Introduction to Systems and Control Theory

Eva Zerz
FB Mathematik
TU Kaiserslautern
67663 Kaiserslautern
Germany

July 2002

Improved version
July 2005

Abstract: These lecture notes give a completely self-contained introduction to the control theory of linear time-invariant systems. No prior knowledge is required apart from linear algebra and some basic familiarity with ordinary differential equations. Thus, the course is suited for students of mathematics in their second or third year, and for theoretically inclined engineering students. Because of its appealing simplicity and elegance, the behavioral approach has been adopted to a large extent. A short list of recommended text books on the subject has been added, as a suggestion for further reading.

Contents

Chapter 1

Introduction

1.1 What is a system?

The word "system" comes from the Greek word $\sigma\upsilon\sigma\tau\eta\mu\alpha$, which originally meant something like "to stand/put/place together."

Here are some "definitions" I found surfing the web:

A system is

- … a thing that has components which may act independently, but are connected somehow.
- … a complex unity formed of many often diverse parts subject to a common plan or serving a common purpose.
- … a collection of parts that interact with each other to function as a whole.
- … a group of interrelated elements involved in a collective entity.
- … a set of interrelated components related by flows of energy, material, or information.
- I will know one when I see it.

If that sounds vague, it's meant to.

We are interested in **dynamical** systems (Greek: $\delta\upsilon\nu\alpha\mu\iota\kappa o\varsigma$; original meaning: powerful/strong; here: pertaining to power in motion, involving or causing action or change, opposed to static).

The components/parts/elements of a dynamical system evolve in time. Mathematically speaking, they are functions of time, and they will be called "**signals**."

We start with some very general, abstract, and comprehensive definitions of a dynamical system. In the following section, we will introduce some important structural properties. Thus we will specialize step by step, until we arrive at a more concrete class of dynamical systems, which will then be investigated in detail.

Definition 1.1 A **dynamical system** Σ is determined by the following data:

- a set T, called the **time set**;
- a set W, called the **signal value set**;
- a set $\mathcal{B} \subset W^T$, called the **behavior**.

The set T is our mathematical model of **time**. We will deal exclusively with the following cases:

- $T = \mathbb{Z}$ or a subinterval, especially $\mathbb{N} = \{0, 1, 2, \dots\}$ (**discrete** time);
- $T = \mathbb{R}$ or a subinterval, especially $\mathbb{R}_+ := [0, \infty)$ (**continuous** time).

A **signal** w is a function of time, taking its values in the signal value set W. We write

$$w : T \to W, t \mapsto w(t).$$

The set W^T is the set of all functions from T to W, therefore it is the **set of all signals**.

Typically, not all signals in W^T can occur in our system (or at least, a system in which *anything* can happen would not be very interesting from the mathematical point of view). Usually, there will be a system **law** which is satisfied only by some signals.

The subset $\mathcal{B}$ of W^T formalizes this law which governs the system. The signals $w \in \mathcal{B}$ are precisely those which are compatible with the system law, that is, they may occur in our system. We also call them **admissible**, and we write

$$\mathcal{B} = \{w \in W^T \mid w \text{ satisfies the system law}\}.$$

Typically, a signal has several, say q, components coming from the same set K (usually, $K = \mathbb{R}$). Then $W = K^q$, and a signal has the form

$$w : T \to K^q, t \mapsto w(t) = \begin{bmatrix} w_1(t) \\ \vdots \\ w_q(t) \end{bmatrix}.$$

In that case, we call w a **signal vector**, and W^T is the **set of all signal vectors**. Each component w_i of w is again called a **signal**.

This leads to a slight modification of our definition.

Definition 1.2 A **dynamical system** Σ is determined by the following data:

- a set T, called the **time set**;
- a set K, called the **signal value set**;
- a positive integer q, called the **number of signals**;
- a set $\mathcal{B} \subset (K^q)^T = (K^T)^q$, called the **behavior.**

Define $\mathcal{A} = K^T$. This will be called a **signal set**. Finally, we arrive at the following definition.

Definition 1.3 A dynamical system Σ is determined by the following data:

- a set $\mathcal{A}$, called the **signal set**;
- a positive integer q, called the **number of signals**;
- a set $\mathcal{B} \subset \mathcal{A}^q$, called the **behavior**.

Here, $\mathcal{A}$ and q define the setting/mathematical framework for our description of the system. $\mathcal{A}$ is the set of all signals, and $\mathcal{A}^q$ is the set of all signal vectors with q components. The signal vectors $w \in \mathcal{B}$ are precisely those which are compatible with the system law, that is,

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid w \text{ satisfies the system law}\}.$$

Remark 1.4 Prototypes of signal sets:

- the set of all functions from $\mathbb{N}$ to $\mathbb{R}$,
$$\mathcal{A} = \mathbb{R}^{\mathbb{N}}$$
(such functions are usually called sequences);
- the set of k times continuously differentiable functions from $\mathbb{R}$ to $\mathbb{R}$ (where $0 \le k \le \infty$)
$$\mathcal{A} = \mathcal{C}^k(\mathbb{R});$$
- the set of generalized functions or distributions
$$\mathcal{A} = \mathcal{D}'(\mathbb{R}).$$

Remark 1.5 The intermediary Definition 1.2 is actually superfluous. Any system according to 1.2 can put into the setting of Definition 1.1 by putting $W = K^q$ and into the setting of Definition 1.3 by putting $\mathcal{A} = K^T$.

A system according to Definition 1.1 can be transformed to Definition 1.3 only if $W = K^q$. However, this is not a serious restriction, because it is true for most systems of interest.

In that case, Definition 1.3 is indeed the most general one, because it encompasses distributions.

Example 1.6 The motion of a planet around the sun according to Kepler's laws. The time set is certainly continuous. Whether you choose $\mathbb{R}$ or $\mathbb{R}_+$ as your time model, depends on your religious and/or scientific beliefs. Let's not touch these delicate issues and choose $T = \mathbb{R}$ for simplicity. The position of a planet in space is determined by its three real coordinates, therefore $K = \mathbb{R}$ and $q = 3$, or $W = \mathbb{R}^3$. If we let $\mathcal{A}$ be the set of all functions from $\mathbb{R}$ to $\mathbb{R}$, then $\mathcal{A}^q = W^T$.

Being inhabitants of the Earth, we may suspect that $\mathcal{A} = \mathcal{C}^k(\mathbb{R})$ for some k which is large enough for comfort. Anyhow, we put

$$\mathcal{B} = \{w \in \mathcal{A}^3 \mid w \text{ satisfies Kepler's laws}\}.$$

Example 1.7 Suppose we have two signals x and b which are linked via

$$\dot{x}(t) = A(t)x(t) + b(t)$$

where $A : \mathbb{R} \to \mathbb{R}^{n\times n}, t \mapsto A(t)$ is a smooth map. From the theory of ordinary differential equations, we know that if $b : \mathbb{R} \to \mathbb{R}^n, t \mapsto b(t)$ is continuous, then the initial value problem (where $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ are arbitrary)

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + b(t) \\ x(t_0) &= x_0 \end{aligned}$$

has a unique solution $x : \mathbb{R} \to \mathbb{R}^n, t \mapsto x(t)$ which is in $\mathcal{C}^1(\mathbb{R})^n$. Therefore, we may put $T = \mathbb{R}$, $W = \mathbb{R}^{2n}$ and

$$\mathcal{B} = \{(x, b) \in W^T \mid (x, b) \in \mathcal{C}^1(\mathbb{R})^n \times \mathcal{C}^0(\mathbb{R})^n \text{ and } \dot{x}(t) = A(t)x(t) + b(t)\}.$$

This behavior has the remarkable property

$$\forall b \in \mathcal{C}^0(\mathbb{R})^n \, \exists x \in \mathcal{C}^1(\mathbb{R})^n : (x, b) \in \mathcal{B}. \tag{1.1}$$

Later on, we will call signals b with this property **inputs**.

In order to avoid having to specify how many times a function is differentiable, it is convenient to work with distributions. We put $\mathcal{A} = \mathcal{D}'(\mathbb{R})$, $q = 2n$, and

$$\mathcal{B} = \{(x, u) \in \mathcal{A}^q \mid \dot{x} = A(t)x(t) + b(t)\}.$$

This is more general because we may now have discontinuous b, such as, e.g., the Heaviside function. Again, we have (compare with (1.1))

$$\forall b \in \mathcal{A}^n \exists x \in \mathcal{A}^n : (x, b) \in \mathcal{B}.$$

Take the scalar example

$$\dot{x}(t) = x(t) + b(t).$$

The classical solutions are

$$x(t) = e^t x_0 + \int_0^t e^{t-\tau} b(\tau) d\tau$$

where $x_0 \in \mathbb{R}$ is arbitrary. If b is the Heaviside function

$$b(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$$

then we obtain

$$x(t) = e^t x_0 + \begin{cases} 0 & \text{if } t < 0 \\ e^t - 1 & \text{if } t \geq 0 \end{cases}$$

which is $\mathcal{C}^0$, but not $\mathcal{C}^1$, and hence not a classical solution.

Example 1.8 Genetics/Gender-linked genes: An allele (a certain form of a gene) is located on the X-chromosome. Females have two X-chromosomes, males have one X- and one Y-chromosome. Let $p^f(i)$ be the frequency of the allele in the female gene pool of the i-th generation and let $p^m(i)$ be the same for the male gene pool. Since a son inherits his X-chromosome from the mother,

$$p^m(i+1) = p^f(i)$$

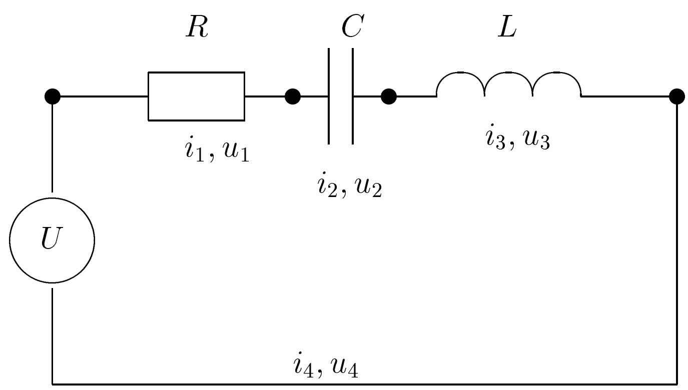
and since a daughter receives one X-chromosome from the father and one from the mother,

$$p^f(i+1) = \tfrac{1}{2}(p^f(i) + p^m(i)).$$

The time set is discrete. Both $T = \mathbb{Z}$ and $T = \mathbb{N}$ are suitable choices. We have two signals p^m and p^f, taking their values in $\mathbb{R}$, hence $W = \mathbb{R}^2$ and

$$\mathcal{B} = \left\{ \begin{bmatrix} p^m \\ p^f \end{bmatrix} \in W^T \mid \begin{bmatrix} p^m(i+1) \\ p^f(i+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} p^m(i) \\ p^f(i) \end{bmatrix} \text{ for all } i \in T \right\}.$$

Example 1.9 An electrical circuit.



This system involves 9 signals: U, $i_1, \dots, i_4$, and $u_1, \dots, u_4$. Kirchhoff's current law says that

$$i_1 = i_2 = i_3 = i_4 =: I$$

and Kirchhoff's voltage law says that

$$u_1 + u_2 + u_3 + u_4 = 0.$$

We have a free voltage source, $u_4 = U$. Moreover, we have the following constituent equations for the resistor (R), capacitor (C), and inductor (L), respectively:

$$u_1 = Ri_1 = RI, \quad C\tfrac{\mathrm{d}u_2}{\mathrm{d}t} = i_2 = I, \quad L\tfrac{\mathrm{d}i_3}{\mathrm{d}t} = L\tfrac{\mathrm{d}I}{\mathrm{d}t} = u_3.$$

Let us describe the system in terms of the signals I, U, u_1, u_2, u_3 (the other 4 signals are easily obtained from them.) We put $\mathcal{A} = \mathcal{D}'(\mathbb{R})$ or $\mathcal{D}'(\mathbb{R}_+)$, $q = 5$ and

$$\mathcal{B} = \{(I, U, u_1, u_2, u_3)^T \in \mathcal{A}^5 | u_1 + u_2 + u_3 + U = 0, u_1 = RI, C\dot{u}_2 = I, L\dot{I} = u_3\}.$$

1.2 System properties: LTID systems

For the definition of linearity, we need to fix an underlying number field K. We will focus on $K = \mathbb{R}$, and therefore we give the definition for that case only.

Definition 1.10 *A dynamical system* $\Sigma = (\mathcal{A}, q, \mathcal{B})$ *is called* **linear** *if* $\mathcal{A}$ *is a real vector space, and* $\mathcal{B}$ *is a subspace of* $\mathcal{A}^q$.

The first requirement means that linear combinations of signals are again signals,

$$a_1, a_2 \in \mathcal{A}, \lambda_1, \lambda_2 \in \mathbb{R} \quad \Rightarrow \quad \lambda_1 a_1 + \lambda_2 a_2 \in \mathcal{A}$$

and the second requirement means that linear combinations of admissible signal vectors are again admissible signal vectors,

$$w_1, w_2 \in \mathcal{B}, \lambda_1, \lambda_2 \in \mathbb{R} \quad \Rightarrow \quad \lambda_1 w_1 + \lambda_2 w_2 \in \mathcal{B}.$$

We call this condition the **superposition principle.**

Remark 1.11 All signal spaces in 1.4 are real vector spaces.

Any $\mathcal{A}$ of the form $\mathcal{A} = \mathbb{R}^T$ is a real vector space. For $a_1, a_2 \in \mathcal{A}$, $\lambda_1, \lambda_2 \in \mathbb{R}$, we have

$$\lambda_1 a_1 + \lambda_2 a_2 : T \to \mathbb{R}, t \mapsto \lambda_1 a_1(t) + \lambda_2 a_2(t).$$

More generally: If W is a real vector space, then so is W^T.

Definition 1.12 *A dynamical system $\Sigma = (T, W, \mathcal{B})$ is called* **linear** *if W is a real vector space and $\mathcal{B}$ is a subspace of W^T.*

Definition 1.13 Let T be such that

$$t_1, t_2 \in T \quad \Rightarrow \quad t_1 + t_2 \in T. \tag{1.2}$$

For $\tau \in T$, we define the shift operator σ_τ by

$$\sigma_\tau : W^T \to W^T, w \mapsto \sigma_\tau w$$

where

$$(\sigma_\tau w)(t) = w(t + \tau).$$

A dynamical system $\Sigma = (T, W, \mathcal{B})$ is called **shift-invariant** (or: time-invariant) if for all $\tau \in T$

$$w \in \mathcal{B} \quad \Rightarrow \quad \sigma_\tau w \in \mathcal{B}.$$

Remark 1.14 Note that the time sets $T = \mathbb{R}$, $T = \mathbb{R}_+$, $T = \mathbb{Z}$, and $T = \mathbb{N}$ all satisfy (1.2). When dealing with time-invariant systems, we will always assume that T is one of these four sets.

Definition 1.15 A dynamical system $\Sigma = (T, W, \mathcal{B})$ is called a **differential (difference) system** if its time set is continuous (discrete) and its system law is given by differential (difference) equations.

This is the class of systems we will study: linear, time-invariant (LTI) differential (difference) systems. The system laws will have the following form:

Differential systems: Systems of linear differential equations with constant coefficients. These can be put in the form

$$(R_d \frac{\mathrm{d}^d}{\mathrm{d}t^d} + \ldots + R_1 \frac{\mathrm{d}}{\mathrm{d}t} + R_0)w = 0 \tag{1.3}$$

where $R_i \in \mathbb{R}^{p \times q}$ are real matrices. We define

$$R := R_d s^d + \ldots + R_1 s + R_0.$$

Then R is a polynomial $p \times q$ matrix, and we may rewrite (1.3) in the concise form

$$R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0.$$

Difference systems: Systems of linear difference equations with constant coefficients. These can be put in the form

$$R_d w(t+d) + \ldots + R_1 w(t+1) + R_0 w(t) = 0 \text{ for all } t \in T$$

where $R_i \in \mathbb{R}^{p \times q}$ are real matrices. Using the shift operator σ_τ, we may write

$$(R_d \sigma_d + \ldots + R_1 \sigma_1 + R_0)w = 0. \tag{1.4}$$

We define $\sigma := \sigma_1$, then $\sigma_k = \sigma^k$ (k-fold application of σ). If we put again

$$R := R_d s^d + \ldots + R_1 s + R_0$$

then we can write (1.4) in the concise form

$$R(\sigma)w = 0.$$

Remark 1.16 The examples discussed so far can be classified as follows.

Example 1.6: Kepler (1571-1630) formulated his laws in a non-differential way. The differential calculus was developed by Newton (1642-1727) and Leibniz (1646-1716). It was in fact Newton who came up with a differential equation for the motion of planets around the sun (the law of gravity). However, it is quite clear that Kepler's laws describe a non-linear and time-invariant system.

Example 1.7: The behavior given by $\dot{x}(t) = A(t)x(t) + b(t)$ is differential and linear. In general, it is not time-invariant. In fact, it is time-invariant if and only if A is a constant function.

Example 1.8: The system given by

$$w(t+1) = Aw(t) \text{ for all } t \in T$$

where

$$A = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

is a linear time-invariant difference system. Using the shift operator, it reads $\sigma w = Aw$ or

$$(\sigma I - A)w = 0.$$

The polynomial matrix R takes the form

$$R = sI - A = \begin{bmatrix} s & -1 \\ -\frac{1}{2} & s - \frac{1}{2} \end{bmatrix}.$$

Example 1.9: The system of the electrical circuit is a linear, time-invariant and differential. The system law reads

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ R & 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & C\frac{\mathrm{d}}{\mathrm{d}t} & 0 \\ L\frac{\mathrm{d}}{\mathrm{d}t} & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} I \\ U \\ u_1 \\ u_2 \\ u_3 \end{bmatrix} = 0.$$

We may rewrite this in the form $\hat{R}(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ by putting

$$w = \begin{bmatrix} I & U & u_1 & u_2 & u_3 \end{bmatrix}^T$$

and

$$\hat{R} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ R & 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & Cs & 0 \\ Ls & 0 & 0 & 0 & -1 \end{bmatrix}.$$

1.3 LTI differential equations and their distributional solutions

Consider $P(\frac{\mathrm{d}}{\mathrm{d}t})y = v$, where P is square and non-singular. Let y_p be one particular solution. Then any solution y has the form $y = y_p + y_h$, where y_h is a solution of the associated homogeneous equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y_h = 0$.

1.3.1 The homogeneous equation

Consider the homogeneous equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$. Assume that

$$P = P_d s^d + \ldots + P_1 s + P_0$$

with coefficient matrices $P_i \in \mathbb{R}^{p\times p}$. After reduction to first order, we have

$$K\dot{\xi} = L\xi$$

where $K, L \in \mathbb{R}^{n\times n}$. Without loss of generality, we may put $n = dp$,

$$\xi = \begin{bmatrix} y \\ \frac{\mathrm{d}}{\mathrm{d}t}y \\ \vdots \\ \frac{\mathrm{d}^{d-1}}{\mathrm{d}t^{d-1}}y \end{bmatrix}, \; K = \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & P_d \end{bmatrix} \text{ and } L = \begin{bmatrix} 0 & I & & \\ \vdots & & \ddots & \\ 0 & & & I \\ -P_0 & -P_1 & \cdots & -P_{d-1} \end{bmatrix}.$$

Note that then $\det(P) = \det(sK - L)$, $\det(K) = \det(P_d)$, and $\det(L) = \pm\det(P_0)$.

Case 1: K is invertible, that is, after putting $A := K^{-1}L$, the system has the explicit form $\dot{\xi} = A\xi$. We say that ξ is a classical solution if it is $\mathcal{C}^1$ and $\dot{\xi} = A\xi$. It is well-known that

$$\Phi(t) := e^{At}$$

is a fundamental matrix for this system, that is, its n columns are a basis of the classical solution space. In other words, any classical solution ξ has the form $\xi(t) = e^{At}\xi_0$ for some $\xi_0 \in \mathbb{R}^n$. Note that the entries of Φ are $\mathcal{C}^\infty$. Hence every classical solution is also $\mathcal{C}^\infty$. It is known that the Wronski determinant $\det(\Phi(t)) = e^{\mathrm{trace}(A)t} \neq 0$ for all $t \in T$, and thus also the entries of Φ^{-1} are in $\mathcal{C}^\infty$, in fact, $\Phi^{-1}(t) = e^{-At}$.

Theorem 1.17 The equation $\dot{\xi} = A\xi$ has no distributional solutions apart from the classical solutions, that is, $\xi(t) = e^{At}\xi_0$, where $\xi_0 \in \mathbb{R}^n$.

Proof: Let $\xi \in \mathcal{D}'(T)^n$ be a distributional solution. Set $\eta := \Phi^{-1}\xi$, then $\Phi\eta = \xi$. Differentiate to get

$$\dot{\xi} = \dot{\Phi}\eta + \Phi\dot{\eta} = A\Phi\eta + \Phi\dot{\eta} = A\xi = A\Phi\eta.$$

Therefore $\Phi\dot{\eta} = 0$ and thus $\dot{\eta} = 0$. According to Lemma A.1, this implies $\eta = \eta_0$, where η_0 is a constant vector. Thus

$$\xi = \Phi\eta = \Phi\eta_0$$

showing that the distributional solution ξ is indeed a classical solution. $\square$

Case 2: K is not invertible. Recall that $\det(sK-L)=\det(P)$ which is non-zero by assumption.

Theorem 1.18 (Kronecker-Weierstraß form) Let $K, L \in \mathbb{R}^{n\times n}$ be such that $\det(sK-L)\neq 0$. Then there exist constant non-singular matrices U, V such that

$$UKV = \begin{bmatrix} I_\nu & 0 \\ 0 & N \end{bmatrix} \quad \text{and} \quad ULV = \begin{bmatrix} A_1 & 0 \\ 0 & I \end{bmatrix}$$

where N is a nilpotent matrix, that is, $N^k = 0$ for some $k \in \mathbb{N}$. The number ν, and the nilpotency index of N, that is, the smallest integer κ such that $N^\kappa = 0$, are uniquely determined by K, L.

Since $K\dot{\xi} = L\xi$ is equivalent to $UKVV^{-1}\dot{\xi} = ULVV^{-1}\xi$, we set $x := V^{-1}\xi$ and obtain

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 \\ N\dot{x}_2 &= x_2. \end{aligned}$$

The second equation implies (by repeated differentiation and multiplication by N) that $x_2 = 0$. Therefore, Case 2 can be reduced to Case 1.

Corollary 1.19 The equation $K\dot{\xi} = L\xi$ has no distributional solutions apart from the classical solutions, that is, $\xi(t) = Vx(t)$, where $x_1(t) = e^{A_1 t}x_{10}$ for some $x_{10} \in \mathbb{R}^\nu$ and $x_2(t) = 0$ for all t.

Summing up the two cases, we have the following result. We use that the entries of $\Phi(t) = e^{At}$ have the form

$$\Phi_{ij}(t) = \sum_\lambda a_\lambda(t) e^{\lambda t}$$

where $\lambda \in \mathbb{C}$ are the eigenvalues of A and $a_\lambda \in \mathbb{C}[t]$ are suitable polynomials. Note that

$$\det(P) = \det(sK-L) = \frac{1}{uv}\det(sI - A_1)\det(sN - I),$$

where $u = \det(U)$ and $v = \det(V)$ are non-zero constants. Since $\det(\lambda N - I) \neq 0$ for all $\lambda \in \mathbb{C}$, the zeros of $\det(P)$ are precisely the eigenvalues of A_1.

Theorem 1.20 The system $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ has no distributional solutions apart from the classical solutions, that is,

$$y(t) = \sum_{\lambda} a_\lambda(t) e^{\lambda t}$$

where $\lambda \in \mathbb{C}$ are the zeros of $\det(P)$, and $a_\lambda \in \mathbb{C}[t]^p$ are suitable polynomial vectors.

1.3.2 The inhomogeneous equation

Let $\mathcal{A} = \mathcal{D}'(T)$. Consider $P(\frac{\mathrm{d}}{\mathrm{d}t})y = v$, where $v \in \mathcal{A}^p$ is given. After reduction to first order, we obtain

$$K\dot{\xi} = L\xi + Mv$$

where $K, L \in \mathbb{R}^{n \times n}$ and $M \in \mathbb{R}^{n \times p}$.

Case 1: K is invertible, that is, after putting $A := K^{-1}L$, $B := K^{-1}M$, the system has the explicit form

$$\dot{\xi} = A\xi + Bv.$$

One uses the well-known "variation of constants" trick and sets

$$\xi = \Phi\psi$$

where $\Phi(t) = e^{At}$ is the fundamental matrix for the homogeneous equation, to obtain

$$\dot{\psi} = \Phi^{-1}Bv.$$

According to Lemma A.2, there exists a distribution ψ which satisfies this equation. Then $\xi = \Phi\psi$ is a particular solution of the inhomogeneous equation. Note that if v is $\mathcal{C}^0$, that is, continuous, then ξ is $\mathcal{C}^1$ and hence it is a classical solution.

Case 2: K is not invertible. Without loss of generality, rewrite $K\dot{\xi} = L\xi + Mv$ as $UKVV^{-1}\dot{\xi} = ULVV^{-1}\xi + UMv$, where U and V are as in Theorem 1.18. Then, putting $\xi = Vx$ and $B = UM$, the equations read

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 + B_1 v \\ N\dot{x}_2 &= x_2 + B_2 v. \end{aligned}$$

Assuming that $N^k = 0$, this implies that

$$x_2 = -(B_2 v + NB_2\dot{v} + \ldots + N^{k-1}B_2 v^{(k-1)}).$$

The equation for x_1 is again as in Case 1.

Theorem 1.21 Let $P \in \mathbb{R}[s]^{p\times p}$ be non-singular. For every $v \in \mathcal{A}^p$, there exists $y \in \mathcal{A}^p$ such that $P(\frac{\mathrm{d}}{\mathrm{d}t})y = v$.

This theorem will be called **fundamental principle** below.

1.4 LTI difference equations

Consider $P(\sigma)y = v$, where P is square and non-singular. Let y_p be one particular solution. Then any solution y has the form $y = y_p + y_h$, where y_h is a solution of the associated homogeneous equation $P(\sigma)y_h = 0$.

1.4.1 The homogeneous equation

Consider the homogeneous equation $P(\sigma)y = 0$. Assume that

$$P = P_d s^d + \ldots + P_1 s + P_0$$

with coefficient matrices $P_i \in \mathbb{R}^{p\times p}$. After reduction to first order, we have

$$K\sigma\xi = L\xi,$$

that is

$$K\xi(t+1) = L\xi(t) \quad \text{for all } t \in T$$

where $K, L \in \mathbb{R}^{n\times n}$. We may choose K, L as in Section 1.3 if we set

$$\xi = \begin{bmatrix} y \\ \sigma y \\ \vdots \\ \sigma^{d-1} y \end{bmatrix}.$$

Time set $T = \mathbb{N}$

Case 1: K is invertible, that is, after putting $A := K^{-1}L$, the system has the explicit form $\sigma\xi = A\xi$. It is well-known that

$$\Phi(t) := A^t$$

is a fundamental matrix for this system, that is, its n columns are a basis of the solution space. In other words, any solution ξ has the form $\xi(t) = A^t \xi_0$ for some $\xi_0 \in \mathbb{R}^n$.

Case 2: K is not invertible. Using Theorem 1.18, the system can be rewritten as

$$\begin{aligned} \sigma x_1 &= A_1 x_1 \\ N \sigma x_2 &= x_2. \end{aligned}$$

The second equation implies that $x_2 = 0$, because $N^k = 0$ and thus $x_2(t) = N x_2(t+1) = \ldots = N^k x_2(t+k) = 0$ for all t. Therefore, Case 2 can be reduced to Case 1.

Lemma 1.22 The equation $K \sigma \xi = L \xi$ has the solutions $\xi(t) = V x(t)$, where $x_1(t) = A_1^t x_{10}$ for some $x_{10} \in \mathbb{R}^\nu$ and $x_2(t) = 0$ for all t.

Summing up the two cases, we have the following result. We use that the entries of $\Phi(t) = A^t$ have the form

$$\Phi_{ij}(t) = \sum_\lambda a_\lambda(t) \lambda^t$$

where $\lambda \in \mathbb{C}$ are the eigenvalues of A and $a_\lambda \in \mathbb{C}[t]$ are suitable polynomials. Note that

$$\det(P) = \det(sK - L) = \frac{1}{uv} \det(sI - A_1) \det(sN - I),$$

where $u = \det(U)$ and $v = \det(V)$ are non-zero constants. Since $\det(\lambda N - I) \neq 0$ for all $\lambda \in \mathbb{C}$, the zeros of $\det(P)$ are precisely the eigenvalues of A_1.

Theorem 1.23 The solutions of the system $P(\sigma) y = 0$ have the form

$$y(t) = \sum_\lambda a_\lambda(t) \lambda^t$$

where $\lambda \in \mathbb{C}$ are the zeros of $\det(P)$, and $a_\lambda \in \mathbb{C}[t]^p$ are suitable polynomial vectors.

Time set $T = \mathbb{Z}$

Case 1: K, L are both invertible. Then $A := K^{-1}L$ is also invertible, and thus $\Phi(t) = A^t$ is again a fundamental matrix (note that we need invertibility of A for $\Phi(t)$ to be defined also for negative values of t).

Case 2: L is invertible. As usual, we may transform the system into Kronecker-Weierstraß form

$$U_1KV_1 = \begin{bmatrix} I & 0 \\ 0 & N_1 \end{bmatrix} \quad \text{and} \quad U_1LV_1 = \begin{bmatrix} A_1 & 0 \\ 0 & I \end{bmatrix}$$

to obtain, for $x = V_1^{-1}\xi$,

$$\begin{aligned} \sigma x_1 &= A_1 x_1 \\ N_1 \sigma x_2 &= x_2 \end{aligned}$$

where A_1 is invertible. Again, the second equation implies that $x_2 = 0$. Thus we have reduced Case 2 to Case 1.

Case 3: K is invertible. We rewrite

$$K\xi(t+1) = L\xi(t) \quad \text{for all } t \in \mathbb{Z}$$

as

$$L\xi(\tau - 1) = K\xi(\tau) \quad \text{for all } \tau \in \mathbb{Z}$$

and compute the Kronecker-Weierstraß form

$$U_2LV_2 = \begin{bmatrix} I & 0 \\ 0 & N_2 \end{bmatrix} \quad \text{and} \quad U_2KV_2 = \begin{bmatrix} A_2 & 0 \\ 0 & I \end{bmatrix}$$

where A_2 is invertible. Then we obtain the new system, for $y = V_2^{-1}\xi$,

$$\begin{aligned} y_1(\tau - 1) &= A_2 y_1(\tau) \\ N_2 y_2(\tau - 1) &= y_2(\tau) \end{aligned}$$

Then $y_2 = 0$ and $y_1(\tau) = A_2^{-\tau} y_{10}$ for some y_{10}. Note that

$$\det(P) = \det(sK - L) = \frac{1}{u_2 v_2} \det(sA_2 - I) \det(sI - N_2)$$

and thus, since zero is the only eigenvalue of the nilpotent matrix N_2,

$$\det(P)(\lambda) = 0 \quad \Leftrightarrow \quad \lambda = 0 \text{ or } \det(\lambda A_2 - I) = 0.$$

Thus the non-zero zeros of $\det(P)$ coincide with the eigenvalues of A_2^{-1}.

