Linear systems

Version: 12 April 2024

Bart Besselink
University of Groningen

Contents

Chapter 1

Introduction

1.1 Linear systems with inputs and outputs

These notes study properties of the set of equations

$$\mathbf{\Sigma}: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{1.1}$$

with $\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t)$, which is referred to as a (linear) *system* $\mathbf{\Sigma}$.

The first equation in (1.1) is a differential equation, where t is the independent variable that generally represents *time*, and $x(t) \in \mathbb{R}^n$ is the dependent variable that is known as the *state*. We also say that the state evolves over the *state space* $\mathbb{R}^n$ and refer to the first equation as the *state equation*. This differential equation is influenced by a function u, which is typically referred to as the *input*. The input, which is assumed to take values in $\mathbb{R}^m$ (i.e., $u(t) \in \mathbb{R}^m$), is assumed to be given outside the system. Note that this makes the differential equation in (1.1) nonhomogeneous. Moreover, the matrix $A \in \mathbb{R}^{n\times n}$ is often referred to as the *system matrix*, whereas $B \in \mathbb{R}^{n\times m}$ is called the *input matrix*.

The second equation in (1.1) is an algebraic equation known as the *output equation*. Namely, the function y with $y(t) \in \mathbb{R}^p$ is called the *output* and generally represents physical quantities of interest or measurements that can be done on a system. The matrices $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ are known as the *output matrix* and *feedthrough matrix*, respectively. In applications, the feedthrough matrix is often zero.

A linear system $\mathbf{\Sigma}$ as in (1.1) is often depicted as a simple block diagram in Figure 1.1, clearly indicating that a system relates inputs and outputs. Even

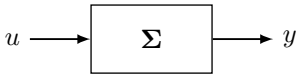


Figure 1.1: Representation of a linear system $\mathbf{\Sigma}$, indicating the input u and output y. The state x can be thought of as living "inside" the block, whereas the equations (1.1) provide the relation between the input and output functions.

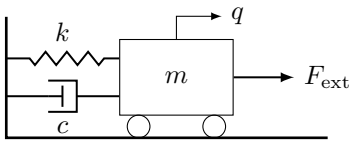


Figure 1.2: Mass-spring-damper system as discussed in Example 1.1.

though such representation seems trivial, it will turn out to be very helpful when considering the interconnection of systems, e.g., for control.

We now turn to a simple example of a linear system.

Example 1.1. Consider the simple mechanical system depicted in Figure 1.2, known as a mass-spring-damper system. If q denotes the position of the mass with mass $m > 0$, it follows from Newton's second law that

$$m\ddot{q}(t) = F(t), \tag{1.2}$$

where F denotes the sum of all forces acting on the mass. In this example, these are given by the spring force $-kq$ with $k > 0$, damper force $-c\dot{q}$, and an external force F_{ext}. This leads to the differential equation

$$m\ddot{q}(t) + c\dot{q}(t) + kq(t) = F_{\text{ext}}(t). \tag{1.3}$$

To write this as a linear system of the form (1.1), define the state x as the vector

$$x = \begin{bmatrix} q \\ \dot{q} \end{bmatrix} \tag{1.4}$$

and the input u as $u = F_{\text{ext}}$. Then, it is easily seen that the dynamics (1.3) can be written as

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u(t). \tag{1.5}$$

If the position q of the mass can be measured, we can model this by selecting the output y to be $y = q$, which can be expressed in terms of the state (1.4) as

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(t). \tag{1.6}$$

Hence, combining (1.5) and (1.6), it is clear that the mass-spring-damper system is a linear system $\mathbf{\Sigma}$ of the form (1.1) with

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D = 0. \tag{1.7}$$

We point out that the state equation is given by the dynamics, whereas the output equation is a choice reflecting the availability of measurements or the physical quantities of interest. ◇

The above example shows that models of mechanical systems can sometimes be expressed as linear systems. The following example is from electrical engineering.

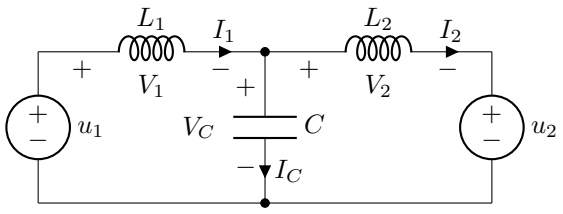


Figure 1.3: A simple electrical circuit with two inductors (with inductances L_1 and L_2), one capacitor (with capacitance C), and two voltage sources, as discussed in Example 1.2.

Example 1.2. Consider the electrical circuit in Figure 1.3, comprising two inductors, one capacitor, and two voltage sources. To obtain a model of this circuit, these components will be considered independently.

First, the dynamics of the inductors is given by

$$\phi_i(t) = L_i I_i(t), \qquad \dot{\phi}_i(t) = V_i(t), \tag{1.8}$$

where ϕ_i, $i = 1, 2$ is the magnetic flux of the inductor. Moreover, V_i and I_i denote the voltage across and current through the inductors, respectively, whereas $L_i > 0$ is the inductance. Similarly, V_C and I_C represent the voltage across and current through the capacitor, which are related by the dynamics

$$q_C(t) = CV_C(t), \qquad \dot{q}_C(t) = I_C(t). \tag{1.9}$$

Here, q_C is the charge of the capacitor with capacitance $C > 0$.

Next, the interconnection of the components can be specified through Kirchhoff's laws. Namely, Kirchhoff's voltage law states that the voltages across each component in a loop should sum to zero. Omitting the arguments t for simplicity of notation, this leads to the equations

$$V_1 + V_C - u_1 = 0, \qquad V_2 + u_2 - V_C = 0, \tag{1.10}$$

for the left and right loop in Figure 1.3, respectively. Note that Kirchhoff's voltage law for the outer loop is implied by the two inner loops. In addition, Kirchhoff's current law states that, for each node of the network, the total incoming and outgoing current should be equal, leading to

$$I_1 = I_2 + I_C. \tag{1.11}$$

As time derivatives appear for the magnetic fluxes ϕ_1 and ϕ_2 and the charge q_C, we define the state x as $x = [\phi_1 \ \phi_2 \ q_C]^{\mathrm{T}}$. Then, combining (1.8)–(1.11), we obtain

$$\begin{aligned}
\dot{x}_1 &= \dot{\phi}_1 = V_1 = -V_C + u_1 &&= -\frac{1}{C} q_C + u_1, \\
\dot{x}_2 &= \dot{\phi}_2 = V_2 = V_C - u_2 &&= \frac{1}{C} q_C - u_2, \\
\dot{x}_3 &= \dot{q}_C = I_C = I_1 - I_2 &&= \frac{1}{L_1}\phi_1 - \frac{1}{L_2}\phi_2,
\end{aligned} \tag{1.12}$$

where it is remarked that the voltages u_1 and u_2 of the two voltage sources are regarded as inputs. Rewriting (1.12), after denoting $u = [\, u_1 \;\; u_2 \,]^{\mathrm{T}}$, leads to

$$\dot{x}(t) = \begin{bmatrix} 0 & 0 & -C^{-1} \\ 0 & 0 & C^{-1} \\ L_1^{-1} & -L_2^{-1} & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} u(t), \tag{1.13}$$

which is a linear system of the form (1.1) without output equation. ◇

Remark 1.1. As a more general perspective on linear systems (1.1), we can allow the state x to evolve on a finite-dimensional vector space $\mathcal{X}$, i.e., $x(t) \in \mathcal{X}$. Similarly, the input and output can be defined to take values on finite-dimensional vector spaces $\mathcal{U}$ and $\mathcal{Y}$, respectively. In these notes, we will generally identify $\mathcal{X}$ with $\mathbb{R}^n$, $\mathcal{U}$ with $\mathbb{R}^m$, and $\mathcal{Y}$ with $\mathbb{R}^p$. Nonetheless, in this more general geometric setting, it is crucial to regard A, B, C, and D as linear maps rather than matrices. Our notation will not distinguish between linear maps and a matrix corresponding to such linear map (for a given basis). ◁

1.2 Nonlinear systems and linearization

In these notes, we will study linear systems of the form (1.1). However, there is a large class of physical systems that can not be described by linear differential equations. In this section, we consider such nonlinear systems and show that they can be *approximated* locally by linear systems.

Consider nonlinear systems of the form

$$\boldsymbol{\Sigma}_{\mathrm{nl}} : \begin{cases} \dot{x}(t) &= f\big(x(t), u(t)\big), \\ y(t) &= h\big(x(t), u(t)\big), \end{cases} \tag{1.14}$$

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}^m$, and output $y(t) \in \mathbb{R}^p$ as in the linear case. We assume that the functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are differentiable.

In order to discuss the approximation of $\boldsymbol{\Sigma}_{\mathrm{nl}}$ in (1.14) by a linear system of the form (1.1), the notion of an equilibrium is introduced as follows.

Definition 1.1 (Equilibrium)**.** *Consider the system $\boldsymbol{\Sigma}_{\mathrm{nl}}$ for constant input $u(t) = \bar{u}$ for all $t \in \mathbb{R}$. Then, $\bar{x} \in \mathbb{R}^n$ is called an equilibrium for $\bar{u}$ if*

$$f(\bar{x}, \bar{u}) = 0. \tag{1.15}$$

Note that an equilibrium point $\bar{x}$ for $\bar{u}$ has the property that the time derivative $f(\bar{x}, \bar{u})$ is zero, which means that $x(t) = \bar{x}$ is a *constant* solution for the differential equation in (1.14) given the constant input $u(t) = \bar{u}$. Note that $\bar{u}$ is chosen a priori and we will sometimes refer to such constant input as a *nominal* input. Moreover, for a given $\bar{u}$, the system $\boldsymbol{\Sigma}_{\mathrm{nl}}$ might have multiple equilibria. We also sometimes write $(\bar{x}, \bar{u})$ for an equilbrium, to explicitly give the corresponding input value $\bar{u}$.

Whereas an equilibrium point $\bar{x} \in \mathbb{R}^n$ is an element of the state space, we also denote the corresponding output value $\bar{y} \in \mathbb{R}^p$ as

$$\bar{y} = h(\bar{x}, \bar{u}). \tag{1.16}$$

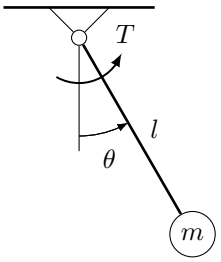


Figure 1.4: A pendulum with mass m and length l.

Example 1.3. A simple example of a nonlinear system is given by the dynamics of a pendulum, see Figure 1.4. After summing moments around the pivot, the dynamics of the pendulum can be found as

$$ml^2\ddot{\theta} + mgl\sin(\theta) = T, \tag{1.17}$$

where θ gives the angular position of the pendulum. Moreover, m and l denote the mass and length, respectively, whereas g is the gravitational constant. The external torque T acts as an input to the system.

After defining the state x as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix}, \tag{1.18}$$

and the input $u = T$, the dynamics (1.17) can be written as

$$\dot{x}(t) = f\big(x(t), u(t)\big) = \begin{bmatrix} x_2 \\ -\frac{g}{l}\sin(x_1) + \frac{1}{ml^2}u \end{bmatrix}, \tag{1.19}$$

which indeed represents nonlinear dynamics. When the output y is chosen as the angular position θ, the corresponding output equation simply becomes

$$y(t) = h\big(x(t), u(t)\big) = x_1. \tag{1.20}$$

such that (1.19) and (1.20) together form a nonlinear system as in (1.14).

Take $\bar{u} = 0$ to be the nominal input. Then, solving (1.15) for f given in (1.19) shows that any $\bar{x} = [\,\bar{x}_1\ \bar{x}_2\,]^{\mathrm{T}}$ satisfying $\bar{x}_2 = 0$ and $\sin(\bar{x}_1) = 0$ is an equilibrium point. This leads to two physical equilibria given as

$$\bar{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \bar{x} = \begin{bmatrix} \pi \\ 0 \end{bmatrix}, \tag{1.21}$$

corresponding the down and up position of the pendulum in Figure 1.4, respectively. ◇

After selecting an equilibrium $(\bar{x}, \bar{u})$, we can consider the dynamics of $\boldsymbol{\Sigma}_{\mathrm{nl}}$ in coordinates that measure the deviation from this equilibrium point. Specifically, introduce $\tilde{x}$ as the deviation from the equilibrium point, i.e.,

$$\tilde{x}(t) = x(t) - \bar{x}. \tag{1.22}$$

After defining deviations from the nominal input $\bar{u}$ and equilibrium output $\bar{y}$ similarly as

$$\tilde{u}(t) = u(t) - \bar{u}, \qquad \tilde{y}(t) = y(t) - \bar{y}, \tag{1.23}$$

we can see that the dynamics in terms of $\tilde{x}$ reads

$$\dot{\tilde{x}}(t) = \dot{x}(t) - 0 = f\big(\bar{x} + \tilde{x}(t), \bar{u} + \tilde{u}(t)\big). \tag{1.24}$$

Here, we have used (1.22) and the fact that $\bar{x} \in \mathbb{R}^n$, i.e., is constant. At this point, (1.24) is just a restatement of the dynamics (1.14), although expressed in terms of $\tilde{x}$. However, if we assume that both $\tilde{x}$ and $\tilde{u}$ are *small*, we can use a Taylor series approximation of f around $(\tilde{x}, \tilde{u}) = (0, 0)$ (or, stated differently, around $(x, u) = (\bar{x}, \bar{u})$) to obtain

$$f\big(\bar{x} + \tilde{x}, \bar{u} + \tilde{u}\big) \approx f(\bar{x}, \bar{u}) + \frac{\partial f}{\partial x}(\bar{x}, \bar{u})\tilde{x} + \frac{\partial f}{\partial u}(\bar{x}, \bar{u})\tilde{u}. \tag{1.25}$$

In the above, we used the notation

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \qquad \frac{\partial f}{\partial u} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_1}{\partial u_2} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \frac{\partial f_2}{\partial u_1} & \frac{\partial f_2}{\partial u_2} & \cdots & \frac{\partial f_2}{\partial u_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \frac{\partial f_n}{\partial u_2} & \cdots & \frac{\partial f_n}{\partial u_m} \end{bmatrix} \tag{1.26}$$

to denote the matrix of partial derivatives of f with respect to x and u, respectively.

Then, the use of (1.25) leads to an *approximation* of the dynamics (1.24) as

$$\dot{\tilde{x}}(t) = \frac{\partial f}{\partial x}(\bar{x}, \bar{u})\tilde{x}(t) + \frac{\partial f}{\partial u}(\bar{x}, \bar{u})\tilde{u}(t), \tag{1.27}$$

by using the definition of an equilibrium point in (1.15).

Similarly, we can use (1.23) and (1.16) to express the deviation from the equilibrium output as

$$\tilde{y}(t) = h\big(\bar{x} + \tilde{x}(t), \bar{u} + \tilde{u}(t)\big) - h(\bar{x}, \bar{u}), \tag{1.28}$$

after which a Taylor series approximation of h around $(\tilde{x}, \tilde{u}) = (0, 0)$ leads to the approximation

$$\tilde{y}(t) = \frac{\partial h}{\partial x}(\bar{x}, \bar{u})\tilde{x}(t) + \frac{\partial h}{\partial u}(\bar{x}, \bar{u})\tilde{u}(t). \tag{1.29}$$

We note that the matrices of partial derivatives in both (1.27) and (1.29) are evaluated at the equilibrium $\bar{x}$ corresponding to $\bar{u}$ and are thus constant matrices. Hence, the equations (1.27) and (1.29) describe a linear system $\boldsymbol{\Sigma}$ as in (1.1) with state $\tilde{x}$, input $\tilde{u}$, and output $\tilde{y}$.

This motivates the following definition.

Definition 1.2 (Linearization)**.** *Let $(\bar{x}, \bar{u})$ be an equilibrium of the nonlinear system $\boldsymbol{\Sigma}_{\mathrm{nl}}$ as in (1.14). Then, the linear system*

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + B\tilde{u}(t), \\ \tilde{y}(t) &= C\tilde{x}(t) + D\tilde{u}(t), \end{aligned} \tag{1.30}$$

with state $\tilde{x}(t) \in \mathbb{R}^n$*, input* $\tilde{u}(t) \in \mathbb{R}^m$*, output* $\tilde{y}(t) \in \mathbb{R}^p$*, and*

$$A = \frac{\partial f}{\partial x}(\bar{x}, \bar{u}), \quad B = \frac{\partial f}{\partial u}(\bar{x}, \bar{u}), \quad C = \frac{\partial h}{\partial x}(\bar{x}, \bar{u}), \quad D = \frac{\partial h}{\partial u}(\bar{x}, \bar{u}), \tag{1.31}$$

is called the linearization of (1.14) around the equilibrium $(\bar{x}, \bar{u})$*.*

It is important to note that the linearization (1.30) is based on the Taylor series approximations of f (in (1.25)) and h, which can only be expected to be accurate for small $\tilde{x}$ and $\tilde{u}$. Consequently, the linearized system is only a good approximation when the deviations from the equilibrium point are small.

We return to the dynamics of the pendulum in Example 1.3.

Example 1.4. Consider the equilibrium $\bar{x} = 0$ for $\bar{u} = 0$, corresponding to the down position of the pendulum. Then, a direct computation gives

$$\frac{\partial f}{\partial x}(x, u) = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l}\cos(x_1) & 0 \end{bmatrix}, \qquad \frac{\partial f}{\partial u}(x, u) = \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix}, \tag{1.32}$$

after which evaluation at $\bar{x} = 0$ yields

$$A = \frac{\partial f}{\partial x}(\bar{x}, \bar{u}) = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l} & 0 \end{bmatrix}, \qquad B = \frac{\partial f}{\partial u}(\bar{x}, \bar{u}) = \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix}. \tag{1.33}$$

Thus, the linearization of (1.19)-(1.20) around $(\bar{x}, \bar{u}) = (0, 0)$ is given by

$$\begin{aligned} \dot{\tilde{x}}(t) &= \begin{bmatrix} 0 & 1 \\ -\frac{g}{l} & 0 \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} \tilde{u}(t), \\ \tilde{y}(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \tilde{x}(t). \end{aligned} \tag{1.34}$$

Similarly, we can easily see that the linearization of (1.19)-(1.20) around the equilibrium $(\bar{x}, \bar{u}) = ([0 \ \pi]^{\mathrm{T}}, 0)$, corresponding to the up position, reads

$$\begin{aligned} \dot{\tilde{x}}(t) &= \begin{bmatrix} 0 & 1 \\ \frac{g}{l} & 0 \end{bmatrix} \tilde{x}(t) + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} \tilde{u}(t), \\ \tilde{y}(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \tilde{x}(t). \end{aligned} \tag{1.35}$$

It is important to observe that the linearized dynamics depends on the equilibrium point around which the linearization is performed. Moreover, the meaning of the state $\tilde{x}$ is different for the linearized systems (1.34) and (1.35), as it clear after recalling that $\tilde{x}$ is defined as the deviation from the equilibrium point, see (1.22). ◇

Finally, we give an example of a satellite in geostationary orbit.

Example 1.5. Consider the motion of a communications satellite in the equator plane as in Figure 1.5. Taking the center of the earth as the origin, we use polar coordinates and denote the radial and angular position of the satellite by r and θ, respectively. In these coordinates, the dynamics is given by the nonlinear differential equations

$$\begin{aligned} \ddot{r} &= r\dot{\theta}^2 - \frac{GM}{r^2} + \frac{F_r}{m}, \\ \ddot{\theta} &= -2\frac{\dot{r}\dot{\theta}}{r} + \frac{F_\theta}{mr}, \end{aligned} \tag{1.36}$$

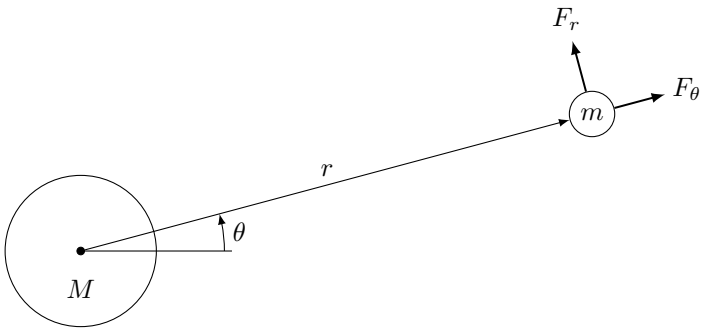


Figure 1.5: A satellite with mass m under the gravitational field of the earth, with mass M.

where we have omitted the arguments t. In (1.36), m and M denote the mass of the satellite and earth, respectively, whereas G is the gravitational constant. Moreover, F_r and F_θ represent the forces that can be exerted on the satellite using thrusters.

The satellite can be used to relay signals across the earth when it has a fixed position in the sky with respect to the earth. Hence, the desired radial position r^{ref} should be constant, whereas the desired angular position θ^{ref} should match the earth's angular velocity Ω, i.e.,

$$r^{\text{ref}}(t) = R_0, \qquad \theta^{\text{ref}}(t) = \theta_0 + \Omega t, \tag{1.37}$$

for some R_0 and θ_0. Moreover, to minimize energy use, it is desired that this reference is achieved without using the thrusters, namely when $F_r^{\text{ref}}(t) = 0$ and $F_\theta^{\text{ref}}(t) = 0$.

