &= \int_{t_0-T}^{t-T} G(t,r+T)u_T(r+T)dr + D(t)u(t-T) \\ &= \int_{t_0}^{t-T} G(t,r+T)u(r)dr + D(t)u(t-T) \end{aligned} \tag{2}$$

Comparing this with (1), time invariance requires that

$$G(t,s+T) = G(t-T,s),\ D(t) = D(t-T),\ \forall T \geq 0.$$

It is easy to see that this will be satisfied if

$$G(t,s) = G(t-s,0)$$

and D is constant.

It is known in the literature that in order to obtain a finite dimensional internal description of the black box, which shall be our next topic, the function G must be factorizable as

$$G(t,s) = H(t)K(s) \tag{3}$$

where H and K take values in $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{n\times k}$ respectively for some integer n. One problem with this description is that it does not allow for modeling of so-called *autonomous dynamics* producing a nonzero output even when the input is identically zero, which would be easily taken care by the state space description as well.

The above discussion can be easily adopted to discrete-time linear systems which can be modeled as

$$y(t) = \sum_{s=t_0}^{t} G(t,s)u(s) + D(t)u(t) \tag{4}$$

1.3. State Space Description of a System

The basic concept of systems theory is that of *state*. Loosely speaking, at each time t, we would like to collect all the *current* information that is relevant for future outputs and is represented by a so-called state vector

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

If at any time t $x(t)$ together with the current input $u(t)$ uniquely determines the output $y(t)$, the time evolution of x defines an *internal* description of the black box. If this can be done so that the number n of state variables, $x_1, x_2, \ldots, x_n$, is finite, we say that the system is finite-dimensional, or lumped. Condition (3) is precisely what is required for this. In fact, as we shall see in Chapter 5, condition (3) allows us to describe the system Σ in *state space form*, i.e. as

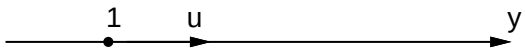
$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{5}$$

in continuous time, and

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{6}$$

in discrete time. Here A, B, C, D take values in $\mathbb{R}^{n\times n}$, $\mathbb{R}^{n\times k}$, $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{m\times k}$ respectively, and it is easy to see that the system Σ is time invariant if and only if A,B, C and D are constant matrices. The problem to determine (5) or (6) from (1) and (4) respectively is the so-called *realization problem* to be discussed in Chapter 5.

Example 1.3.1. A particle of mass 1 moves along the y-axis under influence of the force $u(t)$ for $t \geq 0$.



Newton's law yields

$$\ddot{y} = u$$

so the external description (1) of the system with input u and output y is

$$y(t) = \int_0^t (t-s)u(s)ds$$

i.e. $G(t,s) = t - s$. The system is time invariant. To obtain an internal description define the state

$$x = \begin{bmatrix} y \\ \dot{y} \end{bmatrix}$$

Then $\dot{x}_1 = x_2$ and $\dot{x}_2 = u$, i.e.

$$\begin{cases} \dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ y = (1,0)x \end{cases}$$

However, we could also have chosen the state

$$x = \begin{bmatrix} y + \dot{y} \\ y - \dot{y} \end{bmatrix}$$

yielding the system

$$\begin{cases} \dot{x} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} x + \begin{bmatrix} 1 \\ -1 \end{bmatrix} u \\ y = (1/2, 1/2)x \end{cases}$$

or in infinitely many other ways. The relation between these internal descriptions is merely a simple change of coordinates in the state space $X = \mathbb{R}^2$. □

EXAMPLE 1.3.2. Describe the electrical network in the figure as a linear control system with the voltage u as an input and current y as an output.

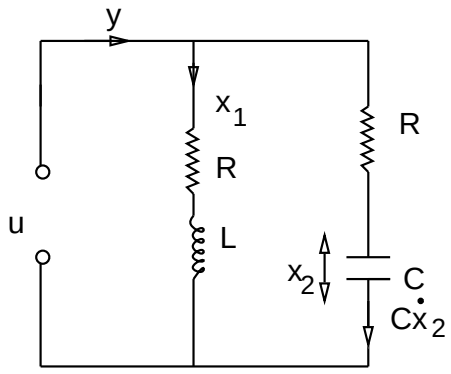
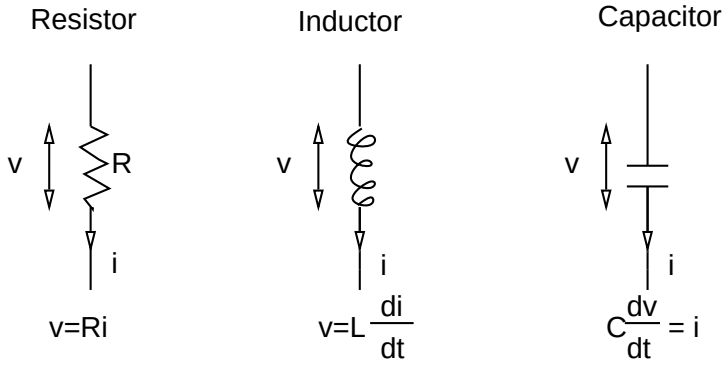


Figure 1.1

From the theory of electricity we know that the elements in Figure 1.1 produce the dynamical behaviors of the following chart



Therefore, choosing state variables as in Figure 1.1, we obtain the state space description

$$\begin{cases} \dot{x} = \begin{bmatrix} -\frac{R}{L} & 0 \\ 0 & -\frac{1}{RC} \end{bmatrix} x + \begin{bmatrix} \frac{1}{L} \\ \frac{1}{RC} \end{bmatrix} u \\ y = (1, -\frac{1}{R})x + \frac{1}{R}u \end{cases}$$

for the system. Clearly the system is time invariant. Let us, for simplicity and without regard to physical reasonability, set $R = L = C = 1$. Then, the system becomes

$$\begin{cases} \dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \\ y = (1, -1)x + u \end{cases} \tag{7}$$

that is,

$$x_1(t) = x_2(t) = \int_{-\infty}^{t} e^{-(t-s)} u(s) ds$$

and

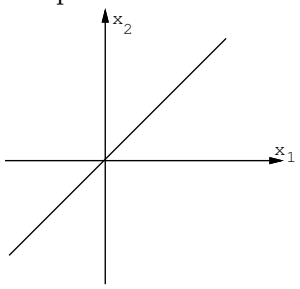
$$y = x_1 - x_2 + u = u.$$

Consequently, $G(t, s) \equiv 0$ and $D = 1$, and hence the two-dimensional external description (7) can be replaced by the zero-dimensional system

$$y = u$$

A natural question then is: what is wrong with the representation (7)? The answer is that the dynamics is confined to the subspace

$$M = \{x \in \mathbb{R}^2 \,|\, x_1 = x_2\}$$

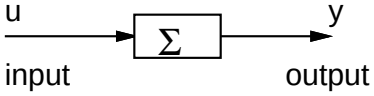


of the state space $X = \mathbb{R}^2$. We say that Σ is not *completely reachability*. Similarly, we cannot observe what is going on in the subspace M, i.e. Σ is not *completely observable*. With a terminology to be established in Chapter 4, we say that M is both the reachable and unobservable subspace of X. □

1.4. The Concept of State in Mathematical Systems Theory*

To spread some further light on the abstract concept of state, let us consider the following experiment.

Kalman's experiment: Control the system



up to time t (after which the control u is identically zero) and then observe the resulting output after time t.

For the purpose of this discussion, define U and Y to be the (possibly infinite-dimensional) vector spaces of the inputs respectively the outputs admissible in this experiment. Then Kalman's experiment defines a linear mapping

$$F : U \to Y.$$

The basic idea of the state space construction is to determine a *state space* X, preferably of finite dimension, over which F can be factored as in the commutative diagram

$$\begin{array}{ccc} U & \xrightarrow{F} & Y \\ & \mathcal{R} \searrow \quad \nearrow \mathcal{O} & \\ & X & \end{array} \tag{8}$$

In other words, find a state space X and two maps $\mathcal{R} : U \to X$ and $\mathcal{O} : X \to Y$ such that

$$F = \mathcal{O}\mathcal{R}.$$

If, for example, we choose $X := U$ so that $\mathcal{R} = I$ (identity) and $\mathcal{O} = F$, we have a valid factorization, but it is of little use, since we obtain no data reduction. Indeed, we need to remember the whole past history of the input. The state space may not even be finite-dimensional.

Therefore, let us reduce the state space by identifying inputs which produce the same output. Hence, we say that $u_1, u_2 \in U$ are equivalent ($u_1 \sim u_2$) if $F(u_1) = F(u_2)$. Then define X to be the set of all equivalence classes under this equivalence. In algebraic terms this is a quotient space denoted $X = U/\sim$, and, in the present context, a vector space. Using this X as a state space it is clear that the factorization (8) has the following properties.

(1) $\mathcal{R}$ is *surjective* (onto), i.e. $\operatorname{Im} \mathcal{R} = X$
(2) $\mathcal{O}$ is *injective* (one-one), i.e. $\mathcal{O}x_1 = \mathcal{O}x_2 \implies x_1 = x_2$, or equivalently, $\ker \mathcal{O} = 0$.

REMARK 1.4.1. We recall that for a linear map $A : X \to Y$, the *range space*(or image) $\operatorname{Im} A$ is defined as $\{Ax \,|\, x \in X\}$ and the *kernel* $\ker A$ as $\{x \in X \,|\, Ax = 0\}$. □

A state space representation is said to be *completely reachable* if property (1) holds and *completely observable* if property (2) holds.

It is not hard to convince oneself that such a construction leads to a state space which is *minimal* in the sense of subspace inclusion, i.e. X contains no proper subspace which can also serve as a state space. This fact can be illustrated (in Venn diagram form) by Figure 1.2.

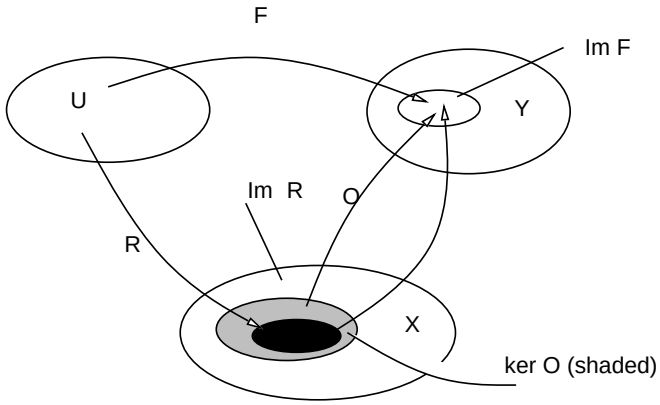


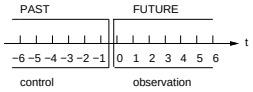
Figure 1.2

In Figure 1.2 we discard from X the part which is not in Im $\mathcal{R}$; it serves no purpose, since it corresponds to unreachable states. Next, from the remaining state space we discard the unobservable part ker $\mathcal{O}$, choosing as the final state space only its complement, the black part of X. (Beware of the fact that the Venn diagram representation is just an illustration and could be misleading! How?) The remaining X satisfies both property (1) and property (2). Compare this with the discussion in Example 1.3.2.

Next, let us illustrate this by means of a time invariant system in discrete time, namely

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{cases}$$

Time invariance implies that the state space construction is identical for all t. Hence, choose the last time of control to be $t = -1$ and first time of observation to be $t = 0$:



Then, the map $F : U \to Y$ can be describe as

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \end{bmatrix} = \begin{bmatrix} CB & CAB & CA^2B & \cdots \\ CAB & CA^2B & CA^3B & \cdots \\ CA^2B & CA^3B & CA^4B & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} u(-1) \\ u(-2) \\ u(-3) \\ \vdots \end{bmatrix}$$

i.e. F has a matrix representation which is a block *Hankel* matrix. Now since

$$\begin{bmatrix} CB & CAB & CA^2B & \cdots \\ CAB & CA^2B & CA^3B & \cdots \\ CA^2B & CA^3B & CA^4B & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix}$$

which we can identify as the factorization

$$F = \mathcal{O}\mathcal{R}.$$

In fact,

$$x(0) = \mathcal{R} \begin{bmatrix} u(-1) \\ u(-2) \\ u(-3) \\ \vdots \end{bmatrix}.$$

In order to insure that $X = \mathbb{R}^n$ is minimal, we should choose

$$\mathcal{R} = [B, AB, A^2B, \cdots]$$

surjective and

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix}$$

injective. This amounts to requiring that both $\mathcal{R}$ and $\mathcal{O}$ are full rank. This whole matter will be discussed in detail in Chapter 3 and Chapter 5. The reader is encouraged to return to the discussion above while studying Chapter 3 and Chapter 5 to reinforce his/her understanding of the concept of state.

CHAPTER 2

Linear Dynamical Systems

2.1. Solutions of linear differential equations

Let us consider an n-dimensional linear time-varying control system:

$$\begin{aligned} \dot{x} &= A(t)x(t) + B(t)u(t) \\ y &= C(t)x(t) + D(t)u(t), \end{aligned} \tag{9}$$

where x, u, y are the state, input and output respectively, and A, B, C, D are respectively $n \times n, n \times m, p \times n, p \times m$ matrices whose entries are continuous functions of time t over $(-\infty, \infty)$.

Consider first the homogeneous system

$$\dot{x}(t) = A(t)x(t) \tag{10}$$

One can show that (10) has a unique solution for every initial state $x(t_0) = a$ on every bounded interval containing t_0 and for all $a \in \mathbb{R}^n$. Then the set of solutions forms a linear space and the dimension of the space is n.

PROPOSITION 2.1.1. *The set of all solutions of* (10) *is a linear space of dimension* n *over the real field.*

Proof: Let $\Phi_k(t, t_0)$, $k = 1, 2, \ldots, n$, be the unique solutions of (10) with linearly independent initial conditions

$$x(t_0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, x(t_0) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, x(t_0) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

If we can show every solution of (10) can be expressed as a linear combination of $\Phi_k(t, t_0)$, then the proposition is proven. Define the $n \times n$ transition matrix

$$\Phi(t, t_0) = [\Phi_1(t, t_0), \Phi_2(t, t_)), \ldots, \Phi_n(t, t_0)].$$

We claim that $\Phi(t, t_0)$ is nonsingular for all t. This can be easily seen by the fact that if $\Phi(t, t_0)$ is singular at some time t_1, it would remain singular for all finite t. Otherwise the uniqueness property of the solution (with respect to a given initial state) will not hold.

Then we have

$$\begin{cases} \frac{\partial \Phi}{\partial t}(t, s) = A(t)\Phi(t, s) \\ \Phi(s, s) = I \end{cases}.$$

Since

$$a = a_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + a_n \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

by the superposition principle, the system (10) with $x(t_0) = a$ has the solution

$$x(t) = a_1\Phi_1(t, t_0) + a_2\Phi_2(t, t_0) + \cdots + a_n\Phi_n(t, t_0)$$

that is

$$x(t) = \Phi(t, t_0)a.$$

□

In the above derivation we have chosen some special initial states to obtain $\Phi(t, t_0)$. Now suppose $\Psi(t)$ be a matrix whose columns are linearly independent solutions to (10). Then the solution $x(t)$ with $x(t_0) = a$ can be expressed

$$x(t) = \Psi(t)w,$$

and $a = \Psi(t_0)w$. Thus $w = \Psi^{-1}(t_0)a$, and

$$x(t) = \Psi(t)\Psi^{-1}(t_0)a,$$

which implies $\Phi(t, t_0) = \Psi(t)\Psi^{-1}(t_0)$.

PROPOSITION 2.1.2. *Let Ψ be an arbitrary $n \times n$ matrix solution of*

$$\dot{\Psi}(t) = A(t)\Psi(t) \quad ; \quad \Psi(t_0) = C$$

where C is nonsingular, then $\Psi(t)$ is nonsingular for all t and

$$\Phi(t, s) = \Psi(t)\Psi(s)^{-1}.$$

Such a $\Psi(t)$ is called a fundamental matrix*, and $\Phi(t, s)$ is called the* state transition matrix.

Let us list some properties of the transition matrix Φ:

(1) $\Phi(t, s) = \Phi(t, \tau)\Phi(\tau, s)$ for all (t, s, τ) as illustrated by the commutative diagram

$$\begin{array}{ccc} X & \overset{\Phi(\tau,s)}{\longrightarrow} & X \\ \Phi(t,s) \searrow & & \downarrow \Phi(t,\tau) \\ & & X \end{array}$$

Proof: Consider the unique solution of

$$\begin{cases} \dot{x}(\sigma) = A(\sigma)x(\sigma) \\ x(s) = a \end{cases}$$

Then, as illustrated in diagram $x(t) = \Phi(t,s)a, \quad x(t) = \Phi(t,\tau)x(\tau)$ and $x(\tau) = \Phi(\tau,s)a$, and hence

$$\Phi(t,\tau)\Phi(\tau,s)a = \Phi(t,s)a.$$

that is

$$[\Phi(t,\tau)\Phi(\tau,s) - \Phi(t,s)]a = 0$$

Since this must hold for all $a \in \mathbb{R}^n$ property (1) follows. □

(2) $\Phi(t,s)$ is nonsingular, and $\Phi(t,s)^{-1} = \Phi(s,t)$.

Proof: This follows immediately from

$$\Phi(t,s)\Phi(s,t) = I \tag{11}$$

which is a consequence of property (1). □

(3) $\frac{\partial\Phi}{\partial s}(t,s) = -\Phi(t,s)A(s)$.

Proof: Differentiating (11) with respect to s yields that

$$\frac{\partial\Phi}{\partial s}(t,s)\Phi(s,t) + \Phi(t,s)A(s)\Phi(s,t) = 0$$

Since $\Phi(s,t)$ is nonsingular, property (3) follows. □

Let us next consider the solution of the control system

$$\begin{cases}\dot{x}(t) = A(t)x(t) + B(t)u(t)\\ x(t_0) = a\end{cases}.$$

Set $z(t) := \Phi(t_0,t)x(t)$, i.e $x(t) = \Phi(t,t_0)z(t)$. Then,

$$\dot{z} = -\Phi(t_0,t)A(t)x(t) + \Phi(t_0,t)\dot{x}(t)$$

$$\text{so that } \begin{cases}\dot{z} = \Phi(t_0,t)B(t)u(t)\\ z(t_0) = a\end{cases}$$

and therefore,

$$z(t) = a + \int_{t_0}^{t} \Phi(t_0,s)B(s)u(s)ds,$$

or, premultiplying by $\Phi(t,t_0)$,

$$x(t) = \Phi(t,t_0)a + \int_{t_0}^{t} \Phi(t,s)B(s)u(s)ds.$$

Notice that this equality also holds for $t \leq t_0$.

2.2. Time-invariant systems

Now consider the time-invariant case:

$$\dot{x} = Ax,$$

where is A is a constant matrix.

For time-invariant systems determining the transition matrix function becomes much simpler in that it can be expressed in terms of the matrix exponential.

DEFINITION 2.2.1. $e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$.

Note that, since $\sum_{k=0}^{N} \frac{|t|^k}{k!} \|A\|^k \leq e^{\|A\||t|} < \infty$ for any finite t, the sum converges.

We collect some properties of matrix exponentials. The proofs are left for readers as exercises.

(1) If $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $e^D = \operatorname{diag}(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n})$;
(2) $e^{P^{-1}AP} = P^{-1}e^A P$;
(3) If $AB = BA$, then, $e^A e^B = e^{A+B}$; *Warning:* In general, $e^A e^B \neq e^{A+B}$.
(4) $(e^A)^{-1} = e^{-A}$;
(5) $\frac{d}{dt} e^{At} = Ae^{At} = e^{At}A$;

It follows from property (5) above that

$$\Psi(t) = e^{At}$$

is a fundamental matrix of the time-invariant system

$$\dot{x} = Ax$$

where A is constant. Then, by Proposition 2.1.2

$$\Phi(t, s) = \Psi(t)\Psi(s)^{-1} = e^{At}(e^{As})^{-1} = e^{At}e^{-As},$$

that is

$$\Phi(t, s) = e^{A(t-s)} \tag{12}$$

Therefore the solution of the time-invariant system

$$\dot{x} = Ax + Bu$$

becomes

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds.$$

REMARK 2.2.2. One might ask whether (12) can be generalized to the time-varying case, i.e. is it true that

$$\Phi(t,s) = \exp\{\int_s^t A(\tau)d\tau\} \tag{13}$$

when $A(t)$ is time varying? The answer is that a sufficient condition for (13) to hold is that $A(t)$ and $\int_s^t A(\tau)d\tau$ commute. □

2.3. Systems of linear difference equations

The corresponding analysis for discrete-time system is quite analogous. In fact, considering the following discrete-time system

$$x(t+1) = A(t)x(t),$$

the transition matrix is generated by

$$\begin{aligned}\Phi(t+1,s) &= A(t)\Phi(t,s)\\ \Phi(t,t) &= I.\end{aligned}$$

It is not hard to see that

$$\Phi(t,s) = A(t-1)A(t-2)\cdots A(s) \quad \text{for } t > s$$

and $\Phi(t,s)$ is defined for $t < s$ only if $\Phi(s,t)$ is invertible, i.e. $A(k)^{-1}$ for $k = s, s+1, \ldots, t-1$. In this case, $\Phi(t,s) = \Phi(s,t)^{-1}$. In addition $\Phi(t,s)$ has the following properties.

(1) $\Phi(t,s) = \Phi(t,\tau)\Phi(\tau,s)$;
(2) $\Phi(t,s-1) = \Phi(t,s)A(s-1)$

Hence, the solution of the control system

$$x(t+1) = A(t)x(t) + B(t)u(t)$$

becomes

$$x(t) = \Phi(t,s)x(s) + \sum_{\sigma=s}^{t-1} \Phi(t,\sigma+1)B(\sigma)u(\sigma).$$

For time-invariant systems, the transfer matrix

$$\Phi(t,s) = A^{t-s}$$

is invertible if and only if A^{-1} exists. Note that e^{At} in continuous time corresponds to A^t in discrete time. Here lies one of the fundamental differences between the continuous-time and discrete-time setting. Indeed e^{At} is never singular, whereas A^t might be.

EXAMPLE 2.3.1 (Sampling a continuous-time system). Sometimes it is important to represent a continuous-time system by means of a discrete-time system, e.g. when implementing a control law on a computer. Let h be the

sampling time and integrate the time-invariant system $\dot{x} = Ax + Bu$ from kh to $kh + h$,

$$x(kh + h) = e^{Ah}x(kh) + \int_{kh}^{kh+h} e^{A(kh+h-s)}Bu(s)\,ds.$$

By restricting the input $u(t)$ to be a piecewise constant signal as

$$u(t) = v(k) \text{ for } t \in [kh, kh + h),\ k \in \mathbb{Z}$$

the integral above can be evaluated as

$$\int_{kh}^{kh+h} e^{A(kh+h-s)}Bu(s)ds = \int_{kh}^{kh+h} e^{A(kh+h-s)}ds\,Bv(k) = \int_{0}^{h} e^{As}ds\,Bv(k).$$

Hence, by letting $z(k) \triangleq x(kh)$ we get the discrete-time system

$$z(k + 1) = Fz(k) + Gv(k),$$

where $F = e^{Ah}$ and $G = \int_0^h e^{As}\,ds\,B$, having the same values as the original system in the sampling points $t = kh$.

As a numerical example, consider the one-dimensional motion of Example 1.3.1 with the state space description

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$

where x_1 is the position and x_2 is the velocity of the particle. Since A is nilpotent, e^{Ah} is easily calculated from the series expansion, which gives

$$F = e^{Ah} = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}.$$

To get G we integrate

$$\int_0^h e^{As}\,ds = \int_0^h \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} ds = \begin{bmatrix} h & \frac{h^2}{2} \\ 0 & h \end{bmatrix},$$

which gives $G = \begin{bmatrix} \frac{h^2}{2} \\ h \end{bmatrix}$. □

CHAPTER 3

Reachability and Observability

In this chapter we shall address two basic issues in systems theory. First, given a control system, it is natural to ask in what way we can control the state of the system, i.e., which states can be reached by choice of a suitable input? This question leads to the concept of reachability.

Second, given the input and the output of a system, can we determine its state? This question leads to the concept of observability.

3.1. Reachability

Consider the system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t); \quad x(t_0) = x_0, \tag{14}$$

where $x \in \mathbb{R}^n$, and $u \in \mathbb{R}^m$. A fundamental issue in systems and control is to understand under what conditions there is always a continuous (in time) input signal u that transfers the state from any given initial state $x(t_0) = x_0$ to any other final state $x(t_1) = x_1$? In the literature this property is called *completely controllable*. If we set $x_1 = 0$, it is called *(null-) controllable*, and it is called *reachable* if we set $x_0 = 0$. As we will show all these notions are equivalent for linear continuous time systems. Note that for nonlinear systems and discrete time systems, these notions are not equivalent anymore.

We shall answer the question by giving necessary and sufficient conditions in terms of the given data. Furthermore, when such a state transfer is possible we calculate the input signal $\hat{u}$ of minimum "energy" that achieves this state transfer.

If we solve the state equations explicitly we obtain

$$x_1 - \Phi(t_1, t_0)x_0 = \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)ds.$$

Let $d \triangleq x_1 - \Phi(t_1, t_0)x_0$ and let $\mathcal{U}$ be the space of input signals. If we define the mapping $L : \mathcal{U} \to \mathbb{R}^n$ as

$$Lu \triangleq \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)ds, \tag{15}$$

we see that the desired state transfer is possible if and only if $d \in \text{Im } L$. Furthermore, if either x_0 or x_1 is arbitrary, d will span the whole $\mathbb{R}^n$. From this we can also see that why those different notions on controllability are equivalent.

Thus, the mathematical problem for reachability is to study under what conditions the equation $Lu = d$ has a solution for any $d \in \mathbb{R}^n$, or the mapping L is onto $\mathbb{R}^n$.

It can be easily verified that L is a linear mapping, but since it does not act between two finite-dimensional vector spaces, L does not have a finite-dimensional matrix representation. As a consequence, the characterization of Im L is not immediate.

Recall that for a constant matrix P, the equation $Pu = d$ has a solution if $d \in$ Im P and P is onto if it has full row rank, namely its row vectors are linearly independent. From linear algebra we also know that Im $P =$ Im PP^T.

Since our L has the form $Lu = \int_{t_0}^{t_1} F(s)u(s)ds$, where $F(t)$ is an $n \times m$ time-varying matrix, it is necessary for us to extend the concept of linear independence to functions of a real variable.

DEFINITION 3.1.1. A set of functions $f_1(t), \cdots, f_N(t)$ is said to be linearly dependent on $[t_0, t_1]$ over the complex field if there exist complex numbers $c_1, \cdots, c_N$ that are not all zero, such that

$$c_1 f_1(t) + \cdots + c_N f_N(t) = 0, \ \forall t \in [t_0, t_1].$$

Otherwise, the functions are said to be *linearly independent* over $[t_0, t_1]$.