Case 4: K, L both singular. Then one uses the Kronecker-Weierstraß form as in Case 2 to obtain the values $\xi(t)$ for $t > 0$ and the Kronecker-Weierstraß form from Case 3 to determine the values $\xi(t)$ for $t < 0$. We omit the details.

Theorem 1.24 The solutions of $P(\sigma)y = 0$ have the form

$$y(t) = \sum_{\lambda} a_\lambda(t)\lambda^t$$

where $\lambda \in \mathbb{C}$ are the non-zero zeros of $\det(P)$ and $a_\lambda \in \mathbb{C}[t]^p$ are suitable polynomial vectors.

1.4.2 The inhomogeneous equation

For simplicity, we treat only the case $T = \mathbb{N}$, that is, $\mathcal{A} = \mathbb{R}^{\mathbb{N}}$. Consider $P(\sigma)y = v$, where $v \in \mathcal{A}^p$ is given. After reduction to first order, we obtain

$$K\sigma\xi = L\xi + Mv$$

where $K, L \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times p}$.

Case 1: K is invertible, that is, after putting $A := K^{-1}L$, $B := K^{-1}M$, the system has the form

$$\sigma\xi = A\xi + Bv.$$

Then

$$\xi(t) = \sum_{i=0}^{t-1} A^{t-i-1}Bv(i)$$

is a particular solution of the inhomogeneous equation.

Case 2: K is not invertible. Without loss of generality, rewrite $K\dot{\xi} = L\xi + Mv$ as $UKVV^{-1}\sigma\xi = ULVV^{-1}\xi + UMv$, where U and V are as in Theorem 1.18. Then, putting $\xi = Vx$ and $B = UM$, the equations read

$$\begin{aligned} \sigma x_1 &= A_1x_1 + B_1v \\ N\sigma x_2 &= x_2 + B_2v. \end{aligned}$$

Assuming that $N^k = 0$, this implies that

$$x_2 = -(B_2v + NB_2\sigma v + \ldots + N^{k-1}B_2\sigma^{k-1}v).$$

The equation for x_1 is again as in Case 1.

Theorem 1.25 Let $P \in \mathbb{R}[s]^{p \times p}$ be non-singular. For every $v \in \mathcal{A}^p$, there exists $y \in \mathcal{A}^p$ such that $P(\sigma)y = v$.

This is the discrete version of the fundamental principle.

Chapter 2

LTID systems: Basic facts

In the following, let R be a $p \times q$ polynomial matrix in the variable s, with real coefficients. We write

$$R \in \mathbb{R}[s]^{p\times q}.$$

From now on, we will restrict our discussion to the following **standard models**:

- The **continuous** standard model is

  $$\mathcal{B} = \{w \in \mathcal{A}^q \mid R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}$$

  where $\mathcal{A} = \mathcal{D}'(T)$ with $T = \mathbb{R}$ or $T = \mathbb{R}_+$.

- The **discrete** standard model is

  $$\mathcal{B} = \{w \in \mathcal{A}^q \mid R(\sigma)w = 0\}$$

  where $\mathcal{A} = \mathbb{R}^T$ with $T = \mathbb{N}$ or $T = \mathbb{Z}$.

2.1 Representations

The polynomial matrix R is called a **representation** of $\mathcal{B}$. Note that once $\mathcal{A}$ and q are fixed, the behavior $\mathcal{B}$ is uniquely determined by R. Conversely, there are many polynomial matrices which represent the same behavior. To see this, we introduce the concept of a unimodular matrix.

Definition 2.1 A square polynomial matrix U is called **unimodular** if its determinant is a non-zero constant, that is, $\det(U) \in \mathbb{R} \setminus \{0\}$. This is equivalent to the existence of a polynomial matrix V such that

$$UV = VU = I.$$

Clearly, V is the inverse of U, which exists because U is non-singular, i.e., $\det(U) \neq 0$. Unimodularity is much stronger than non-singularity, the crucial point is that U possesses a *polynomial* (rather than a rational) inverse.

We observe that pre-multiplication by a unimodular matrix U does not change the behavior represented by R. More precisely, R and $\hat{R} = UR$ represent the same behavior, because

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \quad \Rightarrow \quad U(\tfrac{\mathrm{d}}{\mathrm{d}t})R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = \hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0$$

and

$$\hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \quad \Rightarrow \quad V(\tfrac{\mathrm{d}}{\mathrm{d}t})\hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0$$

where V is the polynomial inverse of U. The same holds if we replace $\frac{\mathrm{d}}{\mathrm{d}t}$ by σ.

Definition 2.2 A polynomial matrix R is called a **minimal** representation of $\mathcal{B}$ if there exists no polynomial matrix which represents the same behavior and has a smaller number of rows.

The most important fact about polynomial matrices is stated next.

Theorem 2.3 (Smith form) For every polynomial matrix $R \in \mathbb{R}[s]^{p \times q}$ there exist unimodular matrices $U \in \mathbb{R}[s]^{p \times p}$ and $V \in \mathbb{R}[s]^{q \times q}$ such that

$$URV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{2.1}$$

where $D \in \mathbb{R}[s]^{r \times r}$ is a non-singular diagonal matrix

$$D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{bmatrix}$$

with $d_1 | d_2 | \ldots | d_r$. This notation means that for $i = 1, \ldots, r-1$, the polynomial d_i divides d_{i+1}, that is, $d_{i+1} = d_i e_i$ for some polynomial e_i. Clearly, the integer r is precisely the rank of the matrix R. The matrix on the right hand side of (2.1) is called **Smith form** of R.

Corollary 2.4 For every polynomial matrix $R \in \mathbb{R}[s]^{p\times q}$ there exists a unimodular matrix $U \in \mathbb{R}[s]^{p\times p}$ such that

$$UR = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

where $R_1 \in \mathbb{R}[s]^{r\times q}$ has full row rank, that is, $\operatorname{rank}(R_1) = r$.

Since R and R_1 represent the same behavior, we have the following conclusion.

Lemma 2.5 Any $\mathcal{B}$ possesses a representation matrix with full row rank. If R is a minimal representation of $\mathcal{B}$, then R has full row rank.

In fact, R is a minimal representation of $\mathcal{B}$ if and only if R has full row rank.

2.2 The fundamental principle

Theorem 2.6 (Fundamental principle, scalar version) Let $p \in \mathbb{R}[s]$ be a non-zero polynomial. Then

$$p(\tfrac{\mathrm{d}}{\mathrm{d}t})y = v \quad \text{or} \quad p(\sigma)y = v$$

possesses a solution $y \in \mathcal{A}$ for every choice of the right hand side signal $v \in \mathcal{A}$.

We can get a matrix version of this theorem immediately.

Corollary 2.7 (Fundamental principle, square matrix version) Let $P \in \mathbb{R}[s]^{p\times p}$ be a non-singular polynomial matrix. Then

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = v \quad \text{or} \quad P(\sigma)y = v$$

possesses a solution $y \in \mathcal{A}^p$ for every choice of the right hand side signal $v \in \mathcal{A}^p$.

Proof: We do only the continuous case, the discrete case is completely analogous. By Theorem 2.3, there exist unimodular matrices U, V such that

$$UPV = D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix}$$

where d_i are non-zero polynomials. Let $v \in \mathcal{A}^p$ be given. Put $\tilde{v} = U(\frac{\mathrm{d}}{\mathrm{d}t})v$. By Theorem 2.6, there exists $\tilde{y}_i \in \mathcal{A}$ such that $d_i(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{y}_i = \tilde{v}_i$ for $i = 1, \dots, p$. In matrix notation,

$$D(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = \tilde{v}.$$

Now we set $y = V(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{y}$. Then

$$U(\tfrac{\mathrm{d}}{\mathrm{d}t})P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = U(\tfrac{\mathrm{d}}{\mathrm{d}t})P(\tfrac{\mathrm{d}}{\mathrm{d}t})V(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = D(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = \tilde{v} = U(\tfrac{\mathrm{d}}{\mathrm{d}t})v$$

and hence, since U is unimodular, $P(\frac{\mathrm{d}}{\mathrm{d}t})y = v$ as desired. $\square$

The following variant for rectangular matrices with full row rank is an easy consequence.

Corollary 2.8 (Fundamental principle, non-square matrix version) Let $R \in \mathbb{R}[s]^{p\times q}$ be a full row rank polynomial matrix, that is, $\mathrm{rank}(R) = p$. Then

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = v \quad \text{or} \quad R(\sigma)w = v$$

possesses a solution $w \in \mathcal{A}^q$ for every choice of the right hand side signal $v \in \mathcal{A}^p$.

The most important consequence of the fundamental principle is stated next.

2.3 Elimination of latent variables

Often, the components of a signal vector can be divided into two classes: a set of components we are truly interested in (called **manifest** variables), and a set of components that were introduced as auxiliary variables during the modelling process (called **latent** variables).

Remark 2.9 For instance, in the electrical circuit of Example 1.9, one is usually interested in the signals U and I because they represent the overall voltage and current that are relevant for this network. On the other hand, the voltages u_1, u_2, u_3 were introduced for determining the network equations, and those quantities are not really interesting by themselves. In fact, one would like to get rid of them in order to describe the relation between I and U alone. This process is called **elimination of latent variables**.

In general, a **system with latent variables** is given by

$$\mathcal{B}_l = \{w \in \mathcal{A}^q \mid \exists l \in \mathcal{A}^r : R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l\}$$

where $R \in \mathbb{R}[s]^{p\times q}$ and $M \in \mathbb{R}[s]^{p\times r}$ (we restrict to continuous time; the discrete case is analogous). Here, w represents the manifest variables, and l represents the latent variables. We are not interested in their precise values, only in their existence. The question arises whether we can write $\mathcal{B}_l$ as a standard model, that is, whether we can find a polynomial matrix $\hat{R}$ such that

$$\mathcal{B}_l = \hat{\mathcal{B}} = \{w \in \mathcal{A}^q \mid \hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}.$$

The answer is yes, and moreover, there is an easy way to obtain the desired $\hat{R}$ from the given R and M.

One solves the linear system of equations $\xi M = 0$, where $\xi \in \mathbb{R}[s]^{1\times p}$. In other words, we compute the left kernel of M, over the polynomial ring. From linear algebra, we know that there are $p - \mathrm{rank}(M)$ linearly independent solutions. Let $\xi_1, \dots, \xi_{p-\mathrm{rank}(M)}$ be a generating system for the left kernel of M. Collecting these row vectors in a matrix X, we have constructed a matrix X which satisfies the following three conditions:

1. $XM = 0$;

2. X has full row rank;

3. any polynomial row vector ξ with $\xi M = 0$ can be written as a polynomial linear combination of the rows of X, that is, $\xi = \eta X$ for some polynomial row vector η.

Lemma 2.10 Let X_1, X_2 be two matrices with the three properties from above. Then we must have $X_1 = UX_2$ for some unimodular matrix U.

Proof: Condition 1 implies $X_1 M = 0$ and $X_2 M = 0$. By Condition 2, each row of X_1 can be written as a polynomial linear combination of the rows of X_2 (and vice versa). This means that there exist polynomial matrices U and V such that

$$X_1 = UX_2 \quad \text{and} \quad X_2 = VX_1.$$

Then

$$(I - UV)X_1 = 0 \quad \text{and} \quad (I - VU)X_2 = 0.$$

Finally, Condition 3 implies that $UV = VU = I$, that is, U has a polynomial inverse, namely V, and thus U is unimodular. □

Theorem 2.11 Let R, M be given polynomial matrices, with the same number of rows. Let X be as described above, and define $\hat{R} := XR$. Then

$$\exists l \in \mathcal{A}^r : R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l \quad \Leftrightarrow \quad \hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0$$

(analogously if $\frac{\mathrm{d}}{\mathrm{d}t}$ is replaced by σ), that is, using the notation from above, $\hat{\mathcal{B}} = \mathcal{B}_l$.

Proof: By Corollary 2.4, there exists a unimodular matrix V such that

$$VM = \begin{bmatrix} M_1 \\ 0 \end{bmatrix}$$

where M_1 has full row rank. Then

$$\begin{aligned} &\exists l : R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l && \Leftrightarrow \\ &\exists l : V(\tfrac{\mathrm{d}}{\mathrm{d}t})R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = V(\tfrac{\mathrm{d}}{\mathrm{d}t})M(\tfrac{\mathrm{d}}{\mathrm{d}t})l && \Leftrightarrow \\ &\exists l : \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = \begin{bmatrix} M_1 \\ 0 \end{bmatrix}(\tfrac{\mathrm{d}}{\mathrm{d}t})l. \end{aligned}$$

Using Corollary 2.8, we obtain

$$\exists l : R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l \quad \Leftrightarrow \quad R_2(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0.$$

It remains to establish a relation between R_2 and $\hat{R}$. Note that $\hat{X} := [0, I]V$ has the three properties given above. According to Lemma 2.10, there exists a unimodular matrix U such that

$$X = U\hat{X}.$$

Thus $\hat{R} = XR = U\hat{X}R = U[0, I]VR = UR_2$, which shows that

$$\hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \quad \Leftrightarrow \quad R_2(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0$$

which completes the proof. □

2.4 Inputs, outputs, and autonomous systems

Definition 2.12 Let the signal vector w be partitioned as

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix},$$

where $w_i \in \mathcal{A}^{q_i}$, where q_1 and q_2 are positive integers with $q_1+q_2 = q$. The subvector w_1 is called a vector of **free variables** (or: inputs) of $\mathcal{B}$ if it is unconstrained by the system law, that is,

$$\forall w_1 \in \mathcal{A}^{q_1} \exists w_2 \in \mathcal{A}^{q_2} : \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathcal{B}.$$

Free variables can be found as follows: We may assume that $R \in \mathbb{R}[s]^{p\times q}$ has full row rank, that is, $\operatorname{rank}(R) = p \leq q$. Then there exists a $p \times p$ non-singular submatrix of R. Assume that $p < q$. We can always permute the columns of R (this just corresponds to renumbering our signal components) such that

$$R = \begin{bmatrix} -Q & P \end{bmatrix}$$

where P is such a non-singular $p \times p$ matrix. Writing

$$w = \begin{bmatrix} u \\ y \end{bmatrix}$$

correspondingly, our system law $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ takes the form

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u. \tag{2.2}$$

From Corollary 2.7, we obtain that this equation has a solution $y \in \mathcal{A}^p$ for every choice of $u \in \mathcal{A}^m$, where $m := q - p$. Therefore, u is a vector of free variables of $\mathcal{B}$.

Definition 2.13 A system law in the form of equation (2.2), where P is square and non-singular, is called an **input-output representation** of $\mathcal{B}$. One calls $p = \operatorname{rank}(R)$ the number of outputs, and $m = q - p$ the number of inputs. The signal subvector u is called input, and the signal subvector y is called output.

It is important to note that in general, one and the same behavior may have several different input-output representations. This is due to the fact that there may be more than one way to select a non-singular $p \times p$ submatrix of R, where $p = \operatorname{rank}(R)$.

Example 2.14 The electrical circuit from Example 1.9 admits 5 different input-output representations. The matrix

$$\hat{R} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ R & 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & Cs & 0 \\ Ls & 0 & 0 & 0 & -1 \end{bmatrix}.$$

has rank 4 and every 4×4 submatrix is non-singular (the parameters R, C, L are supposed to be positive). Each of the 5 signals I, U, u_1, u_2, u_3 can take the role of the input, and then the remaining 4 signals will be outputs. However, from the physical point of view, one would usually consider U as the input and I, u_1, u_2, u_3 as outputs.

Example 2.15 The system from Example 1.8 does not have any free variables. If

$$R(\sigma)w = \begin{bmatrix} \sigma & -1 \\ -\frac{1}{2} & \sigma - \frac{1}{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0$$

then

$$\begin{bmatrix} \sigma - \frac{1}{2} & 1 \\ \frac{1}{2} & \sigma \end{bmatrix} \begin{bmatrix} \sigma & -1 \\ -\frac{1}{2} & \sigma - \frac{1}{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \sigma^2 - \frac{1}{2}\sigma - \frac{1}{2} & 0 \\ 0 & \sigma^2 - \frac{1}{2}\sigma - \frac{1}{2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0$$

and thus for $i = 1, 2$:

$$(\sigma^2 - \tfrac{1}{2}\sigma - \tfrac{1}{2})w_i = 0.$$

This implies that if $w_i(0)$ and $w_i(1)$ are given, all $w_i(t)$ for $t \in T$ are uniquely determined. In particular, w_i is not free. In fact, considering the characteristic equation

$$\lambda^2 - \tfrac{1}{2}\lambda - \tfrac{1}{2} = 0$$

with its solutions $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{2}$, we must have

$$w_i(t) = a_1\lambda_1^t + a_2\lambda_2^t = a_1 + a_2(-\tfrac{1}{2})^t$$

where the constants a_1, a_2 can be computed from $w_i(0)$, $w_i(1)$. In fact, the original equations show that it suffices to know $w_i(0)$ for $i = 1, 2$ to obtain a unique solution.

Systems without inputs are called autonomous (Greek: $\alpha\upsilon\tau o\nu o\mu o\varsigma$; under its own law/self-governed/independent).

Definition 2.16 A system which has no free variables is called **autonomous**.

Example 2.17 A system given by the scalar equation $p(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, where $0 \neq p \in \mathbb{R}[s]$, is autonomous. Any w satisfying $p(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ must be of the form

$$w(t) = \sum_{\lambda} a_\lambda(t) e^{\lambda t}$$

where $\lambda \in \mathbb{C}$ are the zeros of p, and a_λ are polynomials. Thus w is certainly constrained by the system law. Similarly, $p(\sigma)w = 0$ implies

$$w(t) = \sum_{\lambda} a_\lambda(t) \lambda^t$$

in the discrete case.

Remark 2.18 In the subsequent proof, we use the following fact from linear algebra: For a square non-singular matrix P, there exists a matrix Q, called the **adjugate** of P, such that

$$QP = PQ = \det(P) I.$$

Lemma 2.19 *$\mathcal{B}$ is autonomous if and only if it has a square non-singular representation matrix.*

Proof: Any $\mathcal{B}$ possesses a representation R with full row rank. If R is not square, we can construct an input vector u as described above. Conversely, let $\mathcal{B}$ be represented by a square non-singular matrix,

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0.$$

Pre-multiplying this by the adjugate of P, we obtain

$$\det(P)(\tfrac{\mathrm{d}}{\mathrm{d}t})w_i = 0$$

for $i = 1, \ldots, q$, showing that no component of w is unconstrained by the system law. $\square$

2.5 Input-output representations

Let $P(\frac{\mathrm{d}}{\mathrm{d}t})y = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$ be a system law in input-output form and assume that $u \in \mathcal{A}^m$ is given. From the fundamental principle, we know that there exist outputs belonging to this input. What can be said about the set of these outputs? Let y_p be one particular output that belongs to u, that is, $P(\frac{\mathrm{d}}{\mathrm{d}t})y_p = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$. Let y be another output that belongs to u, then $P(\frac{\mathrm{d}}{\mathrm{d}t})(y - y_p) = 0$, that is, $y_h := y - y_p$ is an element of the autonomous system defined by $P(\frac{\mathrm{d}}{\mathrm{d}t})y_h = 0$. Therefore, any output y that belongs to u can be written in the form

$$y = y_p + y_h$$

where y_p is one particular solution of the inhomogeneous equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$ and y_h is an arbitrary solution of the corresponding homogeneous equation.

2.5.1 The continuous homogeneous equation

Consider $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$, where $P \in \mathbb{R}[s]^{p \times p}$ is non-singular. Let $0 \neq \det(P) \in \mathbb{R}[s]$ be its determinant and let

$$\Lambda = \{\lambda \in \mathbb{C} \mid \det(P)(\lambda) = 0\}$$

denote the set of its zeros. Any solution y of $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ has the form

$$y(t) = \sum_{\lambda \in \Lambda} a_\lambda(t) e^{\lambda t}$$

where $a_\lambda \in \mathbb{C}[t]^p$.

In the **scalar** case ($p = 1$), the degree of each polynomial a_λ is given by $\mu(\lambda) - 1$, where $\mu(\lambda)$ is the multiplicity of λ as a zero of P, that is, according to the fundamental theorem of algebra,

$$P = c \prod_{\lambda \in \Lambda} (s - \lambda)^{\mu(\lambda)}$$

where c is the leading coefficient of P. Therefore, each y satisfying $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ is uniquely determined by the

$$\sum_{\lambda \in \Lambda} \mu(\lambda) = \deg(P)$$

coefficients of these polynomials. In particular, the dimension of the solution space of the scalar equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ is precisely the degree of P, that is, the order of the differential equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$.

In the **general** case ($p \geq 1$), we use the Smith form $UPV = D = \operatorname{diag}(d_1, \dots, d_p)$. Since $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ is equivalent to $D(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = 0$, where $y = V(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{y}$, we consider

$$\mathcal{B} = \{y \in \mathcal{A}^p \mid P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = 0\} \quad \text{and} \quad \tilde{\mathcal{B}} = \{\tilde{y} \in \mathcal{A}^p \mid D(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = 0\}$$

and we conclude from the scalar case that

$$\dim \tilde{\mathcal{B}} = \sum_{i=1}^{p} \deg(d_i) = \deg \prod_{i=1}^{p} d_i = \deg \det(D) = \deg \det(P).$$

The map $\tilde{y} \mapsto y = V(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{y}$ provides a vector space isomorphism between $\tilde{\mathcal{B}}$ and $\mathcal{B}$ and thus, their dimensions coincide.

Theorem 2.20 $\mathcal{B}$ is autonomous if and only if it is a finite-dimensional (real or complex) vector space. More precisely, if $\mathcal{B}$ is represented by the square non-singular matrix P, then $\dim \mathcal{B} = \deg \det(P)$.

Remark 2.21 If our signals are complex-valued, then $\mathcal{B}$ is a complex vector space (in general, it makes no sense to speak of the "value" of a distribution, however, the solutions of an autonomous system are classical functions). Then $\dim \mathcal{B}$ is the usual dimension of a complex vector space.

On the other hand, if our signals are supposed to take real values, then $\mathcal{B}$ is a real vector space. However, we still have to admit complex coefficients in the polynomials a_λ, because $\lambda \in \Lambda$ and hence $e^{\lambda t}$ will be complex, in general. However, since the coefficients of P are supposed to be real, we have

$$\lambda \in \Lambda \quad \Rightarrow \quad \overline{\lambda} \in \Lambda,$$

that is, the zeros of $\det(P)$ come in pairs of complex conjugate numbers. If our signals are required to have real values, we must have

$$a_{\overline{\lambda}} = \overline{a_\lambda}$$

and thus the dimension of $\mathcal{B}$ is actually the usual dimension of $\mathcal{B}$ as a real vector space.

2.5.2 The continuous inhomogeneous equation

Suppose that $P(\frac{\mathrm{d}}{\mathrm{d}t})h^{(i)} = Q(\frac{\mathrm{d}}{\mathrm{d}t})\delta^{(i)}$, where $\delta^{(i)} \in \mathcal{A}^m$ has δ in the i-th position and zeros everywhere else. Here, $\delta \in \mathcal{A} = \mathcal{D}'(T)$ denotes the Dirac delta distribution. Define

$$h = \left[\begin{array}{ccc} h^{(1)} & \cdots & h^{(m)} \end{array}\right] \in \mathcal{A}^{p\times m}.$$

Then h is called **impulse response** (or: fundamental solution), because its columns can be seen as the system's response (output) to an input which is an impulse (delta distribution).

Then a particular output y belonging to the input u is given by the continuous **convolution** (if it exists)

$$y = h * u$$

that is, in the classical case (u, h locally integrable, i.e., regular distributions)

$$y(t) = \int_{-\infty}^{\infty} h(t-\tau)u(\tau)\mathrm{d}\tau.$$

The existence of the convolution is guaranteed by any of the following conditions:

- Let $h(t) = 0$ and $u(t) = 0$ for all $t < 0$. Then

  $$y(t) = \int_0^t h(t-\tau)u(\tau)\mathrm{d}\tau,$$

  an integral over a compact interval, which exists due to the assumption of local integrability. Recall that $\mathcal{D}'(\mathbb{R}_+)$ was introduced as the set of all distributions $D \in \mathcal{D}'(\mathbb{R})$ that assign zero to every test function with support in $(-\infty, 0)$. One also says that D vanishes on $(-\infty, 0)$, or that the support of D lies in $[0, \infty)$. For locally integrable $h, u \in \mathcal{D}'(\mathbb{R}_+)$, we may assume that $h(t) = 0$ and $u(t) = 0$ for all $t < 0$. Thus, the convolution is always well-defined if $T = \mathbb{R}_+$ (also in the non-regular case).

- If h or u has compact support, then $h * u$ exists (also in the non-regular case).

- Let u be bounded, and let h be (globally) integrable. Then

  $$\|y\|_\infty \le \|h\|_1 \cdot \|u\|_\infty$$

  and thus, a bounded input $u \in L^\infty$ leads to a bounded output $y \in L^\infty$ provided that $h \in L^1$. This is known as "bounded input, bounded output (BIBO) stability".

Example: Consider $\dot{x}(t) = Ax(t) + b(t)$. We take b as the input, and x as the output. Then $P = sI - A$ and $Q = I$. An impulse response is given by

$$h(t) = \begin{cases} e^{At} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

for both $T = \mathbb{R}$ and $T = \mathbb{R}_+$. If $T = \mathbb{R}$, assume that $u(t) = 0$ for $t < 0$. Then we have

$$y(t) = (h * u)(t) = \int_0^t e^{A(t-\tau)} u(\tau) \mathrm{d}\tau = \int_0^t e^{A\tau} u(t - \tau) \mathrm{d}\tau$$

as a particular output $y = x$ belonging to the input $u = b$.

2.5.3 The discrete homogeneous equation ($T = \mathbb{N}$)

Consider $P(\sigma)y = 0$, where $P \in \mathbb{R}[s]^{p \times p}$ is non-singular. Let $0 \neq \det(P) \in \mathbb{R}[s]$ be its determinant and let

$$\Lambda = \{\lambda \in \mathbb{C} \mid \det(P)(\lambda) = 0\}$$

denote the set of its zeros. Any solution y of $P(\sigma)y = 0$ has the form

$$y(t) = \sum_{\lambda \in \Lambda} a_\lambda(t) \lambda^t$$

where $a_\lambda \in \mathbb{C}[t]^p$.

In the **scalar** case ($p = 1$), the degree of each polynomial a_λ is given by $\mu(\lambda) - 1$, where $\mu(\lambda)$ is the multiplicity of λ as a zero of P, that is,

$$P = c \prod_{\lambda \in \Lambda} (s - \lambda)^{\mu(\lambda)}$$

where c is the leading coefficient of P. Therefore, each y satisfying $P(\sigma)y = 0$ is uniquely determined by the

$$\sum_{\lambda \in \Lambda} \mu(\lambda) = \deg(P)$$

coefficients of these polynomials.

In the **general** case ($p \geq 1$), we use again the Smith form $UPV = D = \mathrm{diag}(d_1, \ldots, d_p)$, to obtain the following result which is completely analogous to its continuous counterpart.

Theorem 2.22 $\mathcal{B}$ is autonomous if and only if it is a finite-dimensional vector space. More precisely, if $\mathcal{B}$ is represented by the square non-singular matrix P, then $\dim \mathcal{B} = \deg \det(P)$.

Remark 2.23 Again, the concept of the dimension of $\mathcal{B}$ used here coincides with the classical notion of the dimension of $\mathcal{B}$ as a real vector space if $\mathcal{A} = \mathbb{R}^{\mathbb{N}}$.

2.5.4 The discrete inhomogeneous equation

Suppose that $P(\sigma)h^{(i)} = Q(\sigma)\delta^{(i)}$, where $\delta^{(i)} \in \mathcal{A}^m$ has δ in the i-th position and zeros everywhere else. Here, $\delta \in \mathcal{A} = \mathbb{R}^T$ denotes the sequence that takes the value 1 at time zero, and the value zero everywhere else. Define

$$h = \left[\begin{array}{ccc} h^{(1)} & \cdots & h^{(m)} \end{array} \right] \in \mathcal{A}^{p \times m}.$$

Then h is called **impulse response** (or: fundamental solution), because its columns can be seen as the system's response (output) to an input which is an impulse (delta distribution).

Then a particular output y belonging to the input u is given by the discrete **convolution** (if it exists)

$$y = h * u$$

that is,

$$y(t) = \sum_{i=-\infty}^{\infty} h(t-i)u(i).$$

We identify $\mathbb{R}^{\mathbb{N}}$ with the set of sequences $a \in \mathbb{R}^{\mathbb{Z}}$ with $a(t) = 0$ for all $t < 0$. The existence of the convolution is guaranteed by any of the following conditions:

- Let $h(t) = 0$ and $u(t) = 0$ for all $t < 0$. Then

  $$y(t) = \sum_{i=0}^{t} h(t-i)u(i),$$

  a finite sum. Thus the convolution always exists for $T = \mathbb{N}$.
- At least one of h, u has compact support.
- u is bounded, and h is summable ("BIBO stability").

Example: Consider $x(t+1) = Ax(t) + b(t)$. We take b as the input, and x as the output. Then $P = sI - A$ and $Q = I$. An impulse response is given by $h(t) = A^{t-1}$ for $t \geq 1$ and $h(t) = 0$ for all $t \leq 0$. If $T = \mathbb{N}$, or $T = \mathbb{Z}$ and $u(t) = 0$ for $t < 0$, we have

$$y(t) = (h * u)(t) = \sum_{i=0}^{t} h(t-i)u(i) = \sum_{i=0}^{t-1} h(t-i)u(i) = \sum_{i=0}^{t-1} A^{t-i-1}u(i)$$

as a particular output $y = x$ belonging to the input $u = b$.

2.6 Reduction to first order

A polynomial matrix $R \in \mathbb{R}[s]^{p \times q}$ can be written in the form

$$R = R_d s^d + \ldots + R_1 s + R_0$$

where $R_i \in \mathbb{R}^{p \times q}$. We may assume that R_d is not the zero matrix. Our system law takes the form

$$(R_d \tfrac{\mathrm{d}^d}{\mathrm{d}t^d} + \ldots + R_1 \tfrac{\mathrm{d}}{\mathrm{d}t} + R_0)w = 0$$

or

$$(R_d \sigma^d + \ldots + R_1 \sigma + R_0)w = 0$$

see (1.3) and (1.4). If we put

$$\xi = \begin{bmatrix} w \\ \frac{\mathrm{d}}{\mathrm{d}t} w \\ \vdots \\ \frac{\mathrm{d}^{d-1}}{\mathrm{d}t^{d-1}} w \end{bmatrix} \quad \text{or} \quad \xi = \begin{bmatrix} w \\ \sigma w \\ \vdots \\ \sigma^{d-1} w \end{bmatrix}$$

in the continuous or discrete case, respectively, then we can rewrite the system law as

$$K \tfrac{\mathrm{d}}{\mathrm{d}t} \xi = L\xi \quad \text{or} \quad K\sigma\xi = L\xi$$

where

$$K = \begin{bmatrix} I_q & & & \\ & \ddots & & \\ & & I_q & \\ & & & R_d \end{bmatrix} \quad \text{and} \quad L = \begin{bmatrix} 0 & I_q & & \\ \vdots & & \ddots & \\ 0 & & & I_q \\ -R_0 & -R_1 & \cdots & -R_{d-1} \end{bmatrix}.$$

Putting $n = dq$ and $k = (d-1)q + p$, we have $K, L \in \mathbb{R}^{k \times n}$ and

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \quad \Leftrightarrow \quad \exists \xi \in \mathcal{A}^n : \left\{ \begin{array}{rcl} K\frac{\mathrm{d}}{\mathrm{d}t}\xi & = & L\xi \\ w & = & [I_q, 0, \cdots, 0]\xi \end{array} \right. \tag{2.3}$$

and similarly in the discrete case. This shows that reduction to first order is nothing but a special way of introducing latent variables. Another way of doing this is discussed in the next section.