We note that θ_0 can be chosen arbitrarily, but that R_0 needs to be determined. However, as the desired positions need to satisfy the dynamics (1.36), it follows from the first equation that

$$0 = r^{\text{ref}}(\dot{\theta}^{\text{ref}})^2 - \frac{GM}{(r^{\text{ref}})^2} = R_0\Omega^2 - \frac{GM}{R_0^2}, \tag{1.38}$$

leading to

$$R_0 = \sqrt[3]{\frac{GM}{\Omega^2}}. \tag{1.39}$$

For analysis, it will be convenient to express the dynamics of the satellite in terms of deviations from the desired position. To this end, introduce the state

$$\begin{aligned} x_1 &= r - r^{\text{ref}} &&= r - R_0, \\ x_2 &= \dot{r} &&= \dot{r}, \\ x_3 &= \theta - \theta^{\text{ref}} &&= \theta - (\theta_0 + \Omega t), \\ x_4 &= \dot{\theta} &&= \dot{\theta} - \Omega, \end{aligned} \tag{1.40}$$

and note that x_1 and x_3 represent the deviations from the desired radial and angular position. The coordinates x_2 and x_4 give the corresponding (radial and

angular) velocities. We will regard the thruster forces as inputs to the satellite system, such that

$$u_1 = F_r, \qquad u_2 = F_\theta. \tag{1.41}$$

In terms of the state (1.40) and input (1.41), the dynamics (1.36) can be written in first-order form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} x_2 \\ (x_1+R_0)(x_4+\Omega)^2 - \frac{GM}{(x_1+R_0)^2} + \frac{u_1}{m} \\ x_4 \\ -2\frac{x_2(x_4+\Omega)}{x_1+R_0} + \frac{u_2}{m(x_1+R_0)} \end{bmatrix}. \tag{1.42}$$

As x_3 represents the deviation from the desired angular position, it can be measured by an observer on earth. Therefore, we choose it as an output, i.e., $y = x_3$. It is clear that the dynamics (1.42) is of the standard form (1.14), where

$$f(x,u) = \begin{bmatrix} x_2 \\ (x_1+R_0)(x_4+\Omega)^2 - \frac{GM}{(x_1+R_0)^2} + \frac{u_1}{m} \\ x_4 \\ -2\frac{x_2(x_4+\Omega)}{x_1+R_0} + \frac{u_2}{m(x_1+R_0)} \end{bmatrix}, \qquad h(x,u) = x_3. \tag{1.43}$$

and where x and u are vectors collecting the states (1.40) and inputs (1.41), respectively.

After recalling that the state x collects the deviations from the desired geostationary orbit and u gives the deviations from the nominal thruster force, it is clear that $(\bar{x}, \bar{u}) = (0, 0)$ is an equilibrium of (1.42). Then, linearization around $(\bar{x}, \bar{u}) = (0, 0)$ leads to the linear system (1.30) with matrices

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\Omega^2 & 0 & 0 & 2\Omega R_0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2\Omega}{R_0} & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ \frac{1}{m} & 0 \\ 0 & 0 \\ 0 & \frac{1}{mR_0} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}, \quad D = 0. \tag{1.44}$$

Here, we emphasize that the matrix B has two columns as the satellite was assumed to have two thrusters that act as control inputs. ◇

1.3 Exercises

Exercise 1.1. Consider the electrical circuit in Figure 1.6. Using the approach of Example 1.2, write the dynamics of the circuit as a linear system. Here, take the voltage of the voltage source u as an input, and the voltage across the capacitor V_C as an output. Use Ohm's law $V_R = RI_R$ to describe the resistor.

Exercise 1.2. Consider the so-called Van der Pol oscillator

$$\ddot{z}(t) - \mu\big(1 - z^2(t)\big)\dot{z}(t) + z(t) = 0$$

where $\mu > 0$.

a. Write the system in the form $\dot{x}(t) = f(x(t))$ by taking $x_1(t) = z(t)$ and $x_2(t) = \dot{z}(t)$.

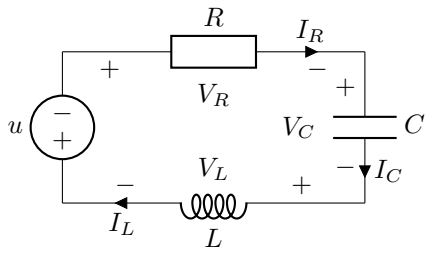


Figure 1.6: A simple electrical circuit with an inductor, capacitor, resistor, and voltage source.

b. Show that $x_1(t) = x_2(t) = 0$ for all t is a solution of $\dot{x} = f(x)$.

c. Determine the linearized system around the point $\bar{x} = [\,0\ 0\,]^{\mathrm{T}}$.

Exercise 1.3. Consider the nonlinear system

$$\begin{aligned}
\dot{x}_1(t) &= -x_1^6(t) - x_2(t),\\
\dot{x}_2(t) &= x_1(t) + u(t),\\
y_1(t) &= x_2^2(t),\\
y_2(t) &= x_1(t).
\end{aligned}$$

a. Show that $x_1(t) = -1$, $x_2(t) = -1$ and $u(t) = 1$ for all t is a solution of the system, i.e., $\bar{x} = [\,-1\ -1\,]^{\mathrm{T}}$ is an equilibrium for $\bar{u} = 1$.

b. Determine the linearized system around this solution.

Exercise 1.4. Consider the nonlinear system

$$\begin{aligned}
\dot{x}_1(t) &= x_2(t)\big(x_2(t) - 1\big) + u_1(t)\cos u_2(t),\\
\dot{x}_2(t) &= u_1(t)\sin u_2(t),\\
y(t) &= \big(x_1(t) + u_2(t)\big)^2 - e^{x_2(t)}.
\end{aligned}$$

a. Let $\bar{u}_1 = 0$ and $\bar{u}_2 = 1$. Show that $\bar{x} = [\,1\ 0\,]^{\mathrm{T}}$ is a corresponding equilibrium point.

b. Determine the linearized system around this equilibrium.

Exercise 1.5. Consider the sattelite in geostationary orbit of Example 1.5.

a. Consider the dynamics (1.42) and verify that $\bar{x} = 0$ is an equilibrium for $\bar{u} = 0$. Is $\bar{x} = 0$ the unique equilibrium for $\bar{u} = 0$?

b. Verify that the linearization of the nonlinear dynamics (1.42) around the equilibrium $(\bar{x}, \bar{u}) = (0, 0)$ is given by (1.30) with matrices (1.44).

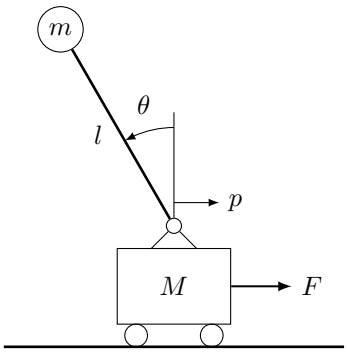


Figure 1.7: A pendulum on a cart.

Exercise 1.6. Consider the pendulum on a cart system as depicted in Figure 1.7. The differential equation relating the position of the cart p and angular position of the pendulum θ is given by

$$\begin{bmatrix} M+m & -ml\cos\theta \\ -ml\cos\theta & ml^2 \end{bmatrix} \begin{bmatrix} \ddot{p} \\ \ddot{\theta} \end{bmatrix} + \begin{bmatrix} c_p\dot{p}(t) + ml\sin(\theta)\dot{\theta}^2 \\ c_\theta\dot{\theta} - mgl\sin\theta \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \tag{1.45}$$

where the arguments t are omitted. Here, M and m are the masses of the cart and pendulum, respectively, whereas l is the length of the pendulum. The gravitational constant is denoted by g, whereas c_p and c_θ denote damping coefficients. The external force on the cart is represented by F.

a. Define the state x and input u as

$$x = \begin{bmatrix} p \\ \dot{p} \\ \theta \\ \dot{\theta} \end{bmatrix}, \qquad u = F, \tag{1.46}$$

respectively. Write the dynamics (1.45) as a system of first-order nonlinear differential equations as in (1.14). This form is referred to as the *state-space* form. For simplicity, use the values $m = 1$, $M = 5$, $l = 1$, $c_p = 1$, $c_\theta = 1$.

b. Show that $(\bar{x}, \bar{u}) = (0, 0)$ is an equilibrium point and linearize the system (obtained in *a.*) around it.

Chapter 2

Solutions of linear systems

This chapter studies the solutions of linear systems. Therefore, in Section 2.1, we first consider systems without inputs, leading to so-called homogeneous linear systems. It will be shown that the matrix exponential plays an important role in describing these solutions and the computation of the matrix exponential will be discussed in detail in Section 2.2. Finally, systems with inputs, also known as nonhomogeneous linear systems, are considered in Section 2.3.

2.1 Homogeneous linear systems

In this section, we consider the linear differential equation

$$\dot{x}(t) = Ax(t), \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ is a matrix. This differential equation is *linear* as the right-hand side Ax is a linear function of x. Moreover, it is *homogeneous* as the right-hand side is zero for $x = 0$ and *time-invariant* as the matrix A is independent of t (we also say that the equation has constant coefficients). We note that (2.1) is a special case of the differential equation $\dot{x} = f(x)$ (for $f(x) = Ax$) as discussed in Appendix A.

Specifically, we would like to obtain a solution to the *initial value problem*[1]

$$\dot{x}(t) = Ax(t), \qquad x(t_0) = x_0, \tag{2.2}$$

i.e., we aim to find a differentiable function $x : J \to \mathbb{R}^n$ (for some interval $J \subset \mathbb{R}$ such that $t_0 \in J$) that satisfies (2.2). Such solution (we will soon show that it exists) will be denoted by $x(\,\cdot\,; t_0, x_0)$. For scalar linear differential equations, i.e., $n = 1$ and $A = a$, it is well-known (see also Example A.1) that

$$x(t; t_0, x_0) = x_0 e^{a(t-t_0)}. \tag{2.3}$$

is the unique solution.

In order to obtain a solution for the general case (2.2), recall that the initial value problem can equivalently be stated as finding a solution to the integral

[1] A precise statement of the initial value problem is given in Problem A.2 in Appendix A.

equation

$$x(t) = x_0 + \int_{t_0}^{t} Ax(\tau)\,\mathrm{d}\tau. \tag{2.4}$$

This leads to the method of successive approximations (see Remark A.4), where we choose the initial function $x^{(0)}(t) = x_0$ for all $t \in \mathbb{R}$ and define the sequence

$$x^{(k+1)}(t) = x_0 + \int_{t_0}^{t} Ax^{(k)}(\tau)\,\mathrm{d}\tau, \tag{2.5}$$

for $k = 1, 2, \ldots$. The first two terms are easily evaluated to be

$$\begin{aligned} x^{(1)}(t) &= x_0 + \int_{t_0}^{t} Ax_0\,\mathrm{d}\tau = \big(I + A(t-t_0)\big)x_0, \\ x^{(2)}(t) &= x_0 + \int_{t_0}^{t} A(I + A(\tau - t_0))x_0\,\mathrm{d}\tau \\ &= \big(I + A(t-t_0) + \tfrac{1}{2}A^2(t-t_0)^2\big)\,x_0. \end{aligned} \tag{2.6}$$

This leads to the conclusion that

$$x^{(k)}(t) = \left(\sum_{l=0}^{k} \frac{A^l(t-t_0)^l}{l!}\right) x_0, \tag{2.7}$$

which can be proven using induction (by exploiting (2.5)).

Thus, we are interested in the (matrix) series

$$\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}, \tag{2.8}$$

where we have changed notation with respect to (2.7) (recall that $A^0 = I$ by definition). Even though we are interested in systems (2.1) for which the matrix A is real, we will consider the series (2.8) for the more general case of complex A.

We would like to show that this matrix series converges for every $t \in \mathbb{R}$, where it is recalled that a matrix series converges if each element converges. We thus aim to show that $\sum_{k=0}^{\infty} \frac{1}{k!} t^k (A^k)_{ij}$ converges for every $i, j \in \{1, 2, \ldots, n\}$ (and each $t \in \mathbb{R}$), where $(A^k)_{ij}$ denotes the element in row i and column j of the matrix A^k. To do so, it will turn out to be convenient to define the norm[2] $\|\cdot\|$ of a matrix $A \in \mathbb{C}^{n\times n}$ as

$$\|A\| = \sup\left\{ \frac{|Ax|}{|x|} \;\middle|\; x \neq 0 \right\} = \sup\big\{ |Ax| \;\big|\; |x| = 1 \big\}, \tag{2.9}$$

where $|x|$ is the norm of $x \in \mathbb{C}^n$ given by $|x| = \sqrt{x^* x}$, with x^* the Hermitian transpose of x.

The relation of the matrix norm (2.9) to matrix elements A_{ij} is given in the following lemma.

[2]It can be verified that this indeed defines a *norm*, i.e., 1) $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$, 2) $\|cA\| = |c|\|A\|$ for all $c \in \mathbb{C}$, and 3) $\|A_1 + A_2\| \leq \|A_1\| + \|A_2\|$.

Lemma 2.1. *Let $A \in \mathbb{C}^{n\times n}$ and $(A)_{ij}$ be any element of A. Then,*

$$|(A)_{ij}| \leq \|A\|. \tag{2.10}$$

Proof. Denote by e_i the i-th column of the identity matrix of size $n \times n$. After noting that

$$Ae_j = \begin{bmatrix} (A)_{1j} \\ (A)_{2j} \\ \vdots \\ (A)_{nj} \end{bmatrix}, \tag{2.11}$$

it is readily seen that

$$|(A)_{ij}| \leq |Ae_j| \leq \|A\|, \tag{2.12}$$

where the final inequality follows from the definition of the matrix norm in (2.9) (note that $|e_j| = 1$). □

The following result is known as the *sub-multiplicative property* of the matrix norm (2.9).

Lemma 2.2. *Let A, $B \in \mathbb{C}^{n\times n}$. Then,*

$$\|AB\| \leq \|A\|\|B\|. \tag{2.13}$$

Proof. Note that the definition in (2.9) implies that $|Ax| \leq \|A\||x|$ for any $x \in \mathbb{C}^n$. Applying this observation to ABx leads to

$$|ABx| \leq \|A\||Bx| \leq \|A\|\|B\||x|, \tag{2.14}$$

for any $x \in \mathbb{C}^n$. This, in turn, implies (2.13). □

Now, we can prove that the matrix series (2.8) converges.

Lemma 2.3. *The matrix series (2.8) converges for every $At \in \mathbb{C}^{n\times n}$.*

Proof. Let $i, j \in \{1, 2, \ldots, n\}$ select any element in the matrix series (2.8). For this (scalar) element, we know that convergence is implied by the stronger notion of *absolute* convergence, such that it is sufficient to show

$$\sum_{k=0}^{\infty} \left| \frac{t^k}{k!} (A^k)_{ij} \right| < \infty \tag{2.15}$$

Considering Lemma 2.1, we have

$$\left| \frac{t^k}{k!} (A^k)_{ij} \right| \leq \left\| \frac{t^k}{k!} A^k \right\| = \frac{\|(At)^k\|}{k!} \leq \frac{\|At\|^k}{k!}, \tag{2.16}$$

where the final inequality is the result of Lemma 2.2. Namely, it follows from this result that $\|A^k\| \leq \|A\|^k$ for all $k = 1, 2, \ldots$, whereas the case $k = 0$ follows immediately as $A^0 = I$. As a consequence, we obtain

$$\sum_{k=0}^{\infty} \left| \frac{t^k}{k!} (A^k)_{ij} \right| \leq \sum_{k=0}^{\infty} \frac{\|At\|^k}{k!} = e^{\|At\|} < \infty. \tag{2.17}$$

Here, we have used the fact that $\|At\|$ is a scalar and the definition of the exponential function, whose series representation is known to (absolutely) converge. This finalizes the proof of the lemma. □

As the matrix series converges, and motivated by the proof of Lemma 2.3, we define the *exponential* of a matrix A as follows.

Definition 2.1 (Matrix exponential)**.** *The exponential of a matrix At with $A \in \mathbb{C}^{n\times n}$, denoted as e^{At}, is defined as*

$$e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}. \tag{2.18}$$

We stress that e^{At} is itself a matrix for each t, which is real if $A \in \mathbb{R}^{n\times n}$. Turning again to such real matrices, and based on the above definition, we immediately obtain the following property.

Lemma 2.4. *Consider the function $t \mapsto e^{At}$ with $A \in \mathbb{R}^{n\times n}$. Then, its time derivative reads*

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{At} = Ae^{At}. \tag{2.19}$$

Proof. Even though the proof of Lemma 2.3 shows pointwise convergence (i.e., for fixed $t \in \mathbb{R}$) of the matrix exponential (2.18), it can be readily extended to show *uniform* convergence[3] (for compact intervals) when considered as the limit of functions $t \mapsto \frac{1}{k!}t^k A^k$.

As a result, we can differentiate on a term-by-term basis and compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} e^{At} &= \frac{\mathrm{d}}{\mathrm{d}t}\left\{ I + \sum_{k=1}^{\infty} \frac{A^k t^k}{k!} \right\} \\
&= \sum_{k=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t}\left\{ \frac{A^k t^k}{k!} \right\} \\
&= \sum_{k=1}^{\infty} \frac{A^k t^{k-1}}{(k-1)!} \\
&= A \sum_{k=1}^{\infty} \frac{A^{k-1} t^{k-1}}{(k-1)!} = Ae^{At},
\end{aligned} \tag{2.20}$$

which proves the desired result. □

Further properties of the matrix exponential will be discussed in Section 2.2.

However, the result of Lemma 2.4 allows us to state the main result of this section, which explicitly gives the solution to the initial value problem (2.2).

Theorem 2.5. *Consider the initial value problem (2.2) with $A \in \mathbb{R}^{n\times n}$. The function*

$$x(t; t_0, x_0) = e^{A(t-t_0)} x_0, \tag{2.21}$$

$t \in \mathbb{R}$, is the unique solution to this initial value problem.

[3] For a discussion on the various notions of convergence, see any text on analysis, e.g., [1].

Proof. First, substitution of $t = t_0$ and the use of the definition of the matrix exponential in (2.18) to show $e^{A\cdot 0} = I$ yields $x(t_0) = x_0$, such that the initial condition is indeed satisfied.

To show that (2.21) is a solution to the differential equation (2.1), use Lemma 2.4 to obtain

$$\tfrac{\mathrm{d}}{\mathrm{d}t}x(t;t_0,x_0) = Ae^{A(t-t_0)}x_0 = Ax(t;t_0,x_0), \tag{2.22}$$

which shows that (2.21) is indeed a solution.

Uniqueness of the solution follows from the theory of differential equations, which guarantees that the sequence of successive approximations (2.5) converges to the solution of the differential equation (see Appendix A). However, we give an independent proof here.

To this end, use the short-hand notation $x(t) = x(t;t_0,x_0)$ and assume that there exists a second solution, denoted x', that satisfies the same initial condition, i.e., $x'(t_0) = x_0$. Then, define $z(t) = x(t) - x'(t)$ and note that it satisfies

$$\dot{z}(t) = \dot{x}(t) - \dot{x}'(t) = A\big(x(t) - x'(t)\big) = Az(t). \tag{2.23}$$

Here, we have used the fact that both x and x' are solutions to (2.1). Now, pre-multiply (2.23) by $2z^{\mathrm{T}}(t)$ to obtain

$$2z^{\mathrm{T}}(t)\dot{z}(t) = 2z^{\mathrm{T}}(t)Az(t), \tag{2.24}$$

and note that the left- and right-hand side are scalar functions of t. In fact, for the left-hand side, we have

$$2z^{\mathrm{T}}(t)\dot{z}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}z^{\mathrm{T}}(t)z(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2, \tag{2.25}$$

where $|\cdot|$ denotes the Euclidean norm. Moreover, after recalling the definition of the matrix norm $\|\cdot\|$ in (2.9), we have that the right-hand side of (2.24) can be bounded as

$$2z^{\mathrm{T}}(t)Az(t) \leq 2\big|z^{\mathrm{T}}(t)Az(t)\big| \leq 2|z(t)|\big|Az(t)\big| \leq 2\|A\||z(t)|^2, \tag{2.26}$$

where the final two inequalities are given by the Cauchy-Schwarz inequality[4] and (2.9), respectively.

After defining the constant $\alpha = 2\|A\|$, the results (2.25) and (2.26), together with the equality (2.24), leads to

$$\tfrac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2 - \alpha|z(t)|^2 \leq 0. \tag{2.27}$$

Pre-multiplication with $e^{-\alpha t}$ (noting that $e^{-\alpha t} > 0$ for all $t \in \mathbb{R}$) gives

$$e^{-\alpha t}\Big(\tfrac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2 - \alpha|z(t)|^2\Big) \leq 0, \tag{2.28}$$

where the left-hand side can be rewritten to obtain

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big\{e^{-\alpha t}|z(t)|^2\big\} \leq 0, \tag{2.29}$$

[4]The Cauchy-Schwarz inequality states that, for two vectors $x, y \in \mathbb{R}^n$, we have that $|x^{\mathrm{T}}y| \leq |x||y|$. The application of this with $x = z(t)$ and $y = Az(t)$ gives the desired inequality.

and we recall that the inequality holds for all $t \in \mathbb{R}$. As a result, integration gives

$$\int_{t_0}^{t} \tfrac{\mathrm{d}}{\mathrm{d}\tau}\big\{e^{-\alpha\tau}|z(\tau)|^2\big\}\,\mathrm{d}\tau = e^{-\alpha t}|z(t)|^2 - e^{-\alpha t_0}|z(t_0)|^2 \leq 0. \tag{2.30}$$

However, we had $z(t_0) = 0$ as both solutions x and x' satisfy the initial condition (t_0, x_0). Thus, (2.30) gives

$$e^{-\alpha t}|z(t)|^2 \leq 0 \tag{2.31}$$

for all $t \in \mathbb{R}$, which implies that $|z(t)|^2 \leq 0$ as $e^{-\alpha t} > 0$. However, as $|\cdot|$ is a norm, this means that $z(t) = 0$ for all $t \in \mathbb{R}$. Moreover, this gives that $x(t) = x'(t)$ for all $t \in \mathbb{R}$, i.e., solutions are unique. □

2.2 Computation of the matrix exponential

In the previous section we have seen that the matrix exponential in Definition 2.1 plays a crucial role in explicitly characterizing the solutions of linear systems. We therefore consider the computation of the matrix exponential in this section.