LEMMA 3.1.2. *A set of real functions $f_1(t), \cdots, f_N(t)$ is linearly independent on $[t_0, t_1]$ if and only if*

$$W(t_0, t_1) := \int_{t_0}^{t_1} F(s)F^T(s)ds$$

is nonsingular, where

$$F(t) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

Proof: We prove first the necessity by contradiction. Suppose $W(t_0, t_1)$ is singular, then there exists a row vector $\alpha \neq 0$ such that $\alpha W(t_0, t_1) = 0$. Thus $\alpha W(t_0, t_1)\alpha^T = 0$, namely

$$\alpha W(t_0, t_1)\alpha^T = \int_{t_0}^{t_1} (\alpha F(s))(\alpha F(s))^T ds = 0.$$

Since $(\alpha F(t))(\alpha F(t))^T$ is continuous and nonnegative, the above identity implies that $\alpha F(t) = 0$ on $[t_0, t_1]$, which contradicts with the linear independence assumption on the rows of $F(t)$.

Proof of the sufficiency is left as an exercise. □

LEMMA 3.1.3. *Consider* $Lu = \int_{t_0}^{t_1} F(s)u(s)ds$, *where*

$$F(t) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix},$$

each $f_i(t)$ *is* $1 \times m$ *matrix valued continuous function over* $[t_0, t_1]$.

The mapping L *is onto* $\mathbb{R}^n$ *if and only if the functions* $f_1(t), \cdots, f_n(t)$ *are linearly independent over* $[t_0, t_1]$.

Proof: Sufficiency: for any given $d \in \mathbb{R}^n$, one can easily verify that

$$u(t) = F^T(t)W^{-1}(t_0, t_1)d$$

solves the equation $Lu = d$.

Necessity: we prove it by contradiction. Suppose there exists a row vector $\alpha \neq 0$ such that $\alpha F(t) = 0$ on $[t_0, t_1]$. Since L is anyway onto, the following equation has a solution:

$$\int_{t_0}^{t_1} F(s)u(s)ds = \alpha^T.$$

Multiplying both sides by α we have $0 = \alpha\alpha^T$, which implies $\alpha = 0$. This is a contradiction. □

Now let us return to the discussion on reachability. For L defined by (15), we have $F(t) = \Phi(t_1, t)B(t)$.

DEFINITION 3.1.4. The *reachability Gramian* $W(t_0, t_1)$ is the matrix

$$W(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi(t_1, s)B(s)B^T(s)\Phi^T(t_1, s)ds.$$

REMARK 3.1.5. $\forall t_0, t_1$ such that $t_0 < t_1 W(t_0, t_1)$ is a symmetric, positive semidefinite matrix. □

We can now state and prove the main theorem of this section.

THEOREM 3.1.6. *Consider the system* (14) *with transition matrix* $\Phi(t, s)$.

(1) *The system is reachable if and only if the reachability Gramian is nonsingular. In this case the minimum energy control that transfers the state from* x_0 *to* x_1 *is*

$$\hat{u}(t) = B^T(t)\Phi^T(t_1, t)W^{-1}(t_0, t_1)[x_1 - \Phi(t_1, t_0)x_0].$$

(2) *If the reachability Gramian is singular, the state transfer from* $x(t_0) = x_0$ *to* $x(t_1) = x_1$ *is possible if and only if*

$$d \triangleq x_1 - \Phi(t_1, t_0)x_0 \in \text{Im } W(t_0, t_1).$$

and the minimum energy solution $\hat{u}$ *is given by*

$$\hat{u} = B^T(t)\Phi^T(t_1, t)a,$$

where a solves $W(t_0,t_1)a = d$ and the energy of u is defined as

$$\|u\| = (\int_{t_0}^{t_1} u^T(s)u(s)ds)^{\frac{1}{2}}.$$

Proof: The first part of the theorem is a direct application of Lemma 3.1.3.

Now we show the second part.

Sufficiency: Suppose first that $d \in \text{Im } W(t_0,t_1)$, then $d = W(t_0,t_1)a$ for some $a \in \mathbb{R}^n$. Let $u \triangleq B^T(t))\Phi^T(t_1,t)a$ and we get $Lu = W(t_0,t_1)a = d$.

Necessity: Suppose now that the state transfer is possible, i.e., $Lu = d$ for some $u(t)$. Furthermore, suppose that $d \notin \text{Im } W(t_0,t_1)$. We show that this will give a contradiction.

Since $rank\ W(t_0,t_1) < n$, there exists a row vector $\alpha \neq 0$, such that $\alpha d \neq 0$ and $\alpha x = 0\ \forall x \in \text{Im } W$. This implies

$$\int_{t_0}^{t_1} \alpha\Phi(t_1,s)B(s)B^T(s)\Phi^T(t_1,s)ds = 0.$$

Thus $\int_{t_0}^{t_1}(\alpha\Phi(t_1,s)B(s))(\alpha\Phi(t_1,s)B(s))^T ds = 0$. Therefore $\alpha Lu = 0$, which implies $\alpha d = 0$. This is a contradiction.

The final step is to prove the optimality of $\hat{u}$. Let u be any solution of $Lu = d$. Then $Lu = L\hat{u}$ so $L(u-\hat{u}) = 0$. This gives

$$0 = \int_{t_0}^{t_1}(a^T\Phi(t_1,s)B(s)(u-\hat{u})ds = \int_{t_0}^{t_1}(\hat{u})^T u ds - \|\hat{u}\|^2.$$

Hence, $\|\hat{u}\|^2 = (u,\hat{u})$. By using the Cauchy-Schwartz inequality we have

$$\|\hat{u}\|^2 \leq \|\hat{u}\|\,\|u\|,$$

which yields that $\|\hat{u}\| \leq \|u\|$. Hence, $\hat{u}$ is optimal. □

EXAMPLE 3.1.7. Consider the system $\dot{x} = Ax + bu$, with

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The system is time invariant and $\phi(t,s) = e^{A(t-s)}$. By series expansion, we get $e^{At} = e^{-t}I$ and the reachability Gramian is $W(t_0,t_1) = \int_{t_0}^{t_1} e^{-2(t_1-s)}ds \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. If $t_1 > t_0$ then the integral is strictly positive and hence

$$\text{Im } W(t_0,t_1) = \{\lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \lambda \in \mathbb{R}\}.$$

Can we control the system from $x_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to $x_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$? Form d as

$$d = x_1 - e^{-(t_1-t_0)}x_0 = \begin{bmatrix} 2 \\ 4 - e^{-(t_1-t_0)} \end{bmatrix}.$$

It is easily seen that $d \notin \text{Im } W(t_0,t_1)$ for every choice of $t_1 > t_0$ and the state transfer is not possible.

Can we control the system from $x_0 = \begin{bmatrix}0\\2\end{bmatrix}$ to $x_1 = \begin{bmatrix}2\\3\end{bmatrix}$? We now get

$$d = x_1 - e^{-(t_1-t_0)}x_0 = \begin{bmatrix}2\\3-2e^{-(t_1-t_0)}\end{bmatrix}.$$

For the particular choice $t_1 - t_0 = \ln 2$ we get $d = \begin{bmatrix}2\\2\end{bmatrix} \in \text{Im } W(t_0, t_1)$ and the state transfer is possible. □

3.2. Reachability for Time-Invariant Systems

In this section we shall consider the special case of time-invariant systems and characterize Im $W(t_0, t_1)$ in terms of the matrix pair (A, B). This characterization will be the first step towards a geometric theory for this class of systems.

We shall mainly analyze continuous-time systems, but almost all theorems obtained are valid for discrete-time systems as well.

DEFINITION 3.2.1. Let (A, B) be a matrix pair, where A is $n \times n$. The *reachability matrix* Γ is defined as

$$\Gamma \triangleq [B, AB, A^2B, .., A^{n-1}B].$$

REMARK 3.2.2. By the Cayley-Hamilton theorem A^{n+j}, $j \geq 0$ is a linear combination of A^j for $j = 0, 1, .., n-1$, which implies that Im Γ equals the image of the infinite matrix $[B, AB, A^2B, A^3B, ...]$. □

The reachability properties of a linear system, in the previous section characterized in terms of $W(t_0, t_1)$, can in the time-invariant case be displayed in a simpler form involving the pair (A, B). The following theorem gives a convenient algebraic characterization of Im $W(t_0, t_1)$.

THEOREM 3.2.3. *Let A be $n \times n$ and B be $n \times k$. Then, for all t_0, t_1 such that $t_0 < t_1$ we have*

$$\text{Im } W(t_0, t_1) = \text{Im } [B, AB, A^2B, .., A^{n-1}B].$$

Proof: Fix (t_0, t_1) and let $W = W(t_0, t_1)$. We first show that Im $\Gamma \subseteq$ Im W by showing that $\ker W^T \subseteq \ker \Gamma^T$. To this end, let $a \in \ker W$ which implies that $0 = a^T W a = \int_{t_0}^{t_1} a^T e^{A(t_1-s)} B B^T e^{A^T(t_1,s)} a \, ds$, i.e.,

$$\int_{t_0}^{t_1} |B^T e^{A^T(t_1-s)} a|^2 \, ds = 0.$$

Since the integrand is continuous and non-negative it follows that

$$B^T e^{A^T(t_1-s)} a \equiv 0 \quad \forall s \in [t_0, t_1],$$

i.e.,

$$\sum_{j=0}^{\infty} \frac{1}{j!}(t_1 - s)^j B^T (A^T)^j a \equiv 0.$$

This implies that $B^T(A^T)^j a = 0$ for all j. Consequently, $[B, AB, A^2B, .., A^{n-1}B]^T a = 0$, i.e., $a \in \ker \Gamma^T$. Hence, Im $\Gamma \subseteq$ Im W.

Conversely, suppose $a \in$ Im W. Then there is an $x \in \mathbb{R}^n$ such that $a = Wx$, and hence

$$a = \sum_{j=0}^{\infty} A^j B \int_{t_0}^{t_1} \frac{1}{j!}(t_1 - s)^j B^T e^{A^T(t_1-s)} x \, ds,$$

from which we see that $a \in \text{Im}\,[B, AB, A^2B, A^3B, ...]$. By Remark 3.2.2. this is equivalent to $a \in$ Im Γ. Hence, Im $W \subseteq$ Im Γ. □

REMARK 3.2.4. Since Im $\Gamma =$ Im $W(t_0, t_1)$ for any interval (t_0, t_1), we see that in the time-invariant case the image of the reachability Gramian is independent of the the interval (t_0, t_1). □

DEFINITION 3.2.5. Let n be the dimension of the state space. The pair (A, B) is said to be *completely reachable (controllable)* if Γ has full rank, i.e.,

$$\text{rank } \Gamma = n.$$

For an arbitrary system, not necessarily completely reachable, Im Γ is a subspace of the state space of great importance, and the following definition will be useful.

DEFINITION 3.2.6. The *reachable subspace* $\mathcal{R}$ is defined as

$$\mathcal{R} \triangleq \text{Im}\,[B, AB, A^2B, .., A^{n-1}B]$$

We easily see that $\mathcal{R}$ is the set of states that can be reached from the origin. An important property of $\mathcal{R}$ is its A-invariance.

LEMMA 3.2.7. *The reachable subspace $\mathcal{R}$ is A-invariant, i.e*

$$A\mathcal{R} \subseteq \mathcal{R}$$

In particular, $e^{At}\mathcal{R} \subseteq \mathcal{R}$ for all $t \in \mathbb{R}^n$.

Proof: Since, by the Cayley-Hamilton theorem, A^n is a linear combination of A^j for $j = 0, 1, .., n-1$ it follows that

$$A\mathcal{R} = \text{Im}\,[AB, A^2B, .., A^nB] \subseteq \text{Im}\,[B, AB, .., A^{n-1}B] = \mathcal{R}.$$

Moreover, by induction we get $A^j\mathcal{R} \subseteq \mathcal{R}$, which implies that

$$e^{At}\mathcal{R} = \sum_{j=0}^{\infty} \frac{t^j}{j!} A^j \mathcal{R} \subseteq \mathcal{R}$$

□ The preceding lemma is very important in the geometric theory of linear systems and as an immediate consequence we have the following theorem.

THEOREM 3.2.8. *Let $\epsilon > 0$. A time-invariant system can be transferred from any $x_0 \in \mathcal{R}$ to any $x_1 \in \mathcal{R}$ in time ϵ.*

Proof: By Lemma 3.2.7, $x_0 \in \mathcal{R}$ implies that $e^{A\epsilon}x_0 \in \mathcal{R}$. Therefore, if in addition, $x_1 \in \mathcal{R}$, $d = x_1 - e^{A\epsilon}x_0 \in \mathcal{R}$, and since $d \in \mathcal{R}$ the state transfer is possible. □

To further clarify the picture we note that if the state of the system is in $\mathcal{R}$ at some instant it is impossible to steer the state out of $\mathcal{R}$, neither is it possible to enter $\mathcal{R}$ from an initial state not in $\mathcal{R}$.

EXAMPLE 3.2.9 (A decomposition theorem). The $A-$invariance of $\mathcal{R}$ allows for a decomposition theorem for time-invariant systems.

Consider the system $\dot{x} = Ax + Bu$, $x(t) \in \mathbb{R}^n$, with the pair (A, B) *not* completely reachable. Let $\mathcal{R}$ be the reachable subspace with $d = \dim \mathcal{R}$ and let V be a complement to $\mathcal{R}$ in $\mathbb{R}^n$, i.e.,

$$\mathbb{R}^n = \mathcal{R} + V.$$

Choose a basis for $\mathbb{R}^n$ such that the d first vectors span $\mathcal{R}$ and the remaining span V and let x be the coordinates in this basis. Partition x as $x = \begin{bmatrix} x_1' & x_2' \end{bmatrix}'$, with x_1 $d-$dimensional. We shall show that in this basis the system has the structure

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} x + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u,$$

where B_1 has d rows.

Hence, we have to show that the blocks A_{21} and B_2 are zero, as indicated. In the chosen basis, the action of A on an arbitrary vector in $\mathcal{R}$ can be written as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}x_1 \\ A_{21}x_1 \end{bmatrix}.$$

Since $\mathcal{R}$ is $A-$invariant it holds that $A_{21}x_1 = 0$ regardless of x_1 and we conclude that $A_{21} = 0$. Moreover, since Im $B \subseteq \mathcal{R}$ it must hold that only the first d rows of B can be nonzero.

It is easily seen that the pair (A_{11}, B_1) is completely reachable and defines a reachable subsystem with state space $\mathcal{R}$.

As a side remark, at this stage, we mention that if the matrix A_{22} is asymptotically stable (see Chapter 4) we say that the pair (A, B) is *stabilizable* . This is a natural definition to make, since $x_2(t)$ is unaffected by the input and tends asymptotically to zero by the stability of A_{22} and by reachability of the pair (A_{11}, B_1) the state $x_1(t)$ can, by choice of suitable input, be steered to the origin.

We illustrate the decomposition theorem with a numerical example. Consider the system

$$\dot{x} = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 2 & 2 \\ 3 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} u.$$

The reachability matrix is

$$\Gamma = \begin{bmatrix} 1 & 2 & 4 \\ -2 & 0 & 8 \\ 1 & 2 & 4 \end{bmatrix},$$

and the set $\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \}$ is a basis for $\mathcal{R}$. A complement can be chosen as $V = \text{Im} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Let the transformation matrix T be defined as $T = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix}$, with inverse $T^{-1} = \begin{bmatrix} 0 & -0.5 & 1 \\ 0 & 0.5 & 0 \\ 1 & 0 & -1 \end{bmatrix}$. In the new basis the system is represented by

$$A = T^{-1} \begin{bmatrix} 2 & 0 & 0 \\ 2 & 2 & 2 \\ 3 & 0 & -1 \end{bmatrix} T = \begin{bmatrix} 0 & -2 & 2 \\ 2 & 4 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

and

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

It is easy to verify that that the pair (A_{11}, B_1) is reachable and since $A_{22} = -1$ the pair (A, B) is stabilizable. □

3.3. Reachability for Discrete-Time Systems

For simplicity we shall only study time-invariant systems.

Consider the discrete-time system

$$(\Sigma) \qquad x(t+1) = Ax(t) + Bu(t); \quad x(t_0) = x_0.$$

For an arbitrary input sequence $\{u(0), u(1), ..., u(t)\}$, we have

$$x(t+1) = A^{t+1}x(0) + \sum_{s=0}^{t} A^{t-s}Bu(s).$$

The question of reachability is whether it is possible to transfer the system from the state x_0 to x_1 in $T+1$ steps. As in the continuous-time case we let d be the difference between the desired state and the state of the uncontrolled motion, i.e., $d \triangleq x(T+1) - A^{T+1}x(0)$. By solving the system equations explicitly, we see that

$$d = [B, AB, A^2B, ..., A^TB] \begin{bmatrix} u(T) \\ u(T-1) \\ . \\ . \\ u(0) \end{bmatrix}.$$

Hence, a necessary and sufficient condition for the desired state transfer to be possible is that $d \in \text{Im}\,[B, AB, A^2B, ..., A^TB]$. We observe that since the rank of this matrix might depend on T we have, in contrast to the continuous-time case, that the state transfer might be possible for some T, but not for $T-1$.

DEFINITION 3.3.1. $\mathcal{R}_t \triangleq \text{Im}\,[B, AB, A^2B, .., A^tB]$

The spaces $\mathcal{R}_t$ forms an increasing sequence of subspaces of the state space $\mathbb{R}^n$ and again by the Cayley-Hamilton theorem we have that $\mathcal{R}_t = \mathcal{R}_{n-1}$ for $t \geq n$. Hence,

$$\mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq .... \subseteq \mathcal{R}_{n-1} = \mathcal{R}_n = \mathcal{R}_{n+1} = ..$$

Consequently, we define the reachable subspace $\mathcal{R}$ as

$$\mathcal{R} \triangleq \text{Im}\,[B, AB, A^2B, .., A^{n-1}B]$$

for discrete time systems as well and we say that the system is *completely reachable* if $\mathcal{R} = \mathbb{R}^n$.

We shall now state and prove an intuitively plausible lemma that will be a major component in the proof of the pole-placement theorem in Chapter 6.

LEMMA 3.3.2. *Let (A, B) be a reachable pair and $b \in \text{Im}\, B$. Then there is an input sequence $\{\tilde{u}(1), \tilde{u}(2), ..., \tilde{u}(n-1)\}$ such that the trajectory $\{x(1), x(2), ..., x(n)\}$ of the system $x(t+1) = Ax(t) + Bu(t)$, $x(1) = b$ spans $\mathbb{R}^n$, i.e.,*

$$\text{span}\,\{x(1), x(2), ..., x(n)\} = \mathbb{R}^n.$$

Proof: Let $\{x(1), x(2), ..., x(k)\}$ be a maximal linearly independent sequence that can be obtained as a trajectory of the system starting at $x(1) = b$. Let the subspace V be defined as $V \triangleq \text{span}\,\{x(1), x(2), ..., x(k)\}$. Clearly, $V \subseteq \mathcal{R}$. Since we want to show that $V = \mathbb{R}^n$, it is enough by reachability to show that $\mathcal{R} \subseteq V$.

Since the sequence is maximal we have that $x(k+1) \in V$ for every $u(k)$. By letting $u(k) = 0$ in the relation $x(k+1) = Ax(k) + Bu(k)$ we see that $Ax(k) \in V$. From $x(k+1) - Ax(k) = Bu(k)$ we conclude that $\text{Im}\, B \subseteq V$.

The next step is to note that V is A-invariant, which follows from the fact that for any j such that $1 \leq j < k$ we have that $Ax(j) = x(j+1) - Bu(j) \in V$.

By applying A to the inclusion $\text{Im}\, B \subseteq V$, using the A-invariance of V, we get that $A\,\text{Im}\, B \subseteq V$. Hence, $\text{Im}\, B + A\,\text{Im}\, B \subseteq V$, and by induction we obtain the inclusion $\mathcal{R} \subseteq V$. □

3.4. Observability

Consider now a system with output:

$$\begin{aligned} \dot{x} &= A(t)x + B(t)u \\ y &= C(t)x + D(t)u, \end{aligned} \tag{16}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$. In this section we shall investigate the problem of determining the state of a system given the input and the output. More specifically, can we determine $\{x(t);\, t_0 \le t \le t_1\}$ from $\{y(t);\, t_0 \le t \le t_1\}$ and $\{u(t);\, t_0 \le t \le t_1\}$? The analysis will be very similar to that of reachability and we shall later show that reachability and observability are in a sense dual concepts.

By solving the system equations (14) explicitly we get

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, s)B(s)u(s)ds.$$

We notice that in order to determine $x(t)$ for all $t \in [t_0, t_1]$, it is enough to determine x_0. The output y is given by the relation

$$y(t) = Cx(t) + D(t)u(t).$$

After rearranging terms we get

$$C(t)\Phi(t, t_0)x_0 = y(t) - D(t)u(t) - \int_{t_0}^{t} C(t)\Phi(t, s)B(s)u(s)ds. \tag{17}$$

The right-hand side of (17) is a known quantity and if we call it $v(t)$ the problem is to determine x_0 from the relation

$$C(t)\Phi(t, t_0)x_0 = v(t) \quad t \in [t_0, t_1]. \tag{18}$$

Now the question is, for a given input u, can different initial states produce the same function $v(t)$? If not, the mapping from initials states to functions v is injective and from a given function v we can determine x_0. If so we say the system is *observable*.

To this end, let $\mathcal{Y}$ be the space of m-dimensional, square-integrable functions on $[t_0, t_1]$ and introduce the mapping $T : \mathbb{R}^n \to \mathcal{Y}$ as

$$(Tx_0)(t) \triangleq C(t)\Phi(t, t_0)x_0. \tag{19}$$

It is easily seen that T is a linear mapping, but since $\mathcal{Y}$ is not a finite-dimensional space, T does not have a finite-dimensional matrix representation.

Recall that in the finite dimensional case, a mapping $\Omega : \mathbb{R}^n \to \mathbb{R}^p$ is injective if and only if Ω has full column rank, where Ω is a constant matrix. This implies that Ω must have more rows than columns. On the other hand, the mapping defined in (19) has the following form:

$$\Omega(t)x = v(t);\ t \in [t_0, t_1], \tag{20}$$

where $\Omega(t)$ has p rows and n columns ($p \le n$)!

LEMMA 3.4.1. *The mapping defined in* (20) *is injective if and only if the column vectors of* $\Omega(t)$ *are linearly independent over* $[t_0, t_1]$.

Proof: Sufficiency: if $x = a$ and $x = b$ produce the same $v(t)$ on $[t_0, t_1]$ we have that $\Omega(t)(a-b) = 0$. Since the columns of $\Omega(t)$ are linearly independent, $a - b = 0$.

Necessity: injection implies that $\Omega(t)b = 0$ over $[t_0, t_1]$ if and only if $b = 0$, which implies the linear independence. □

DEFINITION 3.4.2. The *observability Gramian* $M(t_0, t_1)$ is the matrix

$$\int_{t_0}^{t_1} \Phi^T(t, t_0)C^T(t)C(t)\Phi(t, t_0)dt.$$

THEOREM 3.4.3. *Consider system* (16).

(1) *The system is observable if and only if the observability Gramian* $M(t_0, t_1)$ *is nonsingular.*
(2) *When* $M(t_0, t_1)$ *is singular, for a given input* u *the initial states* $x(t_0) = a$ *and* $x(t_0) = b$ *produce the same output on* $[t_0, t_1]$ *if and only if* $a - b \in \ker M(t_0, t_1)$.

Proof: The first part can be proven by Lemmas 3.1.3 and 3.4.1.

For the second part it is enough to show that $C(t)\Phi(t, t_0)x = 0$ over $[t_0, t_1]$ if and only if $M(t_0, t_1)x = 0$.

Suppose $C(t)\Phi(t, t_0)x = 0$, then we can easily see $M(t_0, t_1)x = 0$. On the other hand, suppose $M(t_0, t_1)x = 0$. This implies that $x^T M(t_0, t_1)x = 0$. Using the same argument we have used before we can draw the conclusion that $C(t)\Phi(t, t_0)x = 0$ over $[t_0, t_1]$. □

Formally, the initial state x_0 can be determined by solving the equation

$$M(t_0, t_1)x_0 = \int_{t_0}^{t_1} \Phi^T(t, t_0)C^T(t)v(t)dt,$$

which has a unique solution if and only if $M(t_0, t_1)$ is positive definite. In the time-invariant case though, state determination is done by means of an *observer* as explained in Chapter 6.

3.5. Observability for Time-Invariant Systems

In this section we shall specialize to the class of time-invariant systems and characterize $\ker M(t_0, t_1)$ in terms of the matrix pair (C, A). As a companion to Theorem 3.2.3 we have the following theorem.

THEOREM 3.5.1. *Let* A *be* $n \times n$ *and* C *be* $m \times n$. *Then for all* t_0, t_1 *such that* $t_0 < t_1$ *we have*

$$\ker M(t_0, t_1) = \ker \begin{bmatrix} C \\ CA \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix}.$$

The proof of this theorem is analogous to that of Theorem 3.2.3.

DEFINITION 3.5.2. The observability matrix Ω is defined as

$$\Omega \triangleq \begin{bmatrix} C \\ CA \\ . \\ . \\ . \\ CA^{n-1} \end{bmatrix}.$$

DEFINITION 3.5.3. Let n be the dimension of the state space. The pair (C, A) is said to be completely observable if Ω has full rank, i.e.

$$\operatorname{rank} \Omega = n.$$

EXAMPLE 3.5.4. Consider the pair (C, A), where $A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & -1 \end{bmatrix}$, describing the electrical network in Example 1.3.2. The observability matrix Ω is

$$\Omega = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Since $\operatorname{rank} \Omega = 1$ the system is not completely observable and

$$\ker \Omega = \{\lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} ; \lambda \in \mathbb{R}\}.$$

Can we observe any difference between the initial states $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 0 \end{bmatrix}$? No, since $\begin{bmatrix} -1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \end{bmatrix} \in \ker \Omega$.
If the initial states are $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 2 \\ 8 \end{bmatrix}$? Yes, since $\begin{bmatrix} 1 \\ -6 \end{bmatrix} \notin \ker \Omega$. □

EXAMPLE 3.5.5. The pair (C, A) in Example 1.3.1 is $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and the observability matrix is $\Omega = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, which is of full rank, so (C, A) is completely observable. □

For a time-invariant system, not completely observable, $\ker \Omega$ constitutes an interesting subspace of the state space, which we will call the *unobservable subspace*. The following important lemma is easy to prove.