2.7 State

Theorem 2.24 Let $R \in \mathbb{R}[s]^{p \times q}$. There exists an integer n and real matrices $K \in \mathbb{R}^{n \times n}$, $L \in \mathbb{R}^{n \times q}$, $M \in \mathbb{R}^{p \times n}$, $N \in \mathbb{R}^{p \times q}$ such that the system law $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ has a first order latent variable representation of the form

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \quad \Leftrightarrow \quad \exists x \in \mathcal{A}^n : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x & = & Kx + Lw \\ 0 & = & Mx + Nw \end{array} \right. \tag{2.4}$$

and similarly for σ instead of $\frac{\mathrm{d}}{\mathrm{d}t}$.

Proof: Let $R = R_d s^d + \ldots + R_1 s + R_0$ with $R_i \in \mathbb{R}^{p \times q}$. We put $n = dp$ and

$$K = \left[\begin{array}{cccc} 0 & \cdots & \cdots & 0 \\ I_p & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & I_p & 0 \end{array} \right] \quad L = \left[\begin{array}{c} R_0 \\ \vdots \\ \vdots \\ R_{d-1} \end{array} \right]$$

$$M = \left[\begin{array}{cccc} 0 & \cdots & 0 & I_p \end{array} \right] \quad N = R_d.$$

Then

$$\left[\begin{array}{c} \frac{\mathrm{d}}{\mathrm{d}t} I_n - K \\ -M \end{array} \right] x = \left[\begin{array}{c} L \\ N \end{array} \right] w$$

can be pre-multiplied by $U(\frac{\mathrm{d}}{\mathrm{d}t})$, where U is the unimodular matrix

$$U = \left[\begin{array}{cccc} I_p & sI_p & \cdots & s^d I_p \\ & \ddots & \ddots & \vdots \\ & & \ddots & sI_p \\ & & & I_p \end{array} \right]$$

to obtain the equivalent equation

$$\begin{bmatrix} 0 \\ -I_n \end{bmatrix} x = \begin{bmatrix} R \\ * \end{bmatrix} (\tfrac{\mathrm{d}}{\mathrm{d}t}) w$$

where the $*$ denotes a polynomial matrix whose precise form is not important here, because we only need that such an x exists if and only if $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$. □

Remark 2.25 Comparing (2.3) and (2.4), we note that in (2.3), we have to deal with an implicit equation $K\dot{\xi} = L\xi$, and moreover, K, L are not square, in general. Thus we cannot use the Kronecker-Weierstraß form from Theorem 1.18, which works only for square matrices. The non-square Kronecker form is much more complicated, and not treated here. On the other hand, we have an explicit equation $\dot{x} = Kx + Lw$ in (2.4). Another difference is the size n of the latent variable ξ or x, respectively: ξ has $n_1 = dq$ components, whereas x has $n_2 = dp$ components. Recall that we may assume, without loss of generality, that R has full row rank $p \leq q$, and thus $n_2 \leq n_1$. Finally, the construction of (2.4) can easily be modified by first applying it to each row of R separately, and then combining the results. This may lead to even "smaller" (with respect to the number n of latent variables x_i) representations. Summing up, we see that (2.4) is usually preferable to the "naive" reduction to first order from (2.3).

For every input-output representation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$, we can find, according to Theorem 2.24, a first order representation of the form

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad \Leftrightarrow \quad \exists x \in \mathcal{A}^n : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x & = & Kx + L_1 u + L_2 y \\ 0 & = & Mx + N_1 u + N_2 y \end{array} \right.$$

and similarly for σ instead of $\frac{\mathrm{d}}{\mathrm{d}t}$.

A particularly important case arises when N_2 is non-singular. Then we can solve the second equation for y to obtain

$$y = -N_2^{-1}(Mx + N_1 u)$$

and plug that into the first equation. We get

$$\tfrac{\mathrm{d}}{\mathrm{d}t}x = (K - L_2 N_2^{-1} M)x + (L_1 - L_2 N_2^{-1} N_1)u.$$

Setting

$$\begin{array}{rcl} A & = & K - L_2 N_2^{-1} M \\ B & = & L_1 - L_2 N_2^{-1} N_1 \\ C & = & -N_2^{-1} M \\ D & = & -N_2^{-1} N_1 \end{array}$$

we have

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad\Leftrightarrow\quad \exists x \in \mathcal{A}^n : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x &=& Ax + Bu \\ y &=& Cx + Du. \end{array} \right.$$

The explicit equations

$$\begin{array}{rcl} \dot{x} &=& Ax + Bu \\ y &=& Cx + Du \end{array} \qquad \text{or} \qquad \begin{array}{rcl} \sigma x &=& Ax + Bu \\ y &=& Cx + Du \end{array}$$

are called **state space representations**. We call u the input, y the output, and x the **state** of this system law. This notion comes from the fact that if $x(t_0)$ is known for some $t_0 \in T$, and if u is known on some interval $T_{01} := [t_0, t_1] \cap T$, where $t_1 > t_0$, then x, and thus y, are uniquely determined everywhere in T_{01}. This is due to the solution formulas

$$x(t) = e^{(t-t_0)A}x(t_0) + \int_{t_0}^{t} e^{(t-\tau)A}Bu(\tau)\mathrm{d}\tau$$

or

$$x(t) = A^{t-t_0}x(t_0) + \sum_{i=t_0}^{t-1} A^{t-1-i}Bu(i)$$

respectively. Thus x represents the system's memory in the sense that $x(t_0)$ contains all the information about the "past" needed for determining the "future" (provided that the future input is given).

Let us write $\varphi(t_1, t_0, x_0, u)$ for the state of the system at time t_1 provided that the state at time t_0 was x_0 and that the input function is u. Then the **state transition map** φ has the following properties for all $t_2 > t_1 > t_0 \in T$, $x_0, x_1, x_2 \in \mathbb{R}^n$, $u, u_1, u_2 \in \mathcal{A}^m$:

Consistency: $\varphi(t_0, t_0, x_0, u) = x_0$.

Causality: If $u_1(t) = u_2(t)$ for all $t \in T_{01}$, then $\varphi(t_1, t_0, x_0, u_1) = \varphi(t_1, t_0, x_0, u_2)$.

Semigroup property: If

$$\varphi(t_1, t_0, x_0, u) = x_1 \quad \text{and} \quad \varphi(t_2, t_1, x_1, u) = x_2$$

then

$$\varphi(t_2, t_0, x_0, u) = x_2.$$

Chapter 3

Stability

Stability is concerned with the behavior of signals on the positive time set

$$T_+ := [0, \infty) \cap T$$

in particular, in the limit as time tends to infinity. In the continuous case, $T_+ = \mathbb{R}_+$, and in the discrete case, $T_+ = \mathbb{N}$.

Let $P(\frac{\mathrm{d}}{\mathrm{d}t})y = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$ or $P(\sigma)y = Q(\sigma)u$ be input-output representations of the system law. Let y_1, y_2 be two outputs belonging to the same input. What can be said about the size of their difference $y_1 - y_2$? Recall that $y_1 - y_2$ satisfies the homogeneous equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y_h = 0$ or $P(\sigma)y_h = 0$. Therefore, it is a smooth function of time in the continuous case, and $y_h(t) \in \mathbb{R}^p$ is well-defined (in general, this makes no sense for distributions). For the following definition, let $\|\cdot\|$ denote a norm on $\mathbb{R}^p$, e.g., the Euclidean norm.

Definition 3.1 Let $\mathcal{B}$ be represented by $P(\frac{\mathrm{d}}{\mathrm{d}t})y = Q(\frac{\mathrm{d}}{\mathrm{d}t})u$ or $P(\sigma)y = Q(\sigma)u$, respectively. The input-output representation is called **stable** if any two outputs y_1, y_2 belonging to the same input u satisfy

$$\|y_1(t) - y_2(t)\| < M \quad \text{for all } t \in T_+$$

for some constant M which is independent of t (but may depend on the specific choice of y_1, y_2). It is called **asymptotically stable** if we have additionally

$$\lim_{t\to\infty} \|y_1(t) - y_2(t)\| = 0.$$

Stability means: $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ implies that y is bounded on T_+. Asymptotic stability means: Moreover, $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ implies that $\lim_{t\to\infty} \|y(t)\| = 0$. Since these notions depend only on P and not on Q, we have the following modified definition.

Definition 3.2 An autonomous system is stable if all its signals are bounded on T_+; and asymptotically stable if additionally, all its signals tend to zero as time tends to infinity.

Let $P \in \mathbb{R}[s]^{p \times p}$ be non-singular. In the scalar case ($p = 1$), the solutions to $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ or $P(\sigma)y = 0$ have the form

$$y(t) = \sum_{\lambda} a_\lambda(t) e^{\lambda t} \quad \text{or} \quad y(t) = \sum_{\lambda} a_\lambda(t) \lambda^t,$$

where $\lambda \in \mathbb{C}$ are the zeros of P, and $a_\lambda \in \mathbb{C}[t]$. Moreover, $\deg(a_\lambda) = \mu(\lambda) - 1$. Thus we can see that stability depends on the location of the zeros λ in the complex plane, and on their multiplicities $\mu(\lambda)$.

Remark 3.3 Let $\lambda \in \mathbb{C}$. Consider the function $f : \mathbb{R}_+ \to \mathbb{R}$, $t \mapsto |a(t)e^{\lambda t}| = |a(t)|e^{\mathrm{Re}(\lambda)t}$, where a is a non-zero polynomial. It grows without bound if and only if we have either $\mathrm{Re}(\lambda) > 0$, or $\mathrm{Re}(\lambda) = 0$ and $\deg(a) \geq 1$. It is constant if and only if $\mathrm{Re}(\lambda) = 0$ and $\deg(a) = 0$. It tends to zero if and only if $\mathrm{Re}(\lambda) < 0$.

Consider the function $f : \mathbb{N} \to \mathbb{R}$, $t \mapsto |a(t)\lambda^t| = |a(t)||\lambda|^t$, where a is a non-zero polynomial. It grows without bound if and only if either $|\lambda| > 1$, or $|\lambda| = 1$ and $\deg(a) \geq 1$. It is constant if and only if $|\lambda| = 1$ and $\deg(a) = 0$. It tends to zero if and only if $|\lambda| < 1$.

Thus we can characterize (asymptotic) stability in the scalar case as follows: The system given by the scalar equation $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ or $P(\sigma)y = 0$, where $0 \neq P \in \mathbb{R}[s]$ and $\Lambda = \{\lambda \in \mathbb{C} \mid P(\lambda) = 0\}$, is

continuous-time asymptotically stable if and only if $\mathrm{Re}(\lambda) < 0$ for all $\lambda \in \Lambda$;

continuous-time stable if and only if $\mathrm{Re}(\lambda) \leq 0$ for all $\lambda \in \Lambda$, and if $\mathrm{Re}(\lambda) = 0$, then λ is simple, that is, $\mu(\lambda) = 1$;

discrete-time asymptotically stable if and only if $|\lambda| < 1$ for all $\lambda \in \Lambda$;

discrete-time stable if and only if $|\lambda| \leq 1$ for all $\lambda \in \Lambda$, and if $|\lambda| = 1$, then λ is simple, that is, $\mu(\lambda) = 1$.

Now we generalize this to the matrix case ($p \geq 1$). For this, we need the notion of semi-simple zeros.

Definition 3.4 Let P be a square and non-singular polynomial matrix. A zero λ of $\det(P)$ is called a **semi-simple** zero of P if the multiplicity of λ as a zero of $\det(P)$ (called the algebraic multiplicity $\mu(\lambda)$ of λ) equals the dimension of the kernel of $P(\lambda)$ (called the geometric multiplicity $\nu(\lambda)$ of λ).

Note that in the scalar case ($p = 1$), the geometric multiplicity is always one, and thus, a zero is semi-simple if and only if it is simple. In the general case, however, we only have that simple zeros are semi-simple (this follows from $1 \leq \nu(\lambda) \leq \mu(\lambda)$ which holds for all zeros). The converse is not necessarily true, as can be seen from the following example.

Example 3.5 Consider

$$P = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \quad \text{and} \quad \hat{P} = \begin{bmatrix} s & 1 \\ 0 & s \end{bmatrix}.$$

Then $\det(P) = \det(\hat{P}) = s^2$. The only zero of s^2 is $\lambda = 0$, and its algebraic multiplicity equals 2 in both cases. However,

$$P(0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \hat{P}(0) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

and hence, the geometric multiplicity of $\lambda = 0$ as a zero of P is 2, and its geometric multiplicity as a zero of $\hat{P}$ equals 1. Thus $\lambda = 0$ is a semi-simple (but not simple) zero of P. On the other hand, $\lambda = 0$ is not a semi-simple zero of $\hat{P}$.

Theorem 3.6 Let P be a square and non-singular polynomial matrix. The autonomous system represented by P is

continuous-time asymptotically stable if and only if the zeros of $\det(P)$ have a negative real part;

continuous-time stable if and only if the zeros of $\det(P)$ have a non-positive real part and moreover, each zero λ with $\mathrm{Re}(\lambda) = 0$ is semi-simple;

discrete-time asymptotically stable if and only if the zeros of $\det(P)$ have modulus less than one;

discrete-time stable if and only if the zeros of $\det(P)$ have modulus less than or equal to one, and moreover, each zero λ with $|\lambda| = 1$ is semi-simple.

Proof: We do only the continuous case; the discrete case is completely analogous. Let $UPV = D = \mathrm{diag}(d_1, \dots, d_p)$ be the Smith form of P. Consider

$$\tilde{\mathcal{B}} = \{\tilde{y} \in \mathcal{A}^p \mid D(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{y} = 0\}.$$

Then $\tilde{\mathcal{B}}$ is (asymptotically) stable if and only if $\mathcal{B}$ is (asymptotically) stable.

Asymptotic stability amounts to the requirement that all zeros of the polynomials d_i have a negative real part. However, these are precisely the zeros of $\det(P)$.

For stability, let λ be a zero of $\det(P)$ with $\mathrm{Re}(\lambda) = 0$. We may assume, without loss of generality, that $d_1 | \dots | d_p$. Then we have

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

with $D_1(\lambda)$ non-singular, and $D_2(\lambda) = 0_k$ (the $k \times k$ zero matrix). Moreover,

$$D_2 = \begin{bmatrix} (s-\lambda)^{l_1} p_1 & & \\ & \ddots & \\ & & (s-\lambda)^{l_k} p_k \end{bmatrix}$$

where $1 \leq l_1 \leq \dots \leq l_k$, and p_i are polynomials with $p_i(\lambda) \neq 0$. Then the algebraic multiplicity of λ equals $l_1 + \dots + l_k$, and the geometric multiplicity of λ equals k. The two multiplicities coincide if and only if $l_i = 1$ for all i, that is, if $s - \lambda$ enters linearly in each diagonal entry of D_2. This means that λ is a simple zero of each entry of D_2. Equivalently, the polynomial coefficient vector a_λ of $e^{\lambda t}$ in the representation $\tilde{y}(t) = \sum a_\lambda e^{\lambda t}$ can only be a constant vector. □

Example 3.7 The Smith forms of the matrices in Example 3.5 are

$$D = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \quad \text{and} \quad \hat{D} = \begin{bmatrix} 1 & 0 \\ 0 & s^2 \end{bmatrix}$$

respectively. The system given by $P(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ is stable (its solutions are precisely the constant vectors $y(t) = y(0)$), whereas the system given by $\hat{P}(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ admits the solution

$$y(t) = \begin{bmatrix} y_1(0) - t y_2(0) \\ y_2(0) \end{bmatrix}$$

which is unbounded whenever $y_2(0) \neq 0$.

3.1 Stability of state space representations

Consider the state space equations

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx + Du \end{array} \quad \text{or} \quad \begin{array}{rcl} \sigma x & = & Ax + Bu \\ y & = & Cx + Du. \end{array}$$

We call them **stable** if two states x_1, x_2 belonging to the same input u satisfy

$$\|x_1(t) - x_2(t)\| < M \quad \text{for all } t \in T_+$$

for some constant M. Then

$$\|y_1(t) - y_2(t)\| = \|C(x_1(t) - x_2(t))\| < M_1 \quad \text{for all } t \in T_+.$$

We call the state space equations **asymptotically stable** if additionally

$$\lim_{t\to\infty} \|x_1(t) - x_2(t)\| = 0.$$

Then

$$\lim_{t\to\infty} \|y_1(t) - y_2(t)\| = 0.$$

Note that here, $\|\cdot\|$ is used to denote both a norm on $\mathbb{R}^n$ and a norm on $\mathbb{R}^p$. One may think of the respective Euclidean norms, for instance.

Stability means that $\dot{x} = Ax$ or $\sigma x = Ax$ imply that x is bounded on T_+. Asymptotic stability means that additionally, $\lim_{t\to\infty} \|x(t)\| = 0$.

The equations $\dot{x} = Ax$ and $\sigma x = Ax$ are special cases of autonomous systems, namely with

$$P = sI - A.$$

Then the zeros of $\det(P)$ are precisely the eigenvalues of A. An eigenvalue λ of A is called semi-simple if it is a semi-simple zero of $sI - A$.

Corollary 3.8 Let A be a square real matrix. The autonomous system represented by $\dot{x} = Ax$ or $\sigma x = Ax$, respectively, is

continuous-time asymptotically stable if and only if the eigenvalues of A have a negative real part;

continuous-time stable if and only if the eigenvalues of A have a non-positive real part and moreover, each eigenvalue λ with $\mathrm{Re}(\lambda) = 0$ is semi-simple;

discrete-time asymptotically stable if and only if the eigenvalues of A have modulus less than one;

discrete-time stable if and only if the eigenvalues of A have modulus less than or equal to one, and moreover, each eigenvalue λ with $|\lambda| = 1$ is semi-simple.

Note that the solution of $\dot{x} = Ax$ or $\sigma x = Ax$ is

$$x(t) = e^{At}x_0 \quad \text{or} \quad x(t) = A^t x_0$$

respectively, where $x(0) = x_0$. Using these solutions formulas, we can sharpen the notion of stability as follows.

Corollary 3.9 The autonomous state space system given by $\dot{x} = Ax$ or $\sigma x = Ax$ is stable if and only if

$$\|x(t)\| < N\|x(0)\| \quad \text{for all } t \in T_+ \text{ and all } x.$$

The crucial point is that the constant N is not only independent of t, but also independent of the choice of the solution x of $\dot{x} = Ax$ or $\sigma x = Ax$, respectively.

This stability notion implies the classical **Lyapunov stability**, which is defined as follows: For all $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x_1(0) - x_2(0)\| < \delta$ implies

$$\|x_1(t) - x_2(t)\| < \varepsilon \quad \text{for all } t \in T_+.$$

Roughly speaking, Lyapunov stability means: If the initial values are close to each other, then the solutions will remain close to each other for all $t > 0$.

For a stable state space system, we can say even more: If $x_1(0)$ and $x_2(0)$ are sufficiently close to each other, say

$$\|x_1(0) - x_2(0)\| < \delta,$$

then the signals evolving from these initial values will stay close to each other for all times after $t = 0$, that is,

$$\|x_1(t) - x_2(t)\| < N\delta \quad \text{for all } t \in T_+.$$

Thus, given $\varepsilon > 0$, we may choose $\delta = \frac{\varepsilon}{N}$ to see that we have Lyapunov stability.

3.2 Test for asymptotic stability

Let $A \in \mathbb{R}^{n\times n}$ be given. We wish to test whether A is (discrete- or continuous-time) asymptotically stable. The naive way to do this is to compute all the eigenvalues λ of A and to check whether they satisfy $|\lambda| < 1$ or $\mathrm{Re}(\lambda) < 0$, respectively. However, when n is large, this is quite a difficult and computationally expensive task. Moreover, it seems to be a waste of time and effort to determine the precise location of the eigenvalues in the complex plane, when all we want to know is whether they are contained in a specific region (the open unit disc, or the open left half plane, respectively).

One way out is to compute the characteristic polynomial

$$\chi_A(s) := \det(sI - A) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0.$$

Then there are certain criteria, in terms of the coefficients a_i of χ_A, which make it possible to determine whether A is asymptotically stable or not. The best known of these criteria is the **Routh-Hurwitz test** for continuous time, and the **Schur-Cohn test** for discrete time.

For instance, if $n = 2$, we have $\chi_A(s) = s^2 + a_1 s + a_0$, where $a_1 = -\mathrm{trace}(A)$ and $a_0 = \det(A)$. The matrix $A \in \mathbb{R}^{2\times 2}$ is continuous-time asymptotically stable if and only if $a_1 > 0$ and $a_0 > 0$. In the general case ($n \geq 3$), however, the Routh-Hurwitz criterion is much more complex.

Here, we will adopt another approach, based on **Lyapunov equations**. Recall that a symmetric matrix $P \in \mathbb{R}^{n\times n}$ is called **positive semi-definite** (written $P \geq 0$) if we have $x^T P x \geq 0$ for all $x \in \mathbb{R}^n$. Note that then also for $x \in \mathbb{C}^n$, the number $x^* P x$ is real and non-negative. Here, x^* denotes the Hermitian transpose of x. This is due to the fact that if $x = a + ib$ for some $a, b \in \mathbb{R}^n$, then

$$x^* P x = (a + ib)^* P(a + ib) = (a^T - ib^T)P(a + ib) = a^T P a + b^T P b,$$

because $a^T P b = (a^T P b)^T = b^T P^T a = b^T P a$.

Theorem 3.10 The following are equivalent:

1. The matrix A is continuous-time asymptotically stable.
2. For every $Q \geq 0$, there exists $P \geq 0$ such that

$$A^T P + PA + Q = 0. \tag{3.1}$$

3. There exists $P \geq 0$ such that
$$A^T P + PA + I = 0.$$

Equation (3.1) is called **continuous-time Lyapunov equation**.

Proof: "1 ⇒ 2": If A is asymptotically stable, it makes sense to define
$$P := \int_0^\infty e^{A^T t} Q e^{At} \mathrm{d}t.$$
This is clearly positive semi-definite if Q is, and it satisfies the Lyapunov equation, because
$$A^T P + PA = \int_0^\infty \tfrac{\mathrm{d}}{\mathrm{d}t}(e^{A^T t} Q e^{At}) \mathrm{d}t = e^{A^T t} Q e^{At}\Big|_0^\infty = -Q.$$

"2 ⇒ 3" is a special case ($Q = I$).

"3 ⇒ 1": Let λ be an eigenvalue of A and let $0 \neq x \in \mathbb{C}^n$ be an associated eigenvector, that is, $Ax = \lambda x$. Pre-multiply $A^T P + PA + I = 0$ by x^* and post-multiply by x to get
$$x^* A^T P x + x^* P A x + x^* x = (\overline{\lambda} + \lambda) x^* P x + x^* x = 0.$$
Thus
$$2\mathrm{Re}(\lambda) x^* P x = -x^* x.$$
The right hand side of this equation is negative, and $x^* P x \geq 0$. This can only be true if $\mathrm{Re}(\lambda) < 0$. □

For the sake of completeness, we give also the discrete version of this theorem.

Theorem 3.11 The following are equivalent:

1. The matrix A is discrete-time asymptotically stable.
2. For every $Q \geq 0$, there exists $P \geq 0$ such that
$$A^T P A - P + Q = 0. \tag{3.2}$$
3. There exists $P \geq 0$ such that
$$A^T P A - P + I = 0.$$

Equation (3.2) is called **discrete-time Lyapunov equation**.

Proof: "1 $\Rightarrow$ 2": If A is asymptotically stable, it makes sense to define

$$P := \sum_{i=0}^{\infty} (A^T)^i Q A^i.$$

This is clearly positive semi-definite if Q is, and it satisfies the Lyapunov equation, because

$$A^T P A - P = \sum_{i=0}^{\infty} (A^T)^{i+1} Q A^{i+1} - \sum_{i=0}^{\infty} (A^T)^i Q A^i = -Q.$$

"2 $\Rightarrow$ 3" is a special case ($Q = I$).

"3 $\Rightarrow$ 1": Let λ be an eigenvalue of A and let $0 \neq x \in \mathbb{C}^n$ be an associated eigenvector, that is, $Ax = \lambda x$. Pre-multiply $A^T P A - P + I = 0$ by x^* and post-multiply by x to get

$$x^* A^T P A x - x^* P x + x^* x = (\overline{\lambda}\lambda - 1) x^* P x + x^* x = 0.$$

Thus

$$(|\lambda|^2 - 1) x^* P x = -x^* x.$$

The right hand side of this equation is negative, and $x^* P x \geq 0$. This can only be true if $|\lambda|^2 - 1 < 0$, that is, $|\lambda| < 1$. $\square$

Remark 3.12 If A is (continuous- or discrete-time) asymptotically stable, then the solutions of (3.1) or (3.2), respectively, are uniquely determined. In the continuous case, this can be seen as follows (the discrete case is analogous): Let P_1, P_2 be two solutions of (3.1), and let $P := P_1 - P_2$. Then $A^T P + P A = 0$. Now consider the matrix-valued function $f(t) := e^{A^T t} P e^{At}$. Since

$$\tfrac{\mathrm{d}}{\mathrm{d}t} f(t) = e^{A^T t} (A^T P + P A) e^{At} = 0$$

for all t, the function f is actually constant, that is, $f(t) = f(0) = P$ for all t. Now letting t tend to infinity, we see that P must be zero, because $\lim_{t \to \infty} f(t) = 0$.

The solution formulas $P = \int_0^\infty e^{A^T t} Q e^{At} \mathrm{d}t$ and $P = \sum_{i=0}^{\infty} (A^T)^i Q A^i$ are only of theoretical interest. In practice, one solves the linear matrix equations

$$A^T P + P A + Q = 0 \quad \text{or} \quad A^T P A - P + Q = 0$$

(these are n^2 linear equations for n^2 unknowns; exploiting symmetry, this reduces to $\frac{1}{2} n(n+1)$ unknowns), and then tests whether $P \geq 0$. In continuous time, we have a positive definite solution P provided that Q is positive definite (to test a symmetric matrix for positive definiteness is easier than to test it for positive semi-definiteness). In discrete time however, we have to additionally assume invertibility of A to guarantee that P is positive definite whenever Q is.

Chapter 4

Reachability and controllability

4.1 Basic notions for state space systems

Consider the state space equations

$$\begin{array}{rclcrcl} \dot{x} & = & Ax + Bu & \quad\text{or}\quad & \sigma x & = & Ax + Bu \\ y & = & Cx + Du & & y & = & Cx + Du, \end{array}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$.

Let $\mathcal{U}$ denote the space of **admissible input functions** (until now, we had $\mathcal{U} = \mathcal{D}'(T)^m$ in the continuous and $\mathcal{U} = (\mathbb{R}^T)^m$ in the discrete case, where $T = \mathbb{R}$, $\mathbb{R}_+$ and $T = \mathbb{N}, \mathbb{Z}$ are our usual time sets). In this section, we restrict to piecewise continuous input functions u in the continuous case, then the state function x is piecewise $\mathcal{C}^1$ (in particular, it is a classical function). One calls $X = \mathbb{R}^n$ the **state space** of these systems. Its elements are called **states**. The state transition map

$$\varphi : \{(t, t_0) \in T^2 \mid t \geq t_0\} \times X \times \mathcal{U} \to X, \quad (t, t_0, x_0, u) \mapsto \varphi(t, t_0, x_0, u)$$

yields the state at time t if the state at time t_0 was x_0, and the input function was u. More concretely, we have

$$\varphi(t, t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau \tag{4.1}$$

and

$$\varphi(t, t_0, x_0, u) = A^{t-t_0}x_0 + \sum_{i=t_0}^{t-1} A^{t-i-1}Bu(i). \tag{4.2}$$

The state transition maps are **consistent**, i.e.,

$$\varphi(t, t, x, u) = x \tag{4.3}$$

for all $t \in T$, $x \in X$, $u \in \mathcal{U}$; and they are **strictly causal**, that is, if $u_1(t) = u_2(t)$ for all $t_0 \leq t < t_1$ (note the strict inequality!), then

$$\varphi(t_1, t_0, x, u_1) = \varphi(t_1, t_0, x, u_2) \tag{4.4}$$

for all $x \in X$. Moreover, we have the **semigroup property**:

$$\varphi(t_2, t_1, \varphi(t_1, t_0, x, u), u) = \varphi(t_2, t_0, x, u) \tag{4.5}$$

for all $t_0 \leq t_1 \leq t_2 \in T$, $x \in X$, $u \in \mathcal{U}$. As a consequence, we get the following **concatenation property**: Let $t_1 \in T$, and let $u_1, u_2 \in \mathcal{U}$ be two admissible inputs. Define their concatenation at time t_1 by

$$u(t) = \begin{cases} u_1(t) & \text{if } t < t_1 \\ u_2(t) & \text{if } t_1 \leq t \end{cases}$$

which is again an admissible input. Then

$$\varphi(t_2, t_0, x, u) = \varphi(t_2, t_1, \varphi(t_1, t_0, x, u_1), u_2)$$

for all $x \in X$, and all $t_0 \leq t_1 \leq t_2 \in T$.

Due to **linearity**, we have

$$\varphi(t, t_0, \lambda_1 x_1 + \lambda_2 x_2, \lambda_1 u_1 + \lambda_2 u_2) = \lambda_1 \varphi(t, t_0, x_1, u_1) + \lambda_2 \varphi(t, t_0, x_2, u_2) \tag{4.6}$$

for all $t \geq t_0 \in T$, $\lambda_1, \lambda_2 \in \mathbb{R}$, $x_1, x_2 \in X$, $u_1, u_2 \in \mathcal{U}$. In particular,

$$\varphi(t, t_0, 0, 0) = 0$$

for all $t \geq t_0 \in T$. One says that the zero state is an **equilibrium** of the system when the zero input function is applied.

Due to **time-invariance**, we have

$$\varphi(t, t_0, x_0, u) = \varphi(t - \tau, t_0 - \tau, x_0, \sigma^\tau u) \tag{4.7}$$

for all $x_0 \in X$, $u \in \mathcal{U}$, and all $t \geq t_0 \in T$ and all $\tau \in T$ with the property that $t - \tau, t_0 - \tau \in T$. Here, σ^τ is the shift operator defined by $\sigma^\tau u(s) = u(s + \tau)$ for all s. In particular,

$$\varphi(t, t_0, x_0, u) = \varphi(t - t_0, 0, x_0, \sigma^{t_0} u).$$

Reachability and controllability are concerned with the following questions: Is it possible to steer the system, by a suitable choice of the input function, from one particular state x_0 (which is determined by the past of the system and may be thought of as "unwanted") to another particular state x_1 (which is prescribed by us and thus "desired")? How long does the transition from x_0 to x_1 take? Can we find a concrete formula for an input function that forces the system to go from x_0 to x_1? To formulate and answer these questions, some new concepts need to be introduced.