We start by two examples in which the matrix exponential can be computed on the basis of its definition directly (recall Definition 2.1).

Example 2.1. Let $A \in \mathbb{R}^{n\times n}$ be the diagonal matrix

$$A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}. \tag{2.32}$$

Then, by Definition 2.1,

$$\begin{aligned} e^{At} &= \begin{bmatrix} \sum_{k=0}^{\infty} \frac{\lambda_1^k t^k}{k!} & 0 & \cdots & 0 \\ 0 & \sum_{k=0}^{\infty} \frac{\lambda_2^k t^k}{k!} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sum_{k=0}^{\infty} \frac{\lambda_n^k t^k}{k!} \end{bmatrix}, \\ &= \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_n t} \end{bmatrix}. \end{aligned} \tag{2.33}$$

Thus, a diagonal matrix A leads to a diagonal matrix e^{At} for all $t \in \mathbb{R}$. ◇

Example 2.2. Consider

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{2.34}$$

Then, a direct computation gives

$$A^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = -A, \quad A^4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = -A^2, \tag{2.35}$$

after which higher powers A^k are easily obtained. Hence, Definition 2.1 gives

$$e^{At} = \begin{bmatrix} \sum_{l=0}^{\infty} \frac{(-1)^l t^{2l}}{(2l)!} & \sum_{l=0}^{\infty} \frac{(-1)^l t^{2l+1}}{(2l+1)!} \\ -\sum_{l=0}^{\infty} \frac{(-1)^l t^{2l+1}}{(2l+1)!} & \sum_{l=0}^{\infty} \frac{(-1)^l t^{2l}}{(2l)!} \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}, \tag{2.36}$$

where the power series of the sin and cos function are recognized. ◇

Even though the above examples exploit the definition of the matrix exponential directly, this is not always a convenient approach. To enable the computation of the matrix exponential for a larger class of matrices, the following lemma is used.

Lemma 2.6. *Let $T \in \mathbb{C}^{n\times n}$ be nonsingular and $A \in \mathbb{C}^{n\times n}$. Then, for all $t \in \mathbb{R}$,*

$$e^{TAT^{-1}t} = Te^{At}T^{-1}. \tag{2.37}$$

Proof. This can directly be concluded from the definition (2.18) after observing that $(TAT^{-1})^k = TA^kT^{-1}$ for any nonnegative integer k. □

The relevance of Lemma 2.6 is that it enables the computation of the matrix exponential for diagonalizable matrices A, as illustrated in the following example.

Example 2.3. Consider the matrix

$$A = \begin{bmatrix} -2 & 1 \\ 0 & 1 \end{bmatrix}. \tag{2.38}$$

It is easy to see that its two eigenvalues are given by $\lambda_1 = -2$ and $\lambda_2 = 1$, with corresponding eigenvectors

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}. \tag{2.39}$$

After defining

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad T = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix}, \tag{2.40}$$

it follows that $A = T\Lambda T^{-1}$, such that

$$\begin{aligned} e^{At} = e^{T\Lambda T^{-1}t} = Te^{\Lambda t}T^{-1} &= \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{t} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{3} \\ 0 & \frac{1}{3} \end{bmatrix} \\ &= \begin{bmatrix} e^{-2t} & \frac{1}{3}(e^{t} - e^{-2t}) \\ 0 & e^{t} \end{bmatrix}, \end{aligned} \tag{2.41}$$

where the result of Example 2.1 is used to compute $e^{\Lambda t}$. ◇

The approach of this example is further elaborated upon in the following remark, where the notation $\sigma(A)$ denotes the *spectrum* of the matrix A, i.e., the set of eigenvalues of A.

Remark 2.1. We recall that $A \in \mathbb{R}^{n\times n}$ is diagonalizable if and only if, for each eigenvalue $\lambda \in \sigma(A)$, its algebraic multiplicity a_λ equals its geometric multiplicity g_λ (for detailed definitions of these concepts, see Appendix B.1). In this case, there exists n linearly independent eigenvectors $v_i \in \mathbb{C}^n$ such that

$$AT = T\Lambda, \tag{2.42}$$

with

$$T = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \tag{2.43}$$

and where the eigenvalues $\lambda_1, \ldots, \lambda_n$ are not necessarily distinct. Using the result of Lemma 2.6, we can write

$$e^{A(t-t_0)} = e^{T\Lambda T^{-1}(t-t_0)} = Te^{\Lambda(t-t_0)}T^{-1}, \tag{2.44}$$

such that

$$x(t; t_0, x_0) = e^{A(t-t_0)}x_0 = \sum_{i=1}^{n} c_i v_i e^{\lambda_i(t-t_0)}. \tag{2.45}$$

Here, c_i are the elements of the vector $c = [c_1 \ c_2 \ \cdots \ c_n]^{\mathrm{T}} \in \mathbb{C}^n$ defined as $c = T^{-1}x_0$. We note that the result (2.45) shows that the solution of a homogeneous linear system can be written solely in terms of the eigenvalues and corresponding eigenvectors. Also, it is stressed that c depends on the initial condition x_0 and can be obtained by solving the linear system $Tc = x_0$. ◁

To compute matrix exponentials for matrices that are not necessarily diagonalizable as in Example 2.3, some more results on the matrix exponential are needed. The first one is a technical lemma.

Lemma 2.7. *Let $A, B \in \mathbb{C}^{n\times n}$ be such that $AB = BA$ (i.e., A and B commute). Then, for all $t \in \mathbb{R}$,*

$$e^{At}B = Be^{At}. \tag{2.46}$$

Proof. See Exercise 2.3. □

The above result enables us to prove the following lemma, which gives important properties of the matrix exponential.

Lemma 2.8. *The following properties hold for any $A, B \in \mathbb{C}^{n\times n}$:*

1. *e^{At} is invertible for any $t \in \mathbb{R}$. Specifically, its inverse reads*

$$\left(e^{At}\right)^{-1} = e^{-At}; \tag{2.47}$$

2. *if $AB = BA$, then, for all $t \in \mathbb{R}$,*

$$e^{At}e^{Bt} = e^{(A+B)t}; \tag{2.48}$$

3. for all $t, s \in \mathbb{R}$,

$$e^{At}e^{As} = e^{A(t+s)}. \tag{2.49}$$

Proof. 1. Using the chain rule and the result of Lemma 2.4, we obtain

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\{e^{At}e^{-At}\} &= Ae^{At}e^{-At} + e^{At}(-Ae^{-At}), && (2.50)\\ &= Ae^{At}e^{-At} - Ae^{At}e^{-At} = 0, && (2.51) \end{aligned}$$

where (2.51) follows from Lemma 2.7. Thus, $e^{At}e^{-At} = C$ for all $t \in \mathbb{R}$ for some constant matrix C. However, for $t = 0$, we immediately obtain $C = I$, after which the result (2.47) follows.

2. A similar procedure as in the proof of statement *1* will be followed. Specifically, consider

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\Big\{e^{(A+B)t}e^{-Bt}e^{-At}\Big\} &= (A+B)e^{(A+B)t}e^{-Bt}e^{-At} \\ &\quad + e^{(A+B)t}(-Be^{-Bt})e^{-At} \\ &\quad + e^{(A+B)t}e^{-Bt}(-Ae^{-At}) && (2.52)\\ &= \big(A+B-B-A\big)e^{(A+B)t}e^{-Bt}e^{-At} = 0. && (2.53) \end{aligned}$$

Here, the result (2.53) follows from Lemma 2.7 and the assumption that A and B commute. We thus obtain

$$e^{(A+B)t}e^{-Bt}e^{-At} = I, \tag{2.54}$$

after which post-multiplication with $e^{At}e^{Bt}$ and the use of statement *1* gives the result (2.48).

3. This follows immediately from statement *2* after choosing $B = A\frac{s}{t}$ (assuming $t \neq 0$; the case $t = 0$ is trivial) and noting that A and B commute. □

This result can be used to compute the matrix exponential for a matrix that cannot be diagonalized.

Example 2.4. Let

$$A = D + N, \tag{2.55}$$

with

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{2.56}$$

The matrix A is not diagonalizable since the only eigenvalue $\lambda = 2$ has algebraic multiplicity 2 and geometric multiplicity 1. However, it is clear that D and N commute (as $D = 2I$), such that statement *2* of Lemma 2.8 gives

$$e^{At} = e^{(D+N)t} = e^{Dt}e^{Nt}. \tag{2.57}$$

However, e^{Dt} is easily obtained (see again Example 2.1) as

$$e^{Dt} = \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{2t} \end{bmatrix} = e^{2t}I. \tag{2.58}$$

Moreover, we have $N^2 = 0$, such that $N^k = 0$ for $k = 2, 3, \ldots$. Then, by the definition of the matrix exponential (Definition 2.1), it follows that

$$e^{Nt} = I + Nt = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \tag{2.59}$$

such that we obtain

$$e^{(D+N)t} = e^{2t} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e^{2t} & te^{2t} \\ 0 & e^{2t} \end{bmatrix}. \tag{2.60}$$

Note that the fact that D and N commute is crucial for this computation. ◇

In the remainder of this section, we will generalize the procedure of Example 2.4 to compute the matrix exponential for an arbitrary matrix A.

To do so, we first need the following definition.

Definition 2.2. *A Jordan block $J_k(\lambda) \in \mathbb{C}^{k\times k}$ is the matrix*

$$J_k(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 & 0 \\ 0 & & & \ddots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}. \tag{2.61}$$

Note that a Jordan block $J_k(\lambda)$ is a matrix of the form $\lambda I + N$, where N has the property that $N^k = 0$. A general matrix N satisfying $N^k = 0$ for some positive integer k is called a *nilpotent* matrix.

Hence, following the procedure of Example 2.4, we immediately obtain the following lemma.

Lemma 2.9. *Consider the Jordan block $J_k(\lambda)$ for some positive integer k and $\lambda \in \mathbb{C}$. Then,*

$$e^{J_k(\lambda)t} = e^{\lambda t} \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{k-2}}{(k-2)!} & \frac{t^{k-1}}{(k-1)!} \\ 0 & 1 & t & \ddots & & \frac{t^{k-2}}{(k-2)!} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & t & \frac{t^2}{2!} \\ 0 & & & \ddots & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{2.62}$$

Proof. See Exercise 2.7. □

The definition of a Jordan block is used in the following fundamental result from linear algebra, which shows that any matrix can be written in a so-called *Jordan canonical form*.

Theorem 2.10. *For any matrix $A \in \mathbb{R}^{n\times n}$, there exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ such that $A = TJT^{-1}$ with*

$$J = \begin{bmatrix} J_{k_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{k_2}(\lambda_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & J_{k_r}(\lambda_r) \end{bmatrix}, \tag{2.63}$$

where $\lambda_i \in \sigma(A)$, $i = 1, 2, \ldots, r$ and $n = k_1 + k_2 + \ldots + k_r$. Conversely, let $\lambda \in \sigma(A)$ be any eigenvalue of A. Then, $\lambda = \lambda_i$ for some $i \in \{1, 2, \ldots, r\}$.

Proof. The full proof is out of the scope of this course, but more details are given in Appendix B. □

Here, we recall that $\sigma(A)$ denotes the *spectrum* (the set of eigenvalues) of the matrix A.

It is important to note that, in the Jordan canonical form, the parameters λ_i equal the eigenvalues of the matrix A. However, the Jordan blocks $J_{k_i}(\lambda_i)$ do not necessarily correspond to distinct eigenvalues. In fact, for a given eigenvalue λ, the number of Jordan blocks for this eigenvalue is equal to the *geometric multiplicity* of λ, whereas the sum of the dimensions of these Jordan blocks equals the *algebraic multiplicity* of λ. A detailed discussion of the Jordan canonical form and its computation is given in Appendix B.

The above results now lead to the following approach for computing the matrix exponential e^{At} for a given matrix A.

1. Compute the Jordan canonical form of A as in Theorem 2.10, i.e., find the matrix J in (2.63) and the transformation matrix T such that $A = TJT^{-1}$.

2. For each Jordan block $J_{k_i}(\lambda_i)$ in J, compute the matrix exponential $e^{J_{k_i}(\lambda_i)t}$ using Lemma 2.9.

3. Compute e^{At} through

$$e^{At} = e^{TJT^{-1}t} = Te^{Jt}T^{-1} = T \begin{bmatrix} e^{J_{k_1}(\lambda_1)t} & 0 & \cdots & 0 \\ 0 & e^{J_{k_2}(\lambda_2)t} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{J_{k_r}(\lambda_r)t} \end{bmatrix} T^{-1}, \tag{2.64}$$

   where subsequently Lemma 2.6 and the block-diagonal structure of J (similar to Example 2.1) are used.

There are various other ways to compute matrix exponentials, see, e.g., [8] for an overview.

2.3 Nonhomogeneous linear systems

Returning attention to finding solutions to linear systems, we recall that the unique solution to the initial value problem for homogeneous linear systems is given in Theorem 2.5. In this section, the nonhomogeneous case is considered.

Specifically, consider the nonhomogeneous linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.65}$$

where $x(t) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ as in the homogeneous case. The nonhomogeneous case features in addition an *input* $u(t) \in \mathbb{R}^m$, whose influence on the dynamics is captured through $B \in \mathbb{R}^{n \times m}$. More specifically, $u : J \to \mathbb{R}^m$ is an

input function defined on some interval J, that is assumed to be continuous. Then, for a given input function, the system (2.65) is of the form $\dot{x}(t) = f(t, x(t))$ with $f(t, x) = Ax + Bu(t)$.

Remark 2.2. In these notes, we restrict attention to *continuous* input functions u, such that $f(t, x) = Ax + Bu(t)$ satisfies the assumptions on continuity of Appendix A. However, the results of this section can easily be generalized to more general classes of input functions such as *piecewise continuous* or *locally integrable* inputs. The class of inputs under consideration is often referred to as the class of *admissible* inputs. At the same time, the exact class from which the inputs are chosen is often not important. ◁

For a given input function u and initial condition x_0, we are interested in the initial value problem

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(t_0) = x_0, \;\; u : J \to \mathbb{R}^m, \tag{2.66}$$

where it is assumed that $t_0 \in J$. The solution will be denoted as $x(\,\cdot\,; t_0, x_0, u)$.

Building on the homogeneous case, the following result can be stated.

Theorem 2.11. *Consider the initial value problem (2.66). Then, the function*

$$x(t; t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau \tag{2.67}$$

defined for $t \in J$, is the unique solution to this initial value problem.

Proof. The result can be proven as in the homogeneous case, i.e., by subsequently showing that the initial condition is satisfied, that the solution satisfies the differential equation (2.65), and that this solution is unique.

Instead, we will *derive* the result (2.67) through *variation of constants*. To this end, define

$$z(t) = e^{-At}x(t), \tag{2.68}$$

which satisfies

$$\begin{aligned} \dot{z}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}\big\{e^{-At}x(t)\big\} &= -Ae^{-At}x(t) + e^{-At}\dot{x}(t) \\ &= -Ae^{-At}x(t) + e^{-At}\big(Ax(t) + Bu(t)\big) \qquad (2.69) \\ &= e^{-At}Bu(t) \qquad (2.70) \end{aligned}$$

The result (2.69) follows from substitution of the dynamics (2.65), whereas the final result is obtained by using Lemma 2.7 (as A and $-A$ commute). In the new coordinates z, the initial value problem thus reads

$$\dot{z}(t) = e^{-At}Bu(t), \qquad z(t_0) = e^{-At_0}x_0, \;\; u : J \to \mathbb{R}^m, \tag{2.71}$$

where the initial condition follows from evaluating (2.68) for $t = t_0$. However, its unique solution is easily obtained by integration, such that

$$z(t) = z(t_0) + \int_{t_0}^{t} e^{-A\tau}Bu(\tau)\,\mathrm{d}\tau. \tag{2.72}$$

Now, by Lemma 2.8 (statement *1*), we obtain $x(t) = e^{At}z(t)$, such that the result (2.67) follows from (2.72) with $z(t_0) = e^{-At_0}x_0$ and the use of Lemma 2.8 (statement *3*). □

In the result of Theorem 2.11, we note that the domain for which the solution $x(\,\cdot\,;t_0,x_0,u)$ is defined equals that of the input function. Thus, for input functions defined for all $t \in \mathbb{R}$, the solution is also defined for all time.

Remark 2.3. When also considering the output equation

$$y(t) = Cx(t) + Du(t) \tag{2.73}$$

in addition to the dynamics (2.65), it is clear that the (unique) output solution for initial condition (t_0, x_0) and input function u denoted by $y(\,\cdot\,;t_0,x_0,u)$ is given as

$$y(t;t_0,x_0,u) = Ce^{A(t-t_0)}x_0 + \int_{t_0}^{t} Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t), \tag{2.74}$$

for all $t \in J$. The relation between inputs and outputs will be studied in detail in Chapter 6. ◁

The system (2.65) was called linear because the differential equation is linear (in x). However, linearity of the differential equation also implies that the solution (2.67) is linear in both the initial condition and input function. This observation is known as the *superposition principle* and formalized in the following theorem.

Theorem 2.12. *Consider the linear system (2.65), let (t_0, x_0) and (t_0, x_0') with x_0, $x_0' \in \mathbb{R}^n$ be two initial conditions and u, $u' : J \to \mathbb{R}^m$ be two input functions with $t_0 \in J$. Then, for any α, $\alpha' \in \mathbb{R}$,*

$$x\big(t;t_0,\alpha x_0 + \alpha' x_0', \alpha u + \alpha' u'\big) = \alpha x(t;t_0,x_0,u) + \alpha' x(t;t_0,x_0',u') \tag{2.75}$$

for all $t \in J$.

Proof. This follows directly from (2.67) in Theorem 2.11. □

In addition, the matrices A, B describing the linear system (2.65) are constant, i.e., independent of time. This implies that the solutions of (2.65) do not explicitly depend on time, but only on the time that has elapsed since the initial time t_0.

This is stated in the following theorem.

Theorem 2.13. *Consider the linear system (2.65). Then, for any (t_0, x_0) with $x_0 \in \mathbb{R}^n$ and any input function $u : \mathbb{R} \to \mathbb{R}^m$, it holds that*

$$x(t;t_0,x_0,u) = x(t-t_0;0,x_0,u_{t_0}), \tag{2.76}$$

where u_{t_0} is the time-shifted input function defined as $u_{t_0}(t) = u(t+t_0)$, $t \in \mathbb{R}$.

Proof. This again directly follows from the explicit expression (2.67) in Theorem 2.11. □

The above property is sometimes referred to as the *time-invariance* of the linear system (2.65), which models the observation that many physical processes do not explicitly depend on time. Moreover, because of time-invariance, we will assume that $t_0 = 0$ in the remainder of these notes. We then also use the shorthand notation $x(\,\cdot\,;x_0,u)$ and $y(\,\cdot\,;x_0,u)$ for the state and output trajectories, respectively.

2.4 Exercises

Exercise 2.1. Consider

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{2.77}$$

Compute e^{At} by diagonalizing A and compare the result to the direct approach of Example 2.2.

Exercise 2.2. Let

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix} \tag{2.78}$$

and show that

$$e^{At} = e^{\sigma t} \begin{bmatrix} \cos \omega t & \sin \omega t \\ -\sin \omega t & \cos \omega t \end{bmatrix}. \tag{2.79}$$

Exercise 2.3. Prove Lemma 2.7.

Exercise 2.4. Assume that $e^{At}B = Be^{At}$ for all $t \in \mathbb{R}$ and some matrices $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times n}$. Prove that this implies that $AB = BA$. Note that this is the converse of Lemma 2.7.

Exercise 2.5. Prove statement *1* of Lemma 2.8 by exploiting the result of statement *2*.

Exercise 2.6. Consider the following questions.

a. Let $A \in \mathbb{C}^{n\times n}$. Show that $\left(e^{At}\right)^* = e^{A^* t}$.

b. Show that e^S is unitary if $S \in \mathbb{C}^{n\times n}$ is skew-symmetric, i.e., if $S^* = -S$.

Exercise 2.7. Prove Lemma 2.9.