LEMMA 3.5.6. *The unobservable subspace is $A-$invariant.*

REMARK 3.5.7. By partitioning the state space as $\mathbb{R}^n = \ker \Omega + V$, where V is any complement, we can derive a decomposition theorem analogous to that of Example 3.2.9. Moreover, by combining these two decompositions and partitioning the state space into four particular subspaces we get the *Kalman decomposition* of Chapter 5. □

3.6. Observability for Discrete-Time Systems

We will only study time-invariant systems. Consider the discrete-time system

$$(\Sigma) \qquad \begin{aligned} x(t+1) &= Ax(t) + Bu(t); \quad x(t_0) = x_0 \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Can we distinguish between a system with $x(0) = a$ and a system with $x(0) = b$ by observing u and y on the interval $[0, T]$? By solving the system equations we get that

$$x(0) = a \Rightarrow y_a(t) = CA^t a + \sum_{s=0}^{t-1} CA^{t-s-1}Bu(s) + Du(t)$$

and

$$x(0) = b \Rightarrow y_b(t) = CA^t b + \sum_{s=0}^{t-1} CA^{t-s-1}Bu(s) + Du(t).$$

The conclusion is now that $y_a(t) = y_b(t)$ for all $t \in [0, T]$ if and only if $CA^t(a - b) = 0$ for $t = 0, 1, ..., T$, i.e.

$$\begin{bmatrix} C \\ CA \\ . \\ . \\ CA^T \end{bmatrix} (a - b) = 0.$$

Consequently, the observability properties of a discrete-time system is determined by Ω, where Ω is defined as in Definition 3.5.2, and we say that the system is completely observable if rank $\Omega = n$.

3.7. Duality between reachability and observability

As we have seen reachability is determined by the pair (A, B) and observability by the pair (C, A). There is an important duality between reachability and observability which amounts to reversing the direction of time and transposing matrices determining the dynamics. We shall illustrate this important principle, to which we shall return in Chapter 9, in the time-invariant case.

Consider the controlled system

$$(21) \qquad \dot{x} = Ax + Bu; \quad x(t_0) = a$$

and the dual observed system

$$(22) \qquad \begin{cases} \dot{z} = -A^T z; \quad z(t_1) = b \\ y = B^T z \end{cases}$$

The first thing to observe is the following

PROPOSITION 3.7.1. *The system* (21) *is completely reachable if and only if* (22) *is completely observable.*

Proof: This is immediately clear from the fact that

$$\operatorname{rank}\begin{bmatrix} B^T \\ -B^TA^T \\ \vdots \\ (-1)^{-1}B^T(A^T)^{n-1} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} B & AB & A^{n-1}B \end{bmatrix}. \qquad \square$$

To illustrate the more general principle let us consider the following example. Can we control the system (21) so that it is transferred from state $x(t_0) = a$ at time t_0 to a state at time t_1 such that $b^Tx(t_1) = 0$?

To answer this question let us form

$$\begin{aligned} \frac{d}{dt}(z^Tx) &= z^T\dot{x} + \dot{z}^Tx \\ &= z^TAx + z^TBu - z^TAx \\ &= y^Tu \end{aligned}$$

which integrated between t_0 and t_1 yields

$$b^Tx(t_1) - z(t_0)^Ta = \int_{t_0}^{t_1} u(t)^Ty(t)dt.$$

Therefore, (21) can be controlled so that $b^Tx(t_1) = 0$ if and only if there is a control u so that

$$a^Tz(t_0) = -\int_{t_0}^{t_1} u(t)^Ty(t)dt.$$

This can be interpreted as a condition for observing $a^Tz(t_0)$, i.e. for obtaining a linear functional of the observed function $y(t)$ so that $a^Tz(t_0)$ is reconstructed. In this sense, $b^Tx(t_1) = 0$ can be reached if and only if $a^Tz(t_0)$ can be observed.

These ideas will be further elaborated upon in Chapter 9.

CHAPTER 4

Stability

Stability is one of the most central concepts in systems and control. In most engineering projects unstable systems are useless. Therefore in system analysis and control design stability (or stabilization) is almost always the first priority to be met.

This chapter deals only with the stability of time-invariant linear systems, a subject which is drastically simplified by the fact that the complete set of solutions of the system $\dot{x} = Ax$ can be displayed explicitly by means of the Jordan form. As a consequence, it is enough to check the eigenvalues of A in order to determine the stability of a system.

4.1. Stability of a dynamical system

Let us study a homogeneous linear system

$$\dot{x} = Ax; \quad x(0) = x_0, \tag{23}$$

where A is **constant**.

Definition 4.1.1. The system (23) is *stable* if the solution is bounded on the interval $[0, \infty)$ for all initial values x_0 and *asymptotically stable* if $x(t) \to 0$ when $t \to \infty$ for all x_0.

Remark 4.1.2. This definition is equivalent to the so-called Lyapunov stability for linear systems. However, for nonlinear systems, being asymptotically stable requires more conditions than just the convergence of all solutions.

Theorem 4.1.3.

(1) *The system* (23) *is asymptotically stable if and only if the real parts of all the eigenvalues of A are negative, i.e., the eigenvalues are all located in the open left half plane.*
(2) *The system* (23) *is unstable if A has at least one eigenvalue in the open right half plane.*

Proof: In this proof we shall use a fundamental result from linear algebra, the *Jordan decomposition theorem*. This theorem guarantees the existence of a basis for $\mathbb{R}^n$ in which the representation of the linear mapping A takes a particularly simple form.

Transform the matrix A to Jordan form $A = TJT^{-1}$, where J is a block-diagonal matrix

$$J = \operatorname{diag}(J_1, J_2, ..., J_r)$$

and each $d_\nu \times d_\nu$ block J_ν has the form

$$J_\nu = \begin{bmatrix} \lambda_\nu & 1 & & 0 \\ & \lambda_\nu & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_\nu \end{bmatrix},$$

λ_ν being an eigenvalue of A. Thus,

$$e^{At} = T \begin{bmatrix} e^{J_1 t} & & & 0 \\ & e^{J_2 t} & & \\ & & \ddots & \\ 0 & & & e^{J_r t} \end{bmatrix} T^{-1},$$

so it remains to analyze each $e^{J_\nu t}$. But J_ν has the form

$$J_\nu = \lambda_\nu I + S_\nu$$

where S_ν is a shift matrix

$$S_\nu = \begin{bmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & 0 & \ddots \\ & & & \ddots & 1 \\ 0 & & & & 0 \end{bmatrix}$$

of dimension $d_\nu \times d_\nu$, having the property that $S^i = 0$ for $i \geq d_\nu$. Consequently,

$$e^{J_\nu t} = e^{\lambda_\nu t} e^{S_\nu t} = e^{\lambda_\nu t}(I + tS + \frac{t^2}{2!}S^2 + ... + \frac{t^{d_\nu - 1}}{(d_\nu - 1)!}S^{d_\nu - 1})$$

and therefore, setting $\sigma_\nu = \operatorname{Re}\lambda_\nu$ and $\omega_\nu = \operatorname{Im}\lambda_\nu$,

$$e^{Jt} = \sum_\nu e^{\sigma_\nu t} P_\nu(t)\,(\cos\omega_\nu t + j \sin\omega_\nu t), \tag{24}$$

where $P_\nu(t)$ is a matrix-valued polynomial of dimension $d_\nu - 1$ in t. From this expression it follows that (1) $e^{At}x_0 \to 0$ for all x_0 if and only if $\sigma_\nu \triangleq \operatorname{Re}\lambda_\nu < 0$ for all ν and that (2) $e^{At}x_0 \to \infty$ for at least one x_0 if some $\sigma_\nu > 0$. □

Corollary 4.1.4. *The system* (23) *is stable if and only if all eigenvalues of A are located in the closed left half plane and any eigenvalues on the imaginary axis correspond to one-dimensional Jordan blocks.*

Proof: By Theorem 4.1.3 (1) we only need to worry about terms in (24) for which $\sigma_\nu = 0$ i.e. $e^{\sigma_\nu t} = 1$. These terms will remain bounded if and only if the degree of P_ν is zero, i.e., $d_\nu = 1$. □

DEFINITION 4.1.5. A is a *stable matrix* if $\operatorname{Re}\lambda(A) < 0$.

EXAMPLE 4.1.6 (Time-varying systems). Consider a time-varying system $\dot{x}(t) = A(t)x(t)$, is it true that $\operatorname{Re}\lambda(A(t)) < 0$ for every t implies stability? The answer is no, as can be seen from the following counterexample.

Let $A(t) = \begin{bmatrix} -1 & e^{2t} \\ 0 & -1 \end{bmatrix}$. The eigenvalues are a priori time varying and can be calculated from

$$\det(\lambda(t)I - A(t)) = \begin{vmatrix} \lambda(t)+1 & -e^{2t} \\ 0 & \lambda(t)+1 \end{vmatrix} = 0,$$

which yields that $\lambda(t) = -1$ for all t.

The next step is to calculate $\phi(t,s)$. It easily seen that $A(t)$ and $\int_s^t A(\tau)d\tau$ commute, so by Remark 2.2.2 the transition matrix is given by

$$\phi(t,s) = e^{\int_s^t A(\tau)d\tau} = \exp(\int_s^t -I + e^{2\tau}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} d\tau).$$

The two terms in the exponent commute, so the product rule for matrix exponentials is applicable and gives

$$\phi(t,s) = e^{s-t}\begin{bmatrix} 1 & \frac{e^{2t}-e^{2s}}{2} \\ 0 & 1 \end{bmatrix}.$$

Since $\phi_{12}(t,s)$ is unbounded for $t > s$, the system is unstable. □

4.2. Input-output stability

Now we consider a linear control system

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx. \end{aligned} \tag{25}$$

We want a bounded input to give a bounded output, which is sometimes abbreviated as *BIBO-stability*.

DEFINITION 4.2.1. The system (25) is *input-output* stable if there is a k such that

$$\left.\begin{matrix} x(0) = 0, \\ \|u(t)\| \le 1 \quad t \in [0,\infty) \end{matrix}\right\} \Rightarrow \|y(t)\| \le k, \quad t \in [0,\infty).$$

Since

$$y(t) = \int_0^t Ce^{A(t-s)}Bu(s)\,ds,$$

defining $G(t) \triangleq Ce^{At}B$,

$$\begin{aligned} \|y(t)\| &\le \int_0^t \|G(t-s)\|\,\|u(s)\|\,ds \qquad (\text{since } \|u(t)\| \le 1) \\ &\le \int_0^t \|G(\sigma)\|\,d\sigma \qquad (26) \\ &\le \|C\|\,\|B\| \int_0^t \|e^{A\sigma}\|\,d\sigma, \end{aligned}$$

i.e. a sufficient for condition for input-output stability is that the integral $\int_0^\infty \|e^{At}\|\,dt$ is convergent.

THEOREM 4.2.2. *If A is a stable matrix then the time-invariant system* (25) *is input-output stable. Moreover, if (A, B) is reachable and (C, A) is observable,* (25) *is input-output stable only if A is stable.*

Proof: If all eigenvalues of A have negative real parts so that all σ_ν in (24) are negative then

$$\int_0^\infty \|e^{At}\|\,dt < \infty$$

and hence, in view of (26) the system is input-output stable. The second assertion can be shown using the Jordan form and is omitted here. □

REMARK 4.2.3. Sometimes one prefers to define the norm of a signal as the energy it contains, i.e., $\|u\|_2^2 := \int_0^\infty \|u(s)\|^2 ds$. The conclusions in Theorem 4.2.2 still hold if we replace $\|\cdot\|$ in Definition 4.2.1 with $\|\cdot\|_2$.

4.3. The Lyapunov equation

The *second method of Lyapunov* utilizes a so-called Lyapunov function to investigate stability of dynamical systems. When specializing the method to the linear case the so-called *Lyapunov equation* appears. Lyapunov equations also occur in many other control problems.

Consider again system (23). Let a quadratic function be defined as $V(x) = x^T P x$, where P is a symmetric matrix.

DEFINITION 4.3.1. $V(x)$ (correspondingly P) is said to be *positive definite* if $V(x) > 0\ \forall x \neq 0$, to be *positive semi-definite* if $V(x) \ge 0\ \forall x \neq 0$.

A positive definite V can be viewed as an energy function and a stable system should dissipate energy all the time, namely $\dot{V}$ should always be negative. Thus for a stable system we expect $\dot{V} = x^T(A^T P + PA)x$ to be negative for all nonzero x with some positive definite P. This rational leads to the following Lyapunov equation

$$A^T P + PA + Q = 0, \qquad \text{(LE)}$$

where both P and Q are symmetric. For given matrices A and Q this is a linear equation in the unknown P, and we have the following lemma.

LEMMA 4.3.2. *Let A be a stable matrix. Then* (LE) *has a unique solution*

$$P = \int_0^\infty e^{A^T t} Q e^{At}\, dt. \tag{27}$$

REMARK 4.3.3. It is obvious that if Q is positive definite ($Q > 0$), then $P > 0$. Moreover, even for some positive semi-definite Q, P can still be positive definite, as our discussion below will show. On the other hand, the sum $A^T M + MA$ may be sign indefinite even if A is stable and M is a positive definite matrix.

Proof: Since $\operatorname{Re}\lambda(A) < 0$, the integral is convergent, for the same reason as in the proof of Theorem 4.2.2. Now, straight-forward differentiation yields

$$\frac{d}{dt}(e^{A^T t} Q e^{At}) = A^T e^{A^T t} Q e^{At} + e^{A^T t} Q e^{At} A.$$

Integrating this from 0 to ∞ yields $0 - Q = A^T P + PA$, i.e. P satisfies (LE). It remains to show that there is only one solution. Define the linear map $L : \mathbb{R}^{n\times\, n} \to \mathbb{R}^{n\times\, n}$ as

$$L(P) \triangleq A^T P + PA$$

Since (LE) has a solution for all Q, $\operatorname{Im} L = \mathbb{R}^{n\times n}$, i.e. L has full rank. Hence $L(P) = -Q$ has a unique solution for each Q. □

By specializing the structure of (LE) we are able to give the main theorem of this section.

THEOREM 4.3.4. *Let (C, A) be an observable pair and consider the equation*

$$A^T P + PA + C^T C = 0. \tag{28}$$

Then the following statements are equivalent

1. *A is a stable matrix*
2. (28) *has a positive definite solution P.*

Proof: First, suppose A is a stable matrix. Then, by Lemma 4.3.2, the unique solution P of (28) is given by

$$P = \int_0^\infty e^{A^T t} C^T C e^{At}\, dt.$$

The observability Gramian of the pair (C, A) on the interval $[0, t]$ is

$$M(0, t) = \int_0^t e^{A^T s} C^T C e^{As}\, ds.$$

By observability, the matrix $M(0, t)$ is positive definite for all $t > 0$ and we conclude that $P > 0$.

Conversely, suppose that P is positive definite and form the Lyapunov function $V(x) \triangleq x^T P x$, which is strictly positive for all $x \neq 0$. Differentiation of V along a solution of $\dot{x} = Ax$ yields

$$\frac{d}{dt}(x^T P x) = \dot{x}^T P x + x^T P \dot{x} = x^T (A^T P + PA) x = -x^T C^T C x.$$

Integrating this from t_0 to t we obtain

$$V(x(t)) - V(x(t_0)) = -\int_{t_0}^{t_1} |Cx(s)|^2 \, ds \leq 0 \tag{29}$$

and hence,

$$0 \leq V(x(t)) \leq V(x(t_0))$$

for all $t \geq t_0$. This implies that all solutions of $\dot{x} = Ax$ remain bounded and from Corollary 4.1.4 it follows that $\mathrm{Re}\lambda(A) \leq 0$ and that imaginary eigenvalues, if any, correspond to one-dimensional Jordan blocks.

We shall now exclude the possibility of such imaginary eigenvalues. It is easily verified, using the Jordan form, that the existence of imaginary eigenvalues is equivalent to the existence of a nontrivial periodic solution of $\dot{x} = Ax$. Hence, it is enough to show that there is no nontrivial periodic solution.

Suppose $x(t)$ is a periodic solution such that $x(t_0) = x(t_1)$. By (29) and the continuity of $x(t)$ it now holds that $Cx(t) \equiv 0$ on the interval $[t_0, t_1]$, and by periodicity it holds for all t. Differentiation of this identity yields

$$C\dot{x}(t) \equiv CAx(t) \equiv 0,$$

and by induction it follows $CA^k x(t) \equiv 0$ for all k. Since (C, A) is observable, it follows that $x(t) \equiv 0$. □

We demonstrate that Theorem 4.3.4 leads to a stability test.

COROLLARY 4.3.5. *Let Q be symmetric and positive definite ($Q > 0$). Then the following statements are equivalent*

(1) *A is a stable matrix*
(2) (LE) *has a positive definite solution P.*

Proof: If $Q > 0$ it can be factored as $C^T C = Q$, with C invertible and it trivially holds that (C, A) is observable. □

By dualizing we get the following corollary, important in stochastic systems theory.

COROLLARY 4.3.6. *Let (A, B) be a reachable pair and consider the equation*

$$AP + PA^T + BB^T = 0. \tag{30}$$

Then the following statements are equivalent

(1) *A is a stable matrix*
(2) (30) *has a positive definite solution P.*

Proof: Use the facts that (A, B) is reachable if and only if (B^T, A^T) is observable, and that A is stable if and only if A^T is stable. □

EXAMPLE 4.3.7.

$$\text{Is } A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \text{ a stable matrix?}$$

Choose $Q = I$ which is positive definite as required and solve the Lyapunov equation (LE) for the symmetric matrix $P = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$ yielding

$$\Rightarrow \begin{cases} -2p_{12} + 1 & = 0 \\ p_{11} - 2p_{12} - p_{22} & = 0 \\ 2p_{12} - 4p_{22} + 1 & = 0 \end{cases} \Rightarrow P = \frac{1}{2}\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} > 0,$$

and A is a stable matrix. □

4.4. Stability of discrete-time systems

Here we give the analogous results for discrete-time systems leaving the proofs as an exercise for the reader. Consider the homogeneous system

$$x(t+1) = Ax(t) \quad x(0) = x_0 \tag{31}$$

where A is an $n \times n$ constant matrix.

THEOREM 4.4.1.

(1) *The system in* (31) *is asymptotically stable if and only if* $|\lambda(A)| < 1$ *i.e. if and only if all the eigenvalues of* A *are located strictly inside the unit circle.*
(2) *The system is unstable if there is at least one eigenvalue outside the unit circle.*

DEFINITION 4.4.2. A is a stable matrix if $|\lambda(A)| < 1$.

Consider the discrete Lyapunov equation

$$P = A^T P A + Q \tag{DLE}$$

LEMMA 4.4.3. *Let* A *be a stable matrix. Then* (DLE) *has a unique solution,*

$$P = \sum_{k=0}^{\infty} (A^T)^k Q A^k$$

THEOREM 4.4.4. *Let* (C, A) *be observable and let* P *be a solution of* (DLE) *with* $Q = C^T C$. *Then the following statements are equivalent*

(1) A *is a stable matrix*
(2) $P > 0$.

EXAMPLE 4.4.5. The discrete Lyapunov equation leads to a stability test.

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{1}{6} & -\frac{5}{6} \end{bmatrix}$$

Choose $Q = I$ and solve the equation $P - A^T P A = I$ for the symmetric matrix P.

$$\begin{cases} p_{11} - \frac{1}{36} p_{22} & = 1 \\ \frac{7}{6} p_{12} - \frac{5}{36} p_{22} & = 0 \\ -p_{11} + \frac{5}{3} p_{12} + \frac{11}{36} p_{22} & = 1 \end{cases} \Rightarrow P = \begin{bmatrix} \frac{67}{60} & \frac{1}{2} \\ \frac{1}{2} & \frac{21}{5} \end{bmatrix}$$

Sylvesters test yields

$$\frac{67}{60} > 0 \qquad \det P = \frac{111}{25} > 0 \Rightarrow P > 0$$

showing that P is positive definite, and hence A is a stable matrix.

CHAPTER 5

Realization theory

Finite-dimensional time-invariant linear systems of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{32}$$

and

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}$$

are of fundamental interest, both in applications and in theory. These representations exhibit the state explicitly as a dynamical memory of the system, making them ideally suited for estimation and control where efficient algorithms can be derived.

The system (32) defines an input/output mapping

$$y(t) = \int_{t_0}^{t} G(t,\tau)u(\tau)\,d\tau + D(t)u(t). \tag{33}$$

Conversely, given an input/output description as in (33) we would like to know if there is a finite-dimensional time-invariant system realizing the input/-output behavior and such a representation is consequently called a *realization* of the input/output representation. Clearly, realizations having as few as possible state variables are of special interest, both from a computational and a theoretical point of view, and are called *minimal*.

A major theme in systems theory is that reachability and observability should be equivalent to minimality. The intuition behind this is the idea that states that cannot be reached or observed should not affect the input/output behavior of a system and therefore account for non-minimality.

DEFINITION 5.0.6. The dimension of a realization (A, B, C, D) is defined as the dimension of A.

DEFINITION 5.0.7. The system (32) is a *minimal* of the input/output representation (33) if there is no other realization of (33) of lower dimension.

Two major tasks of realization theory is to analyze properties of realizations such as minimality, and to investigate the problem of realizing input/output behaviors as state space systems.

In this chapter we shall mainly discuss continuous-time systems. However, many properties are common for continuous- and discrete-time systems. Moreover, many results will be relevant for *stochastic* systems as well.

As it turns out, the three properties linearity, finite dimensionality and time invariance together define a nice class of systems which is sometimes referred to as the class of *rational* systems, a term explained later in this chapter. The attributes continuous/discrete-time and deterministic/stochastic are often interchangeable within the class of rational systems.

5.1. Realizability and rationality

We shall now address the question of existence of a system (32) realizing a given input/output description. Without loss of generality we consider the input/output system

$$y(t) = \int_0^t G(t,\tau)u(\tau)\,d\tau,$$

and consequently we let $D = 0$ in (32). By solving the equations (32) we get

$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)\,d\tau.$$

By linearity we must have $Ce^{At}x(0) = 0$ for all t, which can be achieved by letting $x(0) = 0$. A necessary condition for the system (33) to have a finite-dimensional time-invariant realization is now seen to be that $G(t,\tau)$ be a function of the difference $t-\tau$.

EXAMPLE 5.1.1. Consider the simple delay system $y(t) = u(t-1)$, which can be written $y(t) = \int_{-\infty}^t \delta(\tau-(t-1))u(\tau)\,d\tau$. The weighting function is clearly a function of $t-\tau$, but does the system have a finite-dimensional realization? □

In order to characterize the weighting functions $G(t-\tau)$ having finite-dimensional time-invariant realizations we shall employ the Laplace transform. Consider the system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}$$

Suppose the system is at rest at $t = 0$, i.e. $x(0) = 0$. Then applying the Laplace transform to the system we obtain

$$\begin{cases} s\tilde{x}(s) = A\tilde{x}(s) + B\tilde{u}(s) \\ \tilde{y}(s) = C\tilde{x}(s) + D\tilde{u}(s), \end{cases}$$

and therefore

$$\tilde{y}(s) = \big[C(sI-A)^{-1}B + D\big]\tilde{u}(s) = R(s)\tilde{u}(s),$$

where the matrix-valued function

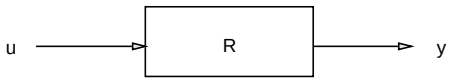
$$R(s) \triangleq C(sI-A)^{-1}B + D \tag{34}$$

is called the *transfer function* of the system (32). By Cramer's rule it is seen that $R(s)$ is a matrix of proper rational functions and that $\mathcal{L}\{G(t)\} =$

$C(SI - A)^{-1}B$ is a matrix of strictly proper rational functions (recall that a rational function is called *proper* if the degree of the numerator is less or equal the degree of the denominator and *strictly proper* if the degree of the numerator is strictly less than the degree of the denominator). Hence, a necessary condition for $G(t)$ to have a realization is that its Laplace transform is a matrix of strictly proper rational functions.

EXAMPLE 5.1.2. Consider again the simple delay system $y(t) = u(t-1)$. Since the Laplace transform of the weighting function is e^{-s}, which is not a rational function, the delay system does not have a finite-dimensional realization. Interpret this result! □

Hence, the system (32) can then be described by the rational matrix function R and we illustrate this as



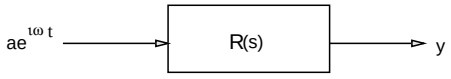
This leads to a natural inverse problem, namely the *realization problem*. Given a proper rational matrix function $R(s)$, determine (A, B, C, D) such that

$$R(s) = C(sI - A)^{-1}B + D.$$

Without loss of generality we may take $D = 0$, because $D = R(\infty)$ can be subtracted from $R(s)$ forming a new $R(s)$ with rational elements such that the degree of the numerator is lower than that of the denominator.

REMARK 5.1.3. There is a very useful interpretation of $R(s) = \mathcal{L}\{G(t)\}$ (for simplicity, assume $R(\infty) = 0$), forming the basis for the frequency domain approach to linear systems.

Suppose that the input/output system is input/output stable, and for simplicity suppose that the system is single-input/single-output (*SISO*). Pass a sine wave with frequency ω through the system and let it go to steady state. More precisely, apply the input $u(t) = \sin(\omega t) = \text{Im } e^{i\omega t}$ to the system starting at $t = -\infty$.



Then

$$\begin{aligned} y(t) &= \text{Im} \int_{-\infty}^{t} G(t-\tau)e^{i\omega\tau}\, d\tau = \{t - \tau = s\} \\ &= \text{Im} \int_{0}^{\infty} e^{i\omega(t-s)}G(s)\, ds = \text{Im } e^{i\omega t}R(i\omega). \end{aligned}$$

Hence, since $y(t) = A\sin(\omega t + \phi)$, where $A = |R(i\omega)|$ and $\phi = \arg R(i\omega)$, the function $R(s)$ tells us how the amplitude and phase of a sinusoid is affected by a linear time-invariant system. For a multivariable system this interpretation holds for each input/output pair separately. Moreover, a direct term, i.e., the case of $R(\infty) = D \neq 0$, can easily be incorporated into this interpretation as well.

Note that this property of $\mathcal{L}\{G(t)\}$ does not depend on finite dimensionality, but only linearity and time invariance. Furthermore, this also leads to a procedure for determining $\tilde{G}(s)$ experimentally. By varying ω we can determine the value of R in a number of points and then estimate $\tilde{G}(s)$ by extrapolation. □

REMARK 5.1.4. In the same way, the $\mathcal{Z}$-transform can be applied to the discrete time system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}$$

to yield

$$\tilde{y}(z) = \big[C(zI - A)^{-1}B + D\big]\tilde{u}(z) = R(z)\tilde{u}(z).$$

where $R(z)$ is a $m \times k$ matrix of rational functions.

Next we show that any $m \times k$ matrix $R(s)$ of a proper rational functions is the transfer function of a system of the form (32) with k inputs and m outputs.