Definition 4.1 Let $t_0 \in T$ be fixed. One says that the state $x_1 \in X$

can be reached from $\boldsymbol{x_0} \in \boldsymbol{X}$ in time $\boldsymbol{\tau} \in \boldsymbol{T}$ ($\tau \geq 0$) if there exists $u \in \mathcal{U}$ such that

$$\varphi(t_0 + \tau, t_0, x_0, u) = x_1.$$

Equivalently, we say that **$\boldsymbol{x_0}$ can be controlled to $\boldsymbol{x_1}$ in time $\boldsymbol{\tau}$**.

can be reached from $\boldsymbol{x_0} \in \boldsymbol{X}$ if this holds for at least one $\tau \geq 0$. Equivalently, **$\boldsymbol{x_0}$ can be controlled to $\boldsymbol{x_1}$**.

We say that the system is

completely reachable from $\boldsymbol{x_0} \in \boldsymbol{X}$ if any $x_1 \in X$ is reachable from x_0.

completely controllable to $\boldsymbol{x_1} \in \boldsymbol{X}$ if any $x_0 \in X$ can be controlled to x_1.

completely reachable (controllable) if x_1 can be reached from x_0 (or: x_0 can be controlled to x_1) for all $x_0, x_1 \in X$.

The choice of the starting time t_0 plays no role since we are dealing with time-invariant systems: We have

$$\varphi(t_0 + \tau, t_0, x_0, u) = \varphi(\tau, 0, x_0, \sigma^{t_0} u).$$

Therefore, if there exists a $t_0 \in T$ such that x_1 is reachable from x_0 (when starting at time t_0) then this is true for any other starting time, e.g., for $t_0 = 0$. Thus we can often choose $t_0 = 0$ for simplicity.

Let $\tau \in T$, $\tau \geq 0$, and $x_0, x_1 \in X$. Define the following sets:

$$\begin{aligned} \mathcal{R}(\tau, x_0) &:= \{x \in X \mid x \text{ is reachable from } x_0 \text{ in time } \tau\} \\ \mathcal{C}(\tau, x_1) &:= \{x \in X \mid x \text{ is controllable to } x_1 \text{ in time } \tau\}. \end{aligned}$$

Moreover, we let $\mathcal{R}(\tau) := \mathcal{R}(\tau, 0)$ denote the set of states that are reachable from $x_0 = 0$ in time τ, and $\mathcal{C}(\tau) := \mathcal{C}(\tau, 0)$ is the set of states that are controllable to $x_1 = 0$ in time τ. Finally,

$$\mathcal{R} := \bigcup_{\tau \geq 0} \mathcal{R}(\tau) \quad \text{and} \quad \mathcal{C} := \bigcup_{\tau \geq 0} \mathcal{C}(\tau)$$

are the set of states that are reachable from zero, and the set of states that are controllable to zero, respectively. The system is completely reachable from zero if and only if $\mathcal{R} = X$, and it is completely controllable to zero if and only if $\mathcal{C} = X$.

Theorem 4.2 Let $s, t \in T$, $0 \leq s \leq t$. We have

1. $\mathcal{R}(s) \subseteq \mathcal{R}(t)$ and $\mathcal{C}(s) \subseteq \mathcal{C}(t)$;
2. $\mathcal{R}(t), \mathcal{C}(t), \mathcal{R}, \mathcal{C}$ are subspaces of $X = \mathbb{R}^n$;
3. There exists $\tau^* \in T$, $\tau^* \geq 0$ such that

$$\mathcal{R} = \mathcal{R}(\tau) \quad \text{and} \quad \mathcal{C} = \mathcal{C}(\tau) \quad \text{for all } \tau \geq \tau^*.$$

Proof:

1. Let $x \in \mathcal{R}(s)$. Then there exists a input function u such that

$$\varphi(t, t-s, 0, u) = x.$$

On the other hand,

$$\varphi(t-s, 0, 0, 0) = 0.$$

Let $\tilde{u}$ be the input function defined by

$$\tilde{u}(\tau) = \begin{cases} 0 & \text{if } \tau < t-s \\ u(\tau) & \text{if } t-s \leq \tau. \end{cases}$$

We have

$$\varphi(t, 0, 0, \tilde{u}) = \varphi(t, t-s, \varphi(t-s, 0, 0, 0), u) = \varphi(t, t-s, 0, u) = x$$

which shows that $x \in \mathcal{R}(t)$. The statement for $\mathcal{C}$ is analogous.

2. We show that $\mathcal{R}(t)$ is a vector space, the statement for $\mathcal{R}$, $\mathcal{C}(t)$, $\mathcal{C}$ is analogous. We have $0 \in \mathcal{R}(t)$, because

$$\varphi(t, 0, 0, 0) = 0.$$

Let $x_1, x_2 \in \mathcal{R}(t)$, and $\lambda_1, \lambda_2 \in \mathbb{R}$. We need to show that $\lambda_1 x_1 + \lambda_2 x_2 \in \mathcal{R}(t)$. By assumption, there exist input functions u_i such that

$$\varphi(t, 0, 0, u_i) = x_i$$

for $i = 1, 2$. Then

$$\varphi(t, 0, 0, \lambda_1 u_1 + \lambda_2 u_2) = \lambda_1 x_1 + \lambda_2 x_2$$

and hence $\lambda_1 x_1 + \lambda_2 x_2 \in \mathcal{R}(t)$.

3. Consider a strictly increasing sequence

$$0 = \tau_0 < \tau_1 < \tau_2 < \dots$$

in T with $\lim_{i \to \infty} \tau_i = \infty$. By part 1,

$$\mathcal{R}(\tau_0) \subseteq \mathcal{R}(\tau_1) \subseteq \mathcal{R}(\tau_2) \subseteq \dots$$

By part 2, this is a sequence of subspaces of $X = \mathbb{R}^n$, with

$$\dim \mathcal{R}(\tau_0) \leq \dim \mathcal{R}(\tau_1) \leq \dim \mathcal{R}(\tau_2) \leq \dots \leq n.$$

This is an increasing sequence of integers less than or equal to n. Such a sequence must become stationary, that is, there exists i_0 such that

$$\dim \mathcal{R}(\tau_i) = \dim \mathcal{R}(\tau_{i_0}) \quad \text{for all } i \geq i_0.$$

We use the following fact from linear algebra: If a vector space is contained in another vector space of the same finite dimension, then the two vector spaces must be the same. Thus

$$\mathcal{R}(\tau_i) = \mathcal{R}(\tau_{i_0}) \quad \text{for all } i \geq i_0.$$

For any $\tau \in T$, $\tau \geq \tau_{i_0}$, there exists $j \geq i_0$ such that $\tau \leq \tau_j$. Then $\mathcal{R}(\tau_{i_0}) = \mathcal{R}(\tau) = \mathcal{R}(\tau_j)$. We conclude that

$$\mathcal{R}(\tau) = \mathcal{R}(\tau_{i_0}) \quad \text{for all } \tau \geq \tau_{i_0}$$

and thus

$$\mathcal{R} = \bigcup_{\tau \geq 0} \mathcal{R}(\tau) = \bigcup_{\tau \geq \tau_{i_0}} \mathcal{R}(\tau) = \mathcal{R}(\tau_{i_0}).$$

Put $\tau^* := \tau_{i_0}$, then for $\tau \geq \tau^*$, we have

$$\mathcal{R} \supseteq \mathcal{R}(\tau) \supseteq \mathcal{R}(\tau^*) = \mathcal{R}$$

and thus $\mathcal{R}(\tau) = \mathcal{R}$ for all $\tau \geq \tau^*$. □

Corollary 4.3 In discrete time,

$$\mathcal{R}(n) = \mathcal{R} \quad \text{and} \quad \mathcal{C}(n) = \mathcal{C},$$

where n is the dimension of the state space. In continuous time,

$$\mathcal{R}(\varepsilon) = \mathcal{R} \quad \text{and} \quad \mathcal{C}(\varepsilon) = \mathcal{C}$$

for every $\varepsilon > 0$.

Remark 4.4 This is probably the first time we encounter a significant difference between the continuous and discrete cases. In a discrete system, if x can be reached from x_0 at all, then it can also be reached in time n, where n is the dimension of the state space. In a continuous system, if x can be reached from x_0 at all, then it can also be reached in an arbitrarily small time ε. This is counterintuitive at first sight: In a "real world" system, it certainly takes "some time" to change from one state to another. The reason is that we admit arbitrarily large input values here, i.e., we make the optimistic assumption that we can put as much "energy" as we like into the system. In a real world system, there are constraints which limit the size of the admissible inputs, and this has the consequence that the transition from one state to another cannot be done arbitrarily fast in practice.

Proof: For discrete time, we use that we know from the previous proof that

$$\mathcal{R}(0) \subseteq \mathcal{R}(1) \subseteq \mathcal{R}(2) \subseteq \dots \tag{4.8}$$

becomes stationary, that is, there exists i_0 such that $\mathcal{R}(i) = \mathcal{R}(i_0)$ for all $i \geq i_0$ and then $\mathcal{R}(i_0) = \mathcal{R}$. We have to show that this happens for some $i_0 \leq n$. Then we are finished, because $\mathcal{R} = \mathcal{R}(i_0) \subseteq \mathcal{R}(n) \subseteq \mathcal{R}$ yields the desired result. Considering the dimensions $d_i := \dim \mathcal{R}(i) \leq n$, we have

$$0 = d_0 \leq d_1 \leq d_2 \leq \dots \leq d_n \leq d_{n+1}.$$

These inequalities cannot all be strict, i.e., we must have $d_i = d_{i+1}$ and hence

$$\mathcal{R}(i) = \mathcal{R}(i+1) \tag{4.9}$$

for some $i \leq n$. The claim is that then we may put $i_0 = i$, that is, the first equality in (4.8) will already yield stationarity. Thus we have to show that (4.9) implies

$$\mathcal{R}(i) = \mathcal{R}(i+k) \quad \text{for all } k \geq 0.$$

We do this by induction on k. The statement is trivial for $k = 0$. Let's assume that the statement is true for k. We need to show it for $k+1$. The inclusion

$$\mathcal{R}(i) \subseteq \mathcal{R}(i+k+1)$$

is clear. For the converse, let $x \in \mathcal{R}(i+k+1)$. This means that there exists an input function u such that

$$\varphi(i+k+1, 0, 0, u) = x.$$

Set

$$x_1 := \varphi(i+k, 0, 0, u).$$

Then

$$x = \varphi(i+k+1, i+k, x_1, u) = \varphi(i+1, i, x_1, \sigma^k u)$$

and $x_1 \in \mathcal{R}(i+k)$, which equals $\mathcal{R}(i)$ by the inductive assumption. Thus there exists an input function u_1 with

$$\varphi(i, 0, 0, u_1) = x_1.$$

Let u_2 be defined by

$$u_2(\tau) = \begin{cases} u_1(\tau) & \text{if } \tau < i \\ (\sigma^k u)(\tau) & \text{if } i \leq \tau. \end{cases}$$

Then

$$\varphi(i+1, 0, 0, u_2) = \varphi(i+1, i, \varphi(i, 0, 0, u_1), \sigma^k u) = \varphi(i+1, i, x_1, \sigma^k u) = x$$

which shows that $x \in \mathcal{R}(i+1)$. Finally, (4.9) implies that $x \in \mathcal{R}(i)$ as desired. For continuous time, let $\varepsilon > 0$ be given. Consider

$$\mathcal{R}(0) \subseteq \mathcal{R}(\frac{\varepsilon}{n}) \subseteq \mathcal{R}(\frac{2\varepsilon}{n}) \subseteq \dots$$

and apply the same argument as for discrete time. $\square$

Corollary 4.5 The following are equivalent: The system is

1. completely reachable (controllable);
2. completely reachable from zero, that is, $\mathcal{R} = X$.

Proof: It is clear that statement 1 implies 2. For the converse direction, let $x_0, x_1 \in X$ be given. We wish to show that x_1 can be reached from x_0.

In continuous time, pick any $\tau^* = \varepsilon > 0$. In discrete time, choose $\tau^* = n$, where n is the dimension of X. Define $x := x_1 - \varphi(\tau^*, 0, x_0, 0)$. By assumption, $x \in \mathcal{R} = \mathcal{R}(\tau^*)$, that is, there exists an input function $u \in \mathcal{U}$ with

$$x = \varphi(\tau^*, 0, 0, u).$$

This can be rewritten as

$$x_1 = \varphi(\tau^*, 0, x_0, 0) + \varphi(\tau^*, 0, 0, u) = \varphi(\tau^*, 0, x_0, u)$$

showing that x_1 can be reached from x_0. □

Until now, we have not used the specific form of the state transition maps from (4.1) and (4.2), but only their properties (4.3)–(4.7). This will change now.

Corollary 4.6 Consider $\dot{x} = Ax + Bu$ or $\sigma x = Ax + Bu$. In the discrete case, assume that A is invertible. Then the following are equivalent: The system is

1. completely reachable (controllable);
2. completely reachable from zero, that is, $\mathcal{R} = X$;
3. completely controllable to zero, that is, $\mathcal{C} = X$.

Proof: We only need to prove "3 ⇒ 1". Let $x_0, x_1 \in X$ be given. We wish to show that x_1 can be reached from x_0.

In continuous time, pick $\varepsilon > 0$ and define $x := x_0 - e^{-A\varepsilon}x_1$. By assumption, $x \in \mathcal{C} = \mathcal{C}(\varepsilon)$, that is, there exists an input function $u \in \mathcal{U}$ with

$$0 = \varphi(\varepsilon, 0, x, u) = e^{A\varepsilon}x + \int_0^\varepsilon e^{A(\varepsilon-\tau)}Bu(\tau)\mathrm{d}\tau.$$

This can be rewritten as

$$x_1 = e^{A\varepsilon}x_0 + \int_0^\varepsilon e^{A(\varepsilon-\tau)}Bu(\tau)\mathrm{d}\tau = \varphi(\varepsilon, 0, x_0, u)$$

showing that x_1 can be reached from x_0.

If A is invertible, an analogous argument can be applied in the discrete case. □

Remark 4.7 Here we have another difference between continuous and discrete systems. The reason is that e^{At} is always an invertible matrix, whereas its discrete counterpart A^t is invertible if and only if A is. Therefore, we have to make this additional assumption in the discrete case. Without it, complete controllability to zero may be strictly weaker than complete controllability. Take for instance $\sigma x = Ax + Bu$ with

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Let $x_0 = x(0) \in \mathbb{R}^2$ be any given initial state. Put $u(0) = -x_1(0)$, then

$$x_1(1) = x_1(0) + u(0) = 0 \quad \text{and} \quad x_2(1) = x_1(0) + u(0) = 0,$$

that is, $x(1) = 0$. This shows that any x_0 can be controlled to zero (in time 1). Hence the system is completely controllable to zero, or $\mathcal{C} = \mathcal{C}(1) = \mathbb{R}^2$. However, it is not completely controllable. If we start in $x_0 = 0$, then every state $x(t)$ will satisfy $x_1(t) = x_2(t)$. This shows that $\mathcal{R} \neq \mathbb{R}^2$, i.e., the system is not completely reachable from zero, and thus it is not completely reachable (controllable).

The next goal is to determine conditions for (complete) reachability/controllability in terms of the matrices A, B. For this, we define the **controllability Gramians** (named after the Danish mathematician J. P. Gram, 1850–1916)

$$W(t) = \int_0^t e^{A\tau} BB^T e^{A^T\tau} \mathrm{d}\tau \quad \text{or} \quad W(t) = \sum_{i=0}^{t-1} A^i BB^T (A^T)^i \quad \in \mathbb{R}^{n\times n} \tag{4.10}$$

and the **Kalman reachability/controllability matrix** (named after the Hungarian control scientist R. E. Kalman, 1930–)

$$K = \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n\times nm}. \tag{4.11}$$

Remark 4.8 The definitions (4.10) and (4.11) are motivated by the following observations: Consider $\sigma x = Ax + Bu$. We know that $\mathcal{R} = \mathcal{R}(n)$. Moreover, $x \in \mathcal{R}$ if and only if there exists $u \in \mathcal{U}$, that is, a sequence of input vectors $u(0), \ldots, u(n-1) \in \mathbb{R}^m$ such that

$$\begin{aligned} x &= \sum_{i=0}^{n-1} A^{n-1-i} Bu(i) = A^{n-1}Bu(0) + \ldots + ABu(n-2) + Bu(n-1) \\ &= \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix} =: Kv. \end{aligned}$$

Therefore, x is reachable from zero if and only if the equation $x = Kv$ possesses a solution $v \in \mathbb{R}^{nm}$. This is the case if and only if $x \in \operatorname{im}(K)$. So we have $\mathcal{R} = \operatorname{im}(K)$. Moreover, note that $W(n) = KK^T$. Since $\operatorname{im}(K) = \operatorname{im}(KK^T)$ holds for any real matrix K, we also have $\mathcal{R} = \operatorname{im}(W(n))$.

If $x \in \mathcal{R}$, we therefore have $x = W(n)z = KK^T z$ for some $z \in \mathbb{R}^n$. Then $v^* := K^T z$ is a special solution of $Kv = x$. This corresponds to

$$\begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix} = \begin{bmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{n-1} \end{bmatrix} z.$$

In other words, $u(t) = B^T (A^T)^{n-t-1} z$ is a special input function that steers the system from state 0 to state x in time n.

Interestingly, the analogous statement is also valid for continuous systems, although the proof is a bit more involved. This is the content of the main theorem of this section, which is stated next.

Theorem 4.9 Consider $\dot{x} = Ax + Bu$ or $\sigma x = Ax + Bu$. In continuous time, let $\tau^* = \varepsilon > 0$ be arbitrary. In discrete time, let $\tau^* = n$, where n is the dimension of the state space. We have

$$\mathcal{R} = \mathcal{R}(\tau^*) = \operatorname{im}(W(\tau^*)) = \operatorname{im}(K).$$

Therefore, the following are equivalent:

1. $\dot{x} = Ax + Bu$ or $\sigma x = Ax + Bu$ is reachable/controllable;
2. $W(\tau^*)$ is non-singular;
3. K has full row rank.

Moreover in that case, an input function which steers the system from state 0 to state x in time τ^* is given by

$$u(t) = B^T e^{A^T(\tau^* - t)} W(\tau^*)^{-1} x \quad \text{or} \quad u(t) = B^T (A^T)^{\tau^* - t - 1} W(\tau^*)^{-1} x.$$

Note that this special input function is smooth in the continuous case (although only piecewise continuity has been required at the beginning).

Since $W(\tau^*)$ is always positive semi-definite due to its form, condition 2 from above is also equivalent to: $W(\tau^*)$ is positive definite, that is, $x^T W(\tau^*) x > 0$ for all $0 \neq x \in \mathbb{R}^n$. Then we write $W(\tau^*) > 0$.

Remark 4.10 It is worth noting that we obtain the same, purely algebraic condition for reachability/controllability both in continuous and in discrete time, namely, $\operatorname{rank}(K) = n$, where n is the dimension of the state space. In that case, it is not ambiguous to simply say that the matrix pair (A, B) is reachable/controllable (this notion is independent of the time set).

The well-known Hamilton-Cayley theorem, i.e.,

$$\chi_A(A) = A^n + a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 I = 0$$

implies the n-th power (and hence all higher powers) of an $n \times n$ matrix A is a linear combination of the first n powers of A, that is, $A^0 = I, A^1 = A, \ldots, A^{n-1}$. Therefore we have

$$\begin{aligned} \operatorname{im}(K) \;=\; K\mathbb{R}^{nm} &= \operatorname{span}\{A^i b_j \mid i = 0, \ldots, n-1, j = 1, \ldots, m\} \\ &= \operatorname{span}\{A^i b_j \mid i \in \mathbb{N}, j = 1, \ldots, m\} \end{aligned}$$

where $b_j \in \mathbb{R}^n$ are the columns of B, that is, $B = [b_1, \ldots, b_m]$. This will be used in the proof of Theorem 4.9.

The result

$$\mathcal{R} = \operatorname{im}(K) = \operatorname{im}(B) + A\operatorname{im}(B) + \ldots + A^{n-1}\operatorname{im}(B)$$

can also be formulated as follows: $\mathcal{R}$ is the smallest A-invariant (that is, $A\mathcal{R} \subseteq \mathcal{R}$) subspace of the state space X that contains $\operatorname{im}(B)$.

Proof: Due to Remark 4.8, we only need to do the continuous case. Then $\tau^* = \varepsilon > 0$ is arbitrary. We start with showing that $\mathcal{R}(\varepsilon) \subseteq \operatorname{im}(K)$: If $x \in \mathcal{R}(\varepsilon)$, then there exists $u \in \mathcal{U}$ such that

$$x = \int_0^\varepsilon e^{A\tau} Bu(\varepsilon - \tau)\mathrm{d}\tau = \int_0^\varepsilon \sum_{i=0}^\infty \frac{\tau^i}{i!} A^i Bu(\varepsilon - \tau)\mathrm{d}\tau \in \operatorname{im}(K).$$

Secondly, we wish to show that $\operatorname{im}(K) = \operatorname{im}(W(\varepsilon))$. From linear algebra, we know that it is equivalent to prove that $\operatorname{im}(K)^\perp = \operatorname{im}(W(\varepsilon))^\perp$. Let $x \in \operatorname{im}(K)^\perp$, that is, $\langle x, y\rangle = 0$ for all $y \in \operatorname{im}(K)$. Then $x^T K z = 0$ for all $z \in \mathbb{R}^{nm}$, which means that $x^T K = 0$ and hence $x^T A^i B = 0$ for all i. Then also $x^T e^{At} B = x^T \sum_{i=0}^\infty \frac{t^i}{i!} A^i B = 0$ and thus $x^T W(\varepsilon) = 0$. This shows that $x \in \operatorname{im}(W(\varepsilon))^\perp$. Conversely, let $x \in \operatorname{im}(W(\varepsilon))^\perp$, then

$$x^T W(\varepsilon) = \int_0^\varepsilon x^T e^{A\tau} BB^T e^{A^T\tau}\mathrm{d}\tau = 0.$$

Post-multiplying this by x, we obtain

$$\int_0^{\varepsilon} \|B^T e^{A^T\tau} x\|^2 \mathrm{d}\tau = 0.$$

We conclude that the smooth function $f(\tau) = B^T e^{A^T\tau} x$ is the constant zero function. Then also all its derivatives are zero. Evaluating them at $\tau = 0$, we obtain

$$B^T x = 0, \quad B^T A^T x = 0, \quad B^T (A^T)^2 x = 0, \quad \ldots$$

that is, $x^T K = 0$ and hence $x \in \operatorname{im}(K)^{\perp}$.

Finally, we need to show that $\operatorname{im}(W(\varepsilon)) \subseteq \mathcal{R}(\varepsilon)$. Let $x \in \operatorname{im}(W(\varepsilon))$, then there exists $z \in \mathbb{R}^n$ such that

$$x = W(\varepsilon) z.$$

Set $u(t) = B^T e^{A^T(\varepsilon - t)} z$. Then

$$\begin{aligned}
\varphi(\varepsilon, 0, 0, u) &= \int_0^{\varepsilon} e^{A(\varepsilon-\tau)} B u(\tau) \mathrm{d}\tau \\
&= \int_0^{\varepsilon} e^{A(\varepsilon-\tau)} B B^T e^{A^T(\varepsilon-\tau)} z \mathrm{d}\tau \\
&= \int_0^{\varepsilon} e^{A\tau} B B^T e^{A^T\tau} z \mathrm{d}\tau = W(\varepsilon) z = x
\end{aligned}$$

which shows that $x \in \mathcal{R}(\varepsilon)$. □

Remark 4.11 One can show that the special input functions given in Theorem 4.9 are optimal in the sense that the **energy** associated with them, that is

$$E(u) = \int_0^{\tau^*} \|u(\tau)\|^2 \mathrm{d}\tau \quad \text{or} \quad E(u) = \sum_{i=0}^{\tau^*-1} \|u(i)\|^2$$

is minimal among the energies of all u with $\varphi(\tau^*, 0, 0, u) = x$. This minimal energy (for controlling the system from 0 to x in time τ^*) is given by $E_{\min}(\tau^*, x) = x^T W(\tau^*)^{-1} x$. This shows that the smaller t is, the more energy is needed to do the transition from 0 to x in time t (compare this with Remark 4.4). More precisely, if $0 < s < t$, then $W(t) - W(s) > 0$, which implies $W(s)^{-1} - W(t)^{-1} > 0$, and hence

$$E_{\min}(s, x) = x^T W(s)^{-1} x > x^T W(t)^{-1} x = E_{\min}(t, x) \quad \text{for all } x \neq 0.$$

Again, the statement is easy to prove in discrete time (then we put $\tau^* = n$): Let v be defined as in Remark 4.8, then $E(u) = \|v\|^2$. We need to find the solution v of $Kv = x$ which makes $\|v\|^2$ minimal. The special input from Theorem 4.9 corresponds to $v^* = K^T(KK^T)^{-1}x$. Let v be another solution of $Kv = x$, then $v = v^* + v_0$, where v_0 solves the homogeneous equation $Kv_0 = 0$. Then

$$\|v\|^2 = \|v^* + v_0\|^2 = \|v^*\|^2 + 2\langle v^*, v_0\rangle + \|v_0\|^2.$$

However, $\langle v^*, v_0\rangle = \langle K^T(KK^T)^{-1}x, v_0\rangle = \langle (KK^T)^{-1}x, Kv_0\rangle = 0$. Hence

$$\|v\|^2 = \|v^*\|^2 + \|v_0\|^2 \geq \|v^*\|^2$$

which shows that v^* is indeed the solution of $Kv = x$ that has minimum norm. Finally,

$$E_{\min}(n, x) = \|v^*\|^2 = \|K^T(KK^T)^{-1}x\|^2 = x^T(KK^T)^{-1}x = x^T W(n)^{-1}x.$$

For continuous time, see Appendix F.

Corollary 4.12 Consider $\dot{x} = Ax + Bu$ or $\sigma x = Ax + Bu$. In the discrete case, assume that A is invertible. Then $\mathcal{C} = \mathcal{R}$.

In general, we only have $\mathcal{R} \subseteq \mathcal{C}$, see for example Remark 4.7.

Proof: Let τ^* be as usual. We have $x \in \mathcal{C} = \mathcal{C}(\tau^*)$ if and only if there exists $u \in \mathcal{U}$ such that

$$\varphi(\tau^*, 0, x, u) = \varphi(\tau^*, 0, x, 0) + \varphi(\tau^*, 0, 0, u) = 0,$$

that is, $\varphi(\tau^*, 0, x, 0) = \varphi(\tau^*, 0, 0, -u)$. Thus $x \in \mathcal{C}$ if and only if $\varphi(\tau^*, 0, x, 0) \in \mathcal{R}$. In continuous time, this means

$$x \in \mathcal{C} \quad \Leftrightarrow \quad e^{A\tau^*}x \in \mathcal{R}. \tag{4.12}$$

Since $e^{A\tau^*}$ is invertible, this shows that $\dim(\mathcal{C}) = \dim(\mathcal{R})$. In discrete time,

$$x \in \mathcal{C} \quad \Leftrightarrow \quad A^{\tau^*}x \in \mathcal{R}. \tag{4.13}$$

If A is invertible, we can argue as in the continuous case to see that $\mathcal{R}$ and $\mathcal{C}$ have the same dimension. Thus it suffices to show that $\mathcal{R} \subseteq \mathcal{C}$. If $x \in \mathcal{R} = \mathcal{R}(\tau^*)$, then $A^{\tau^*}x$ and $e^{A\tau^*}x$ are also in $\mathcal{R}$ (this is due to the A-invariance of $\mathcal{R}$). According to (4.12) and (4.13), this implies $x \in \mathcal{C}$. $\square$

4.2 Controllable matrix pairs

Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. We say that the matrix pair (A, B) is **controllable** if the associated Kalman controllability matrix

$$K = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

has full row rank, that is, $\mathrm{rank}(K) = n$.

If a state space system $\dot{x} = Ax + Bu$ is subject to a coordinate transform $x = Tz$, where $T \in \mathbb{R}^{n\times n}$ is invertible, then we get

$$\dot{z} = T^{-1}ATz + T^{-1}Bu.$$

Discrete systems behave analogously. We say that the matrix pair $(T^{-1}AT, T^{-1}B)$ is **similar** to the matrix pair (A, B). Of course, a coordinate transform should not change structural system properties such as stability and controllability. Indeed, similar matrices have the same eigenvalues, and the ranks of the Kalman controllability matrices of similar matrix pairs coincide.

The following result is limited to the **single-input** case, that is, $m = 1$. Then B is a single column vector. In that case, we simply write b instead of B. The associated Kalman controllability matrix

$$K = \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix}$$

is then a square matrix.

Theorem 4.13 Let $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, and let (A, b) be a controllable matrix pair. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & & & -a_1 \\ & \ddots & & \vdots \\ & & 1 & -a_{n-1} \end{bmatrix} \quad \text{and} \quad \tilde{b} := T^{-1}b = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The numbers a_i are precisely the coefficients of the characteristic polynomial, that is,

$$\chi_A(s) = \chi_{T^{-1}AT}(s) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0.$$

This is called the **controllability form** of (A, b). Moreover, there exists an invertible matrix $T_1 \in \mathbb{R}^{n\times n}$ such that

$$T_1^{-1}AT_1 = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix} \quad \text{and} \quad T_1^{-1}b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

This is called the **controller form** of (A, b). The coefficients a_i are the same as with the controllability form.

Proof: According to the Hamilton-Cayley theorem,

$$A^n = -a_{n-1}A^{n-1} - \ldots - a_1 A - a_0 I.$$

Thus we have

$$\begin{aligned} AK &= \begin{bmatrix} Ab & A^2b & \cdots & A^n b \end{bmatrix} \\ &= \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & & & -a_1 \\ & \ddots & & \vdots \\ & & 1 & -a_{n-1} \end{bmatrix} = K\tilde{A}. \end{aligned}$$

Moreover, we have

$$b = \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = K\tilde{b}.$$

Since (A, b) is controllable, K is invertible. Thus we simply put $T = K$. This completes the proof for the controllability form. For the controller form, the construction is a bit more involved, and it is omitted here. □

If a scalar input-output representation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1\dot{y} + a_0 y = u$$

is reduced to first order in the usual way, i.e., via putting $x = [y, \dot{y}, \ldots, y^{(n-1)}]^T$, then the resulting state space system is precisely in controller form.

We return to the general **multi-input** case, and we give another result about transforming a given matrix pair into some special form via a similarity transform (this corresponds to a coordinate transform in the state space).

Theorem 4.14 (Kalman controllability decomposition) Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Let K be the associated Kalman controllability matrix and let $r := \operatorname{rank}(K)$. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} \quad \text{and} \quad \tilde{B} := T^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{r\times r}$, $B_1 \in \mathbb{R}^{r\times m}$ is a controllable matrix pair.

Remark 4.15 The theorem says that by a suitable coordinate transform, namely $x = Tz$, the given system $\dot{x} = Ax + Bu$ can be put into the form

$$\begin{aligned} \dot{z}_1 &= A_1 z_1 + A_2 z_2 + B_1 u & (4.14) \\ \dot{z}_2 &= A_3 z_2 & (4.15) \end{aligned}$$

where (A_1, B_1) is controllable. The second equation is certainly not controllable (it is completely decoupled from the input, hence we cannot influence z_2 by the choice of the input u). In fact, it is an autonomous equation. If we start in $z_2(0) = 0$, then we have $z_2(t) = 0$ for all t. Then the first equation becomes

$$\dot{z}_1 = A_1 z_1 + B_1 u$$

which is controllable. Thus the states that are reachable from zero in the system (4.14), (4.15) are precisely those of the form $\binom{\zeta}{0}$, where $\zeta \in \mathbb{R}^r$ is arbitrary. In other words, the reachable space of (4.14), (4.15) takes the simple form

$$\{z \in \mathbb{R}^n \mid z \text{ can be reached from } 0\} = \mathbb{R}^r \times \{0\}.$$

This can be used to determine the reachable space of the original system, because

$$\{x \in \mathbb{R}^n \mid x \text{ can be reached from } 0\} = T(\mathbb{R}^r \times \{0\}).$$

Note that if the original (A, B) is controllable, then $r = n$, and the Kalman controllability decomposition becomes trivial. Thus the interesting case arises when (A, B) itself is not controllable.