Exercise 2.8. Compute e^{At} for the following matrices. Avoid the Jordan canonical form whenever possible.

a. $A = \begin{bmatrix} 3 & 6 \\ -2 & -3 \end{bmatrix}$

b. $A = \begin{bmatrix} 8 & 1 \\ -4 & 4 \end{bmatrix}$

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

d. $A = \begin{bmatrix} -1 & 1 & -1 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}$

Exercise 2.9. Compute the Jordan canonical form (i.e., the matrices J and T) for the following matrices. In addition, compute e^{Jt} (the computation of e^{At} is not needed).

a. $A = \begin{bmatrix} 1 & 1 \\ -16 & 9 \end{bmatrix}$

b. $A = \begin{bmatrix} -15 & 9 \\ -25 & 15 \end{bmatrix}$

c. $A = \begin{bmatrix} -6 & 9 & 0 \\ -6 & 6 & -2 \\ 9 & -9 & 3 \end{bmatrix}$

d. $A = \begin{bmatrix} 0 & -4 & 1 \\ 2 & -6 & 1 \\ 4 & -8 & 0 \end{bmatrix}$

Exercise 2.10. Let $A \in \mathbb{R}^{n\times n}$ be given. A subspace $\mathcal{V} \subset \mathbb{R}^n$ is called A-invariant if the implication $v \in \mathcal{V} \implies Av \in \mathcal{V}$ holds. This is also denoted as $A\mathcal{V} \subset \mathcal{V}$.

Let $\mathcal{V}$ be A-invariant. Show that

$$e^{At}\mathcal{V} \subset \mathcal{V} \tag{2.80}$$

for all $t \in \mathbb{R}$. *Hint.* Use that $\mathcal{V}$ is closed.

Exercise 2.11. Let $A \in \mathbb{R}^{n\times n}$. Prove the following two statements:

a. If s is not an eigenvalue of A, then the differential equation

$$\dot{x}(t) = Ax(t) + be^{st}$$

with $b \in \mathbb{R}^n$ has exactly one solution of the form $x(t) = \bar{x}e^{st}$. In particular, if $\det A \neq 0$, then the differential equation $\dot{x}(t) = Ax(t) + b$ with $b \in \mathbb{R}^n$ has exactly one constant solution.

b. If s is not an eigenvalue of A, then the differential equation

$$\dot{x}(t) = Ax(t) + (b_0 + b_1 t + \ldots + b_k t^k)e^{st}$$

with $b_0, \ldots, b_k \in \mathbb{R}^n$ has exactly one solution of the form $x(t) = (\bar{x}_0 + \bar{x}_1 t + \ldots + \bar{x}_k t^k)e^{st}$.

Chapter 3

Stability

In this chapter, the system property known as *stability* is studied. First, stability of linear systems is defined and characterized in Section 3.1, showing a link with the characteristic polynomial of the system matrix. Therefore, stability of polynomials is studied in Sections 3.2 and 3.3, where the latter section considers polynomials with unknown coefficients.

3.1 Stability of linear systems

The concept of stability plays a key role in systems theory as it characterizes the asymptotic behavior of systems. Stability is generally the first requirement for control systems as unstable behavior is often undesired in physical systems.

Consider the homogeneous linear system

$$\dot{x}(t) = Ax(t), \tag{3.1}$$

with $x(t) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ as before. Following the notation of Chapter 2, the unique solution to (3.1) for initial condition $x_0 \in \mathbb{R}^n$ is denoted as $x(\,\cdot\,; x_0)$.

For systems of the form (3.1), (asymptotic) stability is defined as follows.

Definition 3.1. *The system (3.1) is called*

1. *stable if every solution is bounded for all $t \geq 0$, i.e., for any initial condition $x_0 \in \mathbb{R}^n$, there exists $M > 0$ such that*

$$|x(t; x_0)| = \left|e^{At}x_0\right| \leq M, \quad t \geq 0. \tag{3.2}$$

2. *asymptotically stable if every solution tends to zero for $t \to \infty$, i.e., for any initial condition $x_0 \in \mathbb{R}^n$, it holds that*

$$\lim_{t\to\infty} x(t; x_0) = \lim_{t\to\infty} e^{At}x_0 = 0. \tag{3.3}$$

Remark 3.1. Even though we define stability for homogeneous systems of the form (3.1), the same definition applies to general linear systems as in (1.1). In this case, we say that the system is asymptotically stable if the state trajectory converges to zero when the zero input $u(t) = 0$ is applied. This exactly recovers the homogeneous system (3.1). For linear systems with inputs, we therefore sometimes speak of *internal* (asymptotic) stability. ◁

Remark 3.2. For more general systems such as systems with nonlinear or time-varying dynamics, the definition of stability is much more involved and various notions of stability exist. Specifically, Lyapunov stability theory deals with such systems. Here, we just stress that the simple notion of stability in Definition 3.1 is only applicable to linear systems of the form (3.1). ◁

To derive conditions for stability, the following result is required.

Lemma 3.1. *Consider the function $t \mapsto t^k e^{\lambda t}$ for some nonnegative integer k and $\lambda \in \mathbb{C}$. If $\operatorname{Re}(\lambda) < 0$, then*

$$\lim_{t\to\infty} t^k e^{\lambda t} = 0 \tag{3.4}$$

and $|t^k e^{\lambda t}|$ is bounded for all $t \geq 0$.

Proof. See Exercise 3.1. □

The results in Lemma 3.1 can be strengthened by explicitly computing an exponential bound on functions of the form $t^k e^{\lambda t}$. This is formalized in the following lemma.

Lemma 3.2. *Consider the function $t \mapsto t^k e^{\lambda t}$ for some nonnegative integer k and $\lambda \in \mathbb{C}$. For any $\alpha \in \mathbb{R}$ such that $\operatorname{Re}(\lambda) < \alpha$, there exists a positive constant $M \in \mathbb{R}$ such that, for all $t \geq 0$,*

$$\left|t^k e^{\lambda t}\right| \leq M e^{\alpha t}. \tag{3.5}$$

Proof. Consider the function $p(t) := e^{-\alpha t}(t^k e^{\lambda t}) = t^k e^{(\lambda-\alpha)t}$ and note that p is of the form $t^k e^{\lambda' t}$ with $\operatorname{Re}(\lambda') = \operatorname{Re}(\lambda - \alpha) < 0$. Consequently, the conditions of Lemma 3.1 hold for p and, as a result, p is bounded, i.e., there exists $M > 0$ such that $|p(t)| \leq M$ for all $t \geq 0$. Then, $e^{\alpha t}|p(t)| = |t^k e^{\lambda t}| \leq M e^{\alpha t}$, proving the desired result. □

The result of Lemma 3.1 can be exploited for stability analysis after recalling the results on matrix exponentials in Section 2.2 and using those to study the asymptotic behavior of e^{At}. After defining the notation

$$\mathbb{C}_- = \big\{z \in \mathbb{C} \;\big|\; \operatorname{Re}(z) < 0\big\}, \qquad \bar{\mathbb{C}}_- = \big\{z \in \mathbb{C} \;\big|\; \operatorname{Re}(z) \leq 0\big\}, \tag{3.6}$$

for the open and closed left-half complex plane, this approach is made explicit in the following key result.

Theorem 3.3. *Consider the system (3.1). The following statements hold.*

1. *The system (3.1) is stable if and only if*

$$\sigma(A) \subset \bar{\mathbb{C}}_-. \tag{3.7}$$

   and every eigenvalue λ with $\operatorname{Re}(\lambda) = 0$ is semisimple, i.e., has equal algebraic and geometric multiplicity;

2. *The system (3.1) is asymptotically stable if and only if*

$$\sigma(A) \subset \mathbb{C}_-. \tag{3.8}$$

   In this case, there exist scalars $M, \gamma > 0$ such that

$$\left\|e^{At}\right\| \leq M e^{-\gamma t}, \tag{3.9}$$

   for all $t \geq 0$.

Proof. To prove the two statements, we first recall that, by the Jordan canonical form of Theorem 2.10, the matrix exponential can be written as

$$e^{At} = Te^{Jt}T^{-1}, \tag{3.10}$$

see also (2.64). This means that each element of e^{At} is a linear combination of terms of the form $t^k e^{\lambda t}$ for some nonnegative integer k and $\lambda \in \sigma(A)$. We first prove statement *2*.

2. Let (3.8) hold, such that $\mathrm{Re}(\lambda) < 0$ for all $\lambda \in \sigma(A)$. Then, we have from Lemma 3.1 that each term $t^k e^{\lambda t}$ converges to zero. As a result,

$$\lim_{t\to\infty} e^{At} = 0, \tag{3.11}$$

proving asymptotic stability.

To prove the converse by contraposition, let $\sigma(A) \not\subset \mathbb{C}_-$, i.e., there exists an eigenvalue $\lambda \in \sigma(A)$ such that $\mathrm{Re}(\lambda) \geq 0$. Let v be the corresponding eigenvector and note that $x(t) = e^{\lambda t}v$ is a solution of (3.1). Namely,

$$\dot{x}(t) = e^{\lambda t}\lambda v = e^{\lambda t}Av = Ave^{\lambda t} = Ax(t). \tag{3.12}$$

In fact, the solution $x(t) = e^{\lambda t}v$ corresponds to the initial condition $x_0 = v$ and thus equals $x(t; v)$. As $\mathrm{Re}(\lambda) \geq 0$, it can be seen (e.g., using Euler's formula) that $x(t, v)$ does not tend to zero as $t \to \infty$.

The bound on $\|e^{At}\|$ follows from Lemma 3.2 after noting that $\sigma(A) \subset \mathbb{C}_-$ implies that there exists $\gamma > 0$ such that $\mathrm{Re}(\lambda) < -\gamma$ for all $\lambda \in \sigma(A)$.

1. To prove sufficiency of the conditions, we first consider eigenvalues $\lambda \in \sigma(A)$ such that $\mathrm{Re}(\lambda) < 0$. As in the proof of statement *2*, such eigenvalues lead to terms $t^k e^{\lambda t}$ which are bounded by Lemma 3.2. Next, let $\lambda \in \sigma(A)$ satisfy $\mathrm{Re}(\lambda) = 0$ and be semisimple. As the algebraic and geometric multiplicity of λ are the same, this eigenvalue only leads to the trivial scalar Jordan blocks, yielding terms $e^{\lambda t}$. As $\mathrm{Re}(\lambda) = 0$, this term is bounded.

The converse is again shown by considering two cases. First, let λ be such that $\mathrm{Re}(\lambda) > 0$. Then, by the proof of statement *2*, $ve^{\lambda t}$ is a solution which is clearly not bounded. Next, let λ be such that $\mathrm{Re}(\lambda) = 0$, but not semisimple. Then, this leads to terms in e^{At} of the form $t^k e^{i\omega t}$ with $k > 0$ and $\omega = \mathrm{Im}(\lambda)$. By Euler's formula, it is clear that such terms are unbounded. Thus, in the above to cases, no bound as in (3.2) exists, finalizing the proof by contraposition. □

Stability of a linear system is thus completely determined by the eigenvalues of the system matrix. A matrix A satisfying condition (3.8) is called *Hurwitz* (or, sometimes, a *stability matrix*).

Remark 3.3. It is clear from condition (3.8) that the "right-most" eigenvalue in the complex plane determines stability of the system (3.1). This is reflected in an equivalent condition for asymptotic stability. Namely, after defining the so-called *spectral abscis* of the matrix A as

$$\Lambda(A) = \max\big\{\, \mathrm{Re}(\lambda) \;\big|\; \lambda \in \sigma(A)\big\}, \tag{3.13}$$

it is clear that (3.1) is asymptotically stable if and only if $\Lambda(A) < 0$. In this case, the bound (3.9) holds for any $\gamma > 0$ such that $\Lambda(A) < -\gamma$. Since numerical methods exist to compute (bounds on) the spectral abscis, the computation of the entire spectrum could be avoided in determining stability properties. ◁

We will illustrate the result of Theorem 3.3 by means of an example.

Example 3.1. Recall the mass-spring-damper system of Example 1.1 (see also Figure 1.2). For this system, the matrix A reads

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \tag{3.14}$$

where $m > 0$. We will explicitly compute the eigenvalues by considering the characteristic polynomial as

$$\Delta_A(s) = \det(sI - A) = \begin{vmatrix} s & -1 \\ \frac{k}{m} & s + \frac{c}{m} \end{vmatrix} = s^2 + \tfrac{c}{m}s + \tfrac{k}{m}, \tag{3.15}$$

after which the roots are easily found using the quadratic formula as

$$\lambda_1 = -\frac{c}{2m} + \frac{\sqrt{c^2 - 4mk}}{2m}, \qquad \lambda_2 = -\frac{c}{2m} - \frac{\sqrt{c^2 - 4mk}}{2m}. \tag{3.16}$$

If is clear that $\mathrm{Re}(\lambda_i) < 0$ for $i = 1, 2$ if and only if $c > 0$ and $k > 0$, which thus provides a necessary and sufficient condition for asymptotic stability. ◇

3.2 The Routh-Hurwitz criterion

We have seen in Theorem 3.3 that the eigenvalues of the system matrix A determine stability of the linear system (3.1). These eigenvalues are exactly the roots of the characteristic polynomial $\Delta_A(s) = \det(sI - A)$. As a result, we could verify stability of a linear system by studying the roots of polynomials.

We will consider polynomials (in the indeterminate s)

$$p(s) = a_n s^n + a_{n-1} s^{n-1} + \ldots + a_1 s + a_0, \tag{3.17}$$

with $a_0, \ldots, a_n \in \mathbb{R}$ and $a_n \neq 0$. Then, n is called the *degree* of p. The polynomial (3.17) is said to be *monic* if $a_n = 1$. A *root* (or *zero)* of the polynomial p in (3.17) is a (complex) number $\lambda \in \mathbb{C}$ such that

$$p(\lambda) = a_n \lambda^n + a_{n-1} \lambda^{n-1} + \ldots + a_1 \lambda + a_0 = 0, \tag{3.18}$$

and we denote the set of all such roots by $\sigma(p)$, referred to as the spectrum of p.

Now, we can state the following definition of stability of a polynomial.

Definition 3.2. *A polynomial p as in (3.17) is called stable if all its roots have negative real part, i.e., if $\sigma(p) \subset \mathbb{C}_-$.*

The following result provides a recursive algorithm to check the stability of a polynomial.

Theorem 3.4. *Let p be a polynomial as in (3.17) with $a_0, \ldots, a_n \in \mathbb{R}$ and $a_n \neq 0$. Then, p is stable if and only if the following three conditions hold:*

1. *a_{n-1} is nonzero, i.e., $a_{n-1} \neq 0$;*
2. *a_n and a_{n-1} have the same sign, i.e., $a_n a_{n-1} > 0$;*

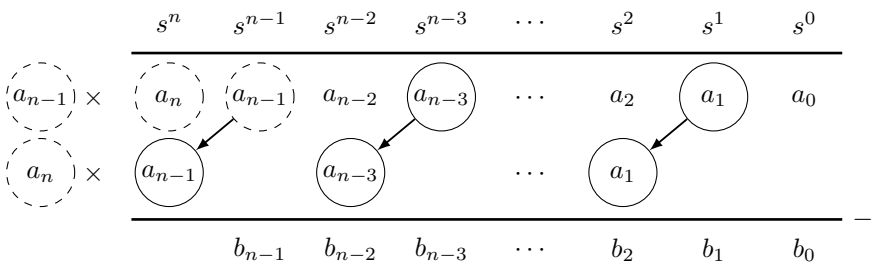


Figure 3.1: Illustration of one step in the Routh table to verify the conditions of Theorem 3.4. Explicit expressions for the coefficients b_i are given in (3.22). Note (from the rightmost three coefficients) that the table is given for n even.

3. the polynomial

$$q(s) = a_{n-1}p(s) - a_n\big(a_{n-1}s^n + a_{n-3}s^{n-2} + a_{n-5}s^{n-4} + \dots\big) \tag{3.19}$$

is stable.

Proof. A proof can be found in [6]. □

It is important to note that the polynomial q in (3.19) is of degree $n-1$, which is one less than the degree of the original polynomial p. This not only simplifies stability analysis, but also allows for the repetitive application of Theorem 3.4. Then, stability of the original polynomial p can be concluded when one of the conditions is violated or we are left with a polynomial of degree 1, whose only root is easily obtained. When used on the characteristic polynomial Δ_A of a system matrix A, we can thus assess stability of a linear system without explicitly computing eigenvalues.

The procedure of verifying the conditions in Theorem 3.4 and constructing the polynomial q in (3.19) can conveniently represented using a table, as illustrated in Figure 3.1. Here, the top row indicates the powers of a polynomial of degree n, whereas the polynomial p (of degree n) is indicated in the first row (below the top line). The coefficients a_n and a_{n-1}, surrounded by dashed circles, then play a role in verifying conditions *1* and *2*.

If these conditions are verified, the polynomial q in (3.19) can be constructed by first forming the polynomial

$$a_{n-1}s^n + a_{n-3}s^{n-2} + a_{n-5}s^{n-4} + \dots, \tag{3.20}$$

which is indicated in the second row in Figure 3.1. Note that only odd or even powers are defined (for odd or even degree n, respectively). The polynomial q is now formed by subtracting these two rows after multiplying them with their respective constants a_{n-1} and a_n (again in dashed circles). Note that $a_n \neq 0$ by definition of the degree n and $a_{n-1} \neq 0$ if statement *1* is verified. Specifically, q is now given as

$$q(s) = b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \dots + b_1 s + b_0, \tag{3.21}$$

where the coefficients $b_0, \ldots, b_{n-1} \in \mathbb{R}$ are easily read from the table to be

$$\begin{aligned} b_{n-1} &= a_{n-1}a_{n-1}, \\ b_{n-2} &= a_{n-1}a_{n-2} - a_n a_{n-3}, \\ b_{n-3} &= a_{n-1}a_{n-3}, \\ &\vdots \\ b_2 &= a_{n-1}a_2 - a_n a_1, \\ b_1 &= a_{n-1}a_n, \\ b_0 &= a_{n-1}a_0. \end{aligned} \tag{3.22}$$

Note that b_{n-1} is nonzero by statement *1*. Also, we stress that the final coefficients depend on whether n is even or odd. However, the expressions are easily found using the Routh table.

The use of the Routh table and Theorem 3.4 are illustrated in the following two examples.

Example 3.2. Consider the polynomial

$$p(s) = s^3 + 2s^2 + 4s + 3. \tag{3.23}$$

To evaluate stability, we will repeatedly apply Theorem 3.4 by forming a Routh table as in Figure 3.1. This leads to the following table:

	s^3	s^2	s	1		
$2\times$	1	2	4	3		(p)
$1\times$	2		3			
		4	5	6	$-$	(q)
$5\times$						
$4\times$		5				
			25	30	$-$	(r)

First, note that the first two statements of Theorem 3.4 are satisfied for p in (3.23) ($a_2 = 2 \neq 0$ and $a_3 a_2 = 1 \cdot 2 > 0$). Thus, we form the polynomial q according to the table, such that

$$q(s) = 4s^2 + 5s + 6. \tag{3.24}$$

Again, the first two statements of Theorem 3.4 (but now for the polynomial q in (3.24)) are easily verified, after which we construct the polynomial in statement *3* as

$$r(s) = 25s + 30, \tag{3.25}$$

which again follows from the table above. As the only root of r in (3.25) reads $\lambda = -\frac{30}{25}$ and thus satisfies $\mathrm{Re}(\lambda) < 0$, r is a stable polynomial. Hence, as we have followed the procedure of Theorem 3.4 repeatedly, we can subsequently conclude that q in (3.24) is stable and that the original polynomial p in (3.23) is stable. ◇

Example 3.3. Consider the polynomial

$$p(s) = s^4 + 2s^3 + 2s^2 + 2s + 3, \tag{3.26}$$

leading to the following Routh table:

	s^4	s^3	s^2	s	1		
$2\times$	1	2	2	2	3		(p)
$1\times$	2		2				
		4	2	4	6	$-$	(q)
$1\times$		2	1	2	3	$\div 2$	(q')
$2\times$		1		3			
			1	-4	3	$-$	(r)

Again, the table follows from repeated application of Theorem 3.4. Specifically, after verifying that the first two statements of this theorem hold for p in (3.26), we form q as in the above table. At this point, it is noted that the roots of a polynomial (and therefore its stability properties) remain unchanged after a scaling of the coefficients. This allows us to simplify the polynomial q to

$$q'(s) = 2s^3 + s^2 + 2s + 3. \tag{3.27}$$

The application of Theorem 3.4 to q' leads (after verifying the first two statements) to

$$r(s) = s^2 - 4s + 3. \tag{3.28}$$

It is clear that the second statement of Theorem 3.4 does not hold for r. Consequently, r is not a stable polynomial, after which we can conclude that neither q', q, nor p is stable. ◇

The Routh-Hurwitz criterion in Theorem 3.4 can be used to immediately derive the following result on stability of quadratic polynomials.

Lemma 3.5. *Let p be a quadratic polynomial*

$$p(s) = a_2 s^2 + a_1 s + a_0, \tag{3.29}$$

with $a_0, a_1, a_2 \in \mathbb{R}$ and $a_2 \neq 0$. Then, p is stable if and only if all coefficients are nonzero and have the same sign.

Proof. This follows from the application of Theorem 3.4 to (3.29). □

This result of Lemma 3.5 thus shows that stability of a quadratic polynomial can be evaluated directly on the basis of its coefficients, allowing for stopping the iterations in a Routh table until a quadratic polynomial is obtained. Specifically, using Lemma 3.5, we could have concluded stability of the polynomial p in Example 3.2 already after computing q, see (3.24).

Whereas Lemma 3.5 gives a necessary and sufficient condition for stability on the basis of the signs of the coefficients of a quadratic polynomial, necessity also holds for general polynomials. This is stated next.