THEOREM 5.1.5. *Given a proper rational matrix function $R(s)$ there are constant matrices (A, B, C, D) such that*

$$R(s) = C(sI - A)^{-1}B + D$$

We shall give two constructive proofs for Theorem 5.1.5, each proof giving a different solution as well as insight into the problem. Without loss of generality we can assume that $R(s)$ is strictly proper, i.e., $R(\infty) = 0$.

Proof:# 1(*the standard reachable realization*) Let

$$\mathcal{X}(s) = s^r + a_1 s^{r-1} + \cdots + a_r \tag{35}$$

be the least common denominator of the elements of $R(s)$. Since R is strictly proper, $\mathcal{X}(s)R(s)$ is then a matrix polynomial of degree less than or equal to $r - 1$, i.e.

$$\mathcal{X}(s)R(s) = N_0 + N_1 s + N_2 s^2 + \cdots + N_{r-1}s^{r-1}.$$

Define now (A, B, C) as follows

$$\underset{rk\times rk}{A} = \begin{bmatrix} 0 & I_k & 0 & \cdots & 0 \\ 0 & 0 & I_k & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & I_k \\ -a_r I_k & -a_{r-1} I_k & -a_{r-2} I_k & \cdots & -a_1 I_k \end{bmatrix} \quad \underset{rk\times k}{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ I_k \end{bmatrix}$$

$$\underset{m\times rk}{C} = [N_0, N_1, N_2, \cdots, N_{r-1}]$$

where I_k is an identity matrix of dimension $k \times k$, and let $X(s) = \begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_r(s) \end{bmatrix}$ be the solution of $(sI - A)X(s) = B$. i.e.

$$\begin{aligned} sX_i &= X_{i+1} \quad \text{for } i = 1, 2, \cdots, r-1 \\ sX_r + a_1X_r + a_2X_{r-1} + \cdots + a_rX_1 &= I_k \end{aligned}$$

From this we readily see that $\mathcal{X}(s)X_1 = I_k$ and consequently

$$X_i = \frac{s^{i-1}}{\mathcal{X}(s)} I_k \quad i = 1, 2, \cdots, r$$

Therefore

$$C(sI - A)^{-1}B = CX = \frac{1}{\mathcal{X}(s)}[N_0 + N_1 s + N_2 s^2 + \cdots + N_{r-1}s^{r-1}] = R(s)$$

and hence (A, B, C) defines a realization as required. □ Note that this realization is completely reachable because

$$[B, AB, \cdots, A^{rk-1}B] = \begin{bmatrix} 0 & 0 & \cdots & 0 & I & * & \cdots & * \\ 0 & 0 & \cdots & I & * & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & I & \cdots & * & * & * & \cdots & * \\ I & * & \cdots & * & * & * & \cdots & * \end{bmatrix}$$

has full rank rk. We will call this realization the *standard reachable realization.*

Proof:#2 (*the standard observable realization*) We note that $R(\infty) = 0$, which implies that $R(s)$ is analytic around $s = \infty$ and thus can be expanded in a Laurent-series

$$R(s) = R_1 s^{-1} + R_2 s^{-2} + R_3 s^{-3} + \ldots \tag{36}$$

around infinity. (This expansion holds outside a circle around zero in the complex plane encircling all the poles of R.) Moreover, for any realization (A, B, C) it holds that

$$\begin{aligned} C(sI - A)^{-1}B &= s^{-1}C(I - As^{-1})^{-1}B \\ &= s^{-1}C[I + As^{-1} + A^2s^{-2} + A^3s^{-3} + \cdots]B \\ &= CBs^{-1} + CABs^{-2} + CA^2Bs^{-3} + \cdots \\ &= R_1s^{-1} + R_2s^{-2} + R_3s^{-1} + \cdots = R(s). \end{aligned}$$

Hence, the transfer function is uniquely determined by the sequence

$$\{R_1, R_2, R_3, \ldots\},$$

which is called the sequence of *Markov parameters* of the transfer function.

Before proceeding with the proof we need a lemma, showing that the Markov parameters of a *rational* are give by finite data.

LEMMA 5.1.6. *The matrix coefficients of the Laurent expansion* (36) *satisfy the recursion*

$$R_{r+i} = -a_1 R_{r+i-1} - a_2 R_{r+i-2} - \cdots - a_r R_i \text{ for } i = 1, 2, 3, ... \tag{37}$$

Proof: Multiplying together (35) and (36) yields

$$\begin{aligned}\mathcal{X}(s)R(s) &= (s^r + a_1 s^{r-1} + a_2 s^{r-2} + \cdots + a_r)(R_1 s^{-1} + R_2 s^{-2} + R_3 s^{-3} + ...) \\ &= \cdots + s^{-1}(R_{r+1} + a_1 R_r + a_2 R_{r-1} + \cdots + a_r R_1) \\ &\quad + s^{-2}(R_{r+2} + a_1 R_{r+1} + a_2 R_r + \cdots + a_r R_2) \\ &\quad + s^{-3}(R_{r+3} + a_1 R_{r+2} + a_2 R_{r+1} + \cdots + a_r R_3) \\ &\quad + \cdots\end{aligned}$$

But $\mathcal{X}(s)R(s)$ is a matrix polynomial and consequently the coefficients of the negative powers of s are zero and (37) follows. □ We shall now continue the proof of the theorem. To this end define

$$\underset{rm\times rm}{A} = \begin{bmatrix} 0 & I_m & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & I_m \\ -a_r I_m & -a_{r-1} I_m & -a_{r-2} I_m & \cdots & -a_1 I_m \end{bmatrix} \quad \underset{rm\times k}{B} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_r \end{bmatrix}$$

$$\underset{m\times rm}{C} = [I_m, 0, \cdots, 0]$$

Then, by Lemma 5.1.6 we have

$$AB = \begin{bmatrix} R_2 \\ R_3 \\ \vdots \\ -a_1 R_r - a_2 R_{r-1} - \cdots - a_r R_1 \end{bmatrix} = \begin{bmatrix} R_2 \\ R_3 \\ \vdots \\ R_{r+1} \end{bmatrix}.$$

Similarly, by Lemma 5.1.6, we obtain by induction

$$A^i B = \begin{bmatrix} R_{i+1} \\ R_{i+2} \\ \vdots \\ R_{i+r} \end{bmatrix} \quad i = 0, 1, 2, 3, ... \quad .$$

and hence $R_i = CA^{i-1}B$, showing that (A, B, C) defines a realization. □

This realization is completely observable, because

$$\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{rm-1} \end{bmatrix} = \begin{bmatrix} I_m & 0 & 0 & \cdots & 0 \\ 0 & I_m & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & I_m \\ * & * & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ * & * & * & \cdots & * \end{bmatrix}$$

has full rank rm and it is therefore named *the standard observable realization.*

EXAMPLE 5.1.7. Consider a system with two inputs and two outputs and having the transfer function

$$R(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{2}{s+1} \\ \frac{-1}{s^2+3s+2} & \frac{1}{s+2} \end{bmatrix} = \frac{1}{s^2+3s+2} \begin{bmatrix} s+2 & 2s+4 \\ -1 & s+1 \end{bmatrix}.$$

Then

$$\mathcal{X}(s) = s^2+3s+2 = (s+1)(s+2), \quad r = 2$$
$$a_1 = 3, a_2 = 2 \quad N_0 = \begin{bmatrix} 2 & 4 \\ -1 & 1 \end{bmatrix} \quad N_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

and therefore the standard reachable realization is

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & -3 & 0 \\ 0 & -2 & 0 & -3 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad C = \begin{bmatrix} 2 & 4 & 1 & 2 \\ -1 & 1 & 0 & 1 \end{bmatrix}.$$

To determine the standard observable realization, we calculate the Laurent-expansion around $s = \infty$

$$\begin{cases} r_{11}(s) & = \frac{1}{s+1} = s^{-1}(1+s^{-1})^{-1} = s^{-1} - s^{-2} + s^{-3} - s^{-4} + \cdots \\ r_{12}(s) & = \frac{2}{s+1} = 2s^{-1} - 2s^{-2} + 2s^{-3} - 2s^{-4} + \cdots \\ r_{22}(s) & = \frac{1}{s+2} = s^{-1}(1+2s^{-1}) = s^{-1} - 2s^{-2} + 4s^{-3} - 8s^{-4} + \cdots \\ r_{21}(s) & = \frac{1}{s+2} - \frac{1}{s+1} = -s^{-2} + 3s^{-3} - 7s^{-4} + \cdots, \end{cases}$$

which yields

$$R_1 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \quad R_2 = \begin{bmatrix} -1 & -2 \\ -1 & -2 \end{bmatrix} \quad R_3 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad R_4 = \begin{bmatrix} -1 & -2 \\ -7 & -8 \end{bmatrix}$$

so the standard observable realization is

$$A \text{ as above} \quad B = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -1 & -2 \\ -1 & -2 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Are these realizations minimal? The answer is that, in general, the observable and reachable realizations are not minimal.

5.2. Minimality and McMillan degree

If a realization (A, B, C, D) of a given transfer function $R(s)$ is not minimal, there are obviously too many states in the realization. It is a natural guess that there should be a way of eliminating some of the states to obtain a minimal realization.

Which states should be eliminated? Since we are only concerned with the input/output behavior of the realization we should try to eliminate the states that are not affected by the input and the states that do not affect the output. Hence, a natural candidate for the part of the state space that should be *kept* is the part of the reachable subspace that is not included in the unobservable subspace.

The *Kalman decomposition* is way of partitioning the state space into four subspaces, of which one is a complement to $\mathcal{R} \cap \ker \Omega$ in $\mathcal{R}$.

THEOREM 5.2.1 (The Kalman decomposition). *Let (A, B, C, D) define an n-dimensional realization and let $\mathcal{R}$ be the reachable subspace and let $\ker \Omega$ be the unobservable subspace. Let $V_{\bar{o}\bar{r}}$ and V_{or} be complements defined by*

$$\ker \Omega = \mathcal{R} \cap \ker \Omega + V_{\bar{o}\bar{r}}$$

and

$$\mathcal{R} = \mathcal{R} \cap \ker \Omega + V_{or}.$$

Finally let $V_{\bar{r}o}$ be the subspace of states that are neither in $\mathcal{R}$ nor in $\ker \Omega$.

$$\mathbb{R}^n = \mathcal{R} \cap \ker \Omega + V_{or} + V_{\bar{o}\bar{r}} + V_{\bar{r}o}. \tag{38}$$

Then the four subspaces in the decomposition (38) are linearly independent and in a basis corresponding to the decomposition the matrices (A, B, C) have the following structure

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & 0 & A_{44} \end{bmatrix}$$

and

$$B = \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 0 & C_2 & 0 & C_4 \end{bmatrix}.$$

Proof: The proof is similar to that of Example 3.2.9 and is left for the reader. □

REMARK 5.2.2. The realization defined by (A_{22}, B_2, C_2) is clearly reachable and observable. □

COROLLARY 5.2.3. *The transfer function* $R(s) = C(sI - A)^{-1}B$ *of a realization* (A, B, C) *depends only on the reachable and observable part, i.e. in the notation of Theorem 5.2.1 it holds that*

$$C(sI - A)^{-1}B = C_2(sI - A_{22})^{-1}B_2.$$

Proof: It is easily verified that $CA^kB = C_2A_{22}^kB_2$, which proves the corollary. □

In particular we have shown the following.

COROLLARY 5.2.4. *A minimal realization is reachable and observable.*

We shall now proceed to show that the converse of Corollary 5.2.4 also holds.

DEFINITION 5.2.5. The *McMillan degree* $\delta(R)$ of R is the dimension of a minimal realization (A, B, C, D) such that

$$R(s) = C(sI - A)^{-1}B + D.$$

In the following discussion we assume, without loss of generality, that $R(\infty) = 0$. The McMillan degree can be related to the rank of the block-Hankel matrix

$$H_r = \begin{bmatrix} R_1 & R_2 & R_3 & \cdots & R_r \\ R_2 & R_3 & R_4 & \cdots & R_{r+1} \\ R_3 & R_4 & R_5 & \cdots & R_{r+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_r & R_{r+1} & R_{r+2} & \cdots & R_{2r-1} \end{bmatrix}$$

where, as before, $R_1, R_2, R_3, ...$, are the matrix coefficients of the Laurent expansion

$$R(s) = R_1s^{-1} + R_2s^{-2} + R_3s^{-3} + ... \tag{39}$$

around $s = \infty$ and $r = \deg \mathcal{X}$, where $\mathcal{X}$ is the least common denominator of the elements of R. Let (A, B, C) be an arbitrary realization of $R(s)$ of dimension n. Then

$$\begin{aligned} R(s) &= C(sI - A)^{-1}B = s^{-1}C(I - As^{-1})^{-1}B \\ &= CBs^{-1} + CABs^{-2} + CA^2Bs^{-3} + ... \end{aligned} \tag{40}$$

By identifying the coefficients of (39) and (40) we see that (A, B, C) is a realization of $R(s)$ if and only if

$$CA^{i-1}B = R_i \quad \text{for every } i = 1, 2, 3, ...$$

The fundamental insight is now that given a realization (A, B, C), the Hankel matrix can be written as the product of an observability matrix and

a reachability matrix, i.e.

$$H_\nu = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\nu-1} \end{bmatrix} [B, AB, \cdots, A^{\nu-1}B] \tag{41}$$

for all $\nu = 1, 2, 3, \ldots$. Moreover, if $\dim A = n$ then rank $H_\nu \leq n$ for all $\nu \geq 1$.

Hence, the sequence of Hankel matrices $\{H_\nu\}$ is an invariant for the set of triplets (A, B, C) realizing a given transfer function $R(s)$. This invariance property will be exploited heavily in the rest of this chapter.

THEOREM 5.2.6. *The realization (A, B, C, D) of a matrix proper rational functions is reachable and observable if and only if it is minimal.*

Proof: Without loss of generality, assume $D = 0$. By Corollary 5.2.4 it is enough to show that reachability and observability together imply minimality. For any realization (A, B, C) we let $\Gamma_\nu \triangleq [B, AB, \cdots, A^{\nu-1}B]$ and

$$\Omega_\nu \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\nu-1} \end{bmatrix}.$$

Suppose now that (A, B, C) is a reachable and observable realization of dimension n. Then it holds that

$$\text{rank } H_n = \text{rank } \Omega_n \Gamma_n \leq \text{rank } \Gamma_n = n.$$

Moreover, since rank $\Omega_n^T \Omega_n =$ rank $\Omega_n = n$ and rank $\Gamma_n \Gamma_n^T =$ rank $\Gamma_n = n$ the matrices $\Omega_n^T \Omega_n$ and $\Gamma_n \Gamma_n^T$ are $n \times n$ and nonsingular. Hence,

$$n = \text{rank } \Omega_n^T \, \Omega_n \Gamma_n \, \Gamma_n^T \leq \text{rank } H_n.$$

Therefore, rank $H_n = n$.

For an arbitrary realization $(\tilde{A}, \tilde{B}, \tilde{C})$ of dimension $\tilde{n}$ it holds that

$$n = \text{rank } H_n = \text{rank } \tilde{\Omega}_n \tilde{\Gamma}_n \leq \text{rank } \tilde{\Gamma}_n \leq \tilde{n},$$

showing that the realization (A, B, C) is minimal. □

REMARK 5.2.7. Given an arbitrary realization Theorem 5.2.6 provides a way of checking whether it is minimal, and in the case of non-minimality we can use the Kalman decomposition to obtain a minimal realization. In particular, $\delta(R) = \dim V_{ro}$. □

There is a way to compute the McMillan degree directly from $R(s)$. We first need a lemma putting a bound on rank H_ν.

LEMMA 5.2.8. *Let r be the degree of the least common denominator $\mathcal{X}$. Then* rank $H_\nu =$ rank H_r *for all $\nu \geq r$.*

Proof: According to Lemma 5.1.6 we can successively extend the block-Hankel matrix H_r with block-rows without increasing the rank. Then we can do the same thing with the columns and, by Lemma 5.1.6, this will not increase the rank either. □

THEOREM 5.2.9. *Let $\delta(R)$ be the McMillan degree of R, and let r be the degree of the least common denominator of the elements of R. Then*

$$\delta(R) = \text{rank } H_r.$$

Proof: Let (A, B, C) be a minimal realization of dimension n of the strictly proper part of $R(s)$. As shown in the proof of Theorem 5.2.4, $n = \text{rank } H_n$. Since $\mathcal{X}(s)$ is a divisor of $\det(sI - A)$, as can be seen from Cramer's rule, it holds that $r \leq n$. Hence, by Lemma 5.2.8 rank $H_r = \text{rank } H_n = \delta(R)$. □

The following corollary is a consequence of Theorem 5.2.9.

COROLLARY 5.2.10. *Let (A, B, C, D) be a realization of $R(s)$ of dimension n, and let r be the degree of the least common denominator $\mathcal{X}(s)$ of $R(s)$. Then $r \leq n$. Furthermore, the realization is minimal if and only if the matrices*

$$\Omega_r = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{r-1} \end{bmatrix} \quad \textit{and } \Gamma_r = [B, AB, \cdots, A^{r-1}B]$$

have full rank.

REMARK 5.2.11. This result can be interpreted in a commutative diagram as in Section 2.1,Chapter 2

$$\begin{array}{ccc} U & \xrightarrow{H_r} & Y \\ & {\scriptstyle\Gamma}\searrow \quad \nearrow{\scriptstyle\Omega} & \\ & X & \end{array}$$

The goal is to find a factorization $H_r = \Omega_r \Gamma_r$ over a state space $X = \mathbb{R}^n$ of minimal dimension n. □

EXAMPLE 5.2.12. Let us apply Theorem 5.2.9 to the system defined in Example 5.1.7. Since $r = 2$, the corresponding Hankel matrix is

$$H_r = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ -1 & -2 & 1 & 2 \\ -1 & -2 & 3 & 4 \end{bmatrix}$$

and rank $H_r = 3$. Hence the McMillan degree $\delta(R) = 3$ while the standard observable and reachable realizations both have dimension 4. □

Suppose that we have a minimal realization of a given transfer function $R(s)$. Is the realization unique or are there other minimal realizations? And if there are, in what way are they related to each other? The answer is given in the *state-space isomorphism theorem.*

THEOREM 5.2.13 (State-space isomorphism theorem). *Let* (A, B, C) *and* $(\tilde{A}, \tilde{B}, \tilde{C})$ *be two minimal realizations of a strictly proper transfer function. Then there is a nonsingular matrix* T *such that*

$$(\tilde{A}, \tilde{B}, \tilde{C}) = (TAT^{-1}, TB, CT^{-1}). \tag{42}$$

Proof: If there is a T such that the theorem holds then we necessarily have that

$$\tilde{\Gamma} = T\,\Gamma.$$

In the multivariable case Γ is not square, so we cannot solve for T by employing the inverse of Γ. However, since (A, B) is reachable Γ has full row rank and therefore $\Gamma\Gamma^T$ is invertible. Hence, as a natural candidate for T pick

$$T \triangleq (\tilde{\Gamma}\Gamma^T)(\Gamma\Gamma^T)^{-1}.$$

We shall now show that this choice of T satisfies (42).

The following relations tie the two realizations together,

$$\Omega\Gamma = \tilde{\Omega}\tilde{\Gamma} \tag{43}$$

and

$$\Omega\,A\,\Gamma = \tilde{\Omega}\,\tilde{A}\,\tilde{\Gamma}. \tag{44}$$

By multiplying (43) from the right with $\Gamma^T(\Gamma\Gamma^T)^{-1}$ we get

$$\Omega = \tilde{\Omega}T \tag{45}$$

and in particular $C = \tilde{C}T$.

Since $(\tilde{C}, \tilde{A})$ is observable, $\tilde{\Omega}$ has full column rank and therefore, $\tilde{\Omega}^T\tilde{\Omega}$ is invertible. By multiplying (45) from the left with $(\tilde{\Omega}^T\tilde{\Omega})^{-1}\tilde{\Omega}^T$ it follows that another expression for T is

$$(\tilde{\Omega}^T\tilde{\Omega})^{-1}\tilde{\Omega}^T\Omega = T,$$

and by multiplying (43) from the left with the same matrix we get

$$T\Gamma = \tilde{\Gamma}$$

and in particular $TB = \tilde{B}$.

Finally, multiplying (44) from the right with $\Gamma^T(\Gamma\Gamma^T)^{-1}$ and from the left with $(\tilde{\Omega}^T\tilde{\Omega})^{-1}\tilde{\Omega}^T$ gives that $TA = \tilde{A}T$, i.e. $\tilde{A} = TAT^{-1}$. □

5.3. Characteristic polynomial and minimal realization

Suppose (A, B, C, D) is a state space realization of $R(s)$. An interesting question is if we can check the minimality without computing reachability and observability?

DEFINITION 5.3.1. The characteristic polynomial $\rho(s)$ of a proper rational matrix $R(s)$ is the least common denominator of all minors of $R(s)$. The degree of $\rho(s)$ is called the degree of $R(s)$, and denoted by *deg* $R(s)$.

EXAMPLE 5.3.2. Consider

$$R(s) = \begin{bmatrix} \frac{a}{s+2} & \frac{1}{s+2} \\ \frac{1}{s+2} & \frac{1}{s+2} \end{bmatrix}.$$

Minors of order one are the entries of $R(s)$, and there is only one minor of order 2: $\frac{a-1}{(s+2)^2}$. Hence the characteristic polynomial is $(s+2)^2$ if $a \neq 1$, and $s+2$ if $a = 1$.

EXAMPLE 5.3.3. Consider

$$R(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{1}{s+1} \\ \frac{1}{s+2} & \frac{1}{s+2} \end{bmatrix}.$$

The only minor of order 2 is 0. The characteristic polynomial is $(s+1)(s+2)$.

Form these examples we see that the characteristic polynomial is in general different from the least common denominator of all the entries, as well as from the denominator of the determinant if $R(s)$ is square.

We state the following result without giving a proof.

THEOREM 5.3.4. *A state space realization* (A, B, C, D) *is minimal if and only if*

$$dim\ A = deg\ R(s).$$

This result provides an alternative way to verify if a realization is minimal. In the following we provide a different method for computing the degree of a rational matrix.

Assume $R(s)$ can be factored as

$$R(s) = N_r(s)D_r(s)^{-1} = D_l^{-1}(s)N_l(s),$$

where $D_r(s)$ and $N_r(s)$ are right coprime, and $D_l(s)$ and $N_l(s)$ are left coprime.

Then,

$$deg\ R(s) = deg\ det(D_r(s)) = deg\ det(D_l(s)),$$

where $det(\cdot)$ stands for the determinant.

In the scalar case where

$$r(s) = \frac{n(s)}{d(s)},$$

coprimeness just implies that there is no zero-pole cancellation. Now let

$$r(s) = \frac{c_{p+1}s^p + \cdots + c_2 s + c_1}{s^n + a_n s^{n-1} + \cdots + a_1}.$$

If there is no zero-pole cancellation, we can easily derive a minimal realization as follows.

Let us introduce a new variable $z(t)$ by letting $z(s) = d^{-1}(s)u(s)$. Then $y(s) = n(s)z(s)$. Define state variables

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \begin{bmatrix} z(t) \\ \vdots \\ z^{(n-1)}(t) \end{bmatrix},$$

or,

$$x(s) = \begin{bmatrix} x_1(s) \\ \vdots \\ x_n(s) \end{bmatrix} = \begin{bmatrix} z(s) \\ \vdots \\ s^{n-1}z(s) \end{bmatrix}.$$

Then we have $\dot{x}_i = x_{i+1}$ $(sx_i(s) = x_{i+1}(s))$, $i < n$, and $\dot{x}_n = -\sum_1^n a_i x_i + u$. In summary we have

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix} x + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u$$

$$y = \begin{bmatrix} c_1 & \cdots & c_{p+1} & 0 & \cdots & 0 \end{bmatrix} x.$$

This is the so-called *reachable canonical form*. We can derive the *observable canonical form* in a similar fashion.

5.4. Ho's algorithm[1]

[1] There is a systematic way to factor H_r over a state space X of dimension $n = \text{rank } H_r$, which enables us to determine a *minimal* realization directly from the Markov parameters. We first recall a result from linear algebra.

REMARK 5.4.1. Every matrix A can be transformed by elementary row and column operations to the form $\left[\begin{smallmatrix} I_r & 0 \\ 0 & 0 \end{smallmatrix}\right]$ where $r = \text{rank } A$. Hence there are nonsingular matrices P and Q such that $PAQ = \left[\begin{smallmatrix} I_r & 0 \\ 0 & 0 \end{smallmatrix}\right]$.

We illustrate this result with a numerical example. Let

$$A = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ -1 & -2 & 1 & 2 \\ -1 & -2 & 3 & 4 \end{bmatrix}.$$

[1] This section is optional.

By first performing elementary row operations and then elementary column operations we obtain

$$\begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ -1 & -2 & 1 & 2 \\ -1 & -2 & 3 & 4 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and rank $A = 3$. The matrices P and Q are determined in the following way.

$$PA = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

$$PAQ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow Q = \begin{bmatrix} 1 & -2 & -1 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

In particular, if $A \geq 0$ and symmetric then $Q = P^T$, i.e. there is a nonsingular matrix P such that

$$PAP^T = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}. \tag{46}$$

□

We now proceed to give the construction of a minimal realization. Transform H_r by nonsingular matrices P, Q so that

$$PH_rQ = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}, \tag{47}$$

and define the shift operator σ as

$$\sigma^i\{H_r\} = \begin{bmatrix} R_{i+1} & R_{i+2} & R_{i+3} & \cdots & R_{i+r} \\ R_{i+2} & R_{i+3} & R_{i+4} & \cdots & R_{i+r+1} \\ R_{i+3} & R_{i+4} & R_{i+5} & \cdots & R_{i+r+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_{i+r} & R_{i+r+1} & R_{i+r+2} & \cdots & R_{i+2r-1} \end{bmatrix}$$

THEOREM 5.4.2 (B.L. Ho's algorithm). *Set $n = \text{rank } H_r$, and let P and Q be defined as in (47). Set*

$$\begin{aligned}
\underset{n\times n}{A} &= [I_n|0]P\sigma\{H_r\}Q\begin{bmatrix}\frac{I_n}{0}\end{bmatrix}\\
\underset{n\times k}{B} &= [I_n|0]PH_r\begin{bmatrix}I_k\\0\\\vdots\\0\end{bmatrix} = [I_n|0]P\begin{bmatrix}R_1\\R_2\\\vdots\\R_r\end{bmatrix}\\
\underset{m\times n}{C} &= [I_m,0,\cdots,0]H_rQ\begin{bmatrix}\frac{I_n}{0}\end{bmatrix} = [R_1,R_2,\cdots,R_r]Q\begin{bmatrix}\frac{I_n}{0}\end{bmatrix}
\end{aligned}$$

Then (A,B,C) is a minimal realization.