Proof: Let $v_1, \ldots, v_r \in \mathbb{R}^n$ be a basis of $\mathcal{R}$. Let $w_1, \ldots, w_{n-r} \in \mathbb{R}^n$ be such that $v_1, \ldots, v_r, w_1, \ldots, w_{n-r}$ is a basis of $\mathbb{R}^n$. Then

$$T := \begin{bmatrix} V & W \end{bmatrix} := \begin{bmatrix} v_1 & \cdots & v_r & w_1 & \cdots & w_{n-r} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is an invertible matrix. Due to the A-invariance of $\mathcal{R}$, the columns of AV are again in $\mathcal{R}$. Thus they can be written as linear combinations of the vectors v_i, that is,

$$AV = VA_1$$

for some matrix $A_1 \in \mathbb{R}^{r \times r}$. On the other hand, the columns of AW are in $\mathbb{R}^n$ and thus they can be written as linear combinations of v_i, w_j, that is,

$$AW = VA_2 + WA_3$$

for some matrices A_2, A_3 of appropriate sizes. Summing up, we have

$$AT = \begin{bmatrix} AV & AW \end{bmatrix} = \begin{bmatrix} VA_1 & VA_2 + WA_3 \end{bmatrix} = \begin{bmatrix} V & W \end{bmatrix} \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} = T\tilde{A}.$$

Since $\operatorname{im}(B) \subseteq \mathcal{R}$, the columns of B are linear combinations of the vectors v_i, that is,

$$B = VB_1 = \begin{bmatrix} V & W \end{bmatrix} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = T\tilde{B}$$

for some matrix $B_1 \in \mathbb{R}^{r \times m}$. It remains to show that (A_1, B_1) is a controllable matrix pair. The Kalman controllability matrix associated to $(\tilde{A}, \tilde{B})$ is

$$\tilde{K} = \begin{bmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}\tilde{B} \end{bmatrix} = \begin{bmatrix} B_1 & A_1B_1 & \cdots & A_1^{n-1}B_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

and this has rank r, just like the Kalman controllability matrix of the original matrix pair. Therefore,

$$r = \operatorname{rank} \begin{bmatrix} B_1 & A_1B_1 & \cdots & A_1^{n-1}B_1 \end{bmatrix}.$$

Due to the Hamilton-Cayley theorem, this implies that

$$r = \operatorname{rank} \begin{bmatrix} B_1 & A_1B_1 & \cdots & A_1^{r-1}B_1 \end{bmatrix}$$

which shows that (A_1, B_1) is controllable. $\square$

Remark 4.16 Let χ_A and $\operatorname{spec}(A)$ denote the characteristic polynomial and the spectrum of A, respectively. In a Kalman controllability decomposition, we clearly have

$$\chi_A = \chi_{A_1} \cdot \chi_{A_3}$$

and thus

$$\operatorname{spec}(A) = \operatorname{spec}(A_1) \cup \operatorname{spec}(A_3).$$

One calls χ_{A_3} the **uncontrollable part** of the characteristic polynomial of A with respect to B, and $\lambda \in \operatorname{spec}(A_3)$ an **uncontrollable mode** of (A, B). Of course, it has to be verified that these notions do not depend on the specific choice of the Kalman decomposition (which is certainly non-unique as its construction involves, e.g., the choice of a particular basis of $\mathcal{R}$).

Example 4.17 Consider the matrices from Remark 4.7,

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The Kalman controllability matrix is

$$K = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

which has rank $r = 1$. We choose the vector $v_1 := B$ as a basis of $\mathcal{R} = \operatorname{im}(K)$. If we choose

$$T := \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

then

$$T^{-1}AT = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad T^{-1}B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Thus we see that $\chi_A(s) = (s-1)s$ with uncontrollable part s, and hence 0 is an uncontrollable mode of (A, B).

There is also a direct way to characterize the uncontrollable modes of a matrix pair.

Theorem 4.18 Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, and $\lambda \in \mathbb{C}$. The following are equivalent:

1. λ is an uncontrollable mode of (A, B);
2. $\operatorname{rank} \begin{bmatrix} \lambda I - A & B \end{bmatrix} < n$.

Proof: Let $\tilde{A} = T^{-1}AT$ and $\tilde{B} = T^{-1}B$. Since

$$\begin{bmatrix} \lambda I - \tilde{A} & \tilde{B} \end{bmatrix} = T^{-1} \begin{bmatrix} \lambda I - A & B \end{bmatrix} \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix},$$

we have

$$\operatorname{rank} \begin{bmatrix} \lambda I - \tilde{A} & \tilde{B} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \lambda I - A & B \end{bmatrix}$$

and thus we may assume, without loss of generality, that a Kalman controllability decomposition has already been performed, i.e.,

$$A = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$

where (A_1, B_1) is controllable. Then

$$H(\lambda) := \begin{bmatrix} \lambda I - A & B \end{bmatrix} = \begin{bmatrix} \lambda I - A_1 & -A_2 & B_1 \\ 0 & \lambda I - A_3 & 0 \end{bmatrix}.$$

If λ is an uncontrollable mode, then it is an eigenvalue of A_3. Thus it makes $\lambda I - A_3$ singular, showing that the rank of $H(\lambda)$ cannot be full, that is, it must be less than n. Conversely, assume that the rank of $H(\lambda)$ is not full. Then there exists a vector $x \neq 0$ such that $xH(\lambda) = 0$, that is,

$$\begin{aligned} x_1(\lambda I - A_1) &= 0 \\ -x_1 A_2 + x_2(\lambda I - A_3) &= 0 \\ x_1 B_1 &= 0. \end{aligned}$$

The first and third equations imply that

$$x_1 B_1 = 0, \quad x_1 A_1 B_1 = 0, \quad x_1 A_1^2 B_1 = 0, \quad \ldots$$

The controllability of (A_1, B_1) yields that $x_1 = 0$. Then $x_2 \neq 0$ and

$$x_2(\lambda I - A_3) = 0$$

which implies that λ is an eigenvalue of A_3, that is, an uncontrollable mode. □

Example 4.19 Consider once more

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Then

$$H(\lambda) = \begin{bmatrix} \lambda - 1 & 0 & 1 \\ -1 & \lambda & 1 \end{bmatrix}$$

which has rank 2 for all $\lambda \neq 0$. However, $\text{rank}(H(0)) = 1$, showing again that 0 is an uncontrollable mode of this system.

As a direct consequence of this, we obtain another characterization of controllable matrix pairs.

Corollary 4.20 (Hautus test for controllability) The following are equivalent:

1. (A, B) is controllable.
2. The matrix $H(\lambda) = \begin{bmatrix} \lambda I - A & B \end{bmatrix}$ has full row rank for all $\lambda \in \mathbb{C}$.

The polynomial matrix $H = \begin{bmatrix} sI - A & B \end{bmatrix} \in \mathbb{R}[s]^{n\times(n+m)}$ is called **Hautus controllability matrix**. Since $\lambda I - A$ is non-singular whenever λ is not an eigenvalue of A, it suffices to check condition 2 from above for $\lambda \in \operatorname{spec}(A)$.

Remark 4.21 Controllability is a **generic** property, that is, if a matrix pair (A, B) is chosen "at random", then it is very likely to be a controllable one. More precisely, the set of controllable matrix pairs (A, B) is open and dense in the set $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$.

4.3 Asymptotic controllability

Sometimes, it is not required that a system should go from one state to another in *finite* time τ. Instead, one is satisfied if this happens asymptotically as $\tau \to \infty$.

Definition 4.22 We say that a state space system $\dot{x} = Ax + Bu$ or $\sigma x = Ax + Bu$ is **asymptotically controllable (to zero)** if for any $x_0 \in X = \mathbb{R}^n$, there exists an input function $u \in \mathcal{U}$ such that

$$\lim_{t\to\infty} \varphi(t, 0, x_0, u) = 0.$$

Clearly, controllability implies asymptotic controllability. The proof of the following theorem will be done in the next chapter.

Theorem 4.23 A state space system is asymptotically controllable if and only if its uncontrollable modes λ are asymptotically stable, that is, $\operatorname{Re}(\lambda) < 0$ in continuous time, and $|\lambda| < 1$ in discrete time.

4.4 Controllable behaviors

The Hautus test gives us an idea about how to generalize the notion of controllability from state space systems $\dot{x} = Ax + Bu$ to general systems $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$

where $R \in \mathbb{R}[s]^{p\times q}$, and $w \in \mathcal{A}^q$. In a state space system,

$$R = \begin{bmatrix} sI - A & -B \end{bmatrix} \in \mathbb{R}[s]^{n\times(n+m)} \quad \text{and} \quad w = \begin{bmatrix} x \\ u \end{bmatrix} \in \mathcal{A}^{n+m}.$$

The polynomial matrix R is recognized as the Hautus controllability matrix (up to the sign of B, which does not influence the rank).

In this section, we restrict to continuous systems, and we return to our original signal spaces, that is, $\mathcal{A} = \mathcal{D}'(T)$, where $T = \mathbb{R}, \mathbb{R}_+$.

Definition 4.24 Let $t_0 \in T$ be fixed. Let $R \in \mathbb{R}[s]^{p\times q}$ and

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}$$

where $\mathcal{A}$ is as described above. Let $w^{(1)}, w^{(2)} \in \mathcal{B}$ be two trajectories. We say that $w^{(2)}$ is reachable from $w^{(1)}$ in time $\tau \geq 0$ (or: $w^{(1)}$ is controllable to $w^{(2)}$ in time τ) if there exists a trajectory $w \in \mathcal{B}$ which coincides with $w^{(1)}$ on $(-\infty, t_0) \cap T$ and with $w^{(2)}$ on $(t_0 + \tau, \infty) \cap T$. If there is any such $\tau \geq 0$, then we say that $w^{(2)}$ is reachable from $w^{(1)}$ (or: $w^{(1)}$ is controllable to $w^{(2)}$). One says that $\mathcal{B}$ is **controllable** if any $w^{(1)} \in \mathcal{B}$ can be controlled to any $w^{(2)} \in \mathcal{B}$.

To say that two distributions coincide on an open set $U \subseteq \mathbb{R}$ means that they assign the same value to each test function whose support lies in U. In the situation described above, one calls w a **connecting trajectory** for $w^{(1)}, w^{(2)}$. For classical functions, this means

$$w(t) = \begin{cases} w^{(1)}(t) & \text{if } t < t_0 \\ w^{(2)}(t) & \text{if } t > t_0 + \tau. \end{cases}$$

The choice of the starting time t_0 makes no difference, because we consider only time-invariant systems. Similarly as with state space systems, the transition time τ can be made arbitrarily small, independently of the choice of the two trajectories to be connected.

Theorem 4.25 (Generalized Hautus test) Let $\mathcal{A} = \mathcal{D}'(T)$ for $T = \mathbb{R}$ or $\mathbb{R}_+$ and let $R \in \mathbb{R}[s]^{p\times q}$. Without loss of generality, let R have full row rank. Then the behavior

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}$$

is controllable if and only if $\mathrm{rank}(R(\lambda)) = p$ for all $\lambda \in \mathbb{C}$.

Proof: Let R have full row rank and let

$$URV = \begin{bmatrix} D & 0 \end{bmatrix}$$

be the Smith form of R, with $D = \mathrm{diag}(d_1, \ldots, d_p)$. Since U, V are unimodular, we have, for all $\lambda \in \mathbb{C}$,

$$\mathrm{rank}(R(\lambda)) = \mathrm{rank}(URV(\lambda)) = \mathrm{rank}(D(\lambda)).$$

Thus, $\mathrm{rank}(R(\lambda)) = p$ for all $\lambda \in \mathbb{C}$ if and only if $\det(D(\lambda)) \neq 0$ for all $\lambda \in \mathbb{C}$, that is, if no d_i has a zero in $\mathbb{C}$. This is true if and only if all d_i are constants. Since we may always assume that the d_i are monic polynomials (i.e., their leading coefficients are equal to one), we have

$$\mathrm{rank}(R(\lambda)) = p \text{ for all } \lambda \in \mathbb{C} \quad \Leftrightarrow \quad D = I.$$

Consider as usual

$$\tilde{\mathcal{B}} = \{\tilde{w} \in \mathcal{A}^q \mid \begin{bmatrix} D & 0 \end{bmatrix} \tilde{w} = 0\}$$

which is related to $\mathcal{B}$ via the isomorphism $\tilde{\mathcal{B}} \to \mathcal{B}$, $\tilde{w} \mapsto w = V(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{w}$. This $\tilde{\mathcal{B}}$ is controllable if and only if $\mathcal{B}$ is controllable. However, if $D = I$, then

$$\tilde{\mathcal{B}} = \{0\} \times \mathcal{A}^{q-p} \subset \mathcal{A}^q$$

which is clearly controllable. Conversely, if $D \neq I$, there exists at least one d_i, say d_1, which is not constant. Then the equation for the first component $\tilde{w}_1$ of $\tilde{w}$ reads

$$d_1(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{w}_1 = 0$$

which has precisely the solutions

$$\tilde{w}_1(t) = \sum_{\lambda} a_\lambda(t) e^{\lambda t} \tag{4.16}$$

where $\lambda \in \mathbb{C}$ are the zeros of d_1 (since d_1 is not a constant, there exists at least one such λ). Now consider two trajectories $\tilde{w}^{(1)}$, $\tilde{w}^{(2)}$ in $\tilde{\mathcal{B}}$, where the first component of $\tilde{w}^{(1)}$ is non-zero, and the first component of $w^{(2)}$ is zero. Then $\tilde{w}^{(2)}$ is not reachable from $\tilde{w}^{(1)}$: A connecting trajectory $\tilde{w}$ would have to satisfy $\tilde{w}_1 = 0$ on $(t_0 + \tau, \infty)$, which implies that $\tilde{w}_1 = 0$ everywhere because of (4.16). This shows that the system is not controllable. $\square$

Polynomial matrices that represent controllable behaviors are characterized by the following theorem.

Theorem 4.26 Let $R \in \mathbb{R}[s]^{p\times q}$ be a polynomial matrix with full row rank. The following are equivalent:

1. $\operatorname{rank}(R(\lambda)) = p$ for all $\lambda \in \mathbb{C}$;
2. The Smith form of R is $\begin{bmatrix} I & 0 \end{bmatrix}$;
3. There exists a matrix $T \in \mathbb{R}[s]^{(q-p)\times q}$ such that $\begin{bmatrix} R \\ T \end{bmatrix}$ is unimodular;
4. There exists a matrix $S \in \mathbb{R}[s]^{q\times p}$ such that $RS = I$;
5. If $R = UR_1$ for some $U \in \mathbb{R}[s]^{p\times p}$, $R_1 \in \mathbb{R}[s]^{p\times q}$, then U must be unimodular.

If the equivalent conditions are satisfied, we say that R is **left prime** (or: left irreducible).

Proof: We have already seen in the previous proof that "1 $\Rightarrow$ 2". The converse is obvious.

For "2 $\Rightarrow$ 3", assume that

$$R = U \begin{bmatrix} I & 0 \end{bmatrix} V = U \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = UV_1$$

where U and V are unimodular. Define $T := V_2$, then

$$\begin{bmatrix} R \\ T \end{bmatrix} = \begin{bmatrix} UV_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

which shows that the matrix is unimodular.

For "3 $\Rightarrow$ 4", let T be a matrix according to assertion 3. Then there exist matrices S_1, S_2 such that

$$\begin{bmatrix} R \\ T \end{bmatrix} \begin{bmatrix} S_1 & S_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

In particular, $RS_1 = I$.

For "4 $\Rightarrow$ 5", let S be such that $RS = I$ and $R = UR_1$. Then $UR_1S = I$, which shows that U is unimodular.

Finally, we show "5 $\Rightarrow$ 2" by negation. Assume that the Smith form is $\begin{bmatrix} D & 0 \end{bmatrix}$ with $D = \mathrm{diag}(d_1, \ldots, d_p)$ and at least one of the d_i is not a constant, say $d_1 \notin \mathbb{R}$. Then

$$R = U \begin{bmatrix} D & 0 \end{bmatrix} V = U_1 R_1$$

with

$$U_1 = U \begin{bmatrix} d_1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \quad \text{and} \quad R_1 = \begin{bmatrix} 1 & & & & \\ & d_2 & & & \\ & & \ddots & & 0 \\ & & & d_m & \end{bmatrix} V.$$

We have $\det(U_1) = \det(U) d_1$ and thus we have found a factorization $R = U_1 R_1$ in which U_1 is not unimodular. $\square$

A polynomial matrix $R \in \mathbb{R}[s]^{p\times q}$ with full row rank is left prime if and only if its $p \times p$ subdeterminants have no common zeros in $\mathbb{C}$ (since we have $\mathrm{rank}(R(\lambda)) < p$ if and only if $\lambda \in \mathbb{C}$ is a common zero of all $p \times p$ minors). By the fundamental theorem of algebra, it is also equivalent to say that the $p \times p$ subdeterminants of R are devoid of common factors, i.e., they are coprime polynomials in $\mathbb{R}[s]$.

Chapter 5

Feedback control

5.1 Static state feedback

The problem with pre-computed control ("open loop control") is that it may be sensitive with respect to noisy data. For instance, let u be such that

$$\varphi(t, 0, x, u) = 0,$$

that is, the control function u steers the system from x to 0 in time t. If we start in a slightly perturbed initial state, say $x(0) = x + \xi$ with $\|\xi\|$ small but non-zero, we obtain

$$\varphi(t, 0, x(0), u) = \varphi(t, 0, \xi, 0)$$

where $\varphi(t, 0, \xi, 0) = e^{tA}\xi$ in continuous time, and $\varphi(t, 0, \xi, 0) = A^t\xi$ in discrete time. If A is not stable, this deviation from 0 (the desired state) can become arbitrarily large. Thus pre-computed control will usually not work with unstable systems.

Feedback control is an alternative approach. Its basic assumption is that we can measure the state x, and that we can use this information for control. For a state space system

$$\dot{x} = Ax + Bu \quad \text{or} \quad \sigma x = Ax + Bu \tag{5.1}$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$, a **static state feedback law** reads

$$u = Fx + v$$

where $F \in \mathbb{R}^{m\times n}$. Combining the given system with the feedback law yields the so-called **closed loop system**

$$\dot{x} = (A + BF)x + Bv \quad \text{or} \quad \sigma x = (A + BF)x + Bv. \tag{5.2}$$

We say that the matrix pair $(A + BF, B)$ results from (A, B) via static state feedback with feedback matrix F. The goal of this chapter is to answer questions like the **stabilization problem**: Given A, B, can we find F such that $A+BF$ is asymptotically stable? If yes, then all the trajectories of the closed loop system with $v = 0$ will tend to zero as $t \to \infty$. Thus the system is asymptotically controllable to zero, and this will work even if the initial state is subject to disturbance.

5.2 Feedback and controllability

Lemma 5.1 The reachable spaces of (5.1) and (5.2) coincide. In particular, $(A + BF, B)$ is controllable if and only if (A, B) is controllable.

Proof: The reachable space of $(A + BF, B)$ is

$$\mathcal{R} = \text{im}(B) + (A + BF)\text{im}(B) + \ldots + (A + BF)^{n-1}\text{im}(B).$$

Let

$$\mathcal{R}_1 = \text{im}(B) \quad \text{and} \quad \mathcal{R}_{i+1} = (A + BF)\mathcal{R}_i + \text{im}(B) \text{ for } i = 1, \ldots , n - 1.$$

Then $\mathcal{R}_n = \mathcal{R}$. Since

$$\text{im}(B) + (A + BF)\mathcal{V} = \text{im}(B) + A\mathcal{V}$$

holds for any vector space $\mathcal{V}$, we conclude that

$$\mathcal{R}_1 = \text{im}(B) \quad \text{and} \quad \mathcal{R}_{i+1} = A\mathcal{R}_i + \text{im}(B).$$

This shows that

$$\mathcal{R} = \mathcal{R}_n = \text{im}(B) + A\text{im}(B) + \ldots + A^{n-1}\text{im}(B)$$

which is the reachable space of (A, B). □

Lemma 5.2 (Heymann) Let (A, B) be controllable, and let $0 \neq b \in \text{im}(B)$. Then there exists a matrix F such that $(A + BF, b)$ is controllable.

Proof: First, we claim that there exist vectors $u^{(1)}, \dots, u^{(n-1)} \in \mathbb{R}^m$ such that the vectors $v^{(i)}$ defined by

$$v^{(1)} := b \quad \text{and} \quad v^{(i+1)} := Av^{(i)} + Bu^{(i)} \text{ for } i = 1, \dots, n-1$$

are a basis of $\mathbb{R}^n$. We prove this by induction. Since $v^{(1)} = b \neq 0$, we have a one-dimensional vector space

$$\mathcal{V}^{(1)} := \operatorname{span}\{v^{(1)}\}.$$

Now assume that $v^{(1)}, \dots, v^{(k)}$ have already been constructed and that

$$\mathcal{V}^{(k)} = \operatorname{span}\{v^{(1)}, \dots, v^{(k)}\}$$

has dimension k. We wish to choose $u^{(k)}$ such that

$$v^{(k+1)} = Av^{(k)} + Bu^{(k)} \notin \mathcal{V}^{(k)}.$$

Then $\dim(\mathcal{V}^{(k+1)}) = k + 1$ as desired. We need to show that such a choice is always possible for $k < n$. Assume conversely that

$$Av^{(k)} + Bu \in \mathcal{V}^{(k)} \quad \text{for all } u \in \mathbb{R}^m.$$

In particular,

$$Av^{(k)} \in \mathcal{V}^{(k)} \tag{5.3}$$

and thus

$$\operatorname{im}(B) \subseteq \mathcal{V}^{(k)}. \tag{5.4}$$

On the other hand, $\mathcal{V}^{(k)}$ is A-invariant, because

$$A\mathcal{V}^{(k)} = \operatorname{span}\{Av^{(1)}, \dots, Av^{(k)}\}.$$

To see that $Av^{(i)} \in \mathcal{V}^{(k)}$, we use (5.3) for the case $i = k$, and for $1 \leq i \leq k-1$, we have $Av^{(i)} = v^{(i+1)} - Bu^{(i)} \in \mathcal{V}^{(k)}$ due to (5.4). Thus $\mathcal{V}^{(k)}$ is an A-invariant subspace of $\mathbb{R}^n$ that contains the image of B. Thus $\mathcal{V}^{(k)}$ contains the reachable space of (A, B) which is the *smallest* such space. Since (A, B) is controllable, this reachable space is all of $\mathbb{R}^n$, and hence $\mathcal{V}^{(k)} = \mathbb{R}^n$ which implies that $k = n$. Thus the construction works whenever $k < n$, and $\mathcal{V}^{(n)} = \operatorname{span}\{v^{(1)}, \dots, v^{(n)}\} = \mathbb{R}^n$.

Now let $F : \mathbb{R}^n \to \mathbb{R}^m$ be such that $Fv^{(i)} = u^{(i)}$ for $i = 1, \dots, n-1$. Then

$$(A + BF)v^{(i)} = Av^{(i)} + Bu^{(i)} = v^{(i+1)}.$$

Since $v^{(1)} = b$, this implies

$$v^{(i)} = (A + BF)^{i-1} b \quad \text{for } i = 1, \dots, n.$$

Thus the Kalman matrix of $(A + BF, b)$ is $K = [v^{(1)}, \dots, v^{(n)}]$ which has rank n by construction. □

5.3 Pole placement

In this section, we investigate the characteristic polynomial and the spectrum of $A + BF$, where A, B are given, and F may be chosen. The goal is to place the eigenvalues of $A + BF$ in some desirable region of the complex plane (e.g., for obtaining asymptotic stability). This is known as pole placement (or: pole shifting, pole assignment).

Definition 5.3 Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given. Let p be a monic polynomial of degree n. We say that p is **assignable** to (A, B) if there exists a matrix $F \in \mathbb{R}^{m\times n}$ such that

$$\chi_{A+BF} = p,$$

that is, the characteristic polynomial of $A + BF$ equals p.

Theorem 5.4 Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. The matrix pair (A, B) is controllable if and only if every monic polynomial of degree n can be assigned to (A, B).

Remark 5.5 This means that if (A, B) is controllable, then the eigenvalues of $A+BF$ can be shifted, by choice of the feedback matrix F, to any desired location in the complex plane. More precisely, let $\Lambda \subset \mathbb{C}$ be a non-empty set, with

$$\lambda \in \Lambda \quad \Rightarrow \quad \overline{\lambda} \in \Lambda.$$

For $\lambda \in \Lambda$, let $\mu(\lambda)$ be a positive integer with $\sum_{\lambda\in\Lambda} \mu(\lambda) = n$ and $\mu(\lambda) = \mu(\overline{\lambda})$. Then there exists a matrix F such that

$$\chi_{A+BF}(s) = \prod_{\lambda\in\Lambda} (s - \lambda)^{\mu(\lambda)}.$$

In particular, $\operatorname{spec}(A + BF) = \Lambda$.

Recall that the Kalman controllability decomposition yields a factorization $\chi_A = \chi_{A_1} \cdot \chi_{A_3}$, where $\chi_u := \chi_{A_3}$ is the uncontrollable part of χ_A with respect to B. If (A, B) is controllable, one puts $A_1 := A$ and $\chi_u := 1$. Thus (A, B) is uncontrollable if and only if the degree of χ_u is at least one.

Proof: Since a similarity transform does not change the set of assignable polynomials, we may assume, without loss of generality, that a Kalman controllability decomposition has already been performed. Then

$$A+BF = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \begin{bmatrix} F_1 & F_2 \end{bmatrix} = \begin{bmatrix} A_1+B_1F_1 & A_2+B_1F_2 \\ 0 & A_3 \end{bmatrix} \tag{5.5}$$

and thus $\chi_{A+BF} = \chi_{A_1+B_1F_1} \cdot \chi_{A_3}$. This shows that χ_{A+BF} will always be a multiple of the polynomial $\chi_u = \chi_{A_3}$. If (A, B) is not controllable, then χ_u is not a constant, and thus not every polynomial can be assigned to (A, B).

Conversely, let (A, B) be controllable. Let $p = s^n + p_{n-1}s^{n-1} + \ldots + p_1 s + p_0$ be given. We wish to construct F such that $\chi_{A+BF} = p$. If we choose any $0 \neq b \in \text{im}(B)$, then there exists $\tilde{F}$ such that $(A+B\tilde{F}, b)$ is controllable, according to Heymann's Lemma. Thus we can transform this matrix pair into controller form, that is, there exists a non-singular matrix T such that

$$T^{-1}(A + B\tilde{F})T = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix} \quad \text{and} \quad T^{-1}b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

where a_i are the coefficients of $\chi_{A+B\tilde{F}}$. Now let

$$f = \begin{bmatrix} a_0 - p_0 & \ldots & a_{n-1} - p_{n-1} \end{bmatrix} \in \mathbb{R}^{1\times n}.$$

Then

$$T^{-1}(A + B\tilde{F})T + T^{-1}bf = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -p_0 & -p_1 & \cdots & -p_{n-1} \end{bmatrix}$$

and therefore

$$\chi_{A+B\tilde{F}+bfT^{-1}} = s^n + p_{n-1}s^{n-1} + \ldots + p_1 s + p_0 = p.$$

Finally, we use that $b \in \text{im}(B)$, that is, $b = Bv$ for some $v \in \mathbb{R}^m$. Thus

$$A + B\tilde{F} + bfT^{-1} = A + B(\tilde{F} + vfT^{-1})$$

which yields the desired result putting $F := \tilde{F} + vfT^{-1}$. □

As a direct consequence, we obtain the main theorem of this chapter.

Theorem 5.6 (Pole shifting theorem) The polynomials that can be assigned to (A, B) are precisely the ones of the form

$$p = p_1 \cdot \chi_u$$

where χ_u is the uncontrollable part of χ_A with respect to B, and p_1 is an arbitrary monic polynomial of degree $r = n - \deg(\chi_u)$.

Proof: It follows from (5.5) that

$$\chi_{A+BF} = \chi_{A_1+B_1F_1} \cdot \chi_u,$$

where $A_1 \in \mathbb{R}^{r\times r}$ and $B_1 \in \mathbb{R}^{r\times m}$ are the usual matrices from the Kalman controllability decomposition, and $F = [F_1, F_2]$. Thus any assignable polynomial must be a multiple of χ_u. The first factor can be any monic polynomial of degree r, because (A_1, B_1) is a controllable matrix pair. □

Remark 5.7 The pole shifting theorem says that the uncontrollable modes of (A, B) cannot be influenced by static state feedback, whereas the other eigenvalues of the system can be moved to any desired location in the complex plane.

5.4 Stabilization

For stabilization, we don't require that the spectrum of $A + BF$ should coincide with some specific set of eigenvalues; we just want the eigenvalues to be contained in some given region of the complex plane. We put

$$\mathbb{C}_g = \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) < 0\} \quad \text{or} \quad \mathbb{C}_g = \{\lambda \in \mathbb{C} \mid |\lambda| < 1\}$$

in the continuous or discrete case, respectively.

Definition 5.8 A matrix pair (A, B) is called **stabilizable** if there exists F such that $A + BF$ is asymptotically stable, that is, $\mathrm{spec}(A + BF) \subset \mathbb{C}_g$.

Theorem 5.9 The following are equivalent:

1. (A, B) is stabilizable;

2. (A, B) is asymptotically controllable to zero;

3. All uncontrollable modes of (A, B) lie in $\mathbb{C}_g$;

4. Any eigenvalue λ of A which is not in $\mathbb{C}_g$ satisfies rank $\begin{bmatrix} \lambda I - A & B \end{bmatrix} = n$.

Remark 5.10 Part of this theorem was already stated (without proof) in Section 4.3.

Proof: Conditions 3 and 4 are logically equivalent.

For "1 $\Rightarrow$ 2", let F be such that $A + BF$ is asymptotically stable. Put $u = Fx$, then the closed loop system reads

$$\dot{x} = (A + BF)x \quad \text{or} \quad \sigma x = (A + BF)x$$

respectively, and thus $\lim_{t\to\infty} x(t) = 0$ for all $x(0) = x_0$. An explicit formula for u is

$$u(t) = Fe^{(A+BF)t}x_0 \quad \text{or} \quad u(t) = F(A + BF)^t x_0.$$

For "2 $\Rightarrow$ 3", we assume, without loss of generality, that a Kalman controllability decomposition has already been performed. Then the system law reads

$$\begin{array}{rcl} \dot{x}_1 & = & A_1x_1 + A_2x_2 + B_1u \\ \dot{x}_2 & = & A_3x_2 \end{array} \qquad \text{or} \qquad \begin{array}{rcl} \sigma x_1 & = & A_1x_1 + A_2x_2 + B_1u \\ \sigma x_2 & = & A_3x_2. \end{array}$$

By assumption, there exists, for any $x(0) = x_0$, an input function u such that $\lim_{t\to\infty} x(t) = 0$. Since $x_2(t) = e^{A_3 t}x_{02}$ or $x_2(t) = A_3^t x_{02}$, this can only be true if A_3 is asymptotically stable, which means that all the uncontrollable modes of (A, B) are asymptotically stable.