Lemma 3.6. *Let p be a polynomial as in (3.17) with $a_0, \ldots, a_n \in \mathbb{R}$. If p is stable, then all its coefficients are nonzero and have the same sign.*

Proof. Without loss of generality, we assume that p is monic. Namely, this can always be achieved by rescaling the coefficients without affecting stability properties. Since the coefficients are real, complex roots of p appear as complex conjugate pairs. Consequently, p can be factorized into products of the form $(s + \mu)$ for a real root $-\mu$ and the form

$$\big(s + \sigma + i\omega\big)\big(s + \sigma - i\omega\big) = (s + \sigma)^2 + \omega^2 \tag{3.30}$$

for complex conjugate roots $-\sigma \pm i\omega$. As $\mu > 0$ and $\sigma > 0$ by stability of p, it can be concluded that p must have only positive coefficients. □

The use of this lemma is illustrated by the following example.

Example 3.4. Consider the polynomial

$$p(s) = s^5 + 2s^4 + 2s^3 + s^2 + s + 3, \tag{3.31}$$

and the corresponding Routh table:

	s^5	s^4	s^3	s^2	s	1	
$2\times$	1	2	2	1	1	3	(p)
$1\times$	2		2		3		
		4	3	2	-1	6	$-$ (q)

Thus, forming the polynomial q in statement *3* of Theorem 3.4 (after verifying the first two statements) leads to

$$q(s) = 4s^4 + 3s^3 + 2s^2 - s + 1, \tag{3.32}$$

which, by Lemma 3.6, is not stable as not all coefficients have the same sign. Hence, p in (3.31) is not stable. ◇

3.3 Interval polynomials

Whereas the Routh-Hurwitz criterion allows one to verify the stability of a given polynomial, there are applications in which the coefficients of the polynomial are not known exactly, e.g., when physical parameters in systems are uncertain. When it is known that the coefficients lie within a certain interval, it is however still possible to efficiently verify stability.

Specifically, let a_i^-, $a_i^+ \in \mathbb{R}^n$, $i = 0, 1, \ldots, n$, satisfy $a_i^- \leq a_i^+$ and define the set of polynomials $\mathcal{P}$ (again in indeterminate s) as

$$\begin{aligned}\mathcal{P}(s) = \big\{ a_n s^n + a_{n-1} s^{n-1} + \ldots + a_1 s + a_0 \;\big|\; \\ a_i^- \leq a_i \leq a_i^+ \text{ for all } i \in \{0, 1, \ldots, n\} \big\}.\end{aligned} \tag{3.33}$$

Then, stability of the set $\mathcal{P}$ is defined as follows.

Definition 3.3. *The set of polynomials $\mathcal{P}$ as in (3.33) is called stable if each polynomial in the set is stable, i.e., if p is stable for all $p \in \mathcal{P}$.*

The following result, known as Kharitonov's theorem, gives a necessary and sufficient condition for stability of a set of polynomials.

Theorem 3.7. *The set of polynomials $\mathcal{P}$ in (3.33) is stable if and only if the following four polynomials are all stable:*

$$p^{++}(s) = a_0^+ + a_1^+ s + a_2^- s^2 + a_3^- s^3 + a_4^+ s^4 + a_5^+ s^5 + a_6^- s^6 + \ldots, \tag{3.34}$$
$$p^{+-}(s) = a_0^+ + a_1^- s + a_2^- s^2 + a_3^+ s^3 + a_4^+ s^4 + a_5^- s^5 + a_6^- s^6 + \ldots, \tag{3.35}$$
$$p^{-+}(s) = a_0^- + a_1^+ s + a_2^+ s^2 + a_3^- s^3 + a_4^- s^4 + a_5^+ s^5 + a_6^+ s^6 + \ldots, \tag{3.36}$$
$$p^{--}(s) = a_0^- + a_1^- s + a_2^+ s^2 + a_3^+ s^3 + a_4^- s^4 + a_5^- s^5 + a_6^+ s^6 + \ldots. \tag{3.37}$$

Proof. The proof of the *only if* part is immediate as the polynomials (3.34)–(3.37) are all elements of the set $\mathcal{P}$. The proof of the converse is more involved and beyond the scope of these notes. Intuitively, the four polynomials (3.34)–(3.37) represent the "corners" of the set of polynomials (3.33). Details can be found in [7]. □

Remark 3.4. Note that in each of the four polynomials (3.34)–(3.37), there is a pattern of two subsequent upper bounds (++) followed by two lower bounds (−−). There are four different ways to start such pattern, which exactly gives the four so-called Kharitonov polynomials.

If a coefficient a_i is fixed, we just set the lower and upper bound to be equal, i.e., $a_i^- = a_i^+$. ◁

The importance of Kharitonov's theorem (Theorem 3.7) follows from its simplicity. Essentially, it reduces checking stability of infinitely many polynomials to checking stability of only four representative polynomials. For the latter, the Routh-Hurwitz criterion in Theorem 3.4 can be used.

We illustrate this result by an example.

Example 3.5. Consider the set of polynomials $\mathcal{P}$ as in (3.33) with parameters

$$\begin{array}{ll} a_0^- = 15, & a_0^+ = 19, \\ a_1^- = 20, & a_1^+ = 24, \\ a_2^- = 2, & a_2^+ = 3, \\ a_3^- = 1, & a_3^+ = 2, \end{array} \tag{3.38}$$

which is sometimes written compactly as

$$\mathcal{P}(s) = [15, 19] + [20, 24]s + [2, 3]s^2 + [1, 2]s^3. \tag{3.39}$$

The four Kharitonov polynomials are given by

$$p^{++}(s) = 19 + 24s + 2s^2 + s^3, \tag{3.40}$$
$$p^{+-}(s) = 19 + 20s + 2s^2 + 2s^3, \tag{3.41}$$
$$p^{-+}(s) = 15 + 24s + 3s^2 + s^3, \tag{3.42}$$
$$p^{--}(s) = 15 + 20s + 3s^2 + 2s^3, \tag{3.43}$$

of which stability can be checked using the Routh-Hurwitz criterion. When doing so, it can be shown that all four polynomials are stable, such that the set of polynomials $\mathcal{P}$ with coefficients (3.38) is stable as in Definition 3.3. ◇

3.4 Exercises

Exercise 3.1. Prove Lemma 3.1.

Exercise 3.2. Let A be given as

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

Show that the system $\dot{x}(t) = Ax(t)$ is asymptotically stable if and only if the systems $\dot{x}_1(t) = A_{11}x_1(t)$ and $\dot{x}_2(t) = A_{22}x_2(t)$ are both asymptotically stable.

Exercise 3.3. Is the linear system

$$\begin{aligned}\dot{x}_1(t) &= -x_2(t) + x_3(t),\\ \dot{x}_2(t) &= x_1(t) + x_4(t),\\ \dot{x}_3(t) &= -x_4(t),\\ \dot{x}_4(t) &= x_3(t),\end{aligned}$$

asymptotically stable?

Exercise 3.4. Consider the linearized model of the satellite in geostationary orbit of Example 1.5. Take $u = 0$, leading to the autonomous system

$$\dot{\tilde{x}}(t) = A\tilde{x}(t), \qquad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\Omega^2 & 0 & 0 & 2\Omega R_0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2\Omega}{R_0} & 0 & 0 \end{bmatrix}. \tag{3.44}$$

Show that this system is not asymptotically stable.

Exercise 3.5. Consider the linear system $\dot{x} = Ax + Bu$ with

$$A = \begin{bmatrix} 1 & -3 \\ 2 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.45}$$

a. Let $u(t) = 0$ for all t and determine whether the resulting system is (asymptotically) stable.

b. Determine a matrix $F \in \mathbb{R}^{1\times 2}$ such that choosing the *feedback* $u(t) = Fx(t)$ makes the resulting system asymptotically stable. Stated differently, find F such that the system $\dot{x} = (A + BF)x$ is asymptotically stable.

Exercise 3.6. Let $A \in \mathbb{R}^{n\times n}$ be such that $\sigma(A) \subset \mathbb{C}_-$. Show that

$$P = \int_0^\infty e^{A^{\mathrm{T}}t} Q e^{At}\, \mathrm{d}t$$

with $Q = Q^{\mathrm{T}}$ satisfies the matrix equation

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

Hint. Use Lemma 2.4.

Exercise 3.7. Give the details of the proof of Lemma 3.5.

Exercise 3.8. Consider the polynomial

$$p(s) = s^3 + a_2 s^2 + a_1 s + a_0.$$

Using the Routh-Hurwitz criterion, give necessary and sufficient conditions on $a_0, a_1, a_2 \in \mathbb{R}$ for p to be stable.

Exercise 3.9. Consider the polynomial

$$p(s) = s^n + 2s^{n-1} + 3s^{n-2} + \ldots + ns + (n+1).$$

Show that p cannot be stable if $n \geq 4$.

Exercise 3.10. Consider the polynomial

$$p(s) = a_n s^n + a_{n-1} s^{n-1} + \ldots + a_1 s + a_0.$$

a. Show that the polynomial p is stable if and only if the polynomial

$$q(s) = a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n$$

is stable.

b. Show that this result can be helpful for verifying stability of

$$p(s) = 8s^6 + \alpha s^5 + 4s^4 + 2s^3 + 3s^2 + s + 1$$

where $\alpha \in \mathbb{R}$.

Exercise 3.11. For which values of $\alpha \in \mathbb{R}$ is the polynomial

$$p(s) = s^3 + 3s^2 + 3s + \alpha$$

stable?

Exercise 3.12. Is the polynomial

$$p(s) = s^3 + 2s^2 + 4s + \alpha$$

stable for some $\alpha \in [1, 9]$?

Exercise 3.13. Is the set of polynomials given as

$$\mathcal{P}(s) = \left\{ s^3 + 3s^2 + a_1 s + a_0 \;\middle|\; 3 \leq a_1 \leq 4,\, 1 \leq a_0 \leq 2 \right\}$$

stable?

Chapter 4

Controllability and observability

In this chapter we focus on two fundamental system theoretic properties: controllability and observability.

Roughly speaking, controllability (Section 4.1) is related to the question to what extent the state trajectories of a linear system can be influenced through the input. On the other hand, observability (Section 4.2) asks for the extent at which the state trajectories influence the output. Given the importance of these concepts, Sections 4.3 and 4.4 give special forms in which controllability and observability properties are easily seen from the system matrices. Finally, Section 4.5 discusses controllability and observability of eigenvalues rather than systems.

4.1 Controllability

In this section, we focus on the relation between the input u and the state x of the linear system

$$\mathbf{\Sigma}(A, B): \quad \dot{x}(t) = Ax(t) + Bu(t) \tag{4.1}$$

where $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$.

We have derived in Chapter 2 that the state trajectory for an initial state $x_0 \in \mathbb{R}^n$ and input function $u(\cdot)$ can be written explicitly as

$$x(t; x_0, u) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau. \tag{4.2}$$

In our analysis of the influence of the input on the state, we are specifically interested in whether any state x_f can be reached from any initial state x_0 by choosing a suitable input function.

We first consider the following example.

Example 4.1. Consider the system $\mathbf{\Sigma}(A, B)$ with

$$A = \begin{bmatrix} -2 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{4.3}$$

In Example 2.3, it was shown that

$$e^{At} = \begin{bmatrix} e^{-2t} & \frac{1}{3}(e^t - e^{-2t}) \\ 0 & e^t \end{bmatrix}, \tag{4.4}$$

such that the general solution (4.2) for $x_0 = 0$ is given as

$$x(t; x_0, u) = \int_0^t \begin{bmatrix} e^{-2(t-\tau)} \\ 0 \end{bmatrix} u(\tau)\, \mathrm{d}\tau. \tag{4.5}$$

From this explicit computation, it is clear that the input u can never influence the second component in the state x. As a result, there does not exist an input $u(\cdot)$ that steers the system $\mathbf{\Sigma}(A, B)$ from $x_0 = 0$ to a state

$$x_f = \begin{bmatrix} x_{f,1} \\ x_{f,2} \end{bmatrix} \tag{4.6}$$

where $x_{f,2} \neq 0$. ◇

Contrary to the system in the above example, we are interested in systems for which the input can steer the system to an arbitrary state. This property is known as *controllability* and is defined as follows.

Definition 4.1. *The system (4.1) is called controllable (at time $T > 0$) if, for any initial state $x(0) = x_0 \in \mathbb{R}^n$ and any final state $x_f \in \mathbb{R}^n$, there exists an input function $u : [0, T] \to \mathbb{R}^m$ such that $x_f = x(T; x_0, u)$.*

In order to find conditions for controllability, we first restrict attention to a more restrictive notion. Specifically, we will characterize all states x_f that can be reached from an initial condition $x(0) = 0$, leading to the following definition.

Definition 4.2. *Consider the system (4.1). A state $x_f \in \mathbb{R}^n$ is called reachable (at time $T > 0$) if there exists an input function $u : [0, T] \to \mathbb{R}^m$ such that $x_f = x(T; 0, u)$. The reachable subspace (at time T) $\mathcal{W}_T$ is the collection of all reachable states, i.e.,*

$$\mathcal{W}_T = \left\{ \int_0^T e^{A(T-\tau)} B u(\tau)\, \mathrm{d}\tau \;\middle|\; u : [0, T] \to \mathbb{R}^m \right\}. \tag{4.7}$$

It is easily verified that the reachable subspace (4.7) is indeed a linear *subspace* (of $\mathbb{R}^n$), which is the result of linearity of the system (4.1). We also note that (4.2) is used to define the reachable subspace.

Now, using the definition of a reachable state and the reachable subspace as above, we can define reachability of the *system* (4.1).

Definition 4.3. *The system (4.1) is called reachable (at time $T > 0$) if any state $x_f \in \mathbb{R}^n$ is reachable, i.e., if $\mathcal{W}_T = \mathbb{R}^n$.*

Comparing the definitions of controllability and reachability in Definition 4.1 and 4.3, respectively, we immediately obtain the following result.

Lemma 4.1. *The system (4.1) is controllable (at time T) if and only if it is reachable (at time T).*

Proof. It is immediate from the definition that controllability implies reachability (as the latter only considers $x_0 = 0$).

To show the converse, let (4.1) be reachable and $T > 0$. The reachable subspace satisfies $\mathcal{W}_T = \mathbb{R}^n$ as a result. Then, for arbitrary x_0, $x_f \in \mathbb{R}^n$, we have that

$$x_f - e^{AT} x_0 \in \mathcal{W}_T, \tag{4.8}$$

such that, by the definition of $\mathcal{W}_T$ in (4.7), there exists an input function $u(\cdot)$ such that (4.2) holds. As x_0 and x_f are arbitrary, the linear system (4.1) is controllable. □

Remark 4.1. In addition to the concepts of controllability in Definition 4.1 and reachability in Definition 4.3, there exists a third notion that deals with the extent to which trajectories of the system $\mathbf{\Sigma}(A, B)$ can be influenced by the control input. Namely, the system is called *null controllable* (at time $T > 0$) if, for every $x_0 \in \mathbb{R}^n$ there exists an input function $u : [0, T] \to \mathbb{R}^m$ such that $x(T; x_0, u) = 0$. Thus, whereas reachability considers trajectories from the origin to an arbitrary state, null controllability reverses this and calls for trajectories from an arbitrary state to the origin. For linear systems $\mathbf{\Sigma}(A, B)$ as in (4.1), it can be shown that null controllability is equivalent to both controllability and reachability. ◁

As controllability and reachability (both at time $T > 0$) are equivalent by Lemma 4.1, the following characterization of the reachable subspace $\mathcal{W}_T$ (recall Definition 4.2) will turn out to be useful.

Theorem 4.2. *Let $v \in \mathbb{R}^n$ and $T > 0$. Then, the following statements are equivalent:*

1. *$v \perp \mathcal{W}_T$ (i.e., $v^{\mathrm{T}} x = 0$ for all $x \in \mathcal{W}_T$),*
2. *$v^{\mathrm{T}} e^{At} B = 0$ for $0 \leq t \leq T$,*
3. *$v^{\mathrm{T}} A^k B = 0$ for $k = 0, 1, 2, \ldots$,*
4. *$v^{\mathrm{T}} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0$.*

Proof. 1 ⇔ 2. Let $v \in \mathbb{R}^n$ be such that $v^{\mathrm{T}} x = 0$ for all $x \in \mathcal{W}_T$. Then, by the definition of $\mathcal{W}_T$ in (4.7), it follows that

$$\int_0^T v^{\mathrm{T}} e^{A(T-\tau)} B u(\tau) \, \mathrm{d}\tau = 0, \tag{4.9}$$

for *every* input function $u(\cdot)$. Choose $u(t) = B^{\mathrm{T}} e^{A^{\mathrm{T}}(T-\tau)} v$ for $t \in [0, T]$, such that (4.9) gives

$$0 = \int_0^T v^{\mathrm{T}} e^{A(T-\tau)} B B^{\mathrm{T}} e^{A^{\mathrm{T}}(T-\tau)} v \, \mathrm{d}\tau = \int_0^{\mathrm{T}} \left| B^{\mathrm{T}} e^{A^{\mathrm{T}}(T-\tau)} v \right|^2 \mathrm{d}\tau. \tag{4.10}$$

Thus, it can be concluded that $B^{\mathrm{T}} e^{A^{\mathrm{T}} t} v = 0$ for all $0 \leq t \leq T$, which implies *2* after taking the transpose. Conversely, after assuming that *2* holds, it immediately follows that (4.9) is satisfied. This implies *1* by the definition of $\mathcal{W}_T$.

2 $\Leftrightarrow$ *3*. This follows immediately after recalling the definition of the matrix exponential e^{At}.

3 $\Rightarrow$ *4*. The evaluation of the product $v^{\mathrm{T}} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$ yields terms of the form $v^{\mathrm{T}}A^kB$ for $k = 0, 1, \ldots, n-1$, which equal zero by *3*.

4 $\Rightarrow$ *3*. By the Cayley-Hamilton theorem, a matrix A satisfies its own characteristic polynomial. Thus, there exist coefficients a_i, $i = 0, 1, \ldots, n$ such that

$$a_n A^n + a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 I = 0, \tag{4.11}$$

where $a_n = 1$, see also Theorem B.7. Stated differently, A^n can be expressed as a linear combination of $I, A, \ldots, A^{n-1}$. By induction, this holds for all A^k, $k \geq n$, such that $v^{\mathrm{T}}A^kB = 0$ for $k = 0, 1, \ldots, n-1$ (as given by condition *4*) implies that $v^{\mathrm{T}}A^kB$ for all $k = 0, 1, 2, \ldots$. □

The result of Theorem 4.2 has two important consequences. First, as conditions *3* and *4* are independent of T, it follows that also the reachable subspace $\mathcal{W}_T$ is independent of T. Thus, if a state x_f is reachable for some $T > 0$, it is in fact reachable for *any* $T > 0$. Second, the condition *4* provides an explicit characterization of the reachable subspace in terms of the matrices A and B.

These observations are important enough to state as a result.

Corollary 4.3. *The reachable subspace $\mathcal{W}_T$ is independent of T for $T > 0$. Specifically, it satisfies*

$$\mathcal{W}_T = \operatorname{im} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}. \tag{4.12}$$

Because of the above, we will often write $\mathcal{W}$ instead of $\mathcal{W}_T$. Here, we recall that the *image* of a matrix $M \in \mathbb{R}^{p \times q}$, denoted $\operatorname{im} M$, is the subspace defined as

$$\operatorname{im} M = \{ y \in \mathbb{R}^p \mid y = Mx \text{ for some } x \in \mathbb{R}^q \}. \tag{4.13}$$

The reachable subspace $\mathcal{W}$ can be further characterized in terms of the system matrices A and B. To this end, we recall the definition of A-invariance.

Definition 4.4. *Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $\mathcal{V} \subset \mathbb{R}^n$ be a subspace. Then, $\mathcal{V}$ is called A-invariant if the following implication holds:*

$$x \in \mathcal{V} \quad \Longrightarrow \quad Ax \in \mathcal{V}. \tag{4.14}$$

The above implication is also written as $A\mathcal{V} \subset \mathcal{V}$.

Now, the following result can be stated.

Theorem 4.4. *The reachable subspace $\mathcal{W}$ is the smallest A-invariant subspace containing* $\operatorname{im} B$.

Proof. Note that the statement of the theorem essentially contains three statements, i.e., $\mathcal{W}$ contains $\operatorname{im} B$, $\mathcal{W}$ is A-invariant, and $\mathcal{W}$ is the smallest subspace with these properties. We will prove these three aspects independently.

$\mathcal{W}$ *contains* $\operatorname{im} B$. Let $x \in \mathbb{R}^n$ be an arbitrary vector in $\operatorname{im} B$, i.e., $x = Bu$ for some $u \in \mathbb{R}^m$. It is immediate that $x = Bu + AB \cdot 0 + \ldots + A^{n-1}B \cdot 0 \in \mathcal{W}$,

where the result (4.12) is recalled. As $x \in \operatorname{im} B$ is arbitrary, it follows that $\operatorname{im} B \subset \mathcal{W}$.