For the proof we need a sequence of lemmas.

LEMMA 5.4.3. *Let A_O and A_R be the A-matrices of the observable and reachable standard realizations respectively. Then*

$$\sigma^i\{H_r\} = A_O^iH_r = H_r(A_R^T)^i \quad \text{for all } i = 1,2,3,...$$

Proof: From Lemma 5.1.6 we see that $A_OH_r = \sigma\{H_r\}$. Repeated multiplications with A_O using Lemma 5.1.6 gives $A_O^iH_r = \sigma^i\{H_r\}$. The relation involving A_R is obtained in the same manner. □

LEMMA 5.4.4. *Define*

$$H_r^\# = Q\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}P \tag{48}$$

Then $H_r^\#$ is a pseudo inverse of H_r in the sense that

$$H_rH_r^\#H_r = H_r.$$

Proof: We have

$$H_r = P^{-1}\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}Q^{-1}$$

and therefore

$$\begin{aligned}
H_rH_r^\#H_r &= P^{-1}\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}Q^{-1}Q\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}PP^{-1}\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}Q^{-1}\\
&= P^{-1}\begin{bmatrix}I_n & 0\\0 & 0\end{bmatrix}Q^{-1} = H_r
\end{aligned}$$

□

LEMMA 5.4.5. *Let A be defined as in Theorem 5.4.2. Then*

$$A^i = [I_n|0]P\sigma^i\{H_r\}Q\begin{bmatrix}\frac{I_n}{0}\end{bmatrix} \quad \text{for all } i = 1,2,3,... \tag{49}$$

Proof: Use induction. Relation (49) holds trivially for $i = 1$. Suppose it holds for $i = j$ and show that it holds for $i = j + 1$.

$$A^{j+1} = AA^j = [I_n|0]P\sigma\{H_r\}Q \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} P\sigma^j\{H_r\}Q \begin{bmatrix} I_n \\ \hline 0 \end{bmatrix}$$

Then replacing $\sigma\{H_r\}$ by $A_O H_r$ and $\sigma^j\{H_r\}$ by $H_r(A_R^T)^j$ Lemma 5.4.3 and observing (48) we obtain

$$A^{j+1} = [I_n|0]PA_O H_r H_r^{\#} H_r (A_R^T)^j Q \begin{bmatrix} I_n \\ \hline 0 \end{bmatrix}$$

In view of Lemma 5.4.4 and Lemma 5.4.3,

$$A_O H_r H_r^{\#} H_r (A_R^T)^j = A_O H_r (A_R^T)^j = H_r (A_R')^{j+1} = \sigma^{j+1}\{H_r\}$$

and consequently.

$$A^{j+1} = [I_n|0]P\sigma^{j+1}\{H_r\}Q \begin{bmatrix} I_n \\ \hline 0 \end{bmatrix}$$

□

Proof of Theorem 5.4.2 Since $\dim(A, B, C) = \delta(R)$ it only remains to show that $R(s) = C(sI - A)^{-1}B$, i.e. $R_i = CA^{i-1}B \quad$ for $i = 1, 2, 3, \ldots$. With (A, B, C) defined as in the theorem and A^{i-1} as given in Lemma 5.4.5, we have

$$CA^{i-1}B = [I_m, 0, \cdots, 0]H_r Q \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} P\sigma^{i-1}\{H_r\}Q \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} PH_r \begin{bmatrix} I_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then, replacing $\sigma^{i-1}\{H_r\}$ by $A_O^{i-1}H_r$ (Lemma 5.4.3), observing the definition (48) of $H_r^{\#}$, and applying Lemma 5.4.4 we have

$$\begin{aligned} CA^{i-1}B &= [I_m, 0, \cdots, 0]H_r H_r^{\#} A_O^{i-1} H_r \begin{bmatrix} I_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= [I_m, 0, \cdots, 0]\sigma^{i-1}\{H_r\} \begin{bmatrix} I_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} = R_i \end{aligned}$$

because, in view of Lemma 5.4.3 and Lemma 5.4.4,

$$H_r H_r^{\#} A_O^{i-1} H_r = H_r H^{\#} H_r (A_R^T)^{i-1} = H_r (A_R^T)^{i-1} = \sigma^{i-1}\{H_r\}.$$

□

EXAMPLE 5.4.6. Apply B.L. Ho's algorithm to $R(s)$ as defined in Example 5.1.7. Note that P and Q are determined in Remark 5.4.1 as

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1 & 0 & 1 & 0 \end{bmatrix} \quad Q = \begin{bmatrix} 1 & -2 & -1 & -1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Then, the algorithm of Theorem 5.4.2 yields a minimal realization with (A, B, C) given by

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} -1 & -2 & 1 & 2 \\ -1 & -2 & 3 & 4 \\ 1 & 2 & -1 & -2 \\ 3 & 4 & -7 & -8 \end{bmatrix} \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 0 & 2 \\ 1 & -1 & -3 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -1 & -2 \\ -1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 2 & -1 & -2 \\ 0 & 1 & -1 & -2 \end{bmatrix} \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

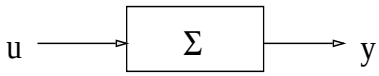
i.e. we have the following minimal realization

$$\begin{cases} \dot{x}_1 = -x_1 + u_1 + 2u_2 \\ \dot{x}_2 = -x_1 + u_2 \\ \dot{x}_3 = x_1 - x_2 - 3x_3 \\ y_1 = x_1 \\ y_2 = x_2 \end{cases}$$

CHAPTER 6

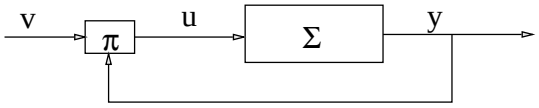
State Feedback and Observers

Suppose that the time-invariant system



$$(\Sigma) \quad \begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases}$$

is unstable. Is it possible to stabilize the system by linear feedback?



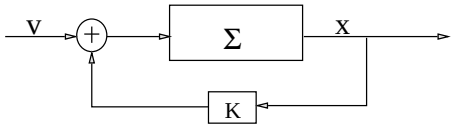
More generally, given any linear system Σ, is it possible to move the poles of the transfer function to any preassigned location? We shall allow the feedback function π to be either multiplication by a constant matrix, or, more generally, itself a linear system of type Σ.

6.1. Feedback with Complete State Information

First let us consider the case when the state x is available for observation, i.e., $y = x$. Then, it is possible to choose a feedback of the type

$$u(t) = Kx(t) + v(t)$$

where K is a constant matrix and v may be a function of time t.



The closed-loop system can be written

$$\dot{x} = (A + BK)x + Bv,$$

Hence (A, B) is changed to $(\hat{A}, B)$ where $\hat{A} = A + BK$.

6.1.1. Invariance of reachability. The reachability properties of the system are not changed by such feedback, i.e., the reachable subspace is invariant under state feedback. Denoted by $\mathcal{R}$ and $\mathcal{R}_K$ the reachable subspaces

of the open-loop and closed-loop systems respectively, i.e.,

$$\begin{aligned}\mathcal{R} &= \langle A|\,\mathrm{Im}\ B\rangle \triangleq \mathrm{Im}\,[B, AB, A^2B, \cdots, A^{n-1}B] \\ &= \mathrm{Im}\ B + A\,\mathrm{Im}\ B + \cdots + A^{n-1}\,\mathrm{Im}\ B \\ \mathcal{R}_K &= \langle A+BK|\,\mathrm{Im}\ B\rangle\end{aligned}$$

REMARK 6.1.1. If U and V are subspaces then $U+V$ is defined as

$$U+V = \{u+v\,|\,u\in U, v\in V\}.$$

□

LEMMA 6.1.2. *For every K we have*

$$\mathcal{R}_K = \mathcal{R}.$$

Proof: Set $\hat{A} = A+BK$. Then, since $A\mathcal{R}\subset\mathcal{R}$ and $\mathrm{Im}\ B\subset\mathcal{R}$,

$$\hat{A}\mathcal{R} = A\mathcal{R} + BK\mathcal{R}\subset\mathcal{R},$$

i.e., $\hat{A}\mathcal{R}\subset\mathcal{R}$ and therefore

$$\mathcal{R}_K = \mathrm{Im}\ B + \mathrm{Im}\ \hat{A}B + \mathrm{Im}\ \hat{A}^2B + \cdots + \mathrm{Im}\ \hat{A}^{n-1}B \subset \mathcal{R}$$

Similarly, noting that $A = \hat{A} - BK$, we show that $\mathcal{R}\subset\mathcal{R}_K$. Hence $\mathcal{R} = \mathcal{R}_K$ as claimed. □

6.1.2. Pole-placement. Let $\mathcal{X}_A(s)$ be the characteristic polynomial

$$\mathcal{X}_A(s) = \det(sI - A)$$

of the system matrix A. By feedback the system matrix is changed to $\hat{A} = A+BK$. The question is now if K can be chosen so that $\hat{A}$ is a stable matrix, i.e. so that the zeros of the characteristic polynomial

$$\mathcal{X}_{A+BK}(s) = \det(sI - A - BK)$$

are all in the open left half plane. The answer is that the zeros of $\mathcal{X}_A(s)$, i.e., the poles of

$$R(s) = C(sI-A)^{-1}B + D,$$

can be moved to an arbitrary location in the complex plane, except that complex poles have to appear in pairs (i.e. $a\pm ib$), if and only if the system is completely reachable.

THEOREM 6.1.3. *(Pole-placement Theorem) For any real polynomial*

$$\varphi(s) = s^n + \gamma_1 s^{n-1} + \cdots + \gamma_n$$

there is a matrix K such that

$$\mathcal{X}_{A+BK} = \varphi$$

if and only if (A,B) is completely reachable.

We first consider the case when u is scalar.

LEMMA 6.1.4. *Let $b \in \mathbb{R}^n$ and let (A, b) be completely reachable, i.e. $\langle A|\,\mathrm{Im}\ b\rangle = \mathbb{R}^n$. Then there is a nonsingular $n \times n$ matrix T such that*

$$TAT^{-1} = F \triangleq \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \dots & -a_1 \end{bmatrix} \quad Tb = h \triangleq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

where $a_1, a_2, \cdots, a_n$ are the coefficients of

$$\mathcal{X}_A(s) = s^n + a_1 s^{n-1} + \cdots + a_n.$$

The transformation T is unique and is given by

$$T = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} \tag{50}$$

where c is the unique row vector solution to

$$c[b, Ab, \cdots, A^{n-1}b] = (0, 0, \cdots, 0, 1). \tag{51}$$

Proof: Since (A, b) is completely reachable, there is a unique solution to (51). The system (51) can be written

$$\begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

or equivalently, $Tb = h$. To see that T is nonsingular, we show that the rows are linearly independent. To this end, let $\alpha_1, \alpha_2, \dots \alpha_n$ be real numbers such that

$$\alpha_1 c + \alpha_2 cA + \cdots + \alpha_n cA^{n-1} = 0,$$

multiply from the right by b and use (51). This yields $\alpha_n = 0$. Then multiply by Ab which yields $\alpha_{n-1} = 0$ and then by A^2b which yields $\alpha_{n-2} = 0$ etc. Since therefore $\alpha_1 = \alpha_2 = ... = \alpha_n = 0$, the rows of T are linearly independent. Set $T^{-1} = [s_1, s_2, \cdots, s_n]$ where s_i are the columns. Then

$$TT^{-1} = \begin{bmatrix} cs_1 & cs_2 & \dots & cs_n \\ cAs_1 & cAs_2 & \dots & cAs_n \\ \vdots & \vdots & \vdots & \\ cA^{n-1}s_1 & cA^{n-1}s_2 & \dots & cA^{n-1}s_n \end{bmatrix}$$

and from this it follows that

$$TAT^{-1} = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} A(s_1, s_2, \cdots, s_n) = \begin{bmatrix} cAs_1 & cAs_2 & \dots & cAs_n \\ cA^2s_1 & cA^2s_2 & \dots & cA^2s_n \\ \vdots & \vdots & \vdots & \\ cA^ns_1 & cA^ns_2 & \dots & cA^ns_n \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ cA^ns_1 & cA^ns_2 & \dots & \dots & cA^ns_n \end{bmatrix}.$$

But, by the Cayley-Hamilton theorem we have

$$cA^ns_i = -\sum_{j=1}^{n} a_j cA^{n-j}s_i = -a_{n+1-i},$$

because $cA^{n-j}s_i = 1$ if $n - j = i - 1$ and 0 otherwise. This completes the proof of the lemma. □

Suppose now that (A, b) is completely reachable and that the system

$$\dot{x} = Ax + bu$$

with a scalar input is to be stabilized by the feedback

$$u = kx + v$$

where K is a constant row vector. Then, the feedback matrix becomes

$$A + bk = T^{-1}FT + T^{-1}hk = T^{-1}[F + hg]T$$

where

$$g = (g_n, g_{n-1}, \cdots, g_1) \triangleq kT^{-1}$$

or $k = gT$. Note the reversed numbering of the elements in the row vector g. Then, since similar matrices have the same characteristic polynomial, we have

$$\mathcal{X}_{A+bk} = \mathcal{X}_{F+hg}$$

However,

$$F + hg = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \dots & -a_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \\ g_n & g_{n-1} & g_{n-2} & \cdots & g_1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ -\gamma_n & -\gamma_{n-1} & -\gamma_{n-2} & \cdots & -\gamma_1 \end{bmatrix}$$

that is $\mathcal{X}_{F+hg} = \varphi$ if

$$g_i = a_i - \gamma_i, \quad i = 1, 2, \cdots, n.$$

EXAMPLE 6.1.5. Consider the system

$$\begin{cases} \dot{x}_1 = x_1 - 3x_2 + u \\ \dot{x}_2 = 4x_1 + 2x_2 + u \end{cases}$$

Then

$$A = \begin{bmatrix} 1 & -3 \\ 4 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathcal{X}_A(s) = s^2 - 3s + 14,$$

i.e., the poles are located at $s = \frac{3}{2} \pm i\frac{\sqrt{47}}{2}$. Is it possible to place the poles at $s = -1$ and $s = -2$ instead? Yes, because

$$[b, Ab] = \begin{bmatrix} 1 & -2 \\ 1 & 6 \end{bmatrix}$$

has full rank. The required closed-loop characteristic polynomial

$$\varphi(s) = (s+1)(s+2) = s^2 + 3s + 2,$$

i.e., $\gamma_1 = 3, \gamma_2 = 2$, and therefore we should choose g as

$$g_1 = a_1 - \gamma_1 = -3 - 3 = -6$$
$$g_2 = a_2 - \gamma_2 = 14 - 2 = 12$$

Next, we determine the transformation T:

$$(c_1, c_2)\begin{bmatrix} 1 & -2 \\ 1 & 6 \end{bmatrix} = (0, 1) \Rightarrow \begin{cases} c_1 + c_2 & = 0 \\ -2c_1 + 6c_2 & = 1 \end{cases} \Rightarrow t = (-\frac{1}{8}, \frac{1}{8})$$

i.e.

$$T = \begin{bmatrix} c \\ cA \end{bmatrix} = \frac{1}{8}\begin{bmatrix} -1 & 1 \\ 3 & 5 \end{bmatrix} \quad T^{-1} = \begin{bmatrix} -5 & 1 \\ 3 & 1 \end{bmatrix}$$

As a check, we compute

$$TAT^{-1} = \begin{bmatrix} 0 & 1 \\ -14 & 3 \end{bmatrix}, \quad Tb = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The required gain is

$$k = gT = (12, -6)\frac{1}{8}\begin{bmatrix} -1 & 1 \\ 3 & 5 \end{bmatrix} = (-\frac{15}{4}, -\frac{9}{4}).$$

Hence, the feedback

$$u = -\frac{15}{4}x_1 - \frac{9}{4}x_2 + v$$

achieves the desired pole placement. □

It remains to prove Theorem 6.1.3 (*Pole-placement Theorem*) for the case that there are more than one input, as well as the necessity of the reachability condition. For this we need another lemma.

LEMMA 6.1.6 (Heymann). *Let $b \in \text{Im } B$ and $b \neq 0$. If (A, B) is completely reachable, there is a matrix K such that $(A + BK, b)$ is completely reachable.*

Proof: We have to show that there is a matrix K such that

$$\text{rank}\,[b, (A+BK)b, (A+BK)^2 b, ..., (A+BK)^{n-1}b] = n. \tag{52}$$

To this end, consider the discrete-time system

$$x(t+1) = Ax(t) + Bu(t),\ x(1) = b. \tag{53}$$

By Lemma 3.3.2 there is an input sequence $\{\tilde{u}(1), \tilde{u}(2), ..., \tilde{u}(n-1)\}$ such that the corresponding trajectory $\{x(1), x(2), ..., x(n)\}$ of the system spans the state space. The idea is now to express the input $\tilde{u}(t)$ as state-feedback $Kx(t)$.

Introduce the matrices $Z \triangleq [x(1), x(2), ..., x(n)]$, which is invertible since $\{x(1), x(2), ..., x(n)\}$ is a basis, and $U \triangleq [\tilde{u}(1), \tilde{u}(2), ..., \tilde{u}(n)]$, where $\tilde{u}(n)$ is arbitrary . Then we have

$$U = UZ^{-1}Z = KZ, \tag{54}$$

where K is defined as $K \triangleq UZ^{-1}$. From (54) we then see that $\tilde{u}(t) = Kx(t)$, and the solution of (53) with input $\{\tilde{u}(1), \tilde{u}(2), ..., \tilde{u}(n)\}$ can be obtained by iterating $x(t+1) = (A+BK)x(t)$, $x(1) = b$. This implies that $x(t) = (A+BK)^{t-1}b$ and (52) is established. □

Proof of Theorem 6.1.3

If: The scalar-input case has already been proved above. Therefore, we shall reduce the general case that B have several columns to the scalar-input case. Let $b \in \text{Im } B$ and $b \neq 0$. Then there is a vector $u_0 \in \mathbb{R}^k$ such that $b = Bu_0$. Since (A, B) is completely reachable, there is, according to Lemma 6.1.6, a $k \times n$ matrix K such that $(A + B\hat{K}, b)$ is completely reachable. Then, in view of the scalar-input result, there is a row vector k such that

$$\mathcal{X}_{A+B\hat{K}+bk}(s) = \varphi(s)$$

But $B\hat{K} + bk = B(\hat{K} + u_0 k)$, and consequently the desired pole placement is achieved through a feedback with the gain $K = \hat{K} + u_0 k$.

Only If: Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be different real numbers such that $\lambda_i > \|A\|$ for $i = 1, 2, \ldots, n$, and suppose that K is such that

$$\mathcal{X}_{A+BK}(s) = (s-\lambda_1)(s-\lambda_2)\cdots(s-\lambda_n).$$

Then, the matrix $A+BK$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Let $\{z_1, z_2 \ldots, z_n\}$ be the corresponding eigenvectors, i.e.

$$(A+BK)z_i = \lambda_i z_i$$

The eigenvectors form a basis in $X = \mathbb{R}^n$. However, for each $i = 1, 2, \ldots, n$,

$$z_i = (\lambda_i I - A)^{-1}BKz_i = \lambda_i^{-1}(I - A\lambda_i^{-1})^{-1}BKz_i = \sum_{j=0}^{\infty} \lambda_i^{-j-1}A^j BKz_i$$

where we have used the fact that $\|A\lambda_i^{-1}\| < 1$ when expanding in series. Therefore, in view of Remark 3.2.2, $z_i \in \mathcal{R}$ for $i = 1, 2, \ldots, n$, where $\mathcal{R}$ is the reachable subspace corresponding to (A, B), and consequently

$$X = \text{span}\,\{z_1, z_2, \ldots, z_n\} \subset \mathcal{R} \subset \mathbb{R}^n.$$

Therefore, since $X = \mathbb{R}^n$, we must have $\mathcal{R} = \mathbb{R}^n$, i.e. (A, B) is completely reachable.

□

6.2. Observers

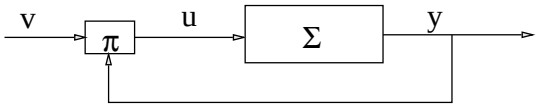
Consider a system where only the output is measurable, i.e.

$$(\Sigma) \quad \begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases}$$

where rank $C < n$. Is it possible to change the poles of

$$R(s) = C(sI - A)^{-1}B$$

by output feedback $u = Ky + v$? The answer is in general no.



However, if dynamical output feedback is used, namely we feed the output as input to an artificially constructed dynamical system and use the state of that system for control design, then pole-placement for the original system is possible when it is viewed as part of this enlarged system. The key idea is that this artificially constructed system will be an estimator of the true state $x(t)$:

$$\frac{d\hat{x}}{dt} = A\hat{x} + Bu + L(y - C\hat{x})$$

where L is a constant matrix to be chosen so that

$$\|x(t) - \hat{x}(t)\| \to 0 \text{ as } t \to \infty$$

We want $\hat{x}$ to adjust asymptotically to the state x. Set $\tilde{x} \triangleq x - \hat{x}$. The estimation error $\tilde{x}$ satisfies the homogeneous system of differential equations

$$\frac{d\tilde{x}}{dt} = (A - LC)\tilde{x}$$

Hence $\tilde{x}(t) \to 0$ when $t \to \infty$ if and only if $(A - LC)$ is a stable matrix, i.e. $\mathcal{X}_{A-LC}$ has all its zeros in the (open) left half plane.

THEOREM 6.2.1. *For any real polynomial*

$$\varphi(s) = s^n + \gamma_1 s^{n-1} + \cdots + \gamma_n$$

there is a matrix L such that

$$\mathcal{X}_{A-LC} = \varphi$$

if and only if (C, A) is completely observable.

Proof: To say that (C, A) is completely observable is equivalent to say that (A', C') is completely reachable which, by Theorem 6.1.3 establishes the existence of a K such that $\mathcal{X}_{A'+C'K} = \varphi$ for any φ, i.e. $\mathcal{X}_{A+K'C} = \varphi$. The theorem follows if we set $L = -K'$. □

In particular, we can choose L so that $A - LC$ is a stable matrix, i.e. $\tilde{x}(t) \to 0$. Using the feedback law $u = K\hat{x} + v$ yields the over-all system

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A & BK \\ LC & A - LC + BK \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + Bv \begin{bmatrix} I \\ I \end{bmatrix}.$$

Let us do a linear coordinate change by letting $\tilde{x} = x - \hat{x}$. Then

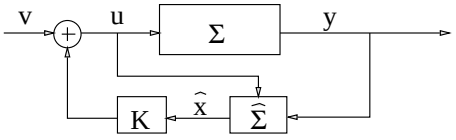
$$\begin{bmatrix} \dot{x} \\ \dot{\tilde{x}} \end{bmatrix} = \begin{bmatrix} A + BK & -BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x \\ \tilde{x} \end{bmatrix} + Bv \begin{bmatrix} I \\ 0 \end{bmatrix},$$

which implies

$$\mathcal{X}_F = \mathcal{X}_{A+BK}\mathcal{X}_{A-LC}$$

since for every matrix M which can be partitioned $M = \left[\begin{smallmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{smallmatrix}\right]$ where M_{11} and M_{22} are quadratic, it holds that $\det M = \det M_{11} \det M_{22}$. The over-all system is then stable if and only if both $A + BK$ and $A - LC$ are stable matrices.

The answer to the original question is then: If Σ is a minimal realization then the system is stabilizable and it can be stabilized as follows



The feedback law, defined by K, and the observer, defined by L, can be determined independently. The dimension of the observer can be further reduced to $n - \operatorname{rank} C$ and yet be made stable, but we shall not pursue this matter here.

REMARK 6.2.2. This result that the poles of the estimator and the system can be assigned independently is usually called the *separation principle*. A more general separation theorem is valid in the context of Optimal (Kalman) filtering and Optimal control. □

We note that the overall system is, however, never minimal. This can easily be verified by calculating the transfer function from v to y. In fact, the overall system is never reachable.

EXAMPLE 6.2.3. Consider Example 6.1.5 with output $y = 2x_1 + 3x_2$. Is it possible to stabilize the system with an observer? Yes, because

$$\begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 14 & 0 \end{bmatrix}$$

is full rank, and we can consequently assign the poles arbitrarily (in pair if they are complex). Construct an observer with poles $s = -3 \pm i$, i.e.,

$$\varphi(s) = (s + 3 + i)(s + 3 - i) = s^2 + 6s + 10, \quad \gamma_1 = 6, \quad \gamma_2 = 10.$$

Then, determine a control law $k = -L^T$ as above for a system for which A^T and C^T play roles of A and b respectively, i.e.,

$$A \leftarrow A^T = \begin{bmatrix} 1 & 4 \\ -3 & 2 \end{bmatrix} \text{ and } b \leftarrow C^T = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

yielding a characteristic polynomial

$$\mathcal{X}_A(s) = s^2 - 3s + 14$$

which is the same as before. Hence,

$$\begin{cases} g_1 = a_1 - \gamma_1 = -3 - 6 = -9 \\ g_2 = a_2 - \gamma_2 = 14 - 10 = 4 \end{cases}$$

$$(c_1, c_2) \begin{bmatrix} 2 & 14 \\ 3 & 0 \end{bmatrix} = (0, 1) \Rightarrow \begin{cases} 2c_1 + 3c_2 = 0 \\ 14c_1 = 1 \end{cases} \Rightarrow c = (\frac{1}{14}, -\frac{1}{21})$$

$$T = \frac{1}{42} \begin{bmatrix} 3 & -2 \\ 9 & 8 \end{bmatrix}$$

$$k = gT = (4, -9)\frac{1}{42} \begin{bmatrix} 3 & -2 \\ 9 & 8 \end{bmatrix} = (-\frac{69}{42}, -\frac{40}{21})$$

$$L = -k^T = \begin{bmatrix} \frac{69}{42} \\ \frac{40}{21} \end{bmatrix}$$

The dynamical feedback is then

$$\begin{cases} u = -\frac{15}{4}\hat{x}_1 - \frac{9}{4}\hat{x}_2 + v \\ \frac{d\hat{x}}{dt} = \begin{bmatrix} 1 & -3 \\ 4 & 2 \end{bmatrix} \hat{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u + \begin{bmatrix} \frac{69}{42} \\ \frac{40}{21} \end{bmatrix} (y - 2\hat{x}_1 - 3\hat{x}_2). \end{cases}$$

□

CHAPTER 7

Linear-Quadratic Optimal Control

Whenever possible we want to control a system in an optimal way. To do this there has to be a criterion defining what is best. Sometimes it is important to minimize the transfer time between two states, sometimes to minimize the control effort (say the energy used), the deviation from some nominal trajectory, or a combination of these. The literature on optimal control is very extensive and this chapter only treats the most fundamental techniques.