For "3 $\Rightarrow$ 1", choose a monic polynomial p_1 of degree $r = n - \deg(\chi_u)$ whose zeros are all in $\mathbb{C}_g$. By the pole shifting theorem, there exists F such that

$$\chi_{A+BF} = p_1 \cdot \chi_u.$$

By assumption, all zeros of χ_u lie in $\mathbb{C}_g$. Hence, all eigenvalues of $A + BF$ are contained in $\mathbb{C}_g$, that is, $A + BF$ is asymptotically stable. $\square$

5.5 Control as interconnection

Feedback control is based on the interconnection of systems: Given a dynamical system (the "plant"), the goal of feedback control is to design another system (a "controller"), in a way that the interconnection of the two systems has certain desired properties. As an example, consider a plant given in classical state space form

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

and let the controller be specified by the feedback law $u = Fx + v$. Then the interconnection (the "closed loop" system) is

$$\dot{x} = (A + BF)x + Bv. \tag{5.6}$$

A typical aim of the controller design in this setting is spectral assignment, that is, a condition is given on the desired location of the eigenvalues of $A+BF$. Note that interconnection means nothing but combining the equations that determine plant and controller, respectively, and to look at their common solutions. For instance, combining the plant given by

$$\begin{bmatrix} \frac{d}{dt}I - A & -B & 0 \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = 0$$

with the controller given by

$$\begin{bmatrix} F & -I & I \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = 0$$

yields the interconnected system

$$\begin{bmatrix} \frac{d}{dt}I - A & -B & 0 \\ \\ F & -I & I \end{bmatrix} \begin{bmatrix} x \\ u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

from which we may eliminate u to get (5.6). It is worth noting that this interconnection is regular in a sense to be defined below.

The behavioral approach to systems theory provides an elegant framework for dealing with such, and more general, interconnection problems. Suppose that the plant is given by $R_1(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$. Similarly, let the controller be determined by $R_2(\frac{d}{dt})w = 0$. Then the interconnection is determined by the system law

$$\begin{bmatrix} R_1 \\ R_2 \end{bmatrix} (\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0.$$

Let $\mathcal{B}_i$ be the plant and the controller,

$$\mathcal{B}_i = \{w \in \mathcal{A}^q \mid R_i(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}$$

respectively ($i = 1, 2$). Then their **interconnection** is defined as

$$\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2 = \{w \in \mathcal{A}^q \mid R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\},$$

where

$$R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}.$$

The interconnection is said to be **regular** if

$$\mathrm{rank}(R_1) + \mathrm{rank}(R_2) = \mathrm{rank}(R).$$

This means that the system laws of plant and controller are independent of each other. Equivalently, the interconnection is regular if $\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{A}^q$.

Let $\mathcal{B}_1$ be a given plant, and let $\mathcal{B}$ be some desired behavior. The goal is to find a controller $\mathcal{B}_2$ such that the interconnection of plant and controller equals this desired behavior, i.e.,

$$\mathcal{B}_1 \cap \mathcal{B}_2 = \mathcal{B}.$$

Obviously, $\mathcal{B} \subseteq \mathcal{B}_1$ is a necessary condition. If it is satisfied, we say that $\mathcal{B}$ is a **subsystem** of $\mathcal{B}_1$. Then the problem is always solvable, because there is the trivial solution $\mathcal{B}_2 = \mathcal{B}$. The question becomes more interesting if we require the interconnection to be regular. Then we say that $\mathcal{B}$ **can be achieved from** $\mathcal{B}_1$ **by regular interconnection**.

Theorem 5.11 Let $\mathcal{B}_1$ be controllable. Then every subsystem of $\mathcal{B}_1$ can be achieved from $\mathcal{B}_1$ by regular interconnection.

Proof: We have seen in the proof of Theorem 4.25 that a controllable behavior $\mathcal{B}_1$ is isomorphic to $\{0\} \times \mathcal{A}^m$, where m is the number of inputs of $\mathcal{B}_1$. Therefore, there exists a behavior $\mathcal{B}_1'$ such that $\mathcal{B}_1 \oplus \mathcal{B}_1' = \mathcal{A}^q$. Now let $\mathcal{B} \subseteq \mathcal{B}_1$ be given. We set $\mathcal{B}_2 := \mathcal{B} + \mathcal{B}_1'$. Then

$$\mathcal{B}_1 \cap \mathcal{B}_2 = \mathcal{B}_1 \cap (\mathcal{B} + \mathcal{B}_1') = \mathcal{B}.$$

To see the last equality, note that the inclusion "$\supseteq$" is obvious, since $\mathcal{B}$ is contained both in $\mathcal{B}_1$ and in $\mathcal{B} + \mathcal{B}_1'$. For the converse inclusion, let $w_1 \in \mathcal{B}_1$ have a decomposition $w_1 = w + w_1'$ with $w \in \mathcal{B} \subseteq \mathcal{B}_1$ and $w_1' \in \mathcal{B}_1'$. Then $w_1' = w_1 - w \in \mathcal{B}_1 \cap \mathcal{B}_1' = \{0\}$. Thus $w_1' = 0$ and $w_1 = w \in \mathcal{B}$ as desired. Finally,

$$\mathcal{B}_1 + \mathcal{B}_2 = \mathcal{B}_1 + \mathcal{B} + \mathcal{B}_1' = \mathcal{A}^q$$

from which it follows that the interconnection is regular. $\square$

Series connection: Let $P_i(\frac{\mathrm{d}}{\mathrm{d}t})y_i = Q_i(\frac{\mathrm{d}}{\mathrm{d}t})u_i$, where $i = 1, 2$, be two input-output representations. The series connection is defined by taking the output of the first system as the input of the second system. Of course, this is only possible if the dimensions match, that is, $p_1 = m_2$, which we assume. We set $u = u_1$, $y_1 = u_2$, and $y = y_2$. The interconnection is therefore described by

$$\begin{bmatrix} -Q_1 & P_1 & 0 \\ 0 & -Q_2 & P_2 \end{bmatrix} \begin{bmatrix} u \\ y_1 \\ y \end{bmatrix} = 0.$$

Since P_i are both square and non-singular, this is a regular interconnection.

Parallel connection: Let $P_i(\frac{\mathrm{d}}{\mathrm{d}t})y_i = Q_i(\frac{\mathrm{d}}{\mathrm{d}t})u_i$, where $i = 1, 2$, be two input-output representations. The parallel connection is defined by giving the same input to both systems, and by summing the outputs. This is only possible if $m := m_1 = m_2$ and $p := p_1 = p_2$. We put $u = u_1 = u_2$ and $y = y_1 + y_2$. The interconnection is given by

$$\begin{bmatrix} -Q_1 & P_1 & 0 & 0 \\ -Q_2 & 0 & P_2 & 0 \\ 0 & I & I & -I \end{bmatrix} \begin{bmatrix} u \\ y_1 \\ y_2 \\ y \end{bmatrix} = 0,$$

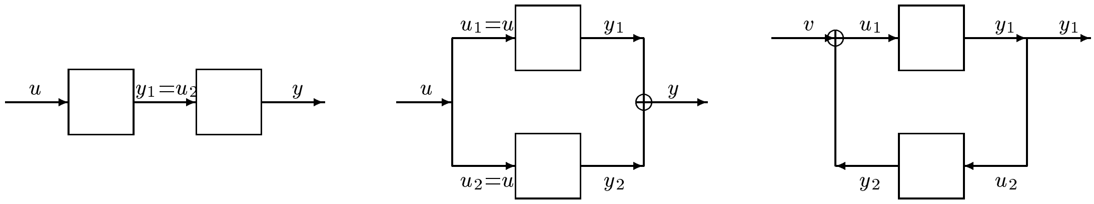
which is actually a regular interconnection of three systems (the two given systems and the summing system).

Feedback connection: Let $P_i(\frac{\mathrm{d}}{\mathrm{d}t})y_i = Q_i(\frac{\mathrm{d}}{\mathrm{d}t})u_i$, where $i = 1, 2$, be two input-output representations. The feedback connection is defined by taking the output of the first system as the input of the second, and vice versa, that is, $y_1 = u_2$ and $y_2 = u_1$. The connection is given by

$$\begin{bmatrix} P_1 & -Q_1 \\ -Q_2 & P_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0.$$

This interconnection is not necessarily regular. It becomes regular if we modify the equation $u_1 = y_2$, using a new input v, say $u_1 = y_2 + v$.

The three standard interconnections are illustrated in the following figure.



Chapter 6

Observability

6.1 Basic notions for state space systems

Consider the state space equations

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx + Du \end{array} \qquad \text{or} \qquad \begin{array}{rcl} \sigma x & = & Ax + Bu \\ y & = & Cx + Du, \end{array}$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$ and $D \in \mathbb{R}^{p \times m}$. Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^p$.

The state transition map

$$\varphi : \{(t, t_0) \in T^2 \mid t \geq t_0\} \times X \times \mathcal{U} \to X, \quad (t, t_0, x_0, u) \mapsto \varphi(t, t_0, x_0, u)$$

gives the state at time t if the state at time t_0 was x_0 and the control function u was applied. We define the **state-to-output map**

$$\eta : \{(t, t_0) \in T^2 \mid t \geq t_0\} \times X \times \mathcal{U} \to Y, \quad (t, t_0, x_0, u) \mapsto \eta(t, t_0, x_0, u)$$

as the map which gives the output at time t if the state at time t_0 was x_0 and the control function was u. For state space systems, we have $\eta(t, t_0, x_0, u) = C\varphi(t, t_0, x_0, u) + Du(t)$, and thus

$$\eta(t, t_0, x_0, u) = Ce^{A(t-t_0)}x_0 + \int_{t_0}^{t} Ce^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau + Du(t) \tag{6.1}$$

and

$$\eta(t, t_0, x_0, u) = CA^{t-t_0}x_0 + \sum_{i=t_0}^{t-1} CA^{t-i-1}Bu(i) + Du(t). \tag{6.2}$$

The state-to-output map is **causal**, i.e., if u_1 and u_2 coincide for all $t_0 \leq t \leq t_1$, then

$$\eta(t_1, t_0, x, u_1) = \eta(t_1, t_0, x, u_2)$$

for all x. We say that it is **strictly causal** if this is already implied by $u_1(t) = u_2(t)$ for $t_0 \leq t < t_1$. In a state space system, strict causality holds if and only if $D = 0$. Due to **linearity**, we have

$$\eta(t, t_0, \lambda_1 x_1 + \lambda_2 x_2, \lambda_1 u_1 + \lambda_2 u_2) = \lambda_1 \eta(t, t_0, x_1, u_1) + \lambda_2 \eta(t, t_0, x_2, u_2)$$

for all $t \geq t_0$, $\lambda_i \in \mathbb{R}$, $x_i \in X$, $u_i \in \mathcal{U}$.

Observability is concerned with the following problem. Usually, we only know the system's manifest variables (in a state space system, the manifest variables are input and output): The input is free and can be chosen by the control engineer, and the output is the system's response which can be measured. The latent variables (in a state space system, the latent variable is the state) are not directly measurable, in general. Remember that they are usually auxiliary variables introduced during modelling, or simply for mathematical convenience, e.g., for reducing a system to first order. Hence the physical meaning of the latent variables may be obscure. The following question arises: If we know the manifest variables of a system, what can we conclude about the latent variables? An observable system is one in which the latent variables can be reconstructed from the manifest variables. Due to the property of state, we only need to reconstruct the state at a specific time t_0, then we know it everywhere in the "future", i.e., for $t \geq t_0$.

Definition 6.1 Let $t_0 \in T$ be fixed. One says that the state $x \in X$

can be distinguished from state $x' \in X$ in time $\tau \in T$ ($\tau \geq 0$) if there exists $u \in \mathcal{U}$ and $t_0 \leq t \leq t_0 + \tau$ such that

$$\eta(t, t_0, x, u) \neq \eta(t, t_0, x', u);$$

then we also say that **u distinguishes between x and x' in time τ.**

can be distinguished from $x' \in X$ if this holds for at least one $\tau \geq 0$.

We say that the system is

observable if for any $x \neq x' \in X$, the state x can be distinguished from the state x'.

As usual, the starting time t_0 is not important as long as we deal only with time-invariant systems, and thus we put $t_0 = 0$. Define the set

$$\mathcal{J}(\tau, x) := \{x' \in X \mid x' \text{ is indistinguishable from } x \text{ in time } \tau\}$$

and let $\mathcal{J}(\tau) := \mathcal{J}(\tau, 0)$ denote the set of states that are indistinguishable from state 0 in time τ. We have

$$x \in \mathcal{J}(\tau) \quad \Leftrightarrow \quad \eta(t, 0, x, u) = \eta(t, 0, 0, u) \text{ for all } 0 \le t \le \tau, u \in \mathcal{U}.$$

Because of linearity,

$$\eta(t, 0, x, u) = \eta(t, 0, x, 0) + \eta(t, 0, 0, u).$$

Therefore we have

$$x \in \mathcal{J}(\tau) \quad \Leftrightarrow \quad \eta(t, 0, x, 0) = 0 \text{ for all } 0 \le t \le \tau.$$

Finally,

$$\mathcal{J} := \bigcap_{\tau \ge 0} \mathcal{J}(\tau)$$

is the set of states that are indistinguishable from state 0. We have

$$x \in \mathcal{J} \quad \Leftrightarrow \quad \eta(t, 0, x, 0) = 0 \text{ for all } t \ge 0.$$

Since $\eta(t, 0, 0, 0) = 0$ for all $t \ge 0$, this means that x cannot be distinguished from zero if and only if the zero input function does not distinguish between x and 0. In other words: If x can be distinguished from 0 at all, then it can also be distinguished from 0 by the zero input function. This shows that the choice of the input function plays no role for the question of observability, i.e., we may put $u = 0$ without loss of generality in most of this chapter.

Theorem 6.2 Let $s, t \in T$, $0 \le s \le t$. We have

1. $\mathcal{J}(t) \subseteq \mathcal{J}(s)$;
2. $\mathcal{J}(t)$, $\mathcal{J}$ are subspaces of $X = \mathbb{R}^n$;
3. There exists $\tau^* \in T$, $\tau^* \ge 0$ such that

$$\mathcal{J} = \mathcal{J}(\tau) \quad \text{for all } \tau \ge \tau^*.$$

Corollary 6.3 In discrete time,

$$\mathcal{J}(n-1) = \mathcal{J},$$

where n is the dimension of the state space. In continuous time,

$$\mathcal{J}(\varepsilon) = \mathcal{J}$$

for every $\varepsilon > 0$.

Corollary 6.4 The following are equivalent:

1. The system is observable;
2. Any non-zero state can be distinguished from zero, that is, $\mathcal{J} = \{0\}$.

We define the **observability Gramians**

$$W(t) = \int_0^t e^{A^T\tau} C^T C e^{A\tau} \mathrm{d}\tau \quad \text{or} \quad W(t) = \sum_{i=0}^{t} (A^T)^i C^T C A^i \quad \in \mathbb{R}^{n\times n}$$

and the **Kalman observability matrix**

$$O = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np\times n}.$$

Theorem 6.5 Consider $\dot{x} = Ax$ or $\sigma x = Ax$, $y = Cx$. In continuous time, let $\tau^* = \varepsilon > 0$ be arbitrary. In discrete time, let $\tau^* = n-1$, where n is the dimension of the state space. We have

$$\mathcal{J} = \mathcal{J}(\tau^*) = \ker(W(\tau^*)) = \ker(O).$$

Therefore, the following are equivalent:

1. $\dot{x} = Ax$ or $\sigma x = Ax$, $y = Cx$ is observable;
2. $W(\tau^*)$ is non-singular;
3. O has full column rank.

Moreover in that case, we have the reconstruction formulas

$$x(0) = W(\tau^*)^{-1} \int_0^{\tau^*} e^{A^T t} C^T y(t) \mathrm{d}t \quad \text{or} \quad x(0) = W(\tau^*)^{-1} \sum_{i=0}^{\tau^*} (A^T)^i C^T y(i).$$

Since $W(\tau^*)$ is always positive semi-definite due to its form, condition 2 from above is also equivalent to: $W(\tau^*) > 0$.

6.2 Observable matrix pairs

Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. We say that the matrix pair (A, C) is **observable** if the associated Kalman observability matrix

$$O = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

has full column rank, that is, $\operatorname{rank}(O) = n$.

If a state space system $\dot{x} = Ax + Bu$, $y = Cx + Du$ is subject to a coordinate transform $x = Tz$, where $T \in \mathbb{R}^{n\times n}$ is invertible, then we get

$$\begin{aligned} \dot{z} &= T^{-1}ATz + T^{-1}Bu \\ y &= CTz + Du. \end{aligned}$$

Discrete systems behave analogously. We say that the matrix pair $(T^{-1}AT, CT)$ is **similar** to the matrix pair (A, C). A coordinate transform does not change structural system properties such as stability and observability.

The following result is limited to the **single-output** case, that is, $p = 1$. Then C is a single row vector. In that case, we simply write c instead of C. The associated Kalman observability matrix is then a square matrix.

Theorem 6.6 Let $A \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^{1\times n}$, and let (A, c) be an observable matrix pair. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & & & -a_1 \\ & \ddots & & \vdots \\ & & 1 & -a_{n-1} \end{bmatrix} \quad \text{and} \quad \tilde{c} := cT = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

The numbers a_i are precisely the coefficients of the characteristic polynomial, that is,

$$\chi_A(s) = \chi_{T^{-1}AT}(s) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0.$$

This is called the **observer form** of (A, c). Moreover, there exists an invertible matrix $T_1 \in \mathbb{R}^{n\times n}$ such that

$$T_1^{-1}AT_1 = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix} \quad \text{and} \quad cT_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}.$$

This is called the **observability form** of (A, c). The coefficients a_i are the same as with the observer form.

If a scalar input-output representation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1\dot{y} + a_0 y = u$$

is reduced to first order in the usual way, i.e., via putting $x = [y, \dot{y}, \ldots, y^{(n-1)}]^T$ and $y = [1, 0, \ldots, 0]x$, then the resulting state space system is precisely in observability form.

We return to the general **multi-output** case, and we give another result about transforming a given matrix pair into some special form via a similarity transform.

Theorem 6.7 (Kalman observability decomposition) Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. Let O be the associated Kalman observability matrix and let $r := \operatorname{rank}(O)$. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{bmatrix} A_1 & 0 \\ A_2 & A_3 \end{bmatrix} \quad \text{and} \quad \tilde{C} := CT = \begin{bmatrix} C_1 & 0 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{r\times r}$, $C_1 \in \mathbb{R}^{p\times r}$ is an observable matrix pair.

Remark 6.8 The theorem says that via a suitable coordinate transform, namely $x = Tz$, the given system $\dot{x} = Ax$, $y = Cx$ can be put into the form

$$\begin{aligned} \dot{z}_1 &= A_1 z_1 \\ \dot{z}_2 &= A_2 z_1 + A_3 z_2 \\ y &= C_1 z_1 \end{aligned}$$

where (A_1, C_1) is observable. The state z_2 is certainly not observable (the output does not depend on z_2 at all, hence there is no chance of reconstructing z_2 by measuring y). On the other hand, since (A_1, C_1) is observable, $z_1(0)$ is reconstructible. Thus

$$\{z \in \mathbb{R}^n \mid z \text{ cannot be distinguished from } 0\} = \{0\} \times \mathbb{R}^{n-r}.$$

This can be used to determine the indistinguishable space of the original system, because

$$\{x \in \mathbb{R}^n \mid x \text{ cannot be distinguished from } 0\} = T(\{0\} \times \mathbb{R}^{n-r}).$$

Note that if the original (A, C) is observable, then $r = n$, and the Kalman observability decomposition becomes trivial. Thus the interesting case arises when (A, C) itself is not observable.

Remark 6.9 In a Kalman observability decomposition, we clearly have

$$\chi_A = \chi_{A_1} \cdot \chi_{A_3}$$

and thus

$$\operatorname{spec}(A) = \operatorname{spec}(A_1) \cup \operatorname{spec}(A_3).$$

One calls χ_{A_3} the **unobservable part** of the characteristic polynomial of A with respect to C, and $\lambda \in \operatorname{spec}(A_3)$ an **unobservable mode** of (A, C).

There is also a direct way to characterize the unobservable modes of a matrix pair.

Theorem 6.10 Let $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{p \times n}$, and $\lambda \in \mathbb{C}$. The following are equivalent:

1. λ is an unobservable mode of (A, C);
2. $\operatorname{rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} < n$.

As a direct consequence of this, we obtain another characterization of observable matrix pairs.

Corollary 6.11 (Hautus test for observability) The following are equivalent:

1. (A, C) is observable.
2. The matrix $H(\lambda) = \begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ has full column rank for all $\lambda \in \mathbb{C}$.

The polynomial matrix $H = \begin{bmatrix} sI - A \\ C \end{bmatrix} \in \mathbb{R}[s]^{(n+p)\times n}$ is called **Hautus observability matrix**. It suffices to check condition 2 from above for $\lambda \in \operatorname{spec}(A)$.

Remark 6.12 Observability is a **generic** property, that is, if a matrix pair (A, C) is chosen "at random", then it is very likely to be an observable one. More precisely, the set of observable matrix pairs (A, C) is open and dense in the set $\mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$.

6.3 Asymptotic observability

Observability means

$$x \in \mathcal{J} \quad \Rightarrow \quad x = 0.$$

An alternative formulation is

$$\eta(t, 0, x, 0) = 0 \text{ for all } t \geq 0 \quad \Rightarrow \quad \varphi(t, 0, x, 0) = 0 \text{ for all } t \geq 0.$$

For asymptotic observability, one is satisfied if this is true in the limit as $t \to \infty$.

Definition 6.13 We say that a state space system $\dot{x} = Ax$ or $\sigma x = Ax$, $y = Cx$, is **asymptotically observable** if $\eta(t, 0, x, 0) = 0$ for all $t \geq 0$ implies that

$$\lim_{t\to\infty} \varphi(t, 0, x, 0) = 0.$$

Clearly, observability implies asymptotic observability.

Theorem 6.14 A state space system is asymptotically observable if and only if its unobservable modes λ are asymptotically stable, that is, $\operatorname{Re}(\lambda) < 0$ in continuous time, and $|\lambda| < 1$ in discrete time.

6.4 Observable latent variable descriptions

The Hautus test gives us an idea about how to generalize the notion of observability from state space systems $\dot{x} = Ax + Bu$, $y = Cx + Du$ to general systems $R(\frac{\mathrm{d}}{\mathrm{d}t})w = M(\frac{\mathrm{d}}{\mathrm{d}t})l$ where $R \in \mathbb{R}[s]^{p\times q}$, $M \in \mathbb{R}[s]^{p\times r}$ and $w \in \mathcal{A}^q$, $l \in \mathcal{A}^r$. In a state space system,

$$R = \begin{bmatrix} B & 0 \\ -D & I \end{bmatrix}, \quad M = \begin{bmatrix} sI - A \\ C \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} u \\ y \end{bmatrix}, \quad l = x.$$

The polynomial matrix M is recognized as the Hautus observability matrix.

In this section, we restrict to continuous systems, and we return to our original signal spaces, that is, $\mathcal{A} = \mathcal{D}'(T)$, where $T = \mathbb{R}, \mathbb{R}_+$.

Definition 6.15 We say that the latent variables l can be observed from the manifest variables w if l is uniquely determined by w, which means that $w_1 = w_2$ implies that $l_1 = l_2$.

Theorem 6.16 (Generalized Hautus test) Let $\mathcal{A} = \mathcal{D}'(T)$ for $T = \mathbb{R}$ or $\mathbb{R}_+$ and let $M \in \mathbb{R}[s]^{p\times r}$. Without loss of generality, let M have full column rank. Then the latent variable description of

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists l \in \mathcal{A}^r : R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l\}$$

is observable if and only if $\operatorname{rank}(M(\lambda)) = r$ for all $\lambda \in \mathbb{C}$. In this case, one calls the matrix M **right prime** (or: right irreducible).

Remark 6.17 In Theorem 4.26, the notion of left primeness was characterized. The connection with right primeness is easy: A matrix is right prime if and only if its transpose is left prime.

Let

$$UMV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

be the Smith form of M. Since

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = M(\tfrac{\mathrm{d}}{\mathrm{d}t})l \quad \Leftrightarrow \quad U(\tfrac{\mathrm{d}}{\mathrm{d}t})R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = U(\tfrac{\mathrm{d}}{\mathrm{d}t})M(\tfrac{\mathrm{d}}{\mathrm{d}t})V(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{l} = \begin{bmatrix} D(\frac{\mathrm{d}}{\mathrm{d}t}) \\ 0 \end{bmatrix} \tilde{l}_1$$

where $l = V(\frac{\mathrm{d}}{\mathrm{d}t})\tilde{l}$, it is not restrictive to admit only full column rank matrices M.

Proof: We need to show that the right primeness of M is equivalent to

$$\mathcal{B}_{\text{unobs}} := \{l \in \mathcal{A}^r \mid M(\tfrac{\mathrm{d}}{\mathrm{d}t})l = 0\} = \{0\}.$$

If M is right prime, there exists a polynomial matrix S such that $SM = I$. Then

$$M(\tfrac{\mathrm{d}}{\mathrm{d}t})l = 0 \quad \Rightarrow \quad S(\tfrac{\mathrm{d}}{\mathrm{d}t})M(\tfrac{\mathrm{d}}{\mathrm{d}t})l = 0 \quad \Rightarrow \quad l = 0$$

and thus $\mathcal{B}_{\text{unobs}} = \{0\}$ as desired. Conversely, if M is not right prime, then its Smith form

$$UMV = \begin{bmatrix} D \\ 0 \end{bmatrix}$$

where $D = \operatorname{diag}(d_1, \ldots, d_r)$ contains at least one d_i which is not a constant. Define

$$\tilde{\mathcal{B}}_{\text{unobs}} = \{\tilde{l} \in \mathcal{A}^r \mid D(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{l} = 0\}$$

which is related to $\mathcal{B}_{\text{unobs}}$ via the isomorphism $\tilde{\mathcal{B}}_{\text{unobs}} \to \mathcal{B}_{\text{unobs}}$, $\tilde{l} \mapsto l = V(\tfrac{\mathrm{d}}{\mathrm{d}t})\tilde{l}$. Since one of the d_i is not a constant, $\tilde{\mathcal{B}}_{\text{unobs}} \neq \{0\}$ and hence $\mathcal{B}_{\text{unobs}} \neq \{0\}$. $\square$

Chapter 7

Observers

7.1 State observers

Feedback control requires the knowledge of the state of the system. However, the state is usually not directly measurable. In an observable system, it can – in principle – be reconstructed from the inputs and outputs. However, the reconstruction procedure is sensitive with respect to noisy data.

Observer design is an alternative approach. Its basic idea is to build another system whose state converges to the state of the given system, independently of the initial conditions. For a state space system (in this chapter, we put $D = 0$ for simplicity),

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx \end{array} \qquad \text{or} \qquad \begin{array}{rcl} \sigma x & = & Ax + Bu \\ y & = & Cx \end{array} \tag{7.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, a **state observer** has the form (in continuous time, the discrete case is analogous)

$$\begin{array}{rcl} \dot{z} & = & Az + Bu + Ld \\ \hat{y} & = & Cz \end{array}$$

where $L \in \mathbb{R}^{n\times p}$. The signal $d := \hat{y} - y$ is the difference between the observer output $\hat{y}$ and the output y of the original system. Thus, the observer is a state space system with state z, inputs u and d, and output $\hat{y}$. Since A, B, C are the same matrices as with the given system, the observer can be seen as a copy of the original system, with the additional input d. The observer equations can be rewritten as

$$\begin{aligned} \dot{z} &= (A + LC)z + Bu - Ly \\ \hat{y} &= Cz. \end{aligned} \tag{7.2}$$

We say that the matrix pair $(A + LC, C)$ results from (A, C) via the observer gain matrix L. In this form, the observer has precisely the input and output of the original system as its inputs. Thus it requires only the knowledge of the manifest variables of the given system. One should think of an observer as a signal processing algorithm rather than a physical system. Its goal is to produce an estimate z of the unknown state x.

Consider the difference between the true state x and the observer state z. For this, put $e := z - x$. Then (7.1) and (7.2) together imply that

$$\dot{e} = (A + LC)e$$

and thus $e(t) = e^{(A+LC)t}e(0)$. The goal of this chapter is to answer questions like the **detection problem**: Given A, C, can we find L such that $A + LC$ is asymptotically stable? If yes, then the error signal e will tend to zero as $t \to \infty$, for all $e(0)$. Thus we have

$$\lim_{t\to\infty} \|z(t) - x(t)\| = 0$$

for all $z(0), x(0)$. This shows that the observer state z will asymptotically approach the true state x, and this will work even if the initial states are subject to disturbance.

Lemma 7.1 The indistinguishable spaces of (7.1) and (7.2) coincide. In particular, $(A + LC, C)$ is observable if and only if (A, C) is observable.

7.2 Pole placement

In this section, we investigate the characteristic polynomial and the spectrum of $A + LC$, where A, C are given, and L may be chosen. The goal is to place the eigenvalues of $A + LC$ in some desirable region of the complex plane.

Definition 7.2 Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$ be given. Let p be a monic polynomial of degree n. We say that p is **assignable** to (A, C) if there exists a matrix $L \in \mathbb{R}^{n\times p}$ such that $\chi_{A+LC} = p$.

Theorem 7.3 Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. The matrix pair (A, C) is observable if and only if every monic polynomial of degree n can be assigned to (A, C).

Theorem 7.4 (Pole shifting theorem) The polynomials that can be assigned to (A, C) are precisely the ones of the form

$$p = p_1 \cdot \chi_u$$

where χ_u is the unobservable part of χ_A with respect to C, and p_1 is an arbitrary monic polynomial of degree $r = n - \deg(\chi_u)$.

7.3 Detection

For detection, we don't require that the spectrum of $A + LC$ should coincide with some specific set of eigenvalues; we just want the eigenvalues to be contained in some given region of the complex plane. We put

$$\mathbb{C}_g = \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) < 0\} \quad \text{or} \quad \mathbb{C}_g = \{\lambda \in \mathbb{C} \mid |\lambda| < 1\}$$

in the continuous or discrete case, respectively.

Definition 7.5 A matrix pair (A, C) is called **detectable** if there exists L such that $A + LC$ is asymptotically stable, that is, $\mathrm{spec}(A + LC) \subset \mathbb{C}_g$.

Theorem 7.6 The following are equivalent:

1. (A, C) is detectable;
2. (A, C) is asymptotically observable;
3. All unobservable modes of (A, C) lie in $\mathbb{C}_g$;
4. Any eigenvalue λ of A which is not in $\mathbb{C}_g$ satisfies $\mathrm{rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n$.

Remark 7.7 Part of this theorem was already stated in Section 6.3.

7.4 Compensators

Finally, we combine feedback control with state observation. Consider the state space system

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

We wish to stabilize the system via a feedback $u = Fx + v$. However, since we do not know x, we build an observer

$$\dot{z} = (A + LC)z + Bu - Ly.$$

We take the observer state z instead of x in the feedback law and put

$$u = Fz + v.$$

Then we obtain the closed loop system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} A & BF \\ -LC & A + LC + BF \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} B \\ B \end{bmatrix} v.$$

If we can make this system asymptotically stable, the combination of feedback and observer outlined above will work nicely.