$\mathcal{W}$ is A-invariant. Take $x \in \mathcal{W}$. Then, there exists $u_0, \ldots, u_{n-1} \in \mathbb{R}^m$ such that

$$x = Bu_0 + ABu_1 + \ldots + A^{n-1}Bu_{n-1}, \tag{4.15}$$

see (4.12). Then, the computation of Ax gives

$$Ax = ABu_0 + A^2Bu_1 + \ldots + A^{n-1}Bu_{n-2} + A^nBu_{n-1}, \tag{4.16}$$

where it is clear that $A^kBu_{k-1} \in \mathcal{W}$ for $k = 1, \ldots, n-1$. However, due to the Cayley-Hamilton theorem, we also have $A^nBu_{n-1} \in \mathcal{W}$. As $\mathcal{W}$ is a subspace, we conclude that $Ax \in \mathcal{W}$, i.e., $\mathcal{W}$ is A-invariant.

$\mathcal{W}$ is the smallest subspace with the above properties. To prove this, let $\mathcal{V} \subset \mathbb{R}^n$ be an A-invariant subspace containing $\operatorname{im} B$, that is, $A\mathcal{V} \subset \mathcal{V}$ and $\operatorname{im} B \subset \mathcal{V}$. We need to show that this implies that $\mathcal{W} \subset \mathcal{V}$.

To this end, take $x \in \operatorname{im} B$. Then, as $\operatorname{im} B \subset \mathcal{V}$ and by A-invariance of $\mathcal{V}$, we have $Ax \in \mathcal{V}$. Stated differently, $A \operatorname{im} B \subset \mathcal{V}$. This process can be repeated to obtain $A^k \operatorname{im} B \subset \mathcal{V}$ for all $k = 0, 1, \ldots$. Then, we also have

$$\begin{aligned} \mathcal{V} &\supset \operatorname{im} B + A \operatorname{im} B + \ldots + A^{n-1} \operatorname{im} B \\ &= \operatorname{im} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \\ &= \mathcal{W}, \end{aligned} \tag{4.17}$$

which finalizes the proof. □

This result thus gives a geometric interpretation of the reachable subspace $\mathcal{W}$, that will be exploited later.

Recalling the results of this section leads to the following summary on controllability of the system $\mathbf{\Sigma}(A, B)$ in (4.1).

Corollary 4.5. *The following statements are equivalent:*

1. *there exists $T > 0$ such that the system $\mathbf{\Sigma}(A, B)$ is controllable at T;*
2. *the system $\mathbf{\Sigma}(A, B)$ is controllable at T for all $T > 0$;*
3. $\operatorname{rank} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = n$;
4. $\mathcal{W} = \mathbb{R}^n$.

Proof. This is a direct consequence of Lemma 4.1 and Corollary 4.3. □

Remark 4.2. Because of the equivalence between statements *1* and *2*, we typically omit the explicit dependence on T in the definition of controllability (see Definition 4.1) and we call a system $\mathbf{\Sigma}(A, B)$ controllable if one of the equivalent conditions of Corollary 4.5 holds. Moreover, as controllability properties only depend on the matrices A and B, we also speak of controllability of the *matrix pair* (A, B). ◁

We can now apply the above results to the system in Example 4.1 as follows.

Example 4.2 (Continuation of Example 4.1). We have already seen in Example 4.1 that certain states can not be influenced through the external input, suggesting that the system $\mathbf{\Sigma}(A, B)$ with matrices (4.3) is not controllable. This can be confirmed by Corollary 4.5 by computing

$$\begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 0 & 0 \end{bmatrix}, \tag{4.18}$$

such that $\operatorname{rank}[B \ AB] = 1 < n = 2$, i.e., the system is indeed not controllable. In fact, the reachable subspace equals

$$\mathcal{W} = \operatorname{im} \begin{bmatrix} B & AB \end{bmatrix} = \operatorname{span} \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \tag{4.19}$$

which corresponds with the intuition from Example 4.1 that only the first state component can be influenced. ◇

In the previous example, the fact that the system is not controllable can also be observed from the structure of the matrices A and B. This is generally not the case, as illustrated next.

Example 4.3. Consider the system $\mathbf{\Sigma}(A, B)$ as in (4.1) with

$$A = \begin{bmatrix} -2 & -6 \\ 2 & 5 \end{bmatrix}, \qquad B = \begin{bmatrix} -3 \\ 2 \end{bmatrix}. \tag{4.20}$$

A direct computation shows that

$$\begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} -3 & -6 \\ 2 & 4 \end{bmatrix}, \tag{4.21}$$

such that

$$\mathcal{W} = \operatorname{im} \begin{bmatrix} -3 & -6 \\ 2 & 4 \end{bmatrix} = \operatorname{span} \left\{ \begin{bmatrix} -3 \\ 2 \end{bmatrix} \right\}, \tag{4.22}$$

and the system is not controllable as $\mathcal{W} \neq \mathbb{R}^2$. This fact can also observed by noting that the reachable subspace $\mathcal{W}$ can also be written as $\mathcal{W} = \{x \in \mathbb{R}^2 \mid v^{\mathrm{T}} x = 0\}$ for $v^{\mathrm{T}} = \begin{bmatrix} 2 & 3 \end{bmatrix}$. Then, after defining $z = v^{\mathrm{T}} x$, it follows that $\dot{z}(t) = z(t)$, i.e., the dynamics of z is independent of the input u. As a result, for initial conditions x_0 such that $v^{\mathrm{T}} x_0 = 0$, we have that $v^{\mathrm{T}} x(t) = 0$ for all $t \geq 0$, indicating that the input does not fully influence the system. Note that the vector v defined here plays the same role as the vector v in Theorem 4.2. ◇

Finally, we return to the mass-spring-damper system of Example 1.1.

Example 4.4. Consider the mass-spring-damper system in Figure 1.2 as a system of the form $\mathbf{\Sigma}(A, B)$ with

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} \tag{4.23}$$

and $m > 0$. Then, we have

$$\begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{m} \\ \frac{1}{m} & -\frac{c}{m^2} \end{bmatrix}, \tag{4.24}$$

which has full rank ($n = 2$) for all values of the spring and damping constants k and c. Thus, the system is controllable. ◇

The above example stresses an important point. Namely, controllability implies (by Definition 4.1), that any state $x_f \in \mathbb{R}^n$ can be reached (from any initial state), but it does not guarantee that one can *stay* at x_f. This is easily seen from the mass-spring-damper example, as one can not stay at a given position for a given (nonzero) velocity (recall that both the position q and the velocity $\dot{q}$ form the state x).

4.2 Observability

Whereas the previous section focused on the relation between the input u and state x of a linear system, we consider the relation between the state x and output y in the current section. Therefore, we consider systems of the form

$$\mathbf{\Sigma}(A, C): \begin{cases} \dot{x}(t) = Ax(t), \\ y(t) = Cx(t), \end{cases} \tag{4.25}$$

with $x(t) \in \mathbb{R}^n$ and $y(t) \in \mathbb{R}^p$. For linear systems representing physical behavior, the output y is generally available for measurements, whereas direct knowledge of the state x is unavailable. We therefore want to investigate the extent to which it is possible to reconstruct the state x in case the output y is known.

To this end, we first recall from Chapter 2 that the output trajectory of (4.25) for a given initial condition $x_0 \in \mathbb{R}^n$ is given by

$$y(t; x_0, 0) = Ce^{At}x_0. \tag{4.26}$$

The following example shows that the reconstruction of the state is not always possible.

Example 4.5. Consider the system $\mathbf{\Sigma}(A, C)$ with

$$A = \begin{bmatrix} -2 & 0 \\ 3 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{4.27}$$

It can be shown that

$$e^{At} = \begin{bmatrix} e^{-2t} & 0 \\ 3(e^{-t} - e^{-2t}) & e^{-t} \end{bmatrix}, \tag{4.28}$$

such that the output trajectory (4.26) with $x_0 = [x_{0,1} \ x_{0,2}]^{\mathrm{T}}$ reads

$$y(t, x_0, 0) = e^{-2t}x_{0,1}. \tag{4.29}$$

Thus, the initial condition for the second state component $x_{0,2}$ does not influence the output trajectory and, therefore, it will be impossible to reconstruct $x_{0,2}$ on the basis of the output trajectory (4.29). We note that this can also be concluded from the structure of the matrix A. Namely, only the first state component is measured and the dynamics of this state component is independent of the second state component. ◇

A different way of expressing the independence of the output trajectory on certain initial conditions (as in the example above) is to say that there are distinct initial conditions that lead to the same output trajectory. Such states will be called indistinguishable, as formalized next.

Definition 4.5. *Consider the system (4.25). Two states x_0 and x_0' in $\mathbb{R}^n$ are called indistinguishable on the interval $[0,T]$ with $T>0$ if they lead to the same output trajectories, i.e.,*

$$y(t;x_0,0) = y(t;x_0',0), \tag{4.30}$$

for all $t \in [0,T]$.

The notion of indistinguishable states can be used to define observability, which is a *system* property.

Definition 4.6. *The system (4.25) is called observable on the interval $[0,T]$ if any two states x_0, $x_0' \in \mathbb{R}^n$ are indistinguishable on $[0,T]$ only if $x_0 = x_0'$.*

Thus, a system is observable if distinct initial conditions lead to distinct output trajectories (on the interval $[0,T]$). Similar to the controllability case, we will first characterize all states that are indistinguishable from the origin.

Definition 4.7. *The unobservable subspace (at time T) $\mathcal{N}_T$ is the collection of all states that are indistinguishable from 0 on the interval $[0,T]$, i.e.,*

$$\mathcal{N}_T = \left\{ x \in \mathbb{R}^n \;\middle|\; Ce^{At}x = 0 \text{ for all } t \in [0,T] \right\} \tag{4.31}$$

It is readily checked that $\mathcal{N}_T$ is indeed a subspace (of $\mathbb{R}^n$). Moreover, the following result gives an explicit characterization of the unobservable subspace.

Theorem 4.6. *The unobservable subspace $\mathcal{N}_T$ is independent of T for $T>0$. Specifically, it satisfies*

$$\mathcal{N}_T = \ker \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}. \tag{4.32}$$

Proof. Let $x \in \mathcal{N}_T$. Then, by the definition in (4.31), x is such that

$$Ce^{At}x = 0, \quad \forall\, t \in [0,T]. \tag{4.33}$$

After taking the transpose, this condition reads $x^{\mathrm{T}} e^{A^{\mathrm{T}} t} C^{\mathrm{T}} = 0$ and we recognize condition *2* in Theorem 4.2 (with $v = x$, $B = C^{\mathrm{T}}$, and A replaced by its transpose). Thus, by Theorem 4.2, (4.33) is equivalent to

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x = 0, \tag{4.34}$$

such that x is contained in the kernel of the above matrix (i.e., we have proven that (4.32) holds when $=$ is replaced by $\subset$).

The converse ($\supset$) also follows from the use of Theorem 4.2. □

As this result shows that the unobservable subspace is independent of T, we will often write $\mathcal{N}$ instead of $\mathcal{N}_T$. In (4.32), we have used $\ker M$ to denote the *kernel* of a matrix $M \in \mathbb{R}^{p\times q}$, i.e.,

$$\ker M = \big\{x \in \mathbb{R}^q \;\big|\; Mx = 0\big\}. \tag{4.35}$$

We note that Theorem 4.6 on the unobservable subspace gives a direct counterpart of Corollary 4.3 on the reachable subspace. In fact, we also have a counterpart of Theorem 4.4, which is stated next without proof.

Theorem 4.7. *The unobservable subspace $\mathcal{N}$ is the largest A-invariant subspace contained in* $\ker C$.

Up till now, we have only characterized states that are indistinguishable from the origin. However, we easily see that two states x_0, $x_0' \in \mathbb{R}^n$ are indistinguishable if and only if $x_0 - x_0'$ is indistinguishable from 0. Namely, for x_0, x_0' indistinguishable (recall Definition 4.5 and (4.26)),

$$Ce^{At}x_0 = Ce^{At}x_0' \quad\Longleftrightarrow\quad Ce^{At}(x_0 - x_0') = 0, \tag{4.36}$$

where we interpret both conditions on the same interval $t \in [0, T]$. Thus, the system $\mathbf{\Sigma}(A, C)$ is observable on the interval $[0, T]$ if and only if $\mathcal{N}_T = \{0\}$.

This observation immediately leads to the following result.

Theorem 4.8. *The following statements are equivalent:*

1. *there exists $T > 0$ such that the system $\mathbf{\Sigma}(A, C)$ is observable at T;*
2. *the system $\mathbf{\Sigma}(A, C)$ is observable at T for all $T > 0$;*
3. $\operatorname{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n$;
4. $\mathcal{N} = \{0\}$.

Proof. This follows from Theorem 4.6 and the equivalence (4.36). □

Similar to the case of controllability (see Remark 4.2), the above result motivates us to omit the dependence on T in the definition of observability and we say that the system $\mathbf{\Sigma}(A, C)$ or the *matrix pair* (A, C) is observable if one of the equivalent conditions in Theorem 4.8 holds.

We now return to Example 4.5.

Example 4.6. In Example 4.5 it was shown that the second state component does not influence the output trajectory, suggesting that the system $\mathbf{\Sigma}(A, C)$ with matrices (4.27) is not observable. After computing

$$\operatorname{rank}\begin{bmatrix} C \\ CA \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 1 & 0 \\ -2 & 0 \end{bmatrix} = 1 < n = 2, \tag{4.37}$$

this is confirmed by Theorem 4.8. The unobservable subspace $\mathcal{N}$ is given as

$$\mathcal{N} = \ker\begin{bmatrix} 1 & 0 \\ -2 & 0 \end{bmatrix} = \operatorname{span}\left\{\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}, \tag{4.38}$$

which indeed shows that the second state component does not influence the output trajectory. ◇

One more example is given as follows.

Example 4.7. Consider the system (4.25) with

$$A = \begin{bmatrix} -11 & 3 \\ -3 & -5 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & -1 \end{bmatrix}. \tag{4.39}$$

Then,

$$\operatorname{rank} \begin{bmatrix} C \\ CA \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 1 & -1 \\ -8 & 8 \end{bmatrix} = 1 < n = 2, \tag{4.40}$$

such that the system is not observable. ◇

By comparing Corollary 4.5 and Theorem 4.8, we observe a close relation between the concepts of controllability and observability. This is known as the *duality* between controllability and observability and is important enough to state as a result.

Theorem 4.9. *The system $\mathbf{\Sigma}(A, B)$ in (4.1) is controllable if and only if the system $\mathbf{\Sigma}(A^{\mathrm{T}}, B^{\mathrm{T}})$ in (4.25) is observable.*

This duality will be exploited frequently in the remainder of these notes.

4.3 Canonical forms for uncontrollable or unobservable systems

In the previous sections, we have analyzed linear systems with inputs (4.1) or outputs (4.25) in a given set of coordinates x. In this section, the effect of a change of coordinates on the properties of controllability and observability will be analyzed.

Specifically, let $T \in \mathbb{R}^{n\times n}$ be nonsingular and consider the change of coordinates

$$\bar{x}(t) = Tx(t). \tag{4.41}$$

Then, by differentiating (4.41) and the use of the dynamics of the system $\mathbf{\Sigma}(A, B)$ in (4.1), it can be seen that the dynamics in the coordinates $\bar{x}$ reads

$$\mathbf{\Sigma}(TAT^{-1}, TB): \quad \dot{\bar{x}}(t) = TAT^{-1}\bar{x}(t) + TBu(t). \tag{4.42}$$

Similarly, we have that the system $\mathbf{\Sigma}(A, C)$ in (4.25) is transformed to

$$\mathbf{\Sigma}(TAT^{-1}, CT^{-1}): \begin{cases} \dot{\bar{x}}(t) = TAT^{-1}\bar{x}(t), \\ y(t) = CT^{-1}\bar{x}(t), \end{cases} \tag{4.43}$$

using the same transformation (4.41).

This motivates the following definition.

Definition 4.8. *Two systems $\mathbf{\Sigma}(A, B)$ and $\mathbf{\Sigma}(\bar{A}, \bar{B})$ of the form (4.1) are called similar if there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$\bar{A} = TAT^{-1}, \qquad \bar{B} = TB. \tag{4.44}$$

Two systems $\mathbf{\Sigma}(A, C)$ and $\mathbf{\Sigma}(\bar{A}, \bar{C})$ of the form (4.25) are called similar if there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\bar{A} = TAT^{-1}, \qquad \bar{C} = CT^{-1}. \tag{4.45}$$

The following theorem shows that similar systems have the same controllability and observability properties, such that these properties are invariant under the choice of coordinates.

Theorem 4.10. *Let $T \in \mathbb{R}^{n\times n}$ be nonsingular. Then, the following hold:*

1. *$\mathbf{\Sigma}(A, B)$ is controllable if and only if $\mathbf{\Sigma}(TAT^{-1}, TB)$ is controllable;*
2. *$\mathbf{\Sigma}(A, C)$ is observable if and only if $\mathbf{\Sigma}(TAT^{-1}, CT^{-1})$ is observable.*

Proof. Only the proof of statement *1* will be given; the proof of *2* follows by duality (see Theorem 4.9).

By Corollary 4.5, controllability of $\mathbf{\Sigma}(TAT^{-1}, TB)$ can be verified by computing the rank of the matrix

$$\begin{aligned}\begin{bmatrix} TB & TAT^{-1}TB & \cdots & (TAT^{-1})^{n-1}TB \end{bmatrix} &= \begin{bmatrix} TB & TAB & \cdots & TA^{n-1}B \end{bmatrix} \\ &= T \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}. \end{aligned} \tag{4.46}$$

As a result, since $T \in \mathbb{R}^{n\times n}$ is nonsingular, we obtain

$$\begin{aligned}&\operatorname{rank} \begin{bmatrix} TB & TAT^{-1}TB & \cdots & (TAT^{-1})^{n-1}TB \end{bmatrix} \\ &\qquad = \operatorname{rank} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}, \end{aligned} \tag{4.47}$$

which proves statement *1* by Corollary 4.5. □

Theorem 4.10 thus shows that the properties of controllability and observability are invariant under similarity transformations. Such invariance can be exploited as various problems are more easily solved in a different set of coordinates. In fact, we have already exploited such similarity in the computation of matrix exponentials in Section 2.2 through the Jordan canonical form. In the remainder of these notes, we will frequently use similarity transformations to solve control problems.

Besides invariance of controllability under coordinate transformations, the proof of Theorem 4.10 gives one other property. Namely, by (4.46) and Corollary 4.3, it is easily seen that

$$T\mathcal{W} = \left\{ \bar{x} \in \mathbb{R}^n \;\middle|\; \bar{x} = Tx,\, x \in \mathcal{W} \right\} \tag{4.48}$$

is the reachable subspace of $\mathbf{\Sigma}(TAT^{-1}, TB)$ (where $\mathcal{W}$ is the reachable subspace of $\mathbf{\Sigma}(A, B)$). For an uncontrollable system, this suggests that a suitable change of coordinates could align the reachable subspace with the first k state components in $\bar{x}$, where $k = \dim \mathcal{W}$. Here, we recall that Example 4.2 has a reachable subspace of this form.

This intuition is confirmed by the following theorem.

Theorem 4.11. *Let $\mathbf{\Sigma}(A, B)$ be uncontrollable and define $k = \dim \mathcal{W} < n$. Then, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$TAT^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad TB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{4.49}$$

where $A_{11} \in \mathbb{R}^{k\times k}$, $B_1 \in \mathbb{R}^{k\times m}$, and the matrix pair (A_{11}, B_1) is controllable.

Proof. To prove the theorem, let $\{q_1, \ldots, q_k\}$ be a basis for $\mathcal{W}$, i.e.,

$$\mathcal{W} = \operatorname{span}\{q_1, \ldots, q_k\}, \tag{4.50}$$

and extend this basis by choosing $q_{k+1}, \ldots, q_n$ such that $\{q_1, \ldots, q_n\}$ is a basis for $\mathbb{R}^n$. We will say that $\{q_1, \ldots, q_n\}$ is a basis for $\mathbb{R}^n$ *adapted to* $\mathcal{W}$.

Now, define $T \in \mathbb{R}^{n\times n}$ such that

$$T^{-1} = \begin{bmatrix} q_1 & q_2 & \cdots & q_k & q_{k+1} & \cdots & q_n \end{bmatrix}. \tag{4.51}$$

Note that, since $\{q_1, \ldots, q_n\}$ is a basis for $\mathbb{R}^n$, the above matrix is nonsingular such that we could indeed define T^{-1}.