7.1. Linear-Quadratic regulator

This method is widely used and implemented in many working control systems such as the space shuttle and missiles. It is often relatively easy to implement and the corresponding theoretical principles are well understood. Given a system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ x(t_0) = x_0 \end{cases} \tag{55}$$

on the interval $[t_0, t_1]$, the goal is to determine the control u which minimizes the functional

$$J(u) = \int_{t_0}^{t_1} \left[x(t)^T Q x(t) + u(t)^T R u(t) \right] dt + x(t_1)^T S x(t_1), \tag{56}$$

where $Q \geq 0,\ R > 0,\ S \geq 0$ are symmetric. Preferably, we would like to determine the optimal u as a *control law*, i.e., a feedback function through which the control signal depends not only on t but also on the state x. This is of great importance in stabilizing a system subject to disturbances. A small perturbation of the state can result in a large error if the control is not taking into account possible deviations from a precomputed 'optimal trajectory'. A control law which depends on the state as well of time is less sensitive to disturbances; it is more robust. As a consequence of the special structure of this problem the optimal solution $\hat{u}$ is in fact expressible in feedback form as $\hat{u} = \hat{u}(x, t)$. In the following sections we present two different approaches for deriving the optimal control.

7.1.1. Completion of the square. Let $P \geq 0$ be a differentiable symmetric $n \times n$ matrix function and put

$$V(x,t) \triangleq x^T P(t) x.$$

The function $V(x,t)$ is to be interpreted as the optimal cost function giving the optimal value of (56) starting at $x_0 = x$ and $t_0 = t$. A rational for choosing a quadratic function could be that this is the simplest form that gives a non-negative function. What follows is to show that this choice is well founded.

Define the function $t \to V(x(t),t)$ where $x(t)$ is the solution of (55) and differentiate with respect to t

$$\begin{aligned}\frac{d}{dt}V(x(t),t) &= \dot{x}^T P x + x^T P \dot{x} + x^T \dot{P} x \\ &= x^T A^T P x + u^T B^T P x + x^T P A x + x^T P B u + x^T \dot{P} x\end{aligned}$$

Then integrating from t_0 to t_1 yields

$$V(x(t_1),t_1) - V(x_0,t_0) = \int_{t_0}^{t_1} (u^T B^T P x + x^T P B u)\, dt + \int_{t_0}^{t_1} x^T (A^T P + P A + \dot{P}) x \, dt.$$

Add this to the functional (56), subtract $V(x(t_1),t_1)$ and complete the square to obtain

$$\begin{aligned}J(u) - V(x_0,t_0) &= \int_{t_0}^{t_1} (u + R^{-1}B^T P x)^T R (u + R^{-1} B^T P x)\, dt \\ &+ \int_{t_0}^{t_1} x^T (A^T P + P A + \dot{P} + Q - P B R^{-1} B^T P) x \, dt + x(t_1)^T [S - P(t_1)] x(t_1)\end{aligned}$$

This rather complicated expression can be considerably simplified if we choose P to satisfy the matrix valued differential equation

$$\text{(RE)} \qquad \begin{cases} \dot{P} & = -A^T P - P A + P B R^{-1} B^T P - Q \\ P(t_1) & = S \end{cases}$$

which is the matrix *Riccati equation* (RE). The following theorem ensures that there really is such a P and that it is unique.

THEOREM 7.1.1. (RE) *has a unique solution P on the interval $[t_0, t_1]$, which is positive semidefinite and bounded.*

We skip the proof which is based on standard results in the theory of differential equations. Note, however, that it is essential that Q and S are positive semidefinite and that R is positive definite.

Choose therefore P to be a solution to (RE), then

$$J(u) = V(x_0,t_0) + \int_{t_0}^{t_1} (u + R^{-1}B^T P x)^T R (u + R^{-1} B^T P x)\, dt \geq V(x_0,t_0)$$

with equality if and only if

$$u = -R^{-1} B^T P x \tag{57}$$

The control gain $K = R^{-1}B^TP$ is called the *Kalman gain*. Hence the optimal value of $J(u)$ is obtained by applying the control law (57) leading to the feedback configuration

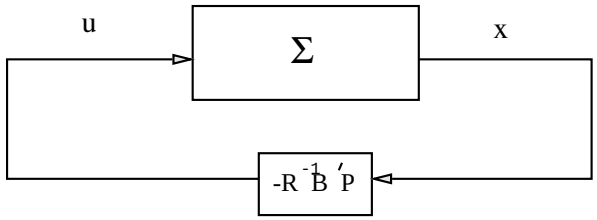


Figure 7.1

and $V(x,t)$ is indeed the optimal cost function. Thus closing the loop results in the new system

$$\begin{cases} \dot{x} = (A + BK)x \\ x(t_0) = x_0 \end{cases} \tag{58}$$

which we may solve for x to yield the optimal control

$$u(t) = -R^{-1}(t)B^TP(t)\Phi_K(t,t_0)x_0 \tag{59}$$

explicitly as a function of time. [Here, of course, Φ_K is the transition matrix function of the closed loop system (58).] Although applying the *open-loop control* (59) to the system (55) theoretically leads to the same results as the feedback (57) implemented as in Figure 7.1, open-loop and closed-loop control may have drastically different performance from the point of view of stability and robustness. The reason for this is better understood by looking at the optimization problem from the point of view of dynamic programming.

7.1.2. Dynamic programming, heuristics. Dynamic programming provides a different approach for optimal control, which is based on Bellman's principle of optimality:

An optimal control has the property that no matter what the previous control has been, the remaining control must constitute an optimal control with regard to the state resulting from the previous control.

This heuristics of the dynamic programming procedure is illustrated in Figure 7.2, where the optimization problem starting from (t_0, x_0) is embedded in the class of problems obtained by varying the initial state and initial time. Thus, let the system start in x at time t and control the system with some control u. At $t + h$ the system is in the corresponding state $x(t + h)$.

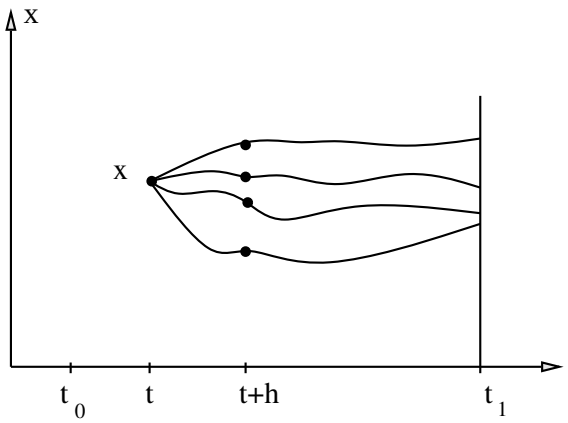


Figure 7.2

Define the optimal cost function as

$$V(a,t) \triangleq \min_u \Big\{ \int_t^{t_1} (x^T Q x + u^T R u)\, ds + x^T(t_1) S x(t_1) \Big\}$$

$$\text{where } \begin{cases} \dot{x} = Ax + Bu \\ x(t) = a \end{cases}$$

Bellman's principle of optimality yields

$$\begin{aligned} V(x,t) &= \min_u \Big\{ \int_{t+h}^{t_1} (x^T Q x + u^T R u)\, ds + x^T(t_1) S x(t_1) + \int_t^{t+h} (x^T Q x + u^T R u)\, ds \Big\} \\ &= \min_u \Big\{ V(x(t+h), t+h) + \int_t^{t+h} (x^T Q x + u^T R u)\, ds \Big\} \end{aligned}$$

from which we have

$$\min_u \Big\{ \frac{1}{h} \big[V(x(t+h), t+h) - V(x(t), t) \big] + \frac{1}{h} \int_t^{t+h} (x^T Q x + u^T R u)\, ds \Big\} = 0.$$

Formally, letting $h \to 0$, we obtain the functional equation for $V(x,t)$

$$\min_{u \in \mathbb{R}^k} \Big\{ \frac{d}{dt} V(x(t), t) + x^T Q x + u^T R u \Big\} = 0, \tag{60}$$

i.e. we get a finite-dimensional optimization problem for each t. Once again we assume

$$V(x,t) = x^T P(t) x$$

for some $P \geq 0$. Then

$$\frac{d}{dt} V(x,t) = \dot{x}^T P x + x^T P \dot{x} + x^T \dot{P} x$$

so (60) may be written

$$\min_{u \in \mathbb{R}^k} \big\{ x^T (A^T P + P A + \dot{P} + Q) x + u^T B^T P x + x^T P B u + u^T R u \big\} = 0.$$

If P satisfies (RE), we get,

$$\min_{u\in\mathbb{R}^k}\left\{(u+R^{-1}B^TPx)^TR(u+R^{-1}B^TPx)\right\}=0$$

where the minimum is attained for $u=-R^{-1}B^TPx$, the optimal control (57) determined above.

REMARK 7.1.2. The equation (60) is a special case of the Hamilton-Jacobi-Bellman equation of optimal control and the derivation presented here is by no means mathematically rigorous. The proper theorem of dynamic programming is the so called "verification theorem", which will be presented in the course "Optimal Control".

Hence, we can regard the feedback control (57) as the optimal control required at time t to minimize *future* cost with the present state as initial condition, i.e. we apply the best possible control given *present conditions*.

7.2. Solving the Riccati equation

As the previous discussion has shown, the key for obtaining an optimal control is to solve the corresponding Riccati equation.

Let $X(t)$ be the regular matrix solution of the feedback system

$$\begin{cases}\dot{X}=(A-BR^{-1}B^TP)X\\ X(t_1)=I\end{cases}$$

and define the matrix function Y as $Y=PX$. Then

$$\begin{aligned}\dot{Y}&=\dot{P}X+P\dot{X}\\&=-A^TY-PAX+PBR^{-1}B^TY-QX+PAX-PBR^{-1}B^TY\\&=-A^TY-QX\end{aligned}$$

Since X^{-1} exists for all t

$$P=YX^{-1}$$

and consequently (RE) can be replaced by the so-called *adjoint system*

$$\begin{cases}\dot{X}=AX-BR^{-1}B^TY;\quad X(t_1)=I\\ \dot{Y}=-QX-A^TY;\quad Y(t_1)=S.\end{cases}\tag{61}$$

EXAMPLE 7.2.1. A missile will intercept an aircraft in time t_1-t_0. Let x_1 be the distance from a nominal trajectory, as depicted in Figure 7.3.

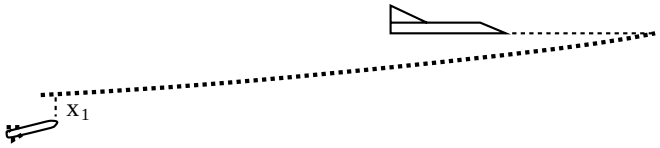


Figure 7.3

Newton's law describes this motion as $\ddot{x}_1 = u$ where the control u is proportional to the force perpendicular to the trajectory. Hence, setting $x_2 = \dot{x}_1$, the system may be written

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

We wish to minimize the square of the distance from the target at the time t_1 of interception added to an integral term representing the fuel burnt.

$$\min\{x_1(t_1)^2 + \int_{t_0}^{t_1} u(t)^2\, dt\} \qquad \int u^2 \sim \text{ used fuel}$$

Setting

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad Q = 0 \quad R = 1 \quad S = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

the Riccati equation becomes

$$\begin{bmatrix} \dot{p}_{11} & \dot{p}_{12} \\ \dot{p}_{21} & \dot{p}_{22} \end{bmatrix} = - \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} - \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} [0, 1] \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

$P(t_1) = S$
which leads to the system of three differential equations

$$\begin{cases} \dot{p}_{11} = p_{12}^2 & p_{11}(t_1) = 1 \\ \dot{p}_{12} = -p_{11} + p_{12}p_{22} & p_{12}(t_1) = 0 \\ \dot{p}_{22} = -2p_{12} + p_{22}^2 & p_{22}(t_1) = 0 \end{cases}$$

It is not easy to solve this system of nonlinear differential equations by brute force. Instead we solve the linear system (61) of eight differential equations.

$$\begin{bmatrix} \dot{x}_{11} & \dot{x}_{12} \\ \dot{x}_{21} & \dot{x}_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \quad X(t_1) = I$$

$$\begin{bmatrix} \dot{y}_{11} & \dot{y}_{12} \\ \dot{y}_{21} & \dot{y}_{22} \end{bmatrix} = - \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \quad Y(t_1) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

i.e.

$$\begin{cases} \dot{x}_{11} = x_{21} & x_{11}(t_1) = 1 \\ \dot{x}_{12} = x_{22} & x_{12}(t_1) = 0 \\ \dot{x}_{21} = -y_{21} & x_{21}(t_1) = 0 \\ \dot{x}_{22} = -y_{22} & x_{22}(t_1) = 1 \\ \dot{y}_{11} = 0 & y_{11}(t_1) = 1 \\ \dot{y}_{12} = 0 & y_{12}(t_1) = 0 \\ \dot{y}_{21} = -y_{11} & y_{21}(t_1) = 0 \\ \dot{y}_{22} = -y_{12} & y_{22}(t_1) = 0 \end{cases} \Rightarrow \begin{cases} x_{11}(t) = 1 - \frac{1}{6}(t_1 - t)^3 \\ x_{12}(t) = -(t_1 - t) \\ x_{21}(t) = \frac{1}{2}(t_1 - t)^2 \\ x_{22}(t) = 1 \\ y_{11}(t) = 1 \\ y_{12}(t) = 0 \\ y_{21}(t) = t_1 - t \\ y_{22}(t) = 0 \end{cases}$$

Therefore we have

$$X(t) = \begin{bmatrix} 1 - \frac{1}{6}(t_1 - t)^3 & -(t_1 - t) \\ \frac{1}{2}(t_1 - t)^2 & 1 \end{bmatrix}$$
$$Y(t) = \begin{bmatrix} 1 & 0 \\ t_1 - t & 0 \end{bmatrix}$$

and

$$X(t)^{-1} = \frac{1}{1 + \frac{1}{3}(t_1 - t)^3} \begin{bmatrix} 1 & t_1 - t \\ -\frac{1}{2}(t_1 - t)^2 & 1 - \frac{1}{6}(t_1 - t)^3 \end{bmatrix}$$

Consequently

$$P(t) = Y(t)X(t)^{-1} = \frac{1}{1 + \frac{1}{3}(t_1 - t)^3} \begin{bmatrix} 1 & t_1 - t \\ t_1 - t & (t_1 - t)^2 \end{bmatrix}$$

and hence we obtain the *optimal control* law

$$u = -R^{-1}B^T Px = -p_{21}x_1 - p_{22}x_2 = \frac{-(t_1 - t)x_1 - (t_1 - t)^2 x_2}{1 + \frac{1}{3}(t_1 - t)^3}.$$

The feedback law is a function of remaining time $t_1 - t$.

7.3. Fixed end-point problems

Consider the problem of transfering the system (55) from state x_0 at time t_0 to state x_1 at time t_1 so that the functional

$$J(u) = \int_{t_0}^{t_1} [x(t)^T Q(t)x(t) + u(t)^T Ru(t)]\, dt$$

is minimized. The parameter matrices are defined as above. Note that no terminal term $x(t_1)^T Sx(t_1)$ is present since the end-point is fixed. By normalization we may set $R = I$. Define $\mathcal{U}$ to be the class of all control signals which achieve the desired transfer. The problem is then to find the u in $\mathcal{U}$ for which $J(u)$ is minimized. The special case when Q is zero was treated in Chapter 3. We shall reduce the present problem to that form.

To this end, let P be the solution to (RE) above. The analysis of Section 7.1.1 yields

$$J(u) = \int_{t_0}^{t_1} (u + B^T Px)^T (u + B^T Px)\, dt + x_0^T P(t_0)x_0 - x^T(t_1)P(t_1)x(t_1)$$

where $P(t_1)$ should be set to zero. Thus the problem is to determine a $u \in \mathcal{U}$ such that

$$\tilde{J}(u) = \int_{t_0}^{t_1} \|u + B^T Px\|^2\, dt$$

is minimized. If $u = -B^T Px \in \mathcal{U}$ we are done. The problem then has the same solution as that with free end-point. In general, of course, this is not true so we set

$$u = -B^T Px + v$$

Consequently, the problem has been reduced to the following: Transfer the system

$$\dot{x} = (A - BB^T P)x + Bv$$

from state x_0 at time t_0 to state x_1 at time t_1 so that

$$\int_{t_0}^{t_1} \|v\|^2 \, dt$$

is minimized. Suppose the system is completely reachable so that the reachability gramian is invertible, i.e. $W(t, t_1)^{-1}$exists , for all $t < t_1$. Then we know from Theorem 3.1.6 that the optimal control is

$$v(t) = B(t)^T \Phi(t_1, t)^T W(t_0, t_1)^{-1}[x_1 - \Phi(t_1, t_0)x_0]$$

According to the principle of optimality this solution is optimal if, instead of (t_0, x_0), we start in the point $(t, x(t)),\ t < t_1$, and $x(t)$ lies on the optimal trajectory. This yields an optimal control law

$$v(t) = K(t)x(t) + w(t)$$

where

$$K(t) = -B(t)^T \Phi(t_1, t)^T W(t, t_1)^{-1} \Phi(t_1, t)$$

and

$$w(t) = B(t)^T \Phi(t_1, t)^T W(t, t_1)^{-1} x_1$$

The original problem posed in the beginning of this section then has the feedback solution

$$u = (K - B^T P)x + w.$$

CHAPTER 8

LQ Control over Infinite Time Interval and ARE

In this chapter we shall consider the linear-quadratic regulator problem for the case that A, B, Q and R are constant and the time interval is infinite. We shall show that, as $t_1 \to \infty$, the solution of the matrix Riccati equation, which is central to the solution of the optimal control problem, tends under certain conditions to a limit P which is the unique symmetric positive definite solution of the *algebraic Riccati equation*, and that the optimal control law is determined by this P.

8.1. Existence of a positive definite solution

Consider the problem of controlling a time-invariant system

$$(\Sigma) \qquad \begin{cases} \dot{x} = Ax + Bu; & x(0) = x_0 \\ y = Cx \end{cases}$$

so that

$$J(u) = \int_0^{t_1} (x^T Q x + u^T R u) dt$$

is minimized, where $Q = C^T C$. In the previous chapter we showed that the optimal solution of this problem is given by the feedback law

$$u(t) = -R^{-1} B^T P(t) x(t)$$

where P is the unique solution of the matrix Riccati differential equation

$$(\text{RDE}) \qquad \begin{cases} \dot{P} = -A^T P - PA + PBR^{-1}B^T P - C^T C \\ P(t_1) = 0 \end{cases}$$

Recall that the minimal value is

$$\min_u J(u) =: V(x_0, t_0) = x_0^T P(t_0) x_0.$$

In this chapter, we consider the situation when the final time t_1 tends to infinity, i.e. we ask what happens if $t_1 \to \infty$; that is, we have to find the solution the stationary problem

$$\min_{u \in \mathcal{U}} \int_0^{\infty} (x^T Q x + u^T R u) dt$$

where $\mathcal{U}$ is the class of u such that $\int_0^{\infty} (x^T Q x + u^T R u) dt < \infty$. Since (RDE) is an autonomous differential equation, $P(t)$ can be expressed as $P(t_1 - t)$ (with an abuse of notation). First, we need to decide under which conditions

there is a limiting solution P_∞ for $P(t_1-t)$ as t_1-t tends to infinity, namely the limiting solution of $P_\infty := P(t_1-0)$ as t_1 tends to infinity.

To this end, we assume that Σ is completely reachable. Then, there exists a control signal $\hat{u}$ that transfers the system from $(0,x_0)$ to $(T,0)$. Now, let $\hat{u}(t)\equiv 0$ for $t\geq T$, which implies that $\hat{y}(t)\equiv 0$ for $t\geq T$.

In order to indicate $V(x_0,0)$ is a function of t_1, similarly we slightly abuse the notation by denoting $V(x_0,0)$ as $V(x_0,t_1)$. Then we know that

$$V(x_0,t_1)\leq \int_0^\infty (x^TQx+u^TRu)dt =: M \quad \text{for all } t_1\in(0,\infty).$$

Moreover, $V(x_0,t_1)$ is monotonically non-decreasing, and therefore,

$$V(x_0,t_1)\to V_\infty(x_0)=x_0^TP_\infty x_0\geq 0$$

for all $x_0\in\mathbb{R}^n$, and therefore $P(t_1-0)\to P_\infty$ as $t_1\to\infty$. But then P_∞ must satisfy the *algebraic Riccati equation*

$$A^TP+PA-PBR^{-1}B^TP+Q=0. \tag{ARE}$$

It is evident that P_∞ is real symmetric positive semidefinite. We note that the ARE can have several solutions That do not need to be symmetric. P_∞ is always a solution to the ARE, but not all solutions of the ARE are limiting solutions of the corresponding RDE.

A natural question is now : when is P_∞ also positive definite, i.e., $P_\infty > 0$? Obviously this is true if $Q>0$. However in our case $Q=C^TC$ that is in general only semi-definite. We know that

$$x_0^TP_\infty x_0\geq V(x_0,t_1)\geq 0 \qquad \forall t_1\geq 0.$$

Thus, if we can prove that, for some $t_1>0$, $V(x_0,t_1)$ is positive for all $x_0\neq 0$, then $P_\infty>0$. Therefore, we assume that $V(x_0,t_1)=0$ although $x_0\neq 0$. Then, for the optimal solution, $u(t)\equiv 0$, and $y(t)\equiv 0$ on $(0,t_1)$, that is

$$\left.\begin{aligned} &Cx\equiv 0\\ &C\dot{x}=CAx\equiv 0\\ &C\ddot{x}=CA^2x\equiv 0\\ &\quad\vdots \end{aligned}\right\} \Longrightarrow \begin{bmatrix} C\\ CA\\ CA^2\\ \vdots \end{bmatrix} x(t)=0$$

where $x(t)=e^{At}x_0\neq 0$. It is clear that this cannot happen if Σ is completely observable. In fact, the observability of the system Σ tells us that

$$x_0\neq 0 \Longrightarrow x_0^TP_\infty x_0\geq V(x_0,t_1)=x_0^TP(t_1)x_0>0, \tag{62}$$

i.e., $P_\infty>0$. Therefore, we have proved the following theorem.

THEOREM 8.1.1. *Suppose that Σ is a minimal realization. Then (ARE) has a real symmetric positive definite solution.*

Note that this does not exclude the possibility that (ARE) may have other solutions which are not positive definite.

COROLLARY 8.1.2. *If Σ is a minimal realization,*

$$x(t) \to 0 \qquad \text{as } t \to \infty$$

for all $u \in \mathcal{U}$.

Proof: Due to time-invariance

$$\int_{\nu}^{\nu+1} (x^T Q x + u^T R u)dt \geq V(x(\nu), 1) = x(\nu)' P(1) x(\nu) \geq \lambda_{\min} \|x(\nu)\|^2$$

where $\lambda_{\min}$ is the smallest eigenvalue of $P(1)$. But according to (62) above, $P(1) > 0$, i.e. $\lambda_{\min} > 0$. Therefore,

$$\sum_{\nu=0}^{\infty} \|x(\nu)\|^2 \leq (\lambda_{\min})^{-1} \int_0^{\infty} (x^T Q x + u^T R u)dt < \infty$$

if $u \in \mathcal{U}$, and consequently, $x(\nu) \to 0$ as $\nu \to \infty$. Now

$$x(t) = e^{A(t-[t])} x([t]) + \int_{[t]}^{t} e^{A(t-\tau)} B u(\tau) d\tau$$

where $[t]$ is the integer part of t. We have just shown that the first term tends to zero as $t \to \infty$. To estimate the second term, we apply Cauchy-Schwartz inequality:

$$\left\| \int_{[t]}^{t} e^{A(t-\tau)} B u(\tau) d\tau \right\|^2 \leq \int_{[t]}^{t} \left\| e^{A(t-\tau)} B \right\|^2 d\tau \int_{[t]}^{t} \|u\|^2 d\tau \to 0, \quad \text{as } t \to \infty,$$

because $u \in \mathcal{U}$ implies that

$\int_{\nu}^{\nu+1} \|u\|^2 d\tau \to 0 \qquad \text{as } \nu \to \infty.$

□

EXAMPLE 8.1.3. Control the system

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

so that

$$\int_0^{\infty} (4x_1^2 + 5x_2^2 + u^2)dt$$

is minimized.

Let $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{5} \end{bmatrix}$, $R = I$. Then $C^T C = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$.

Moreover, $\begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{5} \\ 0 & 2 \\ 0 & 0 \end{bmatrix}$ are of full rank so that (A, B, C) is a minimal realization. Thus, the algebraic Riccati equation becomes

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix} = 0.$$

By solving the resulting algebraic equations

$$\begin{cases} 4 - p_{12}^2 = 0 \\ p_{11} - p_{12}p_{22} = 0 \\ 2p_{12} - p_{22}^2 + 5 = 0 \end{cases}$$

we obtain

$$\begin{aligned} &p_{12} = \pm 2 \\ &p_{12} = 2 \quad \Rightarrow p_{22} = \pm 3 \\ &p_{12} = -2 \Rightarrow p_{22} = \pm 1 \\ &p_{11} = p_{12}p_{22} \end{aligned}$$

Therefore, (ARE) has four solutions:

$$P_1 = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}, \quad P_2 = \begin{bmatrix} -6 & 2 \\ 2 & -3 \end{bmatrix}, \quad P_3 = \begin{bmatrix} -2 & -2 \\ -2 & 1 \end{bmatrix}, \quad P_4 = \begin{bmatrix} 2 & -2 \\ -2 & -1 \end{bmatrix}.$$

It is not hard to see that $P_1 > 0$, $P_2 < 0$, and P_3 and P_4 are indefinite. □

8.2. The optimal control law and the question of uniqueness

We now proceed to determining the optimal control. Let Σ be a *minimal* realization, and let P be an arbitrary solution of (ARE). Then,

$$\begin{aligned} \frac{d}{dt}(x^TPx) &= \dot{x}^TPx + x^TP\dot{x} \\ &= x^T(A^TP + PA)x + u^TB^TPx + x^TPBu \\ &= \|R^{\frac{1}{2}}u + R^{-\frac{1}{2}}B^TPx\|^2 - u^TRu - \|y\|^2 \end{aligned}$$

because $A^TP + PA = PBR^{-1}B^TP - C^TC$ since P is the solution of (ARE). Integrating this yields

(63)

$$x(t)^TPx(t) - x_0^TPx_0 = \int_0^t \|R^{\frac{1}{2}}u + R^{-\frac{1}{2}}B^TPx\|^2 ds - \int_0^t (\|y\|^2 + u^TRu)ds$$

Suppose now that $u \in \mathcal{U}$. Then, $x(t) \to 0$ as $t \to \infty$ (Corollary 8.1.2). Therefore,

$$\begin{aligned} \int_0^\infty (\|y\|^2 + u^TRu)ds &= x_0^TPx_0 + \int_0^\infty \|R^{\frac{1}{2}}u + R^{-\frac{1}{2}}B^TPx\|^2 ds \\ &\geq x_0^TPx_0. \end{aligned}$$

Hence, the control law

$$u = -R^{-1}B^TPx$$

is optimal *provided* that it defines a control signal $u \in \mathcal{U}$. The feedback system becomes

$$\dot{x} = (A - BR^{-1}B^TP)x$$

Setting $\Gamma \triangleq A - BR^{-1}B^TP$ we have $y(t) = Ce^{\Gamma t}x_0$ and $u(t) = -R^{-1}B^TPe^{\Gamma t}x_0$. If Γ is a stable matrix, it follows that $u \in \mathcal{U}$.