Definition 7.8 We call (F, L) a **compensator** for (A, B, C) if

$$K := \begin{bmatrix} A & BF \\ -LC & A + LC + BF \end{bmatrix}$$

is asymptotically stable.

Theorem 7.9 The matrix pair (F, L) is a compensator for (A, B, C) if and only if $A+BF$ and $A+LC$ are both asymptotically stable. Therefore, (A, B, C) possesses a compensator if and only if (A, B) is stabilizable and (A, C) is detectable. If we even have that (A, B) is controllable and (A, C) is observable, then for any monic polynomial p of degree $2n$, there exists (F, L) such that $\chi_K = p$.

Proof: Let

$$T = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}.$$

Then

$$T^{-1}KT = \begin{bmatrix} A + BF & BF \\ 0 & A + LC \end{bmatrix}$$

and thus $\chi_K = \chi_{A+BF} \cdot \chi_{A+LC}$ and $\mathrm{spec}(K) = \mathrm{spec}(A + BF) \cup \mathrm{spec}(A + LC)$. $\square$

Chapter 8

Transfer matrices

An input-output representation has the form

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad \text{or} \quad P(\sigma)y = Q(\sigma)u$$

where $P \in \mathbb{R}[s]^{p\times p}$ is non-singular, and $Q \in \mathbb{R}[s]^{p\times m}$. The rational matrix

$$H := P^{-1}Q \in \mathbb{R}(s)^{p\times m}$$

is called **transfer matrix** (or: transfer function) of the input-output representation.

Lemma 8.1 The transfer matrix of a state space system is

$$H = C(sI - A)^{-1}B + D. \tag{8.1}$$

Proof: We need to eliminate the state from

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx + Du \end{array}$$

where $A \in \mathbb{R}^{n\times n}, \ldots, D \in \mathbb{R}^{p\times m}$ (the discrete case is analogous). Let U be a unimodular matrix such that

$$U \begin{bmatrix} sI - A \\ C \end{bmatrix} = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \tag{8.2}$$

where $R_1 \in \mathbb{R}[s]^{n\times n}$ is non-singular (see Corollary 2.4). Then

$$\exists x : \left\{ \begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx + Du \end{array} \right.$$

is equivalent to

$$\exists x : U(\tfrac{\mathrm{d}}{\mathrm{d}t}) \begin{bmatrix} sI - A \\ C \end{bmatrix} (\tfrac{\mathrm{d}}{\mathrm{d}t})x = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} (\tfrac{\mathrm{d}}{\mathrm{d}t})x = U(\tfrac{\mathrm{d}}{\mathrm{d}t}) \begin{bmatrix} B & 0 \\ -D & I \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix}.$$

According to the fundamental principle, this is also equivalent to

$$(U_3B - U_4D)(\tfrac{\mathrm{d}}{\mathrm{d}t})u + U_4(\tfrac{\mathrm{d}}{\mathrm{d}t})y = 0 \quad \text{where} \quad U = \begin{bmatrix} U_1 & U_2 \\ U_3 & U_4 \end{bmatrix}. \tag{8.3}$$

The matrix $[U_3, U_4]$ has full row rank p, and it follows from (8.2) that

$$U_3(sI - A) + U_4C = 0,$$

that is, $U_3 = -U_4C(sI - A)^{-1}$ which shows that the columns of U_3 are (rational) linear combinations of the columns of U_4. Therefore $p = \operatorname{rank}[U_3, U_4] = \operatorname{rank}(U_4)$. We conclude that $U_4 \in \mathbb{R}[s]^{p\times p}$ is non-singular. Thus (8.3) is an input-output representation with transfer function

$$H = -U_4^{-1}(U_3B - U_4D) = -U_4^{-1}U_3B + D.$$

Using $U_4^{-1}U_3 = -C(sI - A)^{-1}$, we have the desired result. □

8.1 Realization theory

By the previous lemma, it is easy to compute H if A, B, C, D are known. However, one often faces the inverse problem: Given H, find matrices A, B, C, D such that (8.1) holds. This is known as the **realization problem**. If (8.1) is satisfied, the matrix quadruple (A, B, C, D) is called a **realization** of H, and H is called **realizable** if it possesses a realization.

We first observe that any H according to (8.1) will be a **proper** rational matrix, that is, if we write $H = \frac{N}{d}$, where $N \in \mathbb{R}[s]^{p\times m}$ is a polynomial matrix, and $0 \neq d \in \mathbb{R}[s]$ is a scalar polynomial, then

$$\deg(N_{ij}) \leq \deg(d) \quad \text{for all } i, j. \tag{8.4}$$

This follows from $(sI - A)^{-1} = \frac{\operatorname{adj}(sI-A)}{\det(sI-A)}$ using Cramer's rule. In (8.4), strict inequality holds for all i, j if and only if $D = 0$. In this case, one says that H is **strictly proper**. It turns out that properness is not only necessary but also sufficient for realizability.

Theorem 8.2 A rational matrix is realizable if and only if it is proper.

Proof: Let $H \in \mathbb{R}(s)^{p\times m}$ be proper, then we can write $H = D + H_1$ with $D \in \mathbb{R}^{p\times m}$ and H_1 strictly proper. Thus $H_1 = \frac{N}{d}$ with

$$d = s^\nu + d_{\nu-1}s^{\nu-1} + \ldots + d_1 s + d_0 \quad \text{and} \quad N = N_{\nu-1}s^{\nu-1} + \ldots + N_1 s + N_0$$

for some $\nu = \deg(d)$, $d_i \in \mathbb{R}$, $N_i \in \mathbb{R}^{p\times m}$. Put $n = \nu m$ and

$$A = \begin{bmatrix} 0 & I & & \\ \vdots & & \ddots & \\ 0 & & & I \\ -d_0 I & -d_1 I & \cdots & -d_{\nu-1} I \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I \end{bmatrix} \quad C = \begin{bmatrix} N_0 & \cdots & N_{\nu-1} \end{bmatrix}$$

then (A, B, C, D) is a realization of H. To see this, note that

$$(sI - A)\begin{bmatrix} I \\ sI \\ \vdots \\ s^{\nu-1}I \end{bmatrix} = \begin{bmatrix} sI & -I & & & \\ 0 & sI & -I & & \\ \vdots & & \ddots & \ddots & \\ 0 & & & sI & -I \\ d_0 I & d_1 I & \cdots & \cdots & sI + d_{\nu-1} I \end{bmatrix} \begin{bmatrix} I \\ sI \\ \vdots \\ s^{\nu-1}I \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ dI \end{bmatrix}$$

and hence

$$\begin{bmatrix} I \\ sI \\ \vdots \\ s^{\nu-1}I \end{bmatrix} = (sI - A)^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ dI \end{bmatrix} = (sI - A)^{-1} B d.$$

Pre-multiplying this by C, we obtain

$$C \begin{bmatrix} I \\ sI \\ \vdots \\ s^{\nu-1}I \end{bmatrix} = N = C(sI - A)^{-1} B d$$

which yields the desired result, after division by d. □

Thus, any proper rational matrix H is realizable. Let (A, B, C, D) be a realization of H, with $A \in \mathbb{R}^{n\times n}$. We call the number n the **size** of the realization. Of course, it is desirable to have small realizations. We say that a realization of H is **minimal** if there exists no realization of H with a smaller size. The subsequent lemma gives an important relation between two realizations of a transfer function.

Lemma 8.3 If (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are two realizations of the same transfer matrix, then $D = \tilde{D}$ and

$$CA^iB = \tilde{C}\tilde{A}^i\tilde{B}$$

for all $i \in \mathbb{N}$.

Proof: If $H = C(sI - A)^{-1}B + D$, then

$$\lim_{s \to \infty} H(s) = D.$$

This shows that we must have $D = \tilde{D}$. Moreover, we can expand $H - D$ into a Laurent series

$$H - D = C(sI - A)^{-1}B = \sum_{i=0}^{\infty} CA^i B s^{-i-1}$$

and this is convergent on $|s| > \rho(A)$, where $\rho(A)$ is the spectral radius of A. Therefore, by comparing coefficients, $CA^iB = \tilde{C}\tilde{A}^i\tilde{B}$ for all i. □

Now we can give a sufficient condition for minimality which will soon turn out to be also necessary.

Lemma 8.4 Let (A, B, C, D) be such that (A, B) is controllable and (A, C) is observable. Then (A, B, C, D) is a minimal realization of $H = C(sI - A)^{-1}B + D$.

Remark 8.5 The proof of Lemma 8.4 uses **Sylvester's inequality**: If O is a real matrix with n columns, and K is a real matrix with n rows, then

$$\operatorname{rank}(OK) \geq \operatorname{rank}(O) + \operatorname{rank}(K) - n.$$

Proof: Suppose that (A, B, C, D) is a realization of H, with size n, in which (A, B) is controllable and (A, C) is observable. Let $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ be another realization of H, with size $\tilde{n}$. We need to prove that $n \leq \tilde{n}$. Define

$$O = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \qquad K = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

and

$$\tilde{O} = \begin{bmatrix} \tilde{C} \\ \tilde{C}\tilde{A} \\ \vdots \\ \tilde{C}\tilde{A}^{n-1} \end{bmatrix} \qquad \tilde{K} = \begin{bmatrix} \tilde{B} & \tilde{A}\tilde{B} & \cdots & \tilde{A}^{n-1}\tilde{B} \end{bmatrix}.$$

Note that whereas O, K are precisely the Kalman matrices associated to (A, B) and (A, C), this is not true for $\tilde{O}, \tilde{K}$ (we have n instead of $\tilde{n}$ in the highest power of A). Then, because $CA^iB = \tilde{C}\tilde{A}^i\tilde{B}$ for all i,

$$OK = \tilde{O}\tilde{K}.$$

We have

$$\operatorname{rank}(K) \geq \operatorname{rank}(OK) \geq \operatorname{rank}(O) + \operatorname{rank}(K) - n.$$

By assumption, K and O both have rank n. Therefore,

$$n = \operatorname{rank}(OK) = \operatorname{rank}(\tilde{O}\tilde{K}) \leq \operatorname{rank}(\tilde{O}) \leq \tilde{n}$$

as desired. □

Theorem 8.6 (Reduction to minimality) Let (A, B, C, D) be a realization of H.

1. Consider a Kalman controllability decomposition

$$T^{-1}AT = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} \quad T^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \quad CT = \begin{bmatrix} C_1 & C_2 \end{bmatrix}.$$

   Then $H = C(sI - A)^{-1}B + D = C_1(sI - A_1)^{-1}B_1 + D$, that is, (A_1, B_1, C_1, D) is another realization of H, with size $r = \operatorname{rank}(K)$, where K is the Kalman controllability matrix.

2. Consider a Kalman observability decomposition

$$T^{-1}AT = \begin{bmatrix} A_1 & 0 \\ A_2 & A_3 \end{bmatrix} \quad T^{-1}B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad CT = \begin{bmatrix} C_1 & 0 \end{bmatrix}.$$

   Then $H = C(sI - A)^{-1}B + D = C_1(sI - A_1)^{-1}B_1 + D$, that is, (A_1, B_1, C_1, D) is another realization of H, with size $r = \operatorname{rank}(O)$, where O is the Kalman observability matrix.

3. If the two reduction steps are done successively, one ends up with a minimal realization of H.

Proof: One can easily check that a similarity transform does not change the transfer function. Therefore, assume that a Kalman controllability decomposition has already been performed. Then

$$\begin{aligned} H &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} sI - A_1 & -A_2 \\ 0 & sI - A_3 \end{bmatrix}^{-1} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} + D \\ &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} (sI - A_1)^{-1} & * \\ 0 & (sI - A_3)^{-1} \end{bmatrix} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} + D \\ &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} (sI - A_1)^{-1} B_1 \\ 0 \end{bmatrix} + D \ = \ C_1 (sI - A_1)^{-1} B_1 + D. \end{aligned}$$

The second statement is analogous. Recall that after a Kalman controllability decomposition, the matrix pair (A_1, B_1) is controllable. Now if one performs a Kalman observability decomposition with the already reduced system (A_1, B_1, C_1, D), then one obtains

$$T_1^{-1} A_1 T_1 = \begin{bmatrix} A_{11} & 0 \\ A_{12} & A_{13} \end{bmatrix} \quad T_1^{-1} B_1 = \begin{bmatrix} B_{11} \\ B_{12} \end{bmatrix} \quad C_1 T_1 = \begin{bmatrix} C_{11} & 0 \end{bmatrix}$$

in which (A_{11}, C_{11}) is observable. We only need to convince ourselves that the controllability of (A_1, B_1) implies the controllability of (A_{11}, B_{11}) (i.e., the controllability established in the first reduction step is not destroyed by the second reduction step in which we achieve observability). Therefore, the resulting realization $(A_{11}, B_{11}, C_{11}, D)$ is both controllable and observable, and hence minimal according to Lemma 8.4. □

In particular, this theorem shows that if in a realization (A, B) is not controllable or (A, C) is not observable, then the realization can be reduced in size. Moreover, this can be done constructively, using a Kalman decomposition. In other words, a minimal realization will always be both controllable and observable. Combining this result with Lemma 8.4, we obtain the following theorem as a summary.

Theorem 8.7 The matrix quadruple (A, B, C, D) is a minimal realization of $H = C(sI - A)^{-1}B + D$ if and only if (A, B) is controllable and (A, C) is observable.

The next theorem says that minimal realizations are essentially unique (up to similarity transforms).

Theorem 8.8 Any two minimal realizations of a transfer matrix are similar, that is, if (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are two minimal realizations of H, then there exists a non-singular matrix T such that

$$\tilde{A} = T^{-1} A T, \quad \tilde{B} = T^{-1} B, \quad \tilde{C} = C T, \quad \tilde{D} = D.$$

Proof: Since both realizations are minimal, their size must be the same, that is, $n = \tilde{n}$. Moreover, $D = \tilde{D}$, and $CA^iB = \tilde{C}\tilde{A}^i\tilde{B}$ for all i. Let $O, K, \tilde{O}, \tilde{K}$ be the observability and controllability matrices of the two realizations. Then $OK = \tilde{O}\tilde{K}$ and

$$OAK = \tilde{O}\tilde{A}\tilde{K}, \quad OB = \tilde{O}\tilde{B}, \quad CK = \tilde{C}\tilde{K}.$$

By assumption, $\tilde{K}$ has full row rank, and $\tilde{O}$ has full column rank. Therefore, there exist matrices L, N such that

$$\tilde{K}L = I \quad \text{and} \quad N\tilde{O} = I.$$

Put $T := KL \in \mathbb{R}^{n\times n}$, then $T^{-1} = NO$, because

$$NOKL = N\tilde{O}\tilde{K}L = I.$$

We have

$$\begin{aligned} T^{-1}AT &= NOAKL = N\tilde{O}\tilde{A}\tilde{K}L = \tilde{A} \\ T^{-1}B &= NOB = N\tilde{O}\tilde{B} = \tilde{B} \\ CT &= CKL = \tilde{C}\tilde{K}L = \tilde{C} \end{aligned}$$

which completes the proof. □

Starting with an arbitrary realization of H, one can determine the size of a minimal realization of H by successively computing two Kalman decompositions, as outlined in Theorem 8.6. However, there is also a direct way to determine the size of a minimal realization. This will be discussed in the next section.

8.2 Matrix fraction descriptions

Let $H \in \mathbb{R}(s)^{p\times m}$ be given. If $H = P^{-1}Q$ for some $Q \in \mathbb{R}[s]^{p\times m}$, $P \in \mathbb{R}[s]^{p\times p}$ with $\det(P) \neq 0$, we call (P, Q) a **left factorization** (or: left matrix fraction description) of H.

Similarly, if $H = QP^{-1}$ for some polynomial matrices $Q \in \mathbb{R}[s]^{p\times m}$, $P \in \mathbb{R}[s]^{m\times m}$ with $\det(P) \neq 0$, we call (Q, P) a **right factorization** of H.

For example, we have already used the representation $H = \frac{N}{d}$ several times. In other words, (dI_p, N) is a left and (N, dI_m) is a right factorization of H.

In the scalar case, it is desirable to write a rational function $h \in \mathbb{R}(s)$ as the ratio ("fraction") of two coprime polynomials. We wish to do the same with polynomial matrices.

We say that a left factorization (P, Q) is **left coprime** if the matrix $\begin{bmatrix} P & Q \end{bmatrix}$ is left irreducible (see Theorem 4.26). Similarly, a right factorization (Q, P) is called **right coprime** if the matrix $\begin{bmatrix} Q \\ P \end{bmatrix}$ is right irreducible, which means, by definition, that its transpose is left irreducible.

Lemma 8.9 1. Let (P, Q) be a left coprime factorization of H. The degree of the determinant of P is independent of the specific choice of the coprime factorization and therefore

$$d(H) := \deg\det(P) \tag{8.5}$$

is well-defined. If (P, Q) is an arbitrary (not necessarily coprime) left factorization of H, then

$$d(H) \leq \deg\det(P).$$

2. Let (Q, P) be a right coprime factorization of H. The degree of the determinant of P is independent of the specific choice of the coprime factorization and moreover,

$$d(H) = \deg\det(P)$$

where $d(H)$ is the number defined in (8.5). If (Q, P) is an arbitrary (not necessarily coprime) right factorization of H, then

$$d(H) \leq \deg\det(P).$$

Remark 8.10 The degree of the determinant of the "denominator" matrix P is as small as possible if the factorization is coprime. This generalizes the well known fact that the degree of d is minimal if we write a scalar rational function $h = \frac{n}{d} \in \mathbb{R}(s)$ as the ratio of two coprime polynomials. The choice of the coprime factorization does not influence this minimal degree, it does not even matter whether we take a right or left matrix fraction description.

In the proof, we use the formula

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix}$$

which holds provided that A is invertible. Thus

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A)\det(D - CA^{-1}B).$$

This shows that the block matrix is invertible if and only if $D - CA^{-1}B$ is invertible, and we have

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} * & * \\ * & (D - CA^{-1}B)^{-1} \end{bmatrix}.\end{aligned}$$

The matrix $D - CA^{-1}B$ is called **Schur complement** of A.

Proof: 1. Let (P, Q) and (P_1, Q_1) be two left factorizations of H, that is,

$$H = P^{-1}Q = P_1^{-1}Q_1$$

which implies that $Q_1 = UQ$ and $P_1 = UP$, where $U := P_1P^{-1}$. We show that U is polynomial if (P, Q) is left coprime; and even unimodular if both factorizations are coprime. If (P, Q) is left coprime, then there exist, according to Theorem 4.26, polynomial matrices R, S such that

$$PR + QS = I.$$

Then

$$P^{-1} = R + P^{-1}QS = R + HS = R + P_1^{-1}Q_1S$$

and thus $U = P_1P^{-1} = P_1R + Q_1S$. This shows that U is polynomial. Then

$$\det(P_1) = \det(UP) = \det(U)\det(P)$$

means that $\det(P)$ divides $\det(P_1)$, in particular,

$$\deg\det(P_1) \geq \deg\det(P).$$

Similarly, if also (P_1, Q_1) is coprime, there exist polynomial matrices R_1, S_1 such that $P_1R_1 + Q_1S_1 = I$ and then

$$P_1^{-1} = R_1 + P_1^{-1}Q_1S_1 = R_1 + HS_1 = R_1 + P^{-1}QS_1$$

and thus $U^{-1} = PP_1^{-1} = PR_1 + QS_1$ which shows that U^{-1} is polynomial and hence, U is unimodular. Thus we have

$$\det(P_1) = \det(U)\det(P)$$

where $\det(U)$ is a non-zero constant, and hence

$$\deg\det(P_1) = \deg\det(P).$$

2. In view of part 1, it suffices to show that if (P, Q) is a left coprime, and (Q_1, P_1) is a right coprime factorization of H, then $\deg\det(P) = \deg\det(P_1)$. We have $H = P^{-1}Q = Q_1P_1^{-1}$ and thus

$$QP_1 = PQ_1.$$

There exist polynomial matrices R, S, R_1, S_1 such that

$$PR + QS = I \quad \text{and} \quad R_1P_1 + S_1Q_1 = I.$$

Thus

$$\begin{bmatrix} P & -Q \\ S_1 & R_1 \end{bmatrix} \begin{bmatrix} R & Q_1 \\ -S & P_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}$$

where $X = S_1R - R_1S$. Post-multiplying this by

$$\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}$$

yields

$$\begin{bmatrix} P & -Q \\ S_1 & R_1 \end{bmatrix} \begin{bmatrix} \tilde{R} & Q_1 \\ -\tilde{S} & P_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

This shows that the matrices are unimodular, and inverse to each other,

$$\begin{bmatrix} P & -Q \\ S_1 & R_1 \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{R} & Q_1 \\ -\tilde{S} & P_1 \end{bmatrix}.$$

Thus P_1^{-1} is the Schur complement of P and

$$\det(P)\det(P_1^{-1}) = \det \begin{bmatrix} P & -Q \\ S_1 & R_1 \end{bmatrix} = c$$

where $0 \neq c \in \mathbb{R}$. Thus $\det(P) = c\det(P_1)$, in particular, $\deg\det(P) = \deg\det(P_1)$. □

It turns out below that the integer $d(H)$ from (8.5) equals the size of a minimal realization of a proper rational matrix H. For the proof, we need the concept of a row-reduced polynomial matrix.

Definition 8.11 The degree of a non-zero polynomial row vector is defined to be the highest power of s appearing in it with a non-zero coefficient. For a non-singular polynomial matrix $P \in \mathbb{R}[s]^{p\times p}$, let $\delta_i(P)$ be the degree of the i-th row of P, for $i = 1, \ldots, p$. Then P has a unique representation

$$P = SP_{hr} + L$$

where $S = \operatorname{diag}(s^{\delta_1(P)}, \ldots, s^{\delta_p(P)})$, $P_{hr} \in \mathbb{R}^{p\times p}$, and $L \in \mathbb{R}[s]^{p\times p}$ is such that $\delta_i(L) < \delta_i(P)$. One calls P_{hr} the **highest row coefficient matrix**. If P_{hr} is non-singular, we say that P is **row-reduced** (or row-proper).

Lemma 8.12 Let $P \in \mathbb{R}[s]^{p\times p}$ be non-singular.

1. We have
$$\sum_{i=1}^{p} \delta_i(P) \geq \deg\det(P).$$
The matrix P is row-reduced if and only if
$$\sum_{i=1}^{p} \delta_i(P) = \deg\det(P).$$

2. If $H = P^{-1}Q$ is strictly proper, then $\delta_i(Q) < \delta_i(P)$ for $i = 1, \dots, p$; if H is proper, then $\delta_i(Q) \leq \delta_i(P)$ for all i. The converse is also true, provided that P is row-proper.

3. For every non-singular $P \in \mathbb{R}[s]^{p\times p}$ there exists a unimodular matrix $U \in \mathbb{R}[s]^{p\times p}$ such that UP is row-proper.

Proof:

1. Rewrite $P = SP_{hr} + L$ as $P_{hr} = S^{-1}P - S^{-1}L$. Consider the limit as $s \to \infty$. Then we have, since $\lim_{s\to\infty} S^{-1}L = 0$,
$$P_{hr} = \lim_{s\to\infty} S^{-1}P$$
and, putting $\delta(P) := \sum_{i=1}^{p} \delta_i(P)$,
$$\det(P_{hr}) = \lim_{s\to\infty} \frac{\det(P)}{s^{\delta(P)}}.$$
This shows that $\deg\det(P) \leq \delta(P)$ and
$$\det(P_{hr}) = 0 \quad \Leftrightarrow \quad \deg\det(P) < \delta(P).$$

2. Let $H = P^{-1}Q$, then
$$Q_{ij} = \sum_{k=1}^{p} P_{ik}H_{kj}$$
and
$$\frac{Q_{ij}}{s^{\delta_i(P)}} = \sum_{k=1}^{p} \frac{P_{ik}}{s^{\delta_i(P)}} H_{kj}.$$

Consider again the limit as $s \to \infty$. If H is strictly proper, the right hand side tends to zero, and hence all the powers of s appearing in Q_{ij} must be strictly less than $\delta_i(P)$. Since this holds for all j, we obtain $\delta_i(Q) < \delta_i(P)$.

For the converse, write $P = SP_{hr} + L$, and assume that P_{hr} is invertible. Then

$$P^{-1}Q = (SP_{hr} + L)^{-1}Q = (I + P_{hr}^{-1}S^{-1}L)^{-1}P_{hr}^{-1}S^{-1}Q$$

and thus, since $\lim_{s\to\infty} S^{-1}L = 0$ and $\lim_{s\to\infty} S^{-1}Q = 0$, we have

$$\lim_{s\to\infty} P^{-1}Q = 0,$$

that is, $P^{-1}Q$ is strictly proper. The argument for "proper" is similar.

3. If P is row-proper, we are finished. Therefore, assume otherwise, that is, let $\det(P_{hr}) = 0$. We show that there exists a unimodular matrix U such that

$$\deg\det(P) = \deg\det(UP) \leq \sum_{j=1}^{p} \delta_j(UP) < \sum_{j=1}^{p} \delta_j(P).$$

Iteratively, this yields the result. We write $\delta_j := \delta_j(P)$ for simplicity. Since $\det(P_{hr}) = 0$, there exists $0 \neq \alpha = (\alpha_1, \dots, \alpha_p) \in \mathbb{R}^{1\times p}$ with

$$\sum_{j=1}^{p} \alpha_j P_{hr}^{(j)} = 0$$

where $P_{hr}^{(j)}$ denotes the j-th row of P_{hr}. Among the j with $\alpha_j \neq 0$, select j^* with

$$\delta_{j^*} \geq \delta_j \quad \text{for all } j \text{ with } \alpha_j \neq 0.$$

Without loss of generality, let $\alpha_{j^*} = 1$. Then we have

$$P_{hr}^{(j^*)} + \sum_{j\neq j^*} \alpha_j P_{hr}^{(j)} = 0.$$

Now perform the elementary operation: $P^{(j^*)}$ plus $\sum_{j\neq j^*} \alpha_j s^{\delta_{j^*}-\delta_j} P^{(j)}$. The new matrix $P' = UP$ satisfies $\delta_{j^*}(P') < \delta_{j^*}(P)$ and $\delta_j(P') = \delta_j(P)$ for all $j \neq j^*$. This establishes the claim. □

Theorem 8.13 Let H be a proper rational matrix. The size of a minimal realization of H is given by the integer $d(H)$ from (8.5).

Proof: There is no loss of generality in assuming that H is strictly proper. Let $(A, B, C, 0)$ be a realization of H, that is,

$$H = C(sI_n - A)^{-1}B.$$

Define $G := (sI_n - A)^{-1}B$. Then $(sI_n - A, B)$ is a left factorization of G. Thus

$$\deg\det(sI_n - A) = n \geq d(G).$$

On the other hand, if (Q, P) is a right factorization of G, that is, $G = QP^{-1}$, then $H = CG = CQP^{-1}$, that is, (CQ, P) is a right factorization of H. We conclude that

$$d(G) \geq d(H).$$

This shows that $n \geq d(H)$, that is, the size of any realization of H must be at least $d(H)$. Conversely, we show that a strictly proper H possesses a realization of size $d(H)$. Let $H = P^{-1}Q$ be a left coprime factorization. Without loss of generality, let P be row-proper. Now consider the i-th row of P and Q, and denote them by P_i and Q_i, respectively. According to Theorem 2.24, there exist matrices K_i, L_{ij}, M_i, N_{ij} $(j = 1, 2)$ such that

$$P_i(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q_i(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad \Leftrightarrow \quad \exists x : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x_i &=& K_i x_i + L_{i1}u + L_{i2}y \\ 0 &=& M_i x_i + N_{i1}u + N_{i2}y. \end{array} \right.$$

Here K_i can be chosen to be a $\delta_i(P) \times \delta_i(P)$ matrix. Combining these representations via

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \quad K = \begin{bmatrix} K_1 & & \\ & \ddots & \\ & & K_p \end{bmatrix} \quad L_j = \begin{bmatrix} L_{1j} \\ \vdots \\ L_{pj} \end{bmatrix}$$

$$M = \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_p \end{bmatrix} \quad N_j = \begin{bmatrix} N_{1j} \\ \vdots \\ N_{pj} \end{bmatrix}$$

we obtain

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad \Leftrightarrow \quad \exists x : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x &=& Kx + L_1u + L_2y \\ 0 &=& Mx + N_1u + N_2y \end{array} \right.$$

where the size of K is $\sum_{i=1}^{p} \delta_i(P) = \deg\det(P) = d(H)$. Moreover, $N_1 = 0$ (because $P^{-1}Q$ is strictly proper) and $N_2 = P_{hr}$, which is invertible. Thus we can find, as in Section 2.7, matrices A, B, C, D, with A of the same size as K, such that

$$P(\tfrac{\mathrm{d}}{\mathrm{d}t})y = Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u \quad \Leftrightarrow \quad \exists x : \left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}x &=& Ax + Bu \\ y &=& Cx. \end{array} \right.$$

Then $H = P^{-1}Q = C(sI - A)^{-1}B$ which shows that H has a realization of size $d(H)$. □

Remark 8.14 In the final step of the proof, we have used the fact that equivalent representations have the same transfer function. This statement has not been proven. Alternatively, one may give a direct proof of

$$H = P^{-1}Q = C(sI - A)^{-1}B$$

using the special form of A, B, C that results from Section 2.7. The details are omitted.

Theorem 8.15 (McMillan form) For each rational matrix $H \in \mathbb{R}(s)^{p\times m}$ there exist unimodular matrices $U \in \mathbb{R}[s]^{p\times p}$ and $V \in \mathbb{R}[s]^{m\times m}$ such that

$$UHV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{8.6}$$

where $D \in \mathbb{R}(s)^{r\times r}$ is a diagonal matrix

$$D = \begin{bmatrix} \frac{\gamma_1}{\delta_1} & & \\ & \ddots & \\ & & \frac{\gamma_r}{\delta_r} \end{bmatrix}$$

with polynomials $\gamma_i, \delta_i \neq 0$ such that each pair (γ_i, δ_i) is coprime and $\gamma_1 | \ldots | \gamma_r$ and $\delta_r | \ldots | \delta_1$. Clearly, $r = \mathrm{rank}(H)$. The matrix on the right hand side of (8.6) is called **McMillan form** of H and the integer

$$n := \sum_{i=1}^{r} \deg(\delta_i)$$

is called **McMillan-degree** of H.

Theorem 8.16 The size of a minimal realization of H equals the McMillan-degree of H.

Proof: We show that $d(H)$ from (8.5) coincides with the McMillan-degree of H. Let

$$UHV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

be the McMillan form of H, with $D = \operatorname{diag}(\frac{\gamma_1}{\delta_1}, \ldots, \frac{\gamma_r}{\delta_r})$. Define

$$\Gamma = \begin{bmatrix} \operatorname{diag}(\gamma_1, \ldots, \gamma_r) & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \Delta = \begin{bmatrix} \operatorname{diag}(\delta_1, \ldots, \delta_r) & 0 \\ 0 & I \end{bmatrix}.$$

Then $UHV = \Gamma\Delta^{-1}$, that is, (Γ, Δ) is a right factorization of UHV. Since each pair (γ_i, δ_i) is coprime, it is even a right coprime factorization of UHV. Then $(U^{-1}\Gamma, V\Delta)$ is a right coprime factorization of H. Therefore

$$d(H) = \deg\det(V\Delta) = \deg\det(\Delta) = \deg\prod_{i=1}^{r}\delta_i = \sum_{i=1}^{r}\deg(\delta_i)$$

which is precisely the McMillan-degree of H. □

8.3 Poles

A complex number λ is called a **pole** of a rational matrix H if it is a pole of one of its entries. Equivalently, the poles of H are the zeros of the polynomials δ_i in the McMillan form of H. Still equivalently, they are the zeros of $\det(P)$, where P is the denominator matrix in a coprime factorization of H.