Subsequently, consider

$$AT^{-1} = \begin{bmatrix} Aq_1 & Aq_2 & \cdots & Aq_k & Aq_{k+1} & \cdots & Aq_n \end{bmatrix}, \tag{4.52}$$

where we would like to express the right-hand side in the basis $\{q_1, \ldots, q_n\}$. To this end, recall that $\mathcal{W}$ is A-invariant by Theorem 4.4, such that we have $Aq_i \in \mathcal{W}$ for $i = 1, 2, \ldots, k$. In particular, this implies the existence of scalars $\alpha_{ji} \in \mathbb{R}$ such that

$$Aq_i = \sum_{j=1}^{k} \alpha_{ji} q_j, \qquad i = 1, 2, \ldots, k. \tag{4.53}$$

For the terms Aq_i for $i = k+1, \ldots, n$, such property does not hold. However, as $\{q_1, \ldots, q_n\}$ is a basis for $\mathbb{R}^n$, we can write

$$Aq_i = \sum_{j=1}^{n} \alpha_{ji} q_j, \qquad i = k+1, k+2, \ldots, n, \tag{4.54}$$

where again $\alpha_{ji} \in \mathbb{R}$. Rewriting the results (4.53) and (4.54) in matrix form yields

$$\begin{aligned} AT^{-1} &= \begin{bmatrix} Aq_1 & Aq_2 & \cdots & Aq_k & Aq_{k+1} & \cdots & Aq_n \end{bmatrix} \\ &= \begin{bmatrix} q_1 & q_2 & \cdots q_k & q_{k+1} & \cdots & q_n \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1,k} & \alpha_{1,k+1} & \ldots & \alpha_{1,n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2,k} & \alpha_{2,k+1} & \ldots & \alpha_{2,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \alpha_{k,1} & \alpha_{k,2} & \cdots & \alpha_{k,k} & \alpha_{k,k+1} & \ldots & \alpha_{k,n} \\ 0 & 0 & \cdots & 0 & \alpha_{k+1,k+1} & \ldots & \alpha_{k+1,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \alpha_{n,k+1} & \ldots & \alpha_{n,n} \end{bmatrix}. \end{aligned}$$

Thus, after defining

$$A_{11} = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1,k} \\ \vdots & & \vdots \\ \alpha_{k,1} & \cdots & \alpha_{k,k} \end{bmatrix}, \qquad A_{12} = \begin{bmatrix} \alpha_{1,k+1} & \cdots & \alpha_{1,n} \\ \vdots & & \vdots \\ \alpha_{k,k+1} & \cdots & \alpha_{k,n} \end{bmatrix}, \tag{4.55}$$

and

$$A_{22} = \begin{bmatrix} \alpha_{k+1,k+1} & \cdots & \alpha_{k+1,n} \\ \vdots & & \vdots \\ \alpha_{n,k+1} & \cdots & \alpha_{n,n} \end{bmatrix}, \tag{4.56}$$

we obtain

$$AT^{-1} = T^{-1}\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{4.57}$$

which is equivalent to the first statement in (4.49) as T is nonsingular. Note also that $A_{11} \in \mathbb{R}^{k\times k}$.

Next, we show the desired form for TB. To this end, recall again Theorem 4.4, which shows that $\operatorname{im} B \subset \mathcal{W}$. Given the basis $\{q_1, \ldots, q_n\}$ satisfying (4.50), we obtain

$$b_i = \sum_{j=1}^{k} \beta_{ji} q_i, \qquad i = 1, 2, \ldots, m, \tag{4.58}$$

for some $\beta_{ji} \in \mathbb{R}$ and where b_i are the columns of B, i.e., $B = [b_1 \ b_2 \ \cdots \ b_m]$. The result (4.58) can be written as

$$B = \begin{bmatrix} q_1 & q_2 & \cdots & q_k & q_{k+1} & \cdots & q_n \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1,m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2,m} \\ \vdots & \vdots & & \vdots \\ \beta_{k,1} & \beta_{k,2} & \cdots & \beta_{k,m} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{4.59}$$

which is easily observed to be equivalent to the second statement in (4.49) after defining

$$B_1 = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1,m} \\ \vdots & & \vdots \\ \beta_{k,1} & \cdots & \beta_{k,m} \end{bmatrix}. \tag{4.60}$$

Here, we observe that $B_1 \in \mathbb{R}^{k\times m}$.

It remains to be shown that the matrix pair (A_{11}, B_1) is controllable. This follows directly from the fact that $\mathcal{W}$ is the *smallest* A-invariant subspace containing $\operatorname{im} B$ (see Theorem 4.4). To make this explicit, note that

$$k = \operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = \operatorname{rank} T \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}, \tag{4.61}$$

as $T \in \mathbb{R}^{n\times n}$ is nonsingular. After observing that

$$TA^kB = \begin{bmatrix} A_{11}^k B_1 \\ 0 \end{bmatrix} \tag{4.62}$$

due to the block upper-triangular structure of TAT^{-1} and the zero block in TB, we obtain

$$\begin{aligned} k &= \operatorname{rank} T \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \\ &= \operatorname{rank} \begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n-1}B_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \qquad (4.63) \\ &= \operatorname{rank} \begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{k-1}B_1 \end{bmatrix}, \qquad (4.64) \end{aligned}$$

where (4.64) follows from the Cayley-Hamilton theorem. This shows that the pair (A_{11}, B_1) is controllable, finalizing the proof. □

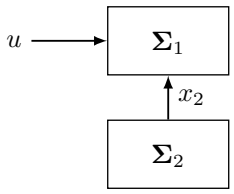


Figure 4.1: Illustration of the system (4.65), where the subsystems $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ represent the dynamics $\dot{\bar{x}}_1(t) = A_{11}\bar{x}_1(t) + A_{12}\bar{x}_2(t) + B_1u(t)$ and $\dot{\bar{x}}_2(t) = A_{22}\bar{x}_2(t)$, respectively.

Remark 4.3. In the statement of Theorem 4.11, note that $B \neq 0$ implies that $k > 0$. Nonetheless, the result also holds for the case $B = 0$, in which case the matrices A_{11} and B_1 are void ($k = 0$). ◁

The result of Theorem 4.11 thus shows that there exists a change of coordinates $\bar{x}(t) = Tx(t)$ such that the dynamics in the coordinates $\bar{x}$ read

$$\mathbf{\Sigma}(TAT^{-1}, TB): \begin{cases} \dot{\bar{x}}_1(t) = A_{11}\bar{x}_1(t) + A_{12}\bar{x}_2(t) + B_1u(t), \\ \dot{\bar{x}}_2(t) = A_{22}\bar{x}_2(t), \end{cases} \tag{4.65}$$

where $\bar{x}_1(t) \in \mathbb{R}^k$ and $\bar{x}_2(t) \in \mathbb{R}^{n-k}$. It is clear from the structure of the dynamics that the input u does not affect the states $\bar{x}_2$. This can also be observed from the reachable subspace of (4.65), which we denote by $\bar{\mathcal{W}}$. Namely, it can be seen that

$$\bar{\mathcal{W}} = \left\{ \bar{x} \in \mathbb{R}^n \;\middle|\; \bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix},\ \bar{x}_2 = 0 \right\}. \tag{4.66}$$

An illustration of the form (4.65) is given in Figure 4.1, indicating that (4.65) can be regarded as the interconnection of two subsystems. Here, the subsystem $\mathbf{\Sigma}_1$ representing the dynamics of $\bar{x}_1$ (under the assumption that $\bar{x}_2 = 0$) is generally referred to as the *controllable subsystem*.

Remark 4.4. It is tempting to call the part describing the dynamics for $\bar{x}_2$ the uncontrollable subsystem, but this is not well-defined. Namely, these dynamics depend on the specific choice of the basis, which is not unique (specifically, it depends on the choice of the vectors $\{q_{k+1}, \ldots, q_n\}$ that extend the basis of $\mathcal{W}$ to a basis of $\mathbb{R}^n$ in the proof of Theorem 4.11). ◁

Remark 4.5. The result of Theorem 4.11 can also be formulated using a more geometric perspective. To this end, recall from Theorem 4.4 that the reachable subspace $\mathcal{W}$ is the smallest A-invariant subspace containing $\operatorname{im} B$. Regarding A as the linear map $A : \mathbb{R}^n \to \mathbb{R}^n$, we can define the linear map $\bar{A} : \mathcal{W} \to \mathcal{W}$ as $\bar{A}x = Ax$ for $x \in \mathcal{W}$ (note that this is well-defined as $\mathcal{W}$ is A-invariant). The linear map $\bar{A}$ is referred to as the *restriction* of A to $\mathcal{W}$ and is sometimes also denoted as $A|_{\mathcal{W}}$. Moreover, as $\operatorname{im} B \subset \mathcal{W}$, we can define $\bar{B} : \mathbb{R}^m \to \mathcal{W}$ satisfying $\bar{B}u = Bu$ for all $u \in \mathbb{R}^m$ as the *codomain restriction* of B to $\mathcal{W}$.

Now, the system

$$\dot{x}(t) = \bar{A}x(t) + \bar{B}u(t), \tag{4.67}$$

with $x(t) \in \mathcal{W}$ represents the controllable subsystem. Note that A_{11} in (4.65) is a matrix representing the linear map $\bar{A}$ for the basis $\{q_1, \ldots, q_k\}$ of $\mathcal{W}$. ◁

We illustrate Theorem 4.11 by means of an example.

Example 4.8. Consider the system $\mathbf{\Sigma}(A, B)$ as in (4.1) with

$$A = \begin{bmatrix} -3 & -1 & 0 \\ -3 & -4 & 2 \\ -10 & -8 & 3 \end{bmatrix}, \qquad B = \begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix}. \tag{4.68}$$

To verify controllability, recall Corollary 4.5 and compute

$$\begin{bmatrix} B & AB & A^2B \end{bmatrix} = \begin{bmatrix} -1 & 1 & -2 \\ 2 & -1 & 1 \\ 2 & 0 & -2 \end{bmatrix}. \tag{4.69}$$

After noting that the columns are not linearly independent (observe that $A^2B + 3AB = -B$), it can be concluded that $\operatorname{rank}[B \ AB \ A^2B] = 2$. Consequently, as $\mathcal{W} = \operatorname{im}\begin{bmatrix} B & AB & A^2B \end{bmatrix}$ by Corollary 4.3, we have from (4.69) that

$$\mathcal{W} = \operatorname{span}\{q_1, q_2\}, \qquad q_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \ q_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \tag{4.70}$$

is a basis for $\mathcal{W}$. Following the proof of Theorem 4.11, we extend this basis to a basis $\{q_1, q_2, q_3\}$ for $\mathbb{R}^3$ by selecting

$$q_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{4.71}$$

Here, note that the choice for q_3 is far from unique. After defining T such that $T^{-1} = [\, q_1 \ q_2 \ q_3 \,]$, a direct computation gives

$$TAT^{-1} = \begin{bmatrix} -2 & 1 & 2 \\ 1 & -1 & 2 \\ 0 & 0 & -1 \end{bmatrix}, \qquad TB = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \tag{4.72}$$

which is indeed of the form (4.49). ◇

Given the duality between controllability and observability, it is no surprise that a result similar to Theorem 4.11 can be stated for unobservable systems. This is formalized in the following theorem, whose proof is omitted as it follows in a similar way as that for the controllability case.

Theorem 4.12. *Let $\mathbf{\Sigma}(A, C)$ be unobservable and define $k = n - \dim\mathcal{N} < n$. Then, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$TAT^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad CT^{-1} = \begin{bmatrix} C_1 & 0 \end{bmatrix}, \tag{4.73}$$

where $A_{11} \in \mathbb{R}^{k\times k}$, $C_1 \in \mathbb{R}^{p\times k}$, and the matrix pair (A_{11}, C_1) is observable.

Thus, there exists a change of coordinates $\bar{x}(t) = Tx(t)$ that transforms the system $\mathbf{\Sigma}(A, C)$ in (4.25) to the form

$$\mathbf{\Sigma}(TAT^{-1}, CT^{-1}) : \begin{cases} \dot{\bar{x}}_1(t) &= A_{11}\bar{x}_1(t), \\ \dot{\bar{x}}_2(t) &= A_{21}\bar{x}_1(t) + A_{22}\bar{x}_2(t), \\ y(t) &= C_1 x_1(t), \end{cases} \tag{4.74}$$

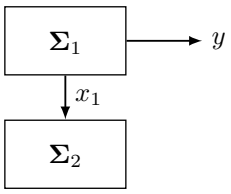


Figure 4.2: Illustration of the system (4.74), where the subsystems $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ represent the dynamics $\dot{\bar{x}}_1(t) = A_{11}\bar{x}_1(t)$ and $\dot{\bar{x}}_2(t) = A_{12}\bar{x}_1(t) + A_{22}\bar{x}_2(t)$, respectively.

with $\bar{x}_1(t) \in \mathbb{R}^k$ and $\bar{x}_2(t) \in \mathbb{R}^{n-k}$. Similar to the case of controllability, this can be regarded as the interconnection between two systems, as illustrated in Figure 4.2. Here, it is clear that the dynamics of $\bar{x}_2$ does not have any influence on the output y, which motivates referring to the system $\mathbf{\Sigma}_2$ (under the assumption that $\bar{x}_1 = 0$) as the *unobservable subsystem*.

Stated more formally, the interconnection in Figure 4.2 shows that initial states for which the $\bar{x}_2$-component is nonzero can not be distinguished from initial states with zero $\bar{x}_2$-component. This is confirmed by computing the unobservable subspace of (4.74) (recall Definition 4.7), denoted by $\bar{\mathcal{N}}$, as

$$\bar{\mathcal{N}} = \left\{ \bar{x} \in \mathbb{R}^n \;\middle|\; \bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}, \, \bar{x}_1 = 0 \right\}. \tag{4.75}$$

4.4 Controllability and observability canonical forms

In the previous section, it is shown that systems that are not controllable or not observable can be put in a form in which the controllable or observable part is easily recognized. In this section, we consider systems that are controllable or observable and present so-called canonical forms that will turn out to be very useful later.

For controllable linear systems with a *single* input, the following result holds.

Theorem 4.13. *Let $\mathbf{\Sigma}(A, B)$ as in (4.1) with $m = 1$ be controllable. Then, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$TAT^{-1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \qquad TB = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{4.76}$$

where $a_0, \ldots, a_{n-1} \in \mathbb{R}$ are the coefficients of the monic characteristic polynomial of A, i.e.,

$$\Delta_A(s) = s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} \ldots + a_1 s + a_0. \tag{4.77}$$

Proof. We will explicitly construct a matrix T that satisfies the desired properties. To do so, introduce the vectors $q_1, q_2, \ldots, q_n \in \mathbb{R}^n$ as

$$\begin{aligned} q_n &= B, && (4.78)\\ q_{n-1} &= AB + a_{n-1}B, && (4.79)\\ q_{n-2} &= A^2B + a_{n-1}AB + a_{n-2}B, && (4.80)\\ &\vdots \\ q_1 &= A^{n-1}B + a_{n-1}A^{n-2}B + \ldots + a_2AB + a_1B, && (4.81) \end{aligned}$$

and note that the above can be written in matrix form as

$$\begin{aligned} &\begin{bmatrix} q_n & q_{n-1} & q_{n-2} & \cdots & q_1 \end{bmatrix} \\ &\quad = \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \begin{bmatrix} 1 & a_{n-1} & a_{n-2} & \cdots & a_1 \\ 0 & 1 & a_{n-1} & & a_2 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \end{aligned} \tag{4.82}$$

Since we consider the single-input case ($m = 1$), we have that the matrix $\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$ is of size $n \times n$. In fact, as the pair (A, B) is controllable, it has full rank (i.e., is nonsingular). Moreover, due to the upper triangular structure of the rightmost matrix in (4.82), this matrix is easily seen to be nonsingular as well, implying that

$$T^{-1} = \begin{bmatrix} q_1 & q_2 & q_3 & \cdots & q_n \end{bmatrix} \tag{4.83}$$

is nonsingular.

To show that the transformation (4.83) leads to the form (4.76), note that

$$Aq_i = q_{i-1} - a_{i-1}B, \tag{4.84}$$

for $i = 2, 3, \ldots, n$. For Aq_1 we obtain from (4.81) that

$$\begin{aligned} Aq_1 &= A^nB + a_{n-1}A^{n-1}B + a_{n-2}A^{n-2}B + \ldots + a_2A^2B + a_1AB \\ &= \big(A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \ldots + a_2A^2 + a_1A + a_0I\big)B - a_0B \\ &= -a_0B, \end{aligned} \tag{4.85}$$

where the result (4.85) follows from recalling that $a_0, \ldots, a_{n-1}$ represent coefficients of the characteristic polynomial of A and the Cayley-Hamilton theorem. Then, after recalling that $q_n = B$, we have

$$\begin{aligned} AT^{-1} &= \begin{bmatrix} Aq_1 & Aq_2 & Aq_3 & \cdots & Aq_{n-1} & Aq_n \end{bmatrix} \\ &= \begin{bmatrix} q_1 & q_2 & q_3 & \cdots & q_{n-1} & q_n \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \end{aligned} \tag{4.86}$$

which is exactly the first equation in (4.76) (recall the definition of T^{-1} in (4.83) and that it is nonsingular).

The second equation in (4.76) follows directly from the definition of T^{-1} in (4.83), finalizing the proof. □

The form (4.76) is parameterized only by the coefficients of the characteristic polynomial, which allows for easily verifying stability properties through the Routh-Hurwitz criterion. Also, we note that it is easily confirmed that the matrix pair given in (4.76) is controllable. Namely, after denoting $\bar{A} = TAT^{-1}$ and $\bar{B} = TB$, we obtain

$$\begin{bmatrix} \bar{B} & \bar{A}\bar{B} & \cdots & \bar{A}^{n-2}\bar{B} & \bar{A}^{n-1}\bar{B} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & & \iddots & 1 & \star \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 0 & 1 & \iddots & & \star \\ 1 & \star & \cdots & \star & \star \end{bmatrix}, \tag{4.87}$$

where a star $\star$ denotes an arbitrary entry. Nonetheless, due to the triangular structure, the matrix above is easily seen to be of full rank, such that controllability is guaranteed by Corollary 4.5.

The following example illustrates the controllable canonical form.

Example 4.9. Consider the linear system $\mathbf{\Sigma}(A, B)$ as in (4.1) with

$$A = \begin{bmatrix} -3 & 2 & -1 \\ -2 & 1 & 0 \\ 3 & -2 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \tag{4.88}$$

and note that $m = 1$, i.e., $\mathbf{\Sigma}(A, B)$ has a single input. Moreover, we have that

$$\operatorname{rank} \begin{bmatrix} B & AB & A^2B \end{bmatrix} = \operatorname{rank} \begin{bmatrix} -1 & 2 & 1 \\ 0 & 2 & -2 \\ 1 & -3 & 2 \end{bmatrix} = 3 = n \tag{4.89}$$

such that $\mathbf{\Sigma}(A, B)$ is controllable. A direct computation shows that

$$\Delta_A(s) = s^3 + 2s^2 + 4s + 1. \tag{4.90}$$

Thus, by Theorem 4.13, the existence of a nonsingular matrix T such that

$$TAT^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -4 & -2 \end{bmatrix}, \qquad TB = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{4.91}$$

is guaranteed. However, we would like to explicitly construct this matrix. To this end, consider the vectors $q_1, q_2, q_3 \in \mathbb{R}^3$ as constructed using (4.78)–(4.81) in the proof of Theorem 4.13. The vectors q_3 and q_2 read

$$q_3 = B = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \tag{4.92}$$

$$q_2 = AB + a_2 B = \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}, \tag{4.93}$$

respectively. We note that the vector AB is already computed in (4.89) and that $a_2 = 2$ follows from the characteristic polynomial (4.90). Similarly,

$$q_1 = A^2B + a_2 AB + a_1 B = \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix} + 2 \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix} + 4 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \tag{4.94}$$

such that the matrix T is defined through its inverse in (4.83) as

$$T^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 2 & 0 \\ 0 & -1 & 1 \end{bmatrix}. \tag{4.95}$$

It can be verified by direct computation that (4.91) holds. ◇

By duality, it is no surprise that a similar form can be found for an observable system $\mathbf{\Sigma}(A, C)$ as in (4.25). As in the controllability case, a *single* output is assumed.

Theorem 4.14. *Let $\mathbf{\Sigma}(A, C)$ as in (4.25) with $p = 1$ be observable. Then, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$TAT^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & -a_0 \\ 1 & 0 & & & 0 & -a_1 \\ 0 & 1 & \ddots & & 0 & -a_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \ddots & 1 & 0 & -a_{n-2} \\ 0 & 0 & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix}, \qquad CT^{-1} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{4.96}$$

where $a_0, \ldots, a_{n-1} \in \mathbb{R}$ are the coefficients of the monic characteristic polynomial of A, i.e.,

$$\Delta_A(s) = s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} \ldots + a_1 s + a_0. \tag{4.97}$$

Proof. This follows immediately from the duality between controllability and observability in Theorem 4.9 after noting that $\Delta_A = \Delta_{A^\mathrm{T}}$. □

4.5 Controllable and observable eigenvalues

In the previous sections, we have studied the notions of controllability and observability as properties of a *system*. In this section, we define notions of controllable and observable *eigenvalues*. These notions will turn out to be useful in the design of stabilizing controllers later.

Specifically, we start with the following definition.

Definition 4.9. *An eigenvalue λ of A is called (A, B)-controllable if*

$$\operatorname{rank} \begin{bmatrix} A - \lambda I & B \end{bmatrix} = n. \tag{4.98}$$

The eigenvalue λ is called (A, C)-observable if

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n. \tag{4.99}$$

We will often, when no confusion can arise, omit the prefix (A, B) and speak of a controllable eigenvalue $\lambda \in \sigma(A)$ instead of an (A, B)-controllable eigenvalue. The same remark holds for observable eigenvalues.

Remark 4.6. The rank conditions above allow for various alternative formulations. For example, instead of (4.98) we can write that for every vector v the implication

$$v^{\mathrm{T}}A = \lambda v^{\mathrm{T}}, \;\; v^{\mathrm{T}}B = 0 \quad \Longrightarrow \quad v = 0. \tag{4.100}$$

Thus, there does not exist a left eigenvector of A corresponding to the eigenvalue λ which is orthogonal to $\operatorname{im} B$. A more geometric alternative is the condition

$$\operatorname{im}(A - \lambda I) + \operatorname{im} B = \mathbb{R}^n. \tag{4.101}$$

Dual interpretations are possible for observable eigenvalues, in which case the geometric condition reads $\ker(A - \lambda I) \cap \ker C = \{0\}$. $\lhd$

The notion of controllable and observable eigenvalues as in Definition 4.9 can directly be related to controllability and observability of systems, see Definitions 4.1 and 4.6. This is stated as follows.