LEMMA 8.2.1. *If Σ is a minimal realization and if P is a real symmetric positive definite solution of* (ARE)*, then*

$$\Gamma = A - BR^{-1}B^TP$$

is a stable matrix.

Proof: The algebraic Riccati equation can be written as

$$\Gamma^TP + P\Gamma + PBR^{-1}B^TP + C^TC = 0.$$

We now proceed as in the proof of Theorem 28. Form the Lyapunov function $V(x) \triangleq x^TPx$, which is strictly positive for all $x \neq 0$. Differentiation of V along a solution of $\dot{x} = \Gamma x$ yields

$$\begin{aligned}\frac{d}{dt}(x^TPx) = \dot{x}^TPx + x^TP\dot{x} = x^T\Gamma^TPx + x^TP\Gamma x\\ =x^T(\Gamma^TP + P\Gamma)x = -x^T(PBR^{-1}B^TP + C^TC)x \leq 0.\end{aligned}$$

Consequently, $V(x(t))$ is non-increasing along $\dot{x} = \Gamma x$, and it follows that the solutions remain bounded. Hence, $\operatorname{Re}\lambda(\Gamma) \leq 0$.

We now show that Γ has no eigenvalue on the imaginary axis. Recall that Γ has a purely imaginary eigenvalue if and only if there is a nontrivial periodic solution of $\dot{x} = \Gamma x$. Therefore, it suffices to show that there is no nontrivial periodic solution.

Suppose that $x(t)$ is a periodic solution such that $x(t_0) = x(t_1)$. Then $V(x(t_0)) = V(x(t_1))$ and

$$\frac{d}{dt}V(x(t)) = |R^{-\frac{1}{2}}B^TPx(t)|^2 + |Cx(t)|^2 \equiv 0$$

on $[t_0, t_1]$, and by periodicity for all t. Hence, $R^{-1}B^TPx \equiv 0$ and

$$\dot{x} = \Gamma x = Ax - BR^{-1}B^TPx = Ax.$$

Moreover, $Cx \equiv 0$ and $\frac{d}{dt}Cx = CAx \equiv 0$. Further differentiation yields $CA^kx(t) \equiv 0$ for all k. Since (C, A) is observable, it follows that $x(t) \equiv 0$. □

LEMMA 8.2.2. *There is at most one real symmetric solution of* (ARE) *such that*

$$\Gamma = A - BR^{-1}B^TP$$

is a stable matrix.

Proof: Assume that there are two such matrices P_1 and P_2. According to (63), we have, for $i = 1, 2$,

$$\int_0^t (\|y\|^2 + u^TRu)ds = x_0^TP_ix_0 + \int_0^t \|R^{\frac{1}{2}}u + R^{-\frac{1}{2}}B^TP_ix\|^2ds - x(t)^TP_ix(t).$$

Set $u = -R^{-1}B^TP_2x$ and let $t \to \infty$. Since $A - BR^{-1}B^TP_2$ is stable, $x(t) \to 0$ and therefore,

$$\int_0^\infty (\|y\|^2 + u^TRu)ds = x_0^TP_1x_0 + \int_0^\infty \|R^{-\frac{1}{2}}B^T(P_1 - P_2)x\|^2ds = x_0^TP_2x_0$$

that is $x_0^TP_2x_0 \geq x_0^TP_1x_0$. The same argument with $u = -R^{-1}B^TP_1x$ tells us that $x_0^TP_1x_0 \geq x_0^TP_2x_0$. Hence, we have shown that

$$x_0^TP_1x_0 = x_0^TP_2x_0, \quad \text{for all } x_0 \in \mathbb{R}^n,$$

which implies that $P_1 = P_2$. □

EXAMPLE 8.2.3. Consider Example 8.1.3. Then

$$\Gamma = A - BB^TP = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -p_{12} & -p_{22} \end{bmatrix}$$

The characteristic polynomial of Γ is

$$\chi_\Gamma(s) = s^2 + p_{22}s + p_{12}$$

which, for each solution of (ARE), takes the form

$$P_1 = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}, \qquad \chi_\Gamma(s) = s^2 + 3s + 2 = (s+1)(s+2) \quad \text{i.e. } \Gamma \text{ is stable.}$$

$$P_2 = \begin{bmatrix} -6 & 2 \\ 2 & -3 \end{bmatrix}, \quad \chi_\Gamma(s) = s^2 - 3s + 2 = (s-1)(s-2) \quad \text{i.e. } \Gamma \text{ is unstable.}$$

$$P_3 = \begin{bmatrix} -2 & -2 \\ -2 & 1 \end{bmatrix}, \quad \chi_\Gamma(s) = s^2 + s - 2 = (s-1)(s+2) \quad \text{i.e. } \Gamma \text{ is unstable.}$$

$$P_4 = \begin{bmatrix} 2 & -2 \\ -2 & -1 \end{bmatrix}, \quad \chi_\Gamma(s) = s^2 - s - 2 = (s+1)(s-2) \quad \text{i.e. } \Gamma \text{ is unstable.}$$

Then, the optimal control law is

$$u = -B^TPx = -p_{12}x_1 - p_{22}x_2$$

where the real symmetric positive definite solution P_1 must be used, i.e.

$$u = -2x_1 - 3x_2.$$

The optimal future cost is

$$x^TPx = [x_1, x_2] \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 6x_1^2 + 4x_1x_2 + 3x_2^2.$$

□

THEOREM 8.2.4. *Let Σ be a minimal realization. Then* (ARE) *has a unique real symmetric positive definite solution P. Moreover,*

$$u = -R^{-1}B^TPx$$

is the unique optimal control law which minimizes $J(u) = \int_0^\infty (\|y\|^2 + u^TRu)dt$ and $\min J(u) = x_0^TPx_0$.

Proof: By Theorem 8.1.1, (ARE) has a solution $P > 0$. But Lemma 8.2.1 and Lemma 8.2.2 show that this is the only solution which is real symmetric positive definite. Since only such a solution corresponds to a matrix Γ which is a stable matrix, it follows from the definition above that $u = -R^{-1}B^TPx$ is an optimal control law and that the optimal value is x^TPx. □
The following simple observations from linear algebra have been used above.

REMARK 8.2.5. If $x^TP(t)x \to x^TPx$ for all $x \in \mathbb{R}^n$, then $P(t) \to P$ as $t \to \infty$.

Proof: First take $x = e_k$, the unit vector with zeros in all positions except position k where there is a one. Then, $p_{kk}(t) \to p_{kk}$, taking care of the diagonal entries. Next take $x = e_k + e_l$. Then,

$$p_{kk}(t) + 2p_{kl}(t) + p_{ll}(t) \to p_{kk} + 2p_{kl} + p_{ll},$$

which shows that $p_{kl}(t) \to p_{kl}$. □

REMARK 8.2.6. If Q is a symmetric matrix, then

$$\lambda_{\min}\|x\|^2 \leq x^TQx \leq \lambda_{\max}\|x\|^2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and the largest eigenvalues of Q, respectively.

Proof: If Q is symmetric, there is an orthogonal matrix S such that $SQS^T = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Then

$$x^TQx = x^TS^TSQS^TSx = \sum_{i=1}^{n} \lambda_i(Sx)_i^2$$

i.e.

$$\lambda_{\min}\sum_{i=1}^{n}(Sx)_i^2 \leq x^TQx \leq \lambda_{\max}\sum_{i=1}^{n}(Sx)_i^2$$

Since $S^TS = I$, $x^TS^TSx = \|x\|^2$, the above inequality is the desired one. The proof is complete. □

CHAPTER 9

Kalman Filtering

In 1960 Rudolf E. Kalman published his famous paper describing the optimal solution to the discrete-data linear filtering problem. Before Kalman's solution Norbert Wiener had already described an optimal finite impulse response filter. However, Wiener filter requires computation of the impulse response, which is not suitable for on-line implementation.

In Kalman's derivation a recursive approach was proposed using state space descriptions, which could be easily implemented on-line. The state space description also enables the filter to be used either as a filter, smoother or predictor.

The recursive nature of Kalman filter has proven to be very useful. Kalman filter is perhaps so far the best known result coming from systems and control since 1960's and it has found a very wide range of applications, in particular in navigation and tracking.

Let us use a simple example here to illustrate the idea of Kalman filter.

EXAMPLE 9.0.7. Two persons make an observation of x (say the height of a building) each.

- Person1's observation is y_1 with confidence (variance) $\sigma_{y_1}^2$.
- Person2's observation is y_2 with confidence (variance) $\sigma_{y_2}^2$.

Person 1's observation arrives first so we use

$$\begin{aligned} \hat{x}_1 &= y_1 \\ \sigma_1^2 &= \sigma_{y_1}^2 \end{aligned}$$

as the best estimation of x since no priori information about x is available. We then update once the second observation is available:

$$\hat{x}_2 = \hat{x}_1 + K(y_2 - \hat{x}_1), \tag{64}$$

Here $K = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{y_2}^2}$ is the so-called *Kalman gain*. Consequently

$$\sigma_2^2 = \frac{\sigma_1^2 \sigma_{y_2}^2}{\sigma_1^2 + \sigma_{y_2}^2}.$$

Or we can directly compute

$$\hat{x}_2 = \frac{\sigma_{y_2}^2 y_1 + \sigma_{y_1}^2 y_2}{\sigma_{y_1}^2 + \sigma_{y_2}^2}. \tag{65}$$

While (64) and (65) give the same estimation, only the first is recursive. The latter is an example of the classical Gauss-Markov parameter estimation.

9.1. The discrete-time filter

Consider a linear system

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)v(t) \\ y(t) &= C(t)x(t) + D(t)w(t) \\ x(0) &= x_0 \ (unknown), \end{aligned} \tag{66}$$

defined on the interval $[0, t_1]$, where the input signals $v(t)$ and $w(t)$ are taken to be random processes. More specifically, we assume that v and w are uncorrelated white noises, i.e.,

$$\begin{aligned} E\, v(t)v(s)^T &= Q\delta_{ts} \\ E\, w(t)w(s)^T &= R\delta_{ts} \\ E\, v(t)w(s)^T &= 0, \end{aligned}$$

and that the initial state x_0 is a stochastic vector, with covariance

$$P_0 = E\, x_0 x_0^T,$$

uncorrelated with both v and w, i.e., $E\, x_0 v(t)^T = E\, x_0 w(t)^T = 0$. Moreover, we assume that x_0, v and w are *centered*, i.e., $Ex_0 = Ev(t) = Ew(t) = 0$.

The problem is now to estimate the state $x(t)$ in the best possible way using the observations $\{y(0), y(1), ..., y(t-1)\}$. By best possible we mean that the estimate shall be optimal in the sense of *least-squares*, i.e., the state estimate $\hat{x}(t)$ should be chosen so that the variance of the error

$$E[x_i(t) - \hat{x}_i(t)]^2 \tag{67}$$

is minimized for each $i = 1, 2, \ldots, n$ and each time t.

We would like the estimator to be practically implementable in an efficient manner. As time proceeds the acquired amount of data $\{y(0), y(1), \ldots, y(t-1)\}$ grows and can become very large. We seek a solution implementable with bounded memory, where the bound is independent of the length of the interval $[0, t_1]$. Moreover, we would like the computations to be recursive in order to be efficient.

Finally, we shall restrict the estimate $\hat{x}_i(t)$ to be a *linear* function of the observed data

$$\{y_1(0), \ldots, y_m(0), y_1(1), \ldots, y_m(1), \ldots, y_1(t-1), \ldots, y_m(t-1)\}, \tag{68}$$

which minimizes the least-square criterion (67). This raises the question whether there really is such an estimate and, if so, whether it is unique.

Before turning to this question, however, let us comment on the assumption of linearity. It is apparently a restriction to require that $\hat{x}_i(t)$ should be a linear function of the data (68), rather than allowing nonlinear functions as well. However there is an important case for which *the optimal least-squares estimate in the larger class of not necessarily linear functions of the data is*

in fact linear, namely when x_0, v, w are jointly Gaussian. In this case all stochastic variables defined by the system (66) become Gaussian, because the Gaussian property is preserved under linear transformations.

In the following we shall show how the estimate $\hat{x}(t)$ can be delivered by a time-varying finite-dimensional linear system driven by the observations $y(t)$. Such a system is called a *Kalman filter* and meets with the specifications of bounded memory and recursive computations.

REMARK 9.1.1. A useful interpretation of the system in (66) is as follows. The state process $x(t)$ describes an airplane with dynamics described by the matrices A and B. The actual control action taken by the pilot is to us unknown and is modeled as a random process v with some known statistics. Moreover, we perform observations of the state as $y(t) = Cx(t)+Dw(t)$. The observations are of reduced dimension, and corrupted by the measurement noise w. □

9.1.1. Linear least-squares estimation and orthogonal projections. The family of all stochastic variables with finite second order moment generates a vector space H whose elements are precisely all linear combinations of these generating stochastic variables. The space H is an inner-product space (Hilbert space) with *inner product*

$$(\xi, \eta) = E\{\xi\eta\}$$

and norm $\|\xi\| = (\xi, \xi)^{1/2}$. When H is generated by stochastic variables in (68), it is finite dimensional. In this H there is a nested family of subspaces

$$H_0(y) \subset H_1(y) \subset H_2(y) \subset \ldots \subset H_{t_1}(y)$$

where $H_t(y)$ is the space of all linear combinations of

$$\{y_1(0), \ldots, y_m(0), y_1(1), \ldots, y_m(1), \ldots, y_1(t), \ldots, y_m(t)\}.$$

Since $E[x_i(t) - \hat{x}_i(t)]^2 = \|x_i(t) - \hat{x}_i(t)\|^2$ the linear least-squares problem is thus reduced to determining an $\hat{x}_i(t)$ in $H_t(y)$ which minimizes $\|x_i(t)-\hat{x}_i(t)\|$ for each $i = 1, 2, \ldots, n$. Kalman filter on the other hand will give a recursive procedure for determining the optimal solution in $H_t(y)$based on the optimal solution in $H_{t-1}(y)$.

The existence of such a minimizer as well as equations for determining it is given by the following *projection theorem.*

LEMMA 9.1.2. *Let M be a subspace of the finite-dimensional inner-product space H. Let $h \in H$ be arbitrary. Then there is a unique $\hat{m} \in M$ such that the distance $||h - m||$ is minimized. The minimizer $\hat{m}$ is characterized by the condition that the error $\tilde{h} \triangleq h - \hat{m}$ is orthogonal to M, i.e., $(h - \hat{m}, m) = 0$ for all $m \in M$.*

REMARK 9.1.3. The condition that $(h - \hat{m}, m) = 0$ for all $m \in M$ is called the *normal equations.* □

REMARK 9.1.4. The linear space $H_{t-1}(y)$ is formed by taking *linear* combinations of the observed random variables. By projecting onto this space we get the best linear estimate of $x_i(t)$. We can also imagine estimators being e.g. polynomials of some degree q. The set of such polynomials, $Q_{t-1}(y)$ say, is also a finite-dimensional linear space (not a subspace of H, though) and the best estimator in $Q_{t-1}(y)$ would also be given by orthogonal projection. Hence, the functional form of the estimate is determined by the structure of the subspace onto which we project. □

Consequently, we may regard the least-squares estimate as the orthogonal projection of h onto the subspace M as depicted in Figure 9.3.

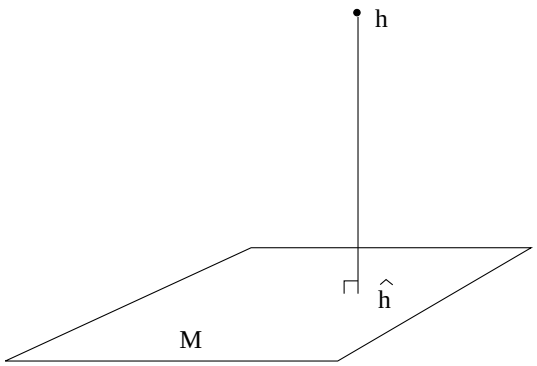


Figure 9.3

The mapping $h \to \hat{m}$ is well defined and called the orthogonal projection of h onto M, which we denote $\hat{m} = E^M h$.

LEMMA 9.1.5. *The mapping $E^M : H \to M$ is linear, i.e., for any $h_1, h_2 \in H$ and $\alpha, \beta \in \mathbb{R}$ it holds that*

$$E^M(\alpha h_1 + \beta h_2) = \alpha E^M h_1 + \beta E^M h_2.$$

9.1.2. Orthogonal projection onto the space generated by a random vector. Let $y = [y_1, ..., y_m]^T$ be a vector of random variables and let x be a random variable. Moreover, let $[y]$ be the space spanned by the components of y, i.e.,

$$[y] \triangleq \text{span}\{y_1, ..., y_m\}.$$

We shall now show how to compute the projection $\hat{x} = E^{[y]}x$. An element of $[y]$ has the form $\sum_1^m k_i y_i = ky$, where $k = [k_1, ..., k_m]$ and therefore we seek a k satisfying the normal equations $(x - ky, y_j) = 0$ for $j = 1, ..., m$. In matrix notation the normal equations become

$$k \begin{bmatrix} Ey_1y_1 & Ey_1y_2 & \cdot & Ey_1y_m \\ \cdot & \cdot & \cdot & \cdot \\ Ey_my_1 & Ey_my_2 & \cdot & Ey_my_m \end{bmatrix} = [Exy_1, \ldots, Exy_m].$$

It follows from the projection theorem that the system $kEyy^T = Exy^T$ always has a solution, and if Eyy^T is invertible, which happens if and only the random variables $\{y_1, ..., y_m\}$ are linearly independent, then

$$k = Exy^T(Eyy^T)^{-1} \text{ and } \hat{x} = E^{[y]}x = Exy^T(Eyy^T)^{-1}y.$$

If the random variables $\{y_1, ..., y_m\}$ are linearly dependent then $\hat{x}$ is still unique but k is not.

When deriving the Kalman filter we shall need to project vectors of random variables component wise onto linear spaces, i.e., if $x = [x_1, ..., x_n]^T$ we define $E^M x$ as $E^M x \triangleq [E^M x_1, ..., E^M x_n]^T$. Moreover, we say that $x \in M$ if $x_i \in M$ for $i = 1, ...n$ and $x \perp M$ if the components of x are orthogonal to M.

When projecting the vector x onto the space $[y]$ the estimate has the form $\hat{x} = E^{[y]}x = Ky$ where K is a matrix. The normal equations can now be written in matrix form as $E(x - Ky)y^T = 0$ and if the random variables $\{y_1, ..., y_m\}$ are linearly independent then $K = Exy^T[Eyy^T]^{-1}$. We state the result as a lemma.

LEMMA 9.1.6. *Let x, y be random vectors and suppose that the components of y are linearly independent. The linear least-squares estimate $\hat{x}$ of x given y is*

$$\hat{x} = Exy^T(Eyy^T)^{-1}y.$$

The following lemmas are needed in the derivation of the Kalman filter.

LEMMA 9.1.7. *Let M be a subspace of the finite-dimensional inner-product space H. Let x be a vector of elements $x_i \in H$ For any matrix A such that the product Ax is well-defined it holds that*

$$E^M Ax = AE^M x.$$

LEMMA 9.1.8. *Let M, N be subspaces of the finite-dimensional inner-product space H. If $M \perp N$ then*

$$E^{M \oplus N} = E^M + E^N.$$

REMARK 9.1.9. Can you find an example showing that if M and N are not orthogonal then the preceding lemma is false? □

9.1.3. The innovation process. Recall that $H_{t-1}(y) \subseteq H_t(y)$. A fundamental idea in the area of random processes is to extract the genuinely new information, with respect to $H_{t-1}(y)$, contained in $y(t)$. To this end, let

$$\tilde{y}(t) \triangleq y(t) - E^{H_{t-1}(y)}y(t).$$

The vector $\tilde{y}(t)$ is by construction orthogonal to the space $H_{t-1}(y)$ and therefore we can perform an orthogonal decomposition of $H_t(y)$ as

$$H_t(y) = H_{t-1}(y) \oplus [\tilde{y}(t)].$$

The stochastic vector $\tilde{y}(t)$ represents the new information about the process y obtained at time t, and $\tilde{y}$ is therefore called the *innovation process*

of y. The innovation process is by construction a white noise and $H_t(y) = H_t(\tilde{y})$ (verify this). Defining the covariance matrix

$$R(t) \triangleq E\tilde{y}(t)\tilde{y}(t)^T$$

we see that if $R(t)$ is positive definite for all t, each component of $y(t)$ carries new information in each step, in that no linear combination of $y_1(t), y_2(t), \ldots, y_m(t)$ lies in the subspace $H_{t-1}(y)$ generated by previous observations. We say that y is *purely nondeterministic* if it has this property.

9.1.4. The Kalman recursions. Returning to the original problem, the linear-least squares estimate of $x(t)$ based on the observations $H_{t-1}(y)$ can be written as $\hat{x}(t) = E^{H_{t-1}(y)}x(t)$. We now derive explicit formulas for the estimator. In the following it will be important to observe the various orthogonality relations that hold, e.g. $v(t) \perp H_t(y)$ and $w(t) \perp H_{t-1}(y)$ (why?).

In order to simplify the notation we suppress the time dependence of the matrices A, B, C, D in the following computations.

The estimate $\hat{x}(t+1)$ satisfies

$$\hat{x}(t+1) = E^{H_t(y)}x(t+1) = E^{H_t(y)}[Ax(t) + Bv(t)] = \{\text{Lemma 9.1.7}\} =$$
$$= AE^{H_t(y)}x(t) + BE^{H_t(y)}v(t) = \{v(t) \perp H_t(y)\} = AE^{H_t(y)}x(t).$$

Decompose $H_t(y)$ as

$$H_t(y) = H_{t-1}(y) \oplus [\tilde{y}(t)]$$

and define $\hat{x}_t(t) = E^{H_t(y)}x(t)$. Using Lemma 9.1.8 we have

$$\hat{x}_t(t) = E^{H_{t-1}(y)}x(t) + E^{[\tilde{y}(t)]}x(t) = \hat{x}(t) + E^{[\tilde{y}(t)]}x(t) = \hat{x}(t) + K(t)\tilde{y}(t).$$

where the time-varying matrix function $K(t)$, called the *Kalman gain*, remains to be determined. Consequently,

$$\hat{x}(t+1) = A\hat{x}(t) + A\,E^{[\tilde{y}(t)]}x(t) = A\hat{x}(t) + AK(t)\tilde{y}(t).$$

At time t, the estimate $\hat{x}(t)$ and $y(t)$ are available and the innovation $\tilde{y}(t)$ is given as

$$\begin{aligned}\tilde{y}(t) &= y(t) - E^{H_{t-1}(y)}y(t)\\ &= y(t) - CE^{H_{t-1}(y)}x(t) - DE^{H_{t-1}(y)}w(t)\\ &= y(t) - C\hat{x}(t),\end{aligned}$$

where the last equality follows from the fact that $w(t) \perp H_{t-1}(y)$. Define the estimation error $\tilde{x}(t) \triangleq x(t) - \hat{x}(t)$ and let $P(t)$ be the covariance matrix of $\tilde{x}(t)$, i.e.,

$$P(t) \triangleq E\,\tilde{x}(t)\tilde{x}(t)^T.$$

The innovation can now be written

$$\tilde{y}(t) = C\tilde{x}(t) + Dw(t). \tag{69}$$

Note that the two terms in the right-hand side of (69) are orthogonal.

We now determine the Kalman gain. Under the assumption that y is purely nondeterministic, Lemma 9.1.6 yields

$$E^{[\tilde{y}(t)]}x(t) = E\,x(t)\tilde{y}(t)^T\,[E\,\tilde{y}(t)\tilde{y}(t)^T]^{-1}\,\tilde{y}(t).$$

Using(69) we get

$$R(t) = E\tilde{y}(t)\tilde{y}(t)^T = CP(t)C^T + DRD^T.$$

Note that $DRD^T > 0$ is a *sufficient* condition for y to be purely nondeterministic.

Moreover, we compute $Ex(t)\tilde{y}(t)^T$ as

$$Ex(t)\tilde{y}(t)^T = Ex(t)[C\tilde{x}(t) + Dw(t)]^T = \{w(t) \perp x(t)\} =$$

$$= Ex(t)\tilde{x}(t)^T C^T = \{\hat{x}(t) \perp \tilde{x}(t)\} = E\tilde{x}(t)\tilde{x}(t)^T C^T = P(t)C^T.$$

Hence, the Kalman gain is

$$K(t) = P(t)C^T[CP(t)C^T + DRD^T]^{-1}. \tag{70}$$

Let $P_t(t) = E(x(t) - \hat{x}_t(t))(x(t) - \hat{x}_t(t))^T$. Then

$$P_t(t) = P(t) - K(t)CP(t),$$

which implies that the covariance is always decreasing after a measurement update.

Finally the state estimate $\hat{x}$ is given by the recursive algorithm

$$\begin{aligned} \hat{x}(t+1) &= A\hat{x}(t) + AK(t)\tilde{y}(t) \\ &= A\hat{x}(t) + AK(t)[y(t) - C\hat{x}(t)]. \end{aligned} \tag{71}$$

The recursion for $\hat{x}(t)$ can also be written

$$\hat{x}(t+1) = [A - AK(t)C]\hat{x}(t) + AK(t)y(t) \tag{72}$$

displaying the filter as a dynamical system driven by the observations $y(t)$. Note that in this form, the Kalman filter has the same structure as the observer in Chapter 6.

Letting $H_{-1}(y) \triangleq \{0\}$ gives the initial value $\hat{x}(0) = 0$, i.e., with no observations available the best estimate of $x(0)$ is the mean.