Theorem 8.17 Let (A, B, C, D) be a realization of H. Any pole of H is an eigenvalue of A. Conversely, an eigenvalue of A which is not a pole of H must be either an uncontrollable mode of (A, B) or an unobservable mode of (A, C). In particular, if (A, B, C, D) is a minimal realization of H, then the eigenvalues of A are precisely the poles of H.

Remark 8.18 This theorem shows that there is a close relation between the eigenvalues of A and the poles of $H = C(sI - A)^{-1}B + D$. This is the reason why one speaks of "pole shifting" in Section 5.3, although "eigenvalue shifting" would probably be more appropriate.

Proof: Without loss of generality, let H be strictly proper. Since

$$H = C(sI - A)^{-1}B = \frac{C\mathrm{adj}(sI - A)B}{\det(sI - A)}$$

any pole of H must be a zero of $\det(sI - A)$, that is, it must be an eigenvalue of A. Now let λ be an eigenvalue of A which is not a pole of H. Assume that λ is not an unobservable mode, that is,

$$\operatorname{rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n. \tag{8.7}$$

We need to show that λ is an uncontrollable mode. Let $G = (sI - A)^{-1}B = QP^{-1}$, where (Q, P) is right coprime. Then we have

$$BP = (sI - A)Q \tag{8.8}$$

and $H = CG = CQP^{-1}$, with

$$\operatorname{rank} \begin{bmatrix} CQ(\lambda) \\ P(\lambda) \end{bmatrix} = m. \tag{8.9}$$

To see this, let v be such that $CQ(\lambda)v = 0$ and $P(\lambda)v = 0$. We need to show that this implies $v = 0$. From (8.8),

$$0 = BP(\lambda)v = (\lambda I - A)Q(\lambda)v$$

and thus

$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} Q(\lambda)v = 0$$

which implies that $Q(\lambda)v = 0$ because of (8.7). But then both $P(\lambda)v = 0$ and $Q(\lambda)v = 0$, which implies that $v = 0$ due to the right coprimeness of (Q, P).

Now we must have $\det(P(\lambda)) \neq 0$. If conversely, $\det(P(\lambda)) = 0$, then there would exist a $v \neq 0$ such that $P(\lambda)v = 0$. Since λ is not a pole of H, the complex matrix $H(\lambda)$ is well-defined, and thus

$$0 = H(\lambda)P(\lambda)v = CQ(\lambda)v$$

and this would be a contradiction to (8.9).

Since λ is an eigenvalue of A, there exists $0 \neq z \in \mathbb{C}^{1\times n}$ such that $z(\lambda I - A) = 0$. Then (8.8) implies $zBP(\lambda) = 0$ and hence, since $P(\lambda)$ is non-singular, we must have $zB = 0$. Then

$$z \begin{bmatrix} \lambda I - A & B \end{bmatrix} = 0$$

which shows that

$$\operatorname{rank} \begin{bmatrix} \lambda I - A & B \end{bmatrix} < n$$

that is, λ is an uncontrollable mode of (A, B). $\square$

Appendix A

Background material from distribution theory

Let $\mathcal{D}$ denote the set of all smooth functions $\varphi : \mathbb{R} \to \mathbb{R}$ which have compact support. Recall that

$$\operatorname{supp}(\varphi) = \operatorname{cl}\{t \in \mathbb{R} \mid \varphi(t) \neq 0\},$$

where $\operatorname{cl}(\cdot)$ denotes the closure with respect to the standard topology of $\mathbb{R}$ (where the term "compact" coincides with "closed and bounded"). The elements of $\mathcal{D}$ are called **test functions**. A **distribution** D is a linear, continuous (in a sense not to be specified here) functional defined on $\mathcal{D}$, that is, it assigns to each test function φ a real number $D(\varphi)$ such that $D(\lambda_1\varphi_1 + \lambda_2\varphi_2) = \lambda_1 D(\varphi_1) + \lambda_2 D(\varphi_2)$ for all $\lambda_1, \lambda_2 \in \mathbb{R}$ and all $\varphi_1, \varphi_2 \in \mathcal{D}$. The set of all distributions is denoted by $\mathcal{D}'$. (We restrict to real-valued test functions and distributions, although the complex case is analogous.)

A function $f : \mathbb{R} \to \mathbb{R}$ is called **locally integrable** if

$$\int_I |f(t)| \mathrm{d}t$$

exists for any compact interval $I \subset \mathbb{R}$. Each locally integrable function f defines a distribution D_f by

$$D_f(\varphi) = \int_{-\infty}^{\infty} f(t)\varphi(t) \mathrm{d}t$$

for all $\varphi \in \mathcal{D}$. We say that f **generates** the distribution D_f. The distributions D_f, where f is locally integrable, are called **regular**, and one often identifies f

with D_f, that is, we can interpret L^1_{loc}, the set of locally integrable functions, as a subset of $\mathcal{D}'$. Note that every continuous function is locally integrable, and hence, it can be seen as an element of $\mathcal{D}'$. However, there are also many non-continuous locally integrable functions, such as, for example, the Heaviside function. An important non-regular distribution is the Dirac delta distribution δ defined by

$$\delta(\varphi) = \varphi(0)$$

for all $\varphi \in \mathcal{D}$.

Linear combinations of distributions are defined by

$$(\lambda_1 D_1 + \lambda_2 D_2)(\varphi) = \lambda_1 D_1(\varphi) + \lambda_2 D_2(\varphi)$$

for $\lambda_1, \lambda_2 \in \mathbb{R}$, and thus $\mathcal{D}'$ becomes a real vector space. One can multiply a distribution D by a smooth function $a : \mathbb{R} \to \mathbb{R}$ via

$$(aD)(\varphi) = D(a\varphi)$$

for all $\varphi \in \mathcal{D}$. Note that we need the smoothness of a to guarantee that $a\varphi$ is again a test function.

The derivative of a distribution is defined by

$$\dot{D}(\varphi) = -D(\dot{\varphi}) \tag{A.1}$$

for all $\varphi \in \mathcal{D}$. For well-definedness, note that the derivative of a test function is again a test function. This definition is motivated by the law of partial integration. Let D_f be a regular distribution, and assume that f is continuously differentiable. Then the distribution $D_{\dot{f}}$ is well-defined, and we certainly want $\dot{D}_f = D_{\dot{f}}$ and thus we put

$$\dot{D}_f(\varphi) = D_{\dot{f}}(\varphi) = \int_{-\infty}^{\infty} \dot{f}(t)\varphi(t)\mathrm{d}t = f\varphi\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(t)\dot{\varphi}(t)\mathrm{d}t = -D_f(\dot{\varphi}).$$

For example, the derivative of the Heaviside function h is the Dirac delta distribution, because

$$\dot{h}(\varphi) = -h(\dot{\varphi}) = -\int_{-\infty}^{\infty} h(t)\dot{\varphi}(t)\mathrm{d}t = -\int_{0}^{\infty} \dot{\varphi}(t)\mathrm{d}t = -\varphi\Big|_{0}^{\infty} = \varphi(0) = \delta(\varphi)$$

for all $\varphi \in \mathcal{D}$. Note that according to (A.1), distributions can be differentiated arbitrarily often: by repeated application of (A.1), we get $D^{(k)}(\varphi) = (-1)^k D(\varphi^{(k)})$ for all $k \in \mathbb{N}$. The distributional derivative (A.1) provides a generalization of the classical concept of differentiability. For instance, any continuous function can be differentiated in the distributional sense, but of course, not necessarily in the classical sense.

We would like to use a similar trick for integration. However, a primitive function of a test function $\varphi \in \mathcal{D}$, for instance,

$$\psi(t) = \int_{-\infty}^{t} \varphi(\tau)\mathrm{d}\tau$$

is *not* a test function, in general. This is because

$$\psi(\infty) = \lim_{t\to\infty} \psi(t) = \int_{-\infty}^{\infty} \varphi(\tau)\mathrm{d}\tau$$

is non-zero in general, and thus, ψ may not have compact support.

Therefore, the set $\mathcal{D}_0$ is introduced as the set of all test functions which have a primitive function that is again a test function. More precisely, let $\mathcal{D}_0 \subset \mathcal{D}$ denote the space of all test functions φ_0 which satisfy the following equivalent conditions:

1. There exists $\psi \in \mathcal{D}$ such that $\varphi_0 = \dot{\psi}$.
2. $\int_{-\infty}^{\infty} \varphi_0(t)\mathrm{d}t = 0$.

In that case, a primitive function $\psi \in \mathcal{D}$ of φ_0 is given by $\psi(t) = \int_{-\infty}^{t} \varphi_0(\tau)\mathrm{d}\tau$, in fact, this is the only primitive in $\mathcal{D}$, because adding a non-zero constant will destroy the compact support property.

Now let α be a fixed test function with $\int_{-\infty}^{\infty} \alpha(t)\mathrm{d}t = 1$. Then for any $\varphi \in \mathcal{D}$, let $\lambda = \int_{-\infty}^{\infty} \varphi(t)\mathrm{d}t$ and set

$$\varphi_0 := \varphi - \lambda\alpha.$$

Then $\varphi_0 \in \mathcal{D}_0$.

Lemma A.1 The equation $\dot{\xi} = 0$ has no distributional solutions apart from the classical solutions, i.e., the constant functions $\xi(t) = \xi_0$ for all t, where ξ_0 is a real number (more precisely, the distributional solutions are the distributions generated by constant functions.)

Proof: Let $\xi \in \mathcal{D}'$ be such that $\dot{\xi} = 0$. Consider first a test function $\varphi \in \mathcal{D}_0$, then

$$\xi(\varphi) = \xi(\dot{\psi}) = -\dot{\xi}(\psi) = 0.$$

Now let $\varphi \in \mathcal{D}$ be arbitrary, and define $\varphi_0 = \varphi - \lambda\alpha$ as above. Then

$$0 = \xi(\varphi_0) = \xi(\varphi) - \lambda\xi(\alpha)$$

and thus

$$\xi(\varphi) = \lambda\xi(\alpha) = \int_{-\infty}^{\infty} c\varphi(t)\mathrm{d}t = D_c(\varphi) = c(\varphi),$$

where the constant c is given by $c = \xi(\alpha)$. Thus $\xi = c$. □

Lemma A.2 Let g be a distribution. Then there exists a distribution G with $\dot{G} = g$. We call G a primitive (distribution) of g.

Proof: Let $g \in \mathcal{D}'$ be given. For $\varphi \in \mathcal{D}$, we define $\varphi_0 \in \mathcal{D}_0$ as above and we set

$$\psi(t) = \int_{-\infty}^{t} \varphi_0(\tau)\mathrm{d}\tau.$$

Then $\psi \in \mathcal{D}$. In particular, if $\varphi = \dot{\phi}$ for some $\phi \in \mathcal{D}$, then $\varphi_0 = \varphi$ and $\psi = \phi$. We define

$$G(\varphi) = -g(\psi)$$

for all $\varphi \in \mathcal{D}$. This is linear and continuous, and for any $\phi \in \mathcal{D}$

$$\dot{G}(\phi) = -G(\dot{\phi}) = g(\phi)$$

as desired. □

According to Lemma A.1, primitives of distributions are unique up to additive constants. If $g = D_f$ is regular, then a primitive of D_f is given by D_F, where

$$F(t) = \int_0^t f(\tau)\mathrm{d}\tau$$

is a primitive function of f.

So far, we have only dealt with distributions in $\mathcal{D}' = \mathcal{D}'(\mathbb{R})$, corresponding to the time set $T = \mathbb{R}$. For $T = \mathbb{R}_+ = [0, \infty)$, one constructs $\mathcal{D}'(\mathbb{R}_+)$ as the subset of $\mathcal{D}'$ consisting of all distributions D which satisfy

$$D(\varphi) = 0$$

for all $\varphi \in \mathcal{D}$ whose support is in $(-\infty, 0)$.

Appendix B

Jordan form

For every matrix $A \in \mathbb{R}^{n\times n}$, there exists an invertible matrix $T \in \mathbb{C}^{n\times n}$ such that

$$T^{-1}AT = J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{bmatrix} \tag{B.1}$$

where each matrix $J_i \in \mathbb{C}^{n_i\times n_i}$ has the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}.$$

The matrix on the right hand side of (B.1) is called **Jordan form** of A. The complex numbers $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of A. The number of matrices J_i having a particular eigenvalue λ on their main diagonal coincides with the geometric multiplicity of that eigenvalue, and the sum of the sizes of these blocks is precisely the algebraic multiplicity of λ.

We have

$$A^t = TJ^tT^{-1} \quad \text{and} \quad e^{At} = Te^{Jt}T^{-1}.$$

Using

$$J^t = \begin{bmatrix} J_1^t & & \\ & \ddots & \\ & & J_k^t \end{bmatrix} \quad \text{and} \quad e^{Jt} = \begin{bmatrix} e^{J_1t} & & \\ & \ddots & \\ & & e^{J_kt} \end{bmatrix}$$

as well as

$$J_i^t = \begin{bmatrix} \lambda_i^t & \binom{t}{1}\lambda_i^{t-1} & \cdots & \binom{t}{n_i-1}\lambda_i^{t-n_i+1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \binom{t}{1}\lambda_i^{t-1} \\ & & & \lambda_i^t \end{bmatrix}$$

and

$$e^{J_i t} = \begin{bmatrix} e^{\lambda_i t} & te^{\lambda_i t} & \cdots & \frac{t^{n_i-1}}{(n_i-1)!}e^{\lambda_i t} \\ & \ddots & \ddots & \vdots \\ & & \ddots & te^{\lambda_i t} \\ & & & e^{\lambda_i t} \end{bmatrix}$$

we can see that the entries of A^t and e^{At} have the form

$$\sum_\lambda a_\lambda(t)\lambda^t \quad \text{and} \quad \sum_\lambda a_\lambda(t)e^{\lambda t}$$

respectively, where λ are the eigenvalues of A and a_λ are polynomials in t.

Strictly speaking, the formula for the discrete case only holds only if A is invertible. The reason is that the eigenvalue zero (which is present if and only if A is not invertible) plays a special role in the discrete case. Jordan blocks with zero on the main diagonal are nilpotent, and thus they contribute nothing to A^t for large enough t. Thus, if A is not invertible, the formula is still correct for t that are large enough. However, this is quite sufficient, e.g., for stability analysis, because this is anyhow concerned with the behavior of A^t for large t, and for $t \to \infty$.

If one restricts to real transformation matrices $T \in \mathbb{R}^{n\times n}$, the real Jordan form can be achieved. Note that since A is real, the non-real eigenvalues come in pairs of complex conjugate numbers. For an eigenvalue pair $a_i \pm \mathrm{i}b_i$, where $a_i, b_i \in \mathbb{R}$, the real Jordan blocks take the form

$$J_i = \begin{bmatrix} \Lambda_i & I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & I_2 \\ & & & \Lambda_i \end{bmatrix}, \quad \text{where } \Lambda_i = \begin{bmatrix} a_i & b_i \\ -b_i & a_i \end{bmatrix}.$$

Appendix C

Kronecker-Weierstraß form

Let $K, L \in \mathbb{R}^{n\times n}$ be matrices with $\det(sK - L) \neq 0$. Then there exist non-singular real matrices U, V such that

$$UKV = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \quad \text{and} \quad ULV = \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix}$$

where N is a nilpotent matrix.

Proof: Let $\alpha \in \mathbb{R}$ be such that $\det(\alpha K - L) \neq 0$. Define

$$\hat{K} = (\alpha K - L)^{-1} K \quad \text{and} \quad \hat{L} = (\alpha K - L)^{-1} L.$$

Then

$$\alpha \hat{K} - \hat{L} = I. \tag{C.1}$$

Using the real Jordan form, there exists a non-singular matrix T such that

$$T^{-1}\hat{K}T = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix}$$

where E_1 is non-singular and E_2 is nilpotent. Then $\alpha E_2 - I$ is non-singular (because zero is the only eigenvalue of a nilpotent matrix). Define

$$U = \begin{bmatrix} E_1^{-1} & 0 \\ 0 & (\alpha E_2 - I)^{-1} \end{bmatrix} T^{-1}(\alpha K - L)^{-1} \quad \text{and} \quad V = T.$$

These matrices are clearly non-singular and

$$UKV = \begin{bmatrix} I & 0 \\ 0 & (\alpha E_2 - I)^{-1}E_2 \end{bmatrix} \quad \text{and} \quad ULV = \begin{bmatrix} E_1^{-1} & 0 \\ 0 & (\alpha E_2 - I)^{-1} \end{bmatrix} T^{-1}\hat{L}T.$$

We set $N := (\alpha E_2 - I)^{-1} E_2$. Since $(\alpha E_2 - I)^{-1}$ and E_2 commute with each other,

$$N^k = (\alpha E_2 - I)^{-k} E_2^k$$

and thus N is nilpotent (since E_2 is nilpotent). On the other hand, (C.1) implies that

$$T^{-1} \hat{L} T = \alpha T^{-1} \hat{K} T - I = \begin{bmatrix} \alpha E_1 - I & 0 \\ 0 & \alpha E_2 - I \end{bmatrix}$$

and thus

$$ULV = \begin{bmatrix} E_1^{-1}(\alpha E_1 - I) & 0 \\ 0 & I \end{bmatrix}.$$

Thus we set $A := E_1^{-1}(\alpha E_1 - I)$ and we are finished. □

Note that the size of $A \in \mathbb{R}^{\nu \times \nu}$ is uniquely determined by K, L, because

$$\nu = \deg(\det(sK - L)).$$

Similarly, the nilpotency index of N is uniquely determined by K, L, because it is equal to the degree of the polynomial part of the rational matrix $(sK - L)^{-1}$ plus one.

Appendix D

Smith form

The ring $\mathcal{P} = \mathbb{R}[s]$ is a **Euclidean domain**. This means that for any $a, b \in \mathcal{P}$, $b \neq 0$, there exist $c, d \in \mathcal{P}$ with

$$a = bc + d \tag{D.1}$$

where either $d = 0$ or $\deg(d) < \deg(b)$. The representation (D.1) is obtained by **division with remainder**.

By an **elementary operation**, we mean one of the following matrix transformations:

- interchanging two rows/columns of a matrix;
- multiplying a row/column by a unit (that is, a non-zero constant);
- adding a multiple of one row/column to another row/column.

It is easy to see that these operations correspond to multiplication by unimodular matrices from the left/right.

Let $R \in \mathcal{P}^{p\times q}$ be a matrix. Then there exist unimodular matrices $U \in \mathcal{P}^{p\times p}$ and $V \in \mathcal{P}^{q\times q}$ such that

$$URV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{D.2}$$

where $D = \mathrm{diag}(d_1, \ldots, d_r)$ is a diagonal matrix, and $d_1 | d_2 | \ldots | d_r$. The matrix on the right hand side of (D.2) is called the **Smith form** of R. The non-zero polynomials $d_1, \ldots, d_r$ are uniquely determined by R (up to multiplication by a non-zero constant).

Proof: Without loss of generality, let $R \neq 0$. It is sufficient to show that by elementary operations, R can be brought into the form

$$R' = \begin{bmatrix} a & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & Q & \\ 0 & & & \end{bmatrix} \tag{D.3}$$

where a divides all entries of Q. Then one applies the same procedure to Q, and the result follows inductively.

Case 1: There exists i, j such that R_{ij} divides all entries of R. By a suitable interchange of rows and columns, this element can be brought into the (1,1) position of the matrix. Therefore without loss of generality, R_{11} divides all entries of R. Now perform the following elementary operations: for all $i \neq 1$, put ith row minus R_{i1}/R_{11} times 1st row; for all $j \neq 1$, put jth column minus R_{1j}/R_{11} times 1st column. Then we are finished.

Case 2: There is no i, j such that R_{ij} divides all entries of R. Let

$$\delta(R) := \min\{\deg(R_{ij}) \mid R_{ij} \neq 0\}.$$

Without loss of generality, $\deg(R_{11}) = \delta(R)$. We show that by elementary operations, we can transform R into R' with $\delta(R') < \delta(R)$. Then we obtain a strictly decreasing sequence $\delta(R) > \delta(R') > \delta(R'') > \ldots \geq 0$. After finitely many steps, we arrive at zero, i.e., we obtain a matrix which has a unit as an entry, and thus we are in Case 1.

Case 2a: R_{11} does not divide all R_{1j}, R_{i1}, say, it does not divide R_{1k}. By the Euclidean algorithm, we can write

$$R_{1k} = R_{11}c + d$$

where $d \neq 0$ and $\deg(d) < \deg(R_{11})$. Perform the elementary operation: kth column minus c times 1st column. Then the new matrix R' has d in the $(1, k)$ position and thus $\delta(R') < \delta(R)$ as desired.

Case 2b: R_{11} divides all R_{1j}, R_{i1}. Similarly as in Case 1, we can transform, by elementary operations, R into the form (D.3). If a divides all entries of Q, then we are finished. If there exists i, j such that a does not divide Q_{ij}, then we perform the elementary operation: 1st row plus $(i + 1)$st row. (Note that the $(i + 1)$st row of R' corresponds to the ith row of Q.) The new matrix has Q_{ij} in the $(1, j + 1)$ position and therefore we are in Case 2a. □

Appendix E

McMillan form

Let $H \in \mathbb{R}(s)^{p\times m}$ be a rational matrix. Then there exist unimodular polynomial matrices $U \in \mathbb{R}[s]^{p\times p}$ and $V \in \mathbb{R}[s]^{m\times m}$ such that

$$UHV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{E.1}$$

where

$$D = \begin{bmatrix} \frac{\gamma_1}{\delta_1} & & \\ & \ddots & \\ & & \frac{\gamma_r}{\delta_r} \end{bmatrix}$$

for some non-zero polynomials γ_i, δ_i, where each pair (γ_i, δ_i) is coprime and $\gamma_1|\gamma_2|\ldots|\gamma_r$ and $\delta_r|\ldots|\delta_2|\delta_1$. Clearly, $r = \operatorname{rank}(H)$. The right hand side of (E.1) is called **McMillan form** of H. The number

$$n := \sum_{i=1}^{r} \deg(\delta_i)$$

is called **McMillan-degree** of H. If $\lambda \in \mathbb{C}$ is a zero of δ_{i+1}, then it is also a zero of δ_i, because $\delta_{i+1}|\delta_i$. Therefore, the **poles** of H are precisely the zeros of δ_1.

Proof: Write $H = \frac{N}{d}$, where $N \in \mathbb{R}[s]^{p\times m}$ and $0 \neq d \in \mathbb{R}[s]$. Compute the Smith form of N, say

$$UNV = \begin{bmatrix} \tilde{D} & 0 \\ 0 & 0 \end{bmatrix} \tag{E.2}$$

where U, V are unimodular and $\tilde{D} = \operatorname{diag}(d_1, \ldots, d_r)$ for some non-zero polynomials d_i. We divide (E.2) by d and put $D := \frac{\tilde{D}}{d} = \operatorname{diag}(\frac{d_1}{d}, \ldots, \frac{d_r}{d})$. Let γ_i, δ_i be coprime polynomials with

$$\frac{d_i}{d} = \frac{\gamma_i}{\delta_i}.$$

Since $d_1 | d_2 | \ldots | d_r$, we have $d_{i+1} = d_i e_i$ for some polynomials e_i, where $i = 1, \ldots, r-1$. This implies

$$\gamma_{i+1}\delta_i = \gamma_i \delta_{i+1} e_i$$

for all i. Since δ_i and γ_i are coprime, γ_{i+1} must be a multiple of γ_i. Since δ_{i+1} and γ_{i+1} are coprime, δ_i must be a multiple of δ_{i+1}. $\square$

Appendix F

An optimal control problem

Consider $\dot{x} = Ax + Bu$, where (A, B) is controllable. Let $\varepsilon > 0$ and $\bar{x} \in \mathbb{R}^n$ be given. We wish to steer the system from state 0 to state $\bar{x}$ in time ε. Moreover, we would like to do this with the smallest possible amount of energy, that is,

$$E(u) = \int_0^\varepsilon \|u(\tau)\|^2 \mathrm{d}\tau \to \min!$$

Define

$$V(t, x) := x^T W(t)^{-1} x$$

where

$$W(t) = \int_0^t e^{A\tau} B B^T e^{A^T \tau} \mathrm{d}\tau$$

is the controllability Gramian. Let us look at the change of $V(t, x(t))$ along a trajectory x of our system. We have

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} V(t, x(t)) &= \tfrac{\mathrm{d}}{\mathrm{d}t} x(t)^T W(t)^{-1} x(t) \\ &= \dot{x}(t)^T W(t)^{-1} x(t) + x(t)^T (\tfrac{\mathrm{d}}{\mathrm{d}t} W(t)^{-1}) x(t) + x(t)^T W(t)^{-1} \dot{x}(t). \end{aligned}$$

Note that for any matrix-valued function W,

$$\tfrac{\mathrm{d}}{\mathrm{d}t} W^{-1} = -W^{-1} \dot{W} W^{-1}.$$

Moreover, we plug in $\dot{x} = Ax + Bu$ and we obtain (omitting the argument t wherever possible)

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} V(t, x) &= (Ax + Bu)^T W^{-1} x - x^T W^{-1} \dot{W} W^{-1} x + x^T W^{-1} (Ax + Bu) \\ &= x^T (A^T W^{-1} + W^{-1} A - W^{-1} \dot{W} W^{-1}) x + 2u^T B^T W^{-1} x. \end{aligned}$$

Consider the matrix

$$X(t) := A^T W(t)^{-1} + W(t)^{-1} A - W(t)^{-1}\dot{W}(t)W(t)^{-1}.$$

We have

$$\begin{aligned} W(t)X(t)W(t) &= W(t)A^T + AW(t) - \dot{W}(t) \qquad \text{(F.1)} \\ &= \int_0^t (e^{A\tau}BB^T e^{A^T\tau}A^T + Ae^{A\tau}BB^T e^{A^T\tau})\mathrm{d}\tau - \dot{W}(t) \\ &= e^{A\tau}BB^T e^{A^T\tau}\Big|_0^t - \dot{W}(t) \\ &= e^{At}BB^T e^{A^T t} - BB^T - \dot{W}(t). \end{aligned}$$

Noting that by the definition of W,

$$\dot{W}(t) = e^{At}BB^T e^{A^T t}$$

we obtain

$$W(t)X(t)W(t) = -BB^T \qquad \text{(F.2)}$$

and hence

$$X(t) = -W(t)^{-1}BB^T W(t)^{-1}.$$

We use this to rewrite our expression for $\frac{\mathrm{d}}{\mathrm{d}t}V(t,x)$ and obtain

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}V(t,x) &= -x^T W^{-1}BB^T W^{-1}x + 2u^T B^T W^{-1}x \\ &= -\|B^T W^{-1}x\|^2 + 2\langle u, B^T W^{-1}x\rangle \\ &= \|u\|^2 - \|u - B^T W^{-1}x\|^2. \end{aligned}$$

Let's integrate this from 0 to ε, exploiting that $x(0) = 0$ and $x(\varepsilon) = \bar{x}$. Then

$$V(\varepsilon,\bar{x}) - V(0,0) = \int_0^\varepsilon \|u(\tau)\|^2\mathrm{d}\tau - \int_0^\varepsilon \|u(\tau) - B^T W(\tau)^{-1}x(\tau)\|^2\mathrm{d}\tau$$

or

$$\bar{x}^T W(\varepsilon)^{-1}\bar{x} = E(u) - \int_0^\varepsilon \|u(\tau) - B^T W(\tau)^{-1}x(\tau)\|^2\mathrm{d}\tau \le E(u). \qquad \text{(F.3)}$$

This shows that

$$E(u) \ge E_{\min}(\varepsilon,\bar{x}) := \bar{x}^T W(\varepsilon)^{-1}\bar{x}.$$

Equality is achieved if and only if the integral in (F.3) vanishes, i.e., if

$$u(t) = B^T W(t)^{-1}x(t). \qquad \text{(F.4)}$$

Plugging that into $\dot{x} = Ax + Bu$, we get

$$\dot{x}(t) = (A + BB^T W(t)^{-1})x(t).$$

Since we know that $x(\varepsilon) = \bar{x}$, the solution of this linear time-varying ordinary differential equation is uniquely determined for all $t > 0$. I claim that this solution is

$$\xi(t) = W(t)e^{A^T(\varepsilon - t)}W(\varepsilon)^{-1}\bar{x}.$$

This can easily be checked: We have $\xi(\varepsilon) = \bar{x}$ and

$$\dot{\xi}(t) = (\dot{W}(t) - W(t)A^T)e^{A^T(\varepsilon - t)}W(\varepsilon)^{-1}\bar{x}.$$

Combining (F.1) with (F.2), we see that

$$\dot{W}(t) = W(t)A^T + AW(t) + BB^T.$$

This implies

$$\begin{aligned}
\dot{\xi}(t) &= (AW(t) + BB^T)e^{A^T(\varepsilon - t)}W(\varepsilon)^{-1}\bar{x} \\
&= (A + BB^T W(t)^{-1})W(t)e^{A^T(\varepsilon - t)}W(\varepsilon)^{-1}\bar{x} \\
&= (A + BB^T W(t)^{-1})\xi(t)
\end{aligned}$$

as desired. Thus $x = \xi$ is the optimal state trajectory. Then, according to (F.4),

$$u(t) = B^T W(t)^{-1}\xi(t) = B^T e^{A^T(\varepsilon - t)}W(\varepsilon)^{-1}\bar{x}$$

is the minimum energy control function that steers the system from 0 to $\bar{x}$ in time ε.

Now let $0 < \varepsilon < \delta$. Then

$$W(\varepsilon) < W(\delta)$$

(we write $P < Q$ if $Q - P$ is positive definite) which implies that

$$W(\varepsilon)^{-1} > W(\delta)^{-1}$$

and hence

$$E_{\min}(\varepsilon, \bar{x}) = \bar{x}^T W(\varepsilon)^{-1}\bar{x} > \bar{x}^T W(\delta)^{-1}\bar{x} = E_{\min}(\delta, \bar{x})$$

for all $\bar{x} \neq 0$. Thus one needs more energy for doing the transition from 0 to $\bar{x}$ in time ε than in time δ. This explains the trade-off between the speed of control on the one hand and the energy consumption of control on the other.

Bibliography

[1] D. Hinrichsen, A. J. Pritchard, *Mathematical Systems Theory I*, Springer, 2005.

[2] T. Kailath, *Linear Systems*, Prentice-Hall, 1980.

[3] H. W. Knobloch, H. Kwakernaak, *Lineare Kontrolltheorie*, Springer, 1985.

[4] J. W. Polderman, J. C. Willems, *Introduction to Mathematical Systems Theory*, Texts in Applied Mathematics 26, Springer, 1998.

[5] H. H. Rosenbrock, *State-space and Multivariable Theory*, Nelson, 1970.

[6] E. D. Sontag, *Mathematical Control Theory*, Texts in Applied Mathematics 6, Springer, 1990.

[7] W. M. Wonham, *Linear Multivariable Control*, Applications of Mathematics 10, Springer, 1979.