Finally, in order to compute the Kalman gain $K(t)$ we need $P(t)$. Fortunately, the sequence of matrices $P(t)$ can be recursively computed by a matrix-valued difference equation. Subtracting (71) from the first of equations (66) and applying (69) we obtain

$$\tilde{x}(t+1) = [A - AK(t)C]\tilde{x}(t) - AK(t)Dw(t) + Bv(t).$$

The three terms at the right-hand side of the previous equation are mutually orthogonal and it follows that

(73)

$$P(t+1) = [A - AK(t)C]P(t)[A - AK(t)C]^T + AK(t)DRD^TK(t)^TA^T + BQB^T$$

which together with (70), after some manipulations, gives

(74)
$$\begin{aligned} P(t+1) &= AP(t)A^T - AP(t)C^T[CP(t)C^T + DRD^T]^{-1}CP(t)A^T + BQB^T \\ P(0) &= P_0. \end{aligned}$$

Despite the fact that this equation is not quadratic it is called the *discrete-time matrix Riccati equation* for reasons that will be explained below. We summarize the results of this section in the following theorem.

THEOREM 9.1.10 (The Kalman filter). *Given a linear stochastic system* (66) *having a purely nondeterministic output process* y*, the linear least-squares estimate* $\hat{x}(t)$ *of the state* $x(t)$ *given the observations* $\{y(0), y(1), .., y(t-1)\}$ *is generated by the Kalman filter* (72) *where the gain* K *is determined from* (70) *and the discrete matrix Riccati equation* (74)*.*

REMARK 9.1.11. We have so far shown that Kalman filter minimizes each $E\|x_i(t) - \hat{x}_i(t)\|^2$. We note that Kalman filter even gives the minimum mean square error (MMSE) solution, and also minimizes the covariance matrix $P_t(t)$.

Let us assume that we use a general filter

$$\begin{aligned} \hat{x}_t(t) &= \hat{x}(t) - L(t)(y(t) - C\hat{x}(t)) \\ &= \hat{x}(t) - (K(t) + \tilde{K}(t))(y(t) - C\hat{x}(t)). \end{aligned}$$

Then,

$$\begin{aligned} P_t(t) &= P(t) - P(t)C^T(CP(t)C^T + DRD^T)^{-1}CP(t) \\ &+ \tilde{K}(CP(t)C^T + DRD^T)\tilde{K}^T. \end{aligned}$$

Clearly $\tilde{K} = 0$ gives the minimal covariance matrix. Since $E\|x(t) - \hat{x}_t(t)\|^2 = tr(P_t(t))$, it is also minimized.

EXAMPLE 9.1.12 (Noisy measurements). Let Z be an unknown scalar quantity, modeled as a random variable with mean $E\,Z = -2$ and variance $\mathrm{Var}Z = 0.5$.

Suppose we are repeatedly performing noisy measurements of Z of the form

$$z(t) = Z + \sigma w(t),$$

where $\{w(t)\}$ is a scalar normalized white-noise sequence. At each instant we would like to make the linear least-squares estimate of Z based on the measurements made so far.

This problem can be embedded in the Kalman filtering set-up in the following way. First define the dynamical system $x(t+1) = x(t)$, $x(0) = Z + 2$ in order to get a zero mean random process to be estimated. This gives $E\,x(0) = 0$ and $E\,x(0)^2 = 0.5$. Then define the observation process as $y(t) = x(t) + \sigma w(t)$. Practically, $y(t)$ is obtained as $y(t) = z(t) - 2$.

We now have the standard setting as in (66), with $A = 1$, $B = 0$, $C = 1$ and $D = \sigma$. The best estimate of $x(t)$ based on $\{y(0), y(1), ..., y(t-1)\}$ is delivered by the dynamical system

$$\hat{x}(t+1) = \hat{x}(t) + k(t)[y(t) - \hat{x}(t)],\ \hat{x}(0) = 0.$$

The Riccati equation for $p(t)$, which in this example is a scalar quantity, can be written

$$p(t+1) = \frac{p(t)\sigma^2}{p(t) + \sigma^2},\ p(0) = 0.5$$

and the Kalman gain $k(t)$ is given by

$$k(t) = \frac{p(t)}{p(t) + \sigma^2}.$$

The asymptotic behavior of $p(t)$ can easily be analyzed. We first look for equilibrium solutions to the Riccati equation

$$p = \frac{p\sigma^2}{p + \sigma^2},$$

with the only solution $p = 0$. Convergence can be established by differentiating the function

$$f(p) \triangleq \frac{p\sigma^2}{p + \sigma^2}.$$

The derivative is

$$f'(p) = \frac{1}{(1 + \frac{p}{\sigma^2})^2}$$

and satisfies $|f'(p)| < 1$ for $p \geq 0$. Consequently, the gain $k(t)$ tends to zero, which is in line with the intuition that further measurements should be less taken into account as time proceeds.

Example 9.1.13 (A scalar problem). Consider the system

$$\begin{cases} x(t+1) = \frac{1}{2}x(t) + b\,v(t) \\ y(t) = x(t) + d\,w(t), \end{cases}$$

where $x(t), v(t), w(t)$ and $y(t)$ are scalars. Moreover, let $E\,x(0) = 0$ and $E\,x(0)^2 = \sigma^2$.

The Riccati equation is, by (74),

$$p(t+1) = \frac{1}{4}p(t) - \frac{1}{4}p(t)^2(p(t) + d^2)^{-1} + b^2.$$

By simplification we get

$$p(t+1) = \frac{1}{4}\frac{p(t)d^2}{p(t) + d^2} + b^2.$$

In order to study the asymptotic behavior of the Riccati equation we define the function

$$f(p) \triangleq \frac{1}{4}\frac{pd^2}{p + d^2} + b^2$$

and look for solutions to $f(p) = p$. After rearranging terms we get the quadratic equation

$$p^2 + (\frac{3}{4}d^2 - b^2)p - (bd)^2 = 0.$$

Since $-(bd)^2 < 0$, we conclude that the roots are real and have opposite signs. Hence, there is a strictly positive stationary solution p to the Riccati equation. The question of convergence can be settled by differentiating $f(p)$,

$$f'(p) = \frac{1}{4}\frac{1}{(1 + \frac{p}{d^2})^2}.$$

Since $|f'(p)| < 1$ for $p \geq 0$, we conclude that the Riccati equation converges to the positive positive stationary solution. Moreover, the iteration converges to the positive stationary solution for all initial values $p(0) \geq 0$, indicating that sometimes in Kalman filtering problems exact knowledge of $E\, x(0)x(0)^T$ might not be so important.

EXAMPLE 9.1.14 (A tracking problem). An important area of application for Kalman filtering techniques is that of tracking systems, such as systems for air traffic control.

As a simple example, we shall consider the problem of tracking a particle performing one-dimensional motion, i.e. $\ddot{y}(t) = f(t)$, where $y(t)$ is the position of the particle and $f(t)$ is the applied force. The tracking system performs noisy measurements of $y(t)$, but $f(t)$ is unknown to the system and is modeled as a random process.

Suppose that the tracking system is working in discrete time with time step h. If h is very small compared to the time constants of the objects being tracked it is reasonable, from the tracking system's point of view, to assume that the input $f(t)$ is a piecewise constant signal $u(k)$, i.e. $f(t) = u(k)$ if $t \in [kh, kh + h)$.

Hence, by sampling a state space representation of $\ddot{y}(t) = f(t)$, as in Example 2.3.1, we get the discrete-time system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} \frac{h^2}{2} \\ h \end{bmatrix} u(k), \tag{75}$$

where $x_1(k)$ is the position of the particle at kh and $x_2(k)$ its velocity.

The unknown input $u(k)$ is now modeled as a white noise sequence with zero mean and variance σ^2, where σ^2 is chosen to reflect the physical performance of the objects being tracked. Moreover, we assume that the tracking system performs observations of the position of the particle at the instants kh corrupted by additive noise, i.e.

$$y(k) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(k) + d\, w(t),$$

where $w(t)$ is a normalized white noise sequence.

We now have the setting as in (66) with

$$A = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} \frac{h^2}{2} \\ h \end{bmatrix} \sigma, C = \begin{bmatrix} 1 & 0 \end{bmatrix} \text{ and } D = d.$$

If we let $h = 0.5$, $\sigma = 1$ and $d = 0.1$ and iterate the Riccati equation with the initial value $P(0) = I$, we reach the stationary value

$$P_\infty = \begin{bmatrix} 0.1532 & 0.2020 \\ 0.2020 & 0.2756 \end{bmatrix} > 0$$

after a few iterations.

Actually, the general theory of the discrete-time Riccati equation gives, under some natural conditions, that the Riccati equation converges to a stationary value $P_\infty > 0$ independent of the positive semi-definite initial value $P(0)$. Moreover, the corresponding Kalman gain K_∞ stabilizes the estimator, i.e. the matrix $A - AK_\infty C$ has its eigenvalues in the open unit disc.

We now verify stability of the estimator in our numerical example. In stationarity we have

$$AK_\infty = \begin{bmatrix} 1.5576 \\ 1.2378 \end{bmatrix}$$

and the closed-loop system matrix $A - AK_\infty C$ has eigenvalues $\{0.22 + i\,0.11, 0.22 - i\,0.11\}$ with magnitude 0.25 so the estimator

$$\hat{x}(k+1) = (A - AK_\infty C)\hat{x}(k) + AK_\infty y(k)$$

is indeed an input/output stable system.

If we reduce the observation noise by letting $d = 0.05$ the stationary Kalman gain increases to

$$AK_\infty = \begin{bmatrix} 1.6326 \\ 1.2974 \end{bmatrix}.$$

This is in accordance with the intuition that the innovations $y(k) - C\hat{x}(k)$ should be trusted more when there is less observation noise.

9.1.5. Kalman filter and classical parameter estimation. Consider the problem of estimating parameter x from the observations

$$y = Cx + v,$$

where $E\{vv^T\} = V$. We wish to find the *linear, unbiased, minimum variance* estimator $\hat{x}^*$. Namely, in the class of $\hat{x} = Ky$, and $E\{\hat{x}\} = E\{x\}$, we have

$$E\{(x - \hat{x}^*)^T(x - \hat{x}^*)\} \to \min.$$

The classical Gauss-Markov theorem tells us

$$\hat{x}^* = \mathcal{I}^{-1}C^T V^{-1} y,$$

where $\mathcal{I} = C^T V^{-1} C$ is called the *information matrix*.

Now an interesting question is how this compares with Kalman filter:

$$\hat{x}(t+1) = \hat{x}(t) + P(t)C^T(CP(t)C^T + V)^{-1}(y - C\hat{x}(t)).$$

We can view $\hat{x}(t)$ and $P(t)$ as the priori information we have on x. Rewrite

$$\begin{aligned}\hat{x}(t+1) &= P(t)C^T(CP(t)C^T+V)^{-1}y + [I-P(t)C^T(CP(t)C^T+V)^{-1}C]\hat{x}(t)\\ &= [P(t)^{-1}+\mathcal{I}]^{-1}C^TV^{-1}y + [P(t)^{-1}+\mathcal{I}]^{-1}P(t)^{-1}\hat{x}(t).\end{aligned}$$

Here we have used the equalities

$$PC^T[CPC^T+V]^{-1} = [I+PV^{-1}C]^{-1}PC^TV^{-1}$$

and

$$I - PC^T(CPC^T+V)^{-1}C = [I+PC^TV^{-1}C]^{-1}.$$

Conclusion: When $P_0^{-1} = 0$, Kalman filter is the same as Gauss-Markov estimation!

9.1.6. Duality between estimation and control. Just as there is a duality between reachability and observability (see Section 3.7, Chapter 3), there is a duality between control and estimation. This should not be surprising since in fact estimation is a problem of observation. We now proceed to demonstrate this principle of duality for the Kalman filter.

To this end, form the adjoint

$$z(t) = A(t)^Tz(t+1) + C(t)^Tu(t+1); \quad z(T) = a \tag{76}$$

of the system (66), formed as usual by reversing time and transposing systems matrices. In analogy to the corresponding construction in Chapter 3, we form the scalar product z^Tx and take differences, i.e.

$$\begin{aligned}&z(t+1)^Tx(t+1) - z(t)^Tx(t)\\ =&z(t+1)^TA(t)x(t) + z(t+1)^TB(t)v(t) - z(t+1)^TA(t)x(t) - u(t+1)^TC(t)x(t)\\ =&z(t+1)^TB(t)v(t) + u(t+1)^TD(t)w(t) - u(t+1)^Ty(t)\end{aligned}$$

where we have also used the second of equations (66). Summing this from $t=0$ to $t=T-1$, we obtain

$$a^Tx(T) - z(0)^Tx_0 = \sum_{t=0}^{T-1}[z(t+1)^TB(t)v(t) + u(t+1)^TD(t)w(t)] - \sum_{t=0}^{T-1}u(t+1)^Ty(t)$$

and consequently

$$\begin{aligned}&E\{[a^Tx(T) + \sum_{t=0}^{T-1}u(t+1)^Ty(t)]^2\}\\ =&z(0)^TP_0z(0) + \sum_{s=1}^{T}[z(s)^TQ(s)z(s) + u(s)^TR(s)u(s)]\end{aligned} \tag{77}$$

where

$$\begin{cases}Q(t) = B(t-1)B(t-1)^T\\ R(t) = D(t-1)D(t-1)^T\end{cases} \tag{78}$$

Recall now that finding a vector sequence $\{u(1), u(2), \dots, u(T)\}$ so that (77) is minimized amounts to determining the least squares estimate of $a^T x(T)$ given the data

$$\{y_1(0), \dots, y_m(0), y_1(1), \dots, y_m(1), \dots, y_1(T-1), \dots, y_m(T-1)\}.$$

In other words,

$$a^T \hat{x}(T) = -\sum_{t=0}^{T-1} u^*(t+1)^T y(t)$$

where u^* is the optimal choice of u.

On the other hand, u^* is also the optimal control minimizing the quadratic cost criterion in the second member of (77) under the dynamics of system (76). This is a *linear-quadratic regulator problem* of the type discussed in Chapter 7 in continuous time, one difference being that time has been reversed and matrices transposed. It can be shown that this control problems has an optimal feedback solution

$$u(t) = K(t)^T x(t)$$

just as in continuous time, and that K is precisely the Kalman gain. Consequently this linear-quadratic regulator problem and the Kalman filtering problem have dual discrete-time matrix Riccati equations. Since we have developed linear-quadratic optimal control in continuous time in Chapter 7 and 8, we shall pursuit this line of investigation further in the context of continuous-time Kalman filtering.

9.2. The continuous-time Kalman filter

The continuous-time counterpart of the linear stochastic system (66) may be written

$$\begin{cases} \dot{x}(t) = A(t)x(t) + B(t)v(t); \quad x(0) = x_0 \\ y(t) = C(t)x(t) + D(t)w(t) \end{cases} \tag{79}$$

where x_0, v and w are centered and pairwise uncorrelated, and where v and w are white noise processes, i.e.

$$E\Big\{ \begin{bmatrix} v(t) \\ w(t) \end{bmatrix} \begin{bmatrix} v(s)^T, w(s)^T \end{bmatrix} \Big\} = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \delta(t-s)$$

where δ is the *Dirac delta function* having the property

$$\int_{-\infty}^{\infty} f(t)\delta(t)dt = f(0).$$

As should be well-known this is not a function but a generalized function, and therefore we enter into a somewhat grey area when it comes to considering the stochastic system (79). In fact, v, w and y must be regarded as *generalized stochastic processes* for which there is a literature. However, have no fear! As long as we handle the Dirac function appropriately (as we have been taught) and stay out of nonlinear filtering (as we shall) we are

OK. Furthermore, we set $P_0 \triangleq E\{x_0 x_0^T\}$ and assume that $DRD^T > 0$ for all t.

The problem at hand, now as before, is to find a linear state estimator

i.e., in external systems description,

$$\hat{x}(t) = \int_0^t G(t,s)y(s)ds,$$

where G takes values in $\mathbb{R}^{n\times m}$, such that

$$E[(x(t) - \hat{x}(t)])(x(t) - \hat{x}(t)])^T]$$

is minimized. The most natural way of derivation is perhaps to extend the discrete-time case by formally letting the sampling time tend to zero. We will also discuss an alternative approach by proceeding along the lines of Section 9.1.6.

9.2.1. Continuous time Kalman filter derived from the discrete case. A key issue in the extension to the continuous case is how to handle the noise. Heuristically (we do not intend to be very rigorous!), we can understand the white noises as the derivatives of some *Brownian motion* (although this derivative does not exist in a conventional sense). For example,

$$\int_s^t w(r)dr = \beta(t) - \beta(s),$$

where

$$E(\beta(t) - \beta(s)) = 0,\ E(\beta(t) - \beta(s))(\beta(t) - \beta(s))^T = R(t-s),\ t > s.$$

Thus,

$$\int_s^t Rdr = E\int_s^t w(r)dr \int_s^t w^T(\tau)d\tau.$$

Then,

$$\int_s^t (\int_s^t E\{w(r)w^T(\tau)d\tau - R\})dr = 0.$$

Since this is true for any interval, we have

$$\int_s^t E\{w(r)w^T(\tau)\}d\tau = R,\ \forall r \in [t,s].$$

Thus,

$$E\{w(r)w^T(\tau)\} = R\delta(r-\tau).$$

We can derive $E\{w(t)\} = 0$ similarly.

Now we use the discrete Kalman filter to derive the continuous one by letting $\Delta t \to 0$.

Let $x(t+1) = x(t+\Delta t)$, when Δt is very small, we have ("$\approx$" means equal up to $\mathcal{O}(\Delta t^2)$)

$$A_d = e^{A\Delta t} \approx I + A\Delta t,\ C_d = C.$$

Since $v_d(t) = \int_t^{t+\Delta t} e^{A(t+\Delta t-s)} v(s)ds \approx \beta(t+\Delta t) - \beta(t)$,

$$Q_d \approx Q\Delta t,\ R_d \approx R/\Delta t.$$

Then, $K_d(t) \approx P(t)C^T(DRD^T)^{-1}\Delta t$. We have

$$\hat{x}(t+1) \approx (I + A\Delta t)(\hat{x}(t) + K_d(t)(y(t) - C\hat{x}(t)),$$

or,

$$\hat{x}(t+1) - \hat{x}(t) \approx A\hat{x}(t)\Delta t + P(t)C^T(DRD^T)^{-1}(y(t) - C\hat{x}(t))\Delta t.$$

Thus, by dividing both sides with Δt and taking the limit, we have

$$\dot{\hat{x}}(t) = A\hat{x}(t) + K(t)(y(t) - C\hat{x}(t)),$$

where $K(t) = P(t)C^T(DRD^T)^{-1}$.

Since

$$\begin{aligned} P(t+\Delta t) &\approx (I + A\Delta t)(I - K_d C)P(t)(I + A\Delta t)^T + BQ\Delta t B^T \\ &\approx P(t) + (AP(t) + P(t)A^T)\Delta t - P(t)C^T(DRD^T)^{-1}CP(t)\Delta t + BQB^T\Delta t \end{aligned}$$

Similarly, we obtain

$$\dot{P}(t) = AP(t) + P(t)A^T - P(t)C^T(DRD^T)^{-1}CP(t) + BQB^T,$$

and we assume $P(0) = P_0$ is known.

The Riccati equation looks very similar to that we studied in Chapter 7, thus it is interesting to find out the duality between optimal control and optimal filtering.

9.2.2. The dual control problem. Define the dual control system

$$\dot{z}(t) = -A(t)^T z(t) + C(t)^T u(t); \quad z(T) = a \tag{80}$$

evolving backward in time on the interval $[0, T]$ and form as before

$$\begin{aligned} \frac{d}{dt}(z^T x) &= z^T\dot{x} + \dot{z}^T x \\ &= z^T Ax + z^T Bv - z^T Ax + u^T Cx \\ &= z^T Bv - u^T Dw + u^T y \end{aligned}$$

where (79) has been needed. Integrating this over the interval $[0, T]$, we obtain

$$a^T x(T) - z(0)^T x_0 = \int_0^T (z^T Bv - u^T Dw)dt + \int_0^T u^T y dt$$

and consequently

$$E\{[a^T x(T) - \int_0^T u^T y dt]^2 = z(0)^T P_0 z(0) + \int_0^T (z^T\tilde{Q}z + u^T\tilde{R}u)dt \tag{81}$$

where $\tilde{Q} \triangleq BQB^T$ and $\tilde{R} \triangleq DRD^T > 0$. Therefore the linear least-squares problem is equivalent to minimizing the quadratic cost criterion in the second member of (81) subject to the constraint of the systems equations (80).

Modulo time reversal, this is a linear-quadratic regulator problem of the type discussed in Chapter 7, and it has a feedback solution

$$u(t) = K(t)^T z(t) \tag{82}$$

where

$$K = PC^T \tilde{R}^{-1} \tag{83}$$

where P is the solution of the matrix Riccati equation

$$\begin{cases} \dot{P} = AP + PA^T - PC^T \tilde{R}^{-1} CP + \tilde{Q} \\ P(0) = P_0 \end{cases} \tag{84}$$

To see this, apply the time reversal operation $t \to T - t$ to the present linear-quadratic problem (changing the sign of all derivatives) and then apply the results of Chapter 7. Then reverse time again to obtain (82)-(84).

The optimal control law (82) yields the closed-loop system

$$\dot{z}(t) = -[A(t) - K(t)C(t)]^T z(t); \quad z(T) = a \tag{85}$$

i.e., as follows from property (3) in Chapter 2, $z(t) = \Psi(T, s)^T a$, where Ψ is the transition matrix

$$\frac{\partial \Psi}{\partial t}(t, s) = [A(t) - K(t)C(t)]\Psi(t, s); \quad \Psi(s, s) = I \tag{86}$$

and consequently the optimal open-loop control is

$$u^*(t) = K(t)^T \Psi(T, t)^T a \tag{87}$$

This will be used to derive the Kalman filter.

9.2.3. The Kalman filter revisited. The optimal control u^* of the linear-quadratic control problem (80)-(81) also provides us with the best linear least-squares estimates

$$\int_0^T u^*(s)^T y(s) ds$$

of $a^T x(T)$ given the data $\{y(t);\, t \in [0, T]\}$. Inserting (87), we obtain

$$\int_0^T u^*(s)^T y(s) ds = a^T \int_0^T \Psi(T, s) \hat{y}(s) ds$$

Since this holds for an arbitrary $a \in \mathbb{R}^n$ and an arbitrary $T > 0$, we must have

$$\hat{x}(t) = \int_0^t \Psi(t, s) \hat{y}(s) ds$$

which, in view of (86), satisfies

$$\begin{cases} \dfrac{d\hat{x}}{dt} = [A(t) - K(t)C(t)]\hat{x}(t) + K(t)y(t) \\ \hat{x}(0) = 0 \end{cases} \tag{88}$$

Another way of writing this, which exhibits the way in which the state estimate $\hat{x}$ is updated, is as follows.

$$\begin{cases} \dfrac{d\hat{x}}{dt} = A(t)\hat{x}(t) + K(t)[y(t) - C(t)\hat{x}(t)] \\ \hat{x}(0) = 0 \end{cases} \tag{89}$$

In fact, just as in discrete time, we can show that

$$\tilde{y}(t) = y(t) - C(t)\hat{x}(t)$$

is a white noise process, and it is called the innovation process. The filter (89) together with the matrix Riccati equation (84) determining the gain (83) is known as the *Kalman* (or the *Kalman-Bucy*) *filter*.

Another way to introduce the Kalman filter is as follows. Suppose that we define the process $\hat{x}$ by means of (88) without connecting it to the optimal estimation problem and then form

$$\begin{aligned} \frac{d}{dt}(z^T\hat{x}) &= z^T\dot{\hat{x}} + \dot{z}^T\hat{x} \\ &= z^T(A - KC)\hat{x} + z^TKy - z^T(A - KC)\hat{x} \\ &= u^Ty \end{aligned}$$

where we have used (85) and (82). Integrating this we obtain

$$a^T\hat{x}(t) = \int_0^t u(s)^Ty(s)ds$$

which is the Kalman estimate if K, and hence u, is chosen in an optimal fashion.

9.2.4. The steady-state Kalman filter. Consider a time-invariant stochastic system

$$\begin{cases} \dot{x} = Ax + Bv \\ y = Cx + Dw \end{cases}$$

where A, B, C and D are constant, and suppose that (A, B) is completely reachable and (C, A) completely observable. Let us consider what happens with the Kalman filter when the interval of observation becomes large.

It follows from the theory developed in Chapter 8 that $P(t)$ tends to a limit P as $t \to \infty$, where P is the unique positive definite symmetric solution of the algebraic Riccati equation

$$AP + PA^T - PC^T(DRD^T)^{-1}CP + BQB^T = 0, \tag{90}$$

and consequently the Kalman gain tends to

$$K = PC^T(DRD^T)^{-1} \tag{91}$$

Therefore, for all practical purpose, we may as soon as the observation interval is sufficiently large to take care of the transient behavior of the Kalman filter use the steady-state Kalman filter

$$\frac{d\hat{x}}{dt} = A\hat{x} + K(y - C\hat{x}) \tag{92}$$

where K is constant and given by (90)-(91).

EXAMPLE 9.2.1. Determine the steady state Kalman filter for

$$\Sigma \quad \begin{cases} \dot{x}_1 = 2v \\ \dot{x}_2 = x_1 \\ y = x_2 + w \end{cases}$$

where $E\{v(t)v(s)\} = E\{w(t)w(s)\} = I\delta(t-s)$ and $E\{v(t)w(s)\} = 0$.

Since $A = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right]$, $C = [0, 1]$ and $D = 1$, (A, B) is completely reachable and (C, A) is completely observable (check yourself!) and consequently the algebraic Riccati equation (90) has a unique real positive definite symmetric solution

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

from which we have the three nonlinear equations

$$\begin{cases} -p_{12}^2 + 4 = 0 \\ p_{11} - p_{12}p_{22} = 0 \\ 2p_{12} - p_{22}^2 = 0. \end{cases}$$

The first equation yields $p_{12} = \pm 2$. From the third equation we see that only $p_{12} = 2$ yields real solutions $p_{22} = \pm 2$. Hence, we have two solutions

$$P_1 = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, \quad P_2 = \begin{bmatrix} -4 & 2 \\ 2 & -2 \end{bmatrix}.$$

Here P_1 is the required positive definite solution whereas P_2 is negative definite. Thus the optimal gain is

$$K = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

so the steady-state Kalman filter becomes

$$\begin{cases} \dot{\hat{x}}_1 = 2(y - \hat{x}_2) \\ \dot{\hat{x}}_2 = \hat{x}_1 + 2(y - \hat{x}_2). \end{cases}$$

Writing the filtering equations on the form

$$\frac{d\hat{x}}{dt} = \begin{bmatrix} 0 & -2 \\ 1 & -2 \end{bmatrix} \hat{x} + \begin{bmatrix} 2 \\ 2 \end{bmatrix} y$$

we see that the coefficient matrix $\left[\begin{smallmatrix} 0 & -2 \\ 1 & -2 \end{smallmatrix}\right]$ is a stable matrix as required by the theory of Chapter 8, its eigenvalues being $-1 \pm i$.

Index