Theorem 4.15. *Consider the systems $\mathbf{\Sigma}(A, B)$ in (4.1) and $\mathbf{\Sigma}(A, C)$ in (4.25).*

1. *$\mathbf{\Sigma}(A, B)$ is controllable if and only if every eigenvalue of A is (A, B)-controllable, i.e., if and only if*

$$\operatorname{rank} \begin{bmatrix} A - \lambda I & B \end{bmatrix} = n \qquad \forall\, \lambda \in \sigma(A). \tag{4.102}$$

2. *$\mathbf{\Sigma}(A, C)$ is observable if and only if every eigenvalue of A is (A, C)-observable, i.e., if and only if*

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n \qquad \forall\, \lambda \in \sigma(A). \tag{4.103}$$

Proof. A proof for statement *2* will be given, after which statement *1* follows by duality (note that λ is (A, B)-controllable if and only if it is $(A^{\mathrm{T}}, B^{\mathrm{T}})$-observable, as follows directly from Definition 4.9).

if. To prove by contraposition, let the pair (A, C) be unobservable. Then, by Theorem 4.8, the unobservable subspace $\mathcal{N}$ is nontrivial (i.e., its dimension satisfies $\dim \mathcal{N} > 0$). As $\mathcal{N}$ is A-invariant, we can define the linear map $A|_{\mathcal{N}} : \mathcal{N} \to \mathcal{N}$ by $x \mapsto Ax$ for $x \in \mathcal{N}$. The linear map has an eigenvalue λ (as $\mathcal{N}$ is nontrivial) and a corresponding eigenvalue $v \in \mathcal{N}$. Thus, we have

$$Av = \lambda v \tag{4.104}$$

for some $v \in \mathcal{N}$ with $v \neq 0$. Moreover, as $\mathcal{N} \subset \ker C$ (see again Theorem 4.7), we also have that $Cv = 0$. Together, this implies

$$\begin{bmatrix} A - \lambda I \\ C \end{bmatrix} v = 0. \tag{4.105}$$

As $v \neq 0$, this means that

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} < n, \tag{4.106}$$

which violates (4.103) and finalizes the proof.

only if. To again set up a proof by contraposition, assume that (4.103) is not satisfied. Then, there exists and eigenvalue $\lambda \in \sigma(A)$ and a corresponding eigenvector $v \neq 0$ such that (4.106) holds, and thus

$$Av = \lambda v, \qquad Cv = 0. \tag{4.107}$$

However, this also implies that $A^k v = \lambda^k v$ and hence $CA^k v = 0$ for all $k = 0, 1, 2, \ldots$, such that the pair (A, C) is not controllable (see the reasoning in the proof of Theorem 4.6)). □

At this point, the result of Theorem 4.15 merely provides an alternative way of verifying whether a system is controllable or observable: the conditions (4.102) and (4.103) are known as the *Hautus test* (or Popov-Belevitch-Hautus test) for controllability and observability, respectively. The main relevance of this result is however that it allows for a clear link to later generalizations (see the notions of stabilizability and detectability in Corollary 5.4 and Theorem 5.8).

It is well-known that eigenvalues are invariant under a similarity transformation, and we have also seen that the notions of controllability and observability (for a system) are invariant under such transformations (see Theorem 4.10). It is therefore no surprise that a similar result holds for controllable and observable eigenvalues.

Theorem 4.16. *Let $T \in \mathbb{R}^{n\times n}$ be nonsingular. Then, the following hold:*

1. *λ is (A, B)-controllable if and only if λ is (TAT^{-1}, TB)-controllable;*
2. *λ is (A, C)-observable if and only if λ is (TAT^{-1}, CT^{-1})-observable;*

Proof. We only prove statement *1.* First, note that $\lambda \in \sigma(A)$ if and only if $\lambda \in \sigma(TAT^{-1})$, which is a property of the similarity transformation. Moreover, we have that

$$\begin{bmatrix} TAT^{-1} - \lambda I & TB \end{bmatrix} = T \begin{bmatrix} A - \lambda I & B \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}, \tag{4.108}$$

such that (as T is nonsingular)

$$\operatorname{rank} \begin{bmatrix} A - \lambda I & B \end{bmatrix} = \operatorname{rank} \begin{bmatrix} TAT^{-1} - \lambda I & TB \end{bmatrix}. \tag{4.109}$$

This finalizes the proof. □

Finally, we relate controllable eigenvalues to the partitioned matrices in the form (4.49) of Theorem 4.11.

Theorem 4.17. *Let $\mathbf{\Sigma}(A, B)$ as in (4.1) be of the form*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{4.110}$$

where $A_{11} \in \mathbb{R}^{k\times k}$, $B_1 \in \mathbb{R}^{k\times m}$ and the matrix pair (A_{11}, B_1) is controllable. Then, $\lambda \in \sigma(A)$ is (A, B)-controllable if and only if $\lambda \notin \sigma(A_{22})$.

Proof. if. We prove by contraposition. Thus, let $\lambda \in \sigma(A)$ be not (A, B)-controllable. Then, there exists a nonzero vector $v \in \mathbb{R}^n$ such that $v^{\mathrm{T}}(\lambda I - A) = 0$ and $v^{\mathrm{T}}B = 0$. After partitioning v as $v^{\mathrm{T}} = [v_1^{\mathrm{T}} \; v_2^{\mathrm{T}}]$, the structure in (4.110) shows that

$$v_1^{\mathrm{T}}(A_{11} - \lambda I) = 0, \qquad v_1^{\mathrm{T}}B_1 = 0, \tag{4.111}$$

$$v_1^{\mathrm{T}}A_{12} + v_2^{\mathrm{T}}(A_{22} - \lambda I) = 0. \tag{4.112}$$

However, as (A_{11}, B_1) is controllable, we have by Theorem 4.15 that (4.111) implies $v_1 = 0$, after which (4.112) reduces to $v_2^{\mathrm{T}}(A_{22} - \lambda I) = 0$. Note also that $v \neq 0$ ensures that $v_2 \neq 0$, such that $\lambda \in \sigma(A_{22})$. This contradicts with the given that $\lambda \notin \sigma(A_{22})$ and hence $\lambda \in \sigma(A)$ is (A, B)-controllable.

only if. To again prove the statement by contraposition, let $\lambda \in \sigma(A_{22})$ and let $v_2 \in \mathbb{R}^{n-k}$ be a corresponding left eigenvector, i.e., $v_2^{\mathrm{T}}A_{22} = \lambda v_2^{\mathrm{T}}$. Then, after defining

$$v = \begin{bmatrix} 0 \\ v_2 \end{bmatrix}, \tag{4.113}$$

we observe that $v \neq 0$ and, due to the structure in (4.110), that

$$v^{\mathrm{T}} \begin{bmatrix} A - \lambda I & B \end{bmatrix} = 0, \tag{4.114}$$

which shows that λ is not (A, B)-controllable by Definition 4.9. Here, we note that $\lambda \in \sigma(A_{22})$ implies that $\lambda \in \sigma(A)$. □

The counterpart on observability is stated without proof.

Theorem 4.18. *Let $\mathbf{\Sigma}(A, C)$ as in (4.25) be of the form*

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad C = \begin{bmatrix} C_1 & 0 \end{bmatrix}, \tag{4.115}$$

where $A_{11} \in \mathbb{R}^{k \times k}$, $C_1 \in \mathbb{R}^{p \times k}$ and the matrix pair (A_{11}, C_1) is observable. Then, $\lambda \in \sigma(A)$ is (A, C)-observable if and only if $\lambda \notin \sigma(A_{22})$.

4.6 Exercises

Exercise 4.1. For each of the following matrix pairs (A, B), determine whether it is controllable. When parameters a_i or b_i are given, investigate the influence of the parameters on controllability.

a. $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

b. $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

c. $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

d. $A = \begin{bmatrix} a_1 & 0 \\ a_2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

e. $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

f. $A = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

g. $A = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

Exercise 4.2. Consider the linearized model of the satellite in geostationary orbit of Example 1.5, i.e., the linear model characterized by the matrices (1.44).

a. Is the system controllable?

b. Is the system observable?

Exercise 4.3. Consider the matrix pair (A, B) with

$$A = \begin{bmatrix} 4 & -4 & 2 \\ 3 & -3 & 2 \\ -3 & 2 & -3 \end{bmatrix}, \qquad B = \begin{bmatrix} 5 \\ 2 \\ -2 \end{bmatrix}.$$

a. Is the pair (A, B) controllable?

b. Give a basis for the controllable subspace $\mathcal{W}$.

c. Verify that the controllable subspace is A-invariant.

Exercise 4.4. Consider the matrices

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

a. Is the pair (A, B) controllable?

b. Is the pair (C, A) observable?

Exercise 4.5. Consider the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \cdots & a_{n-1,n} \\ a_{n,1} & a_{n,2} & a_{n,3} & \cdots & a_{n,n} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Show that the matrix pair (A, B) is controllable if and only if $a_{i,i+1} \neq 0$ for $i = 1, 2, \ldots, n-1$.

Exercise 4.6. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$.

a. Show that (A, B) is controllable if and only if $(A + BF, B)$ is controllable for any matrix $F \in \mathbb{R}^{m\times n}$.

b. Let $R \in \mathbb{R}^{m\times m}$ be nonsingular. Show that (A, B) is controllable if and only if (A, BR) is controllable.

Exercise 4.7. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$.

a. Prove that if

$$\operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{l-1}B \end{bmatrix} = \operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{l}B \end{bmatrix}$$

for some positive integer l, then

$$\operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{l}B \end{bmatrix} = \operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{l+1}B \end{bmatrix}.$$

b. Let $\operatorname{rank} B = r$. Prove that (A, B) is controllable if and only if

$$\operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{n-r}B \end{bmatrix} = n.$$

Exercise 4.8. Prove Theorem 4.9.

Exercise 4.9. Let α, $\beta \in \mathbb{R}$ and consider the matrices

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 0 & 0 & \alpha \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \\ \beta \end{bmatrix}.$$

a. Determine all values of α and β for which (A, B) is controllable.

b. For those values of α and β for which (A, B) is not controllable, determine the uncontrollable eigenvalues.

Exercise 4.10. For the following matrix pairs, find the transformation T^{-1} that puts the system in the form of Theorem 4.11.

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

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

Exercise 4.11. Prove Theorem 4.12 by following the ideas of the proof of Theorem 4.11.

Exercise 4.12. Consider the matrices

$$A = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}.$$

a. Is the matrix pair (A, B) controllable?

b. Is the matrix pair (A, C) observable?

c. Find a transformation that puts the pair (A, B) in the form of Theorem 4.11.

d. Find a transformation that puts the pair (A, C) in the form of Theorem 4.12.

Exercise 4.13. Prove that the characteristic equation for the matrix

$$M = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -m_0 & -m_1 & -m_2 & \cdots & -m_{n-2} & -m_{n-1} \end{bmatrix}$$

is given as

$$\Delta_M(s) = s^n + m_{n-1}s^{n-1} + \ldots + m_1 s + m_0.$$

Chapter 5

Stabilization by feedback

Stability is generally the most important desirable property of a system, as we have seen in Chapter 3. However, as physical systems are not necessarily (asymptotically) stable as autonomous systems, an important question is whether we can make such systems stable by the use of feedback controllers. This chapter addresses this question by designing controllers that rely on measurements of the state (Section 5.1) or output (Section 5.1), respectively. For the latter, so-called state observers play an important role and they are discussed in Section 5.2.

5.1 Stabilization by static state feedback

In this section, the simplest version of the stabilization problem is discussed. Specifically, we will consider systems without outputs and assume that the full state x can be measured and is thus available for control.

Thus, consider the system $\mathbf{\Sigma}(A, B)$ as in (4.1), whose representation is repeated for convenience as

$$\mathbf{\Sigma}(A, B): \quad \dot{x}(t) = Ax(t) + Bu(t), \tag{5.1}$$

where $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$.

Noting that we can choose the input u and have measurements of the state x available by assumption, we let u depend directly on x by introducing the

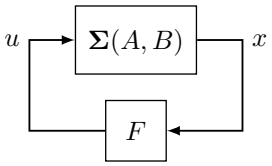


Figure 5.1: Interconnection of the system $\mathbf{\Sigma}(A, B)$ in (5.1) and the state feedback (5.2). Such interconnection is often referred to as a *closed-loop* system.

so-called *static feedback* controller

$$u(t) = Fx(t) \tag{5.2}$$

with $F \in \mathbb{R}^{m\times n}$. Here, the name feedback is motivated by the observation that measurements (of x) are fed back through the input. This is also apparent from Figure 5.1. Then, substitution of (5.2) in the dynamics (5.1) leads to

$$\dot{x}(t) = \big(A + BF\big)x(t), \tag{5.3}$$

which is often referred to as the *closed-loop* system. Note that (5.3) is an autonomous (homogeneous) linear system, i.e., there are no external inputs acting on (5.3).

Note that the stability properties of (5.3) are determined by the eigenvalues of $A + BF$ according to Theorem 3.3. As a result, the problem of stabilization by state feedback can be formulated as follows.

Problem 5.1. *Given matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, find a matrix $F \in \mathbb{R}^{m\times n}$ such that $\sigma(A + BF) \subset \mathbb{C}_-$.*

Here, we recall that $\mathbb{C}_- = \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}$, see also (3.6).

In the remainder of this section, a full solution to the stabilization problem will be provided. After recalling the definition of the characteristic polynomial of a matrix as

$$\Delta_A(s) = \det(sI - A), \tag{5.4}$$

we can state the following celebrated result, which will turn out to be crucial in solving Problem 5.1.

Theorem 5.1 (Pole placement theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given. For every monic*[1] *polynomial p of degree n, there exists $F \in \mathbb{R}^{m\times n}$ such that*

$$\Delta_{A+BF}(s) = p(s) \tag{5.5}$$

if and only if the matrix pair (A, B) is controllable.

Proof. Sufficiency and necessity are proven separately.

if. We will not give a full proof, but only consider the case $m = 1$, i.e., the case of a single input.

First, note that the characteristic polynomial of a matrix is invariant under similarity transformations, i.e., we have

$$\Delta_{A+BF}(s) = \Delta_{T(A+BF)T^{-1}}(s) \tag{5.6}$$

for any nonsingular $T \in \mathbb{R}^{n\times n}$. After denoting

$$\bar{F} = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{n-2} & f_{n-1} \end{bmatrix} = FT^{-1}, \tag{5.7}$$

as well as $\bar{A} = TAT^{-1}$ and $\bar{B} = TB$, we see that the characteristic polynomials of $A + BF$ and $\bar{A} + \bar{B}\bar{F}$ are the same. This motivates us to perform a similarity transformation in which the pole placement problem can be solved easily.

[1] Recall that a polynomial is *monic* if its leading coefficient equals one, see Section 3.2.

In particular, by the controllability canonical form from Theorem 4.13, we can choose T such that

$$\bar{A} = TAT^{-1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \quad \bar{B} = TB = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{5.8}$$

Then, given the structure of $\bar{A}$ and $\bar{B}$, it can easily concluded that

$$\bar{A} + \bar{B}\bar{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ f_0 - a_0 & f_1 - a_1 & f_2 - a_2 & \cdots & f_{n-2} - a_{n-2} & f_{n-1} - a_{n-1} \end{bmatrix}, \tag{5.9}$$

where the notation (5.7) is used. As the matrix (5.9) is in so-called companion form, its characteristic polynomial is given as

$$\Delta_{\bar{A}+\bar{B}\bar{F}}(s) = s^n + (a_{n-1} - f_{n-1})s^{n-1} + \ldots + (a_1 - f_1)s + (a_0 - f_0). \tag{5.10}$$

If the desired (monic) polynomial p reads

$$p(s) = s^n + p_{n-1}s^{n-1} + \ldots + p_1 s + p_0, \tag{5.11}$$

we see that the choice $f_i = a_i - p_i$, $i = 0, 1, \ldots, n-1$ achieves

$$\Delta_{T(A+BF)T^{-1}}(s) = \Delta_{\bar{A}+\bar{B}\bar{F}}(s) = p(s). \tag{5.12}$$

By (5.6), this is the desired result (5.5). Note that, in the original coordinates, the feedback matrix F reads $\bar{F}T$, as follows from (5.7).

only if. Assume that the pair (A, B) is not controllable. Then, by Theorem 4.15, there exists an uncontrollable eigenvalue $\lambda \in \sigma(A)$. This means that there exists a nonzero vector $v \in \mathbb{C}^n$ such that

$$v^{\mathrm{T}} \begin{bmatrix} \lambda I - A & B \end{bmatrix} = 0. \tag{5.13}$$

Then, we also have that $v^{\mathrm{T}}(A + BF) = \lambda v^{\mathrm{T}}$ for *any* matrix F, such that λ is an eigenvalue of $A + BF$ for any $F \in \mathbb{R}^{m \times n}$. Consequently, if p is a monic polynomial for which λ is not a root, i.e., $p(\lambda) \neq 0$, then there does not exist a feedback F such that (5.5) holds. □

The proof of necessity (the *only if*) of Theorem 5.1 shows an important result that we state explicitly as follows.

Corollary 5.2. *If $\lambda \in \sigma(A)$ is not (A, B)-controllable, then λ is an eigenvalue of $A + BF$ for all F.*

The result of Theorem 5.1 actually gives a much stronger result than is needed for solving Problem 5.1. Namely, the pole placement theorem gives conditions under which the eigenvalues of $A + BF$ can be placed *arbitrarily.*

Note that the desired eigenvalue locations are implicitly specified as they are the roots of the desired polynomial p in the statement of Theorem 5.1 (As we consider real-valued matrices A, B and polynomials p with real coefficients, this implies that complex eigenvalues appear as complex conjugate pairs).

Also, we stress that the proof of Theorem 5.1 is constructive and can be used to explicitly find the feedback matrix F that achieves desired eigenvalue locations. This is illustrated in the following example.

Example 5.1. Consider the system $\mathbf{\Sigma}(A, B)$ in (5.1) with

$$A = \begin{bmatrix} -3 & 2 & -1 \\ -2 & 1 & 0 \\ 3 & -2 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \tag{5.14}$$

and assume that the desired closed-loop eigenvalues are given as

$$\{-2, -3\} \tag{5.15}$$

where -3 has multiplicity 2. We note that the above system is the same as the one in Example 4.9, in which it was shown that the transformation $T \in \mathbb{R}^{3\times 3}$ defined by

$$T^{-1} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 2 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \tag{5.16}$$

transforms the system to the controllable canonical form. Specifically, we repeat (4.91) as

$$\bar{A} = TAT^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -4 & -2 \end{bmatrix}, \qquad \bar{B} = TB = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{5.17}$$

Thus, after denoting $\bar{F} = [f_0 \ f_1 \ f_2]$, the closed-loop system matrix $\bar{A} + \bar{B}\bar{F}$ has the characteristic polynomial

$$\Delta_{\bar{A}+\bar{B}\bar{F}}(s) = s^3 + (2 - f_2)s^2 + (4 - f_1)s + (1 - f_0), \tag{5.18}$$

see (5.10) in the proof of Theorem 5.1

The desired characteristic polynomial needs to have the desired closed-loop eigenvalues (5.15) as its roots and it thus easily obtained as

$$p(s) = (s + 2)(s + 3)^2 = s^3 + 8s^2 + 21s + 18. \tag{5.19}$$

Then, equating the polynomials (5.18) and (5.19) leads to $f_0 = -17$, $f_1 = -17$, $f_2 = -6$. Thus, we have $\bar{F} = [-17 \ -17 \ -6]$, which can be transformed to the original coordinates to yield

$$F = \bar{F}T = \begin{bmatrix} 3 & -10 & 3 \end{bmatrix}, \tag{5.20}$$

as the feedback that solves the pole placement problem for the desired eigenvalues (5.15). By direct computation it can be verified that indeed $\sigma(A + BF) = \{-2, -3\}$ (with -3 having multiplicity two). $\diamond$

We recall that, even though Theorem 5.1 gives conditions under which the closed-loop eigenvalues can be placed arbitrarily, the stabilization problem in Problem 5.1 only asks for eigenvalues to be in the open left-half complex plane, suggesting the condition of controllability is stronger than required. This turns out to be the case, as expressed in the following theorem.

Theorem 5.3. *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given. There exists a matrix $F \in \mathbb{R}^{m\times n}$ such that $\sigma(A+BF) \subset \mathbb{C}_-$ if and only if every eigenvalue $\lambda \in \sigma(A)$ satisfying $\lambda \notin \mathbb{C}_-$ is (A,B)-controllable.*

Proof. The *only if* part is a direct consequence of Corollary 5.2, so we only focus on the *if* part and assume that every $\lambda \notin \mathbb{C}_-$ is controllable.

First, consider the case $B = 0$. Then, every eigenvalue $\lambda \in \sigma(A)$ is uncontrollable and therefore (by assumption) satisfies $\lambda \in \mathbb{C}_-$. This implies $\sigma(A) \subset \mathbb{C}_-$ such that $F = 0$ provides a solution.

Second, if the matrix pair (A,B) is controllable, we can choose a polynomial p whose roots are all in $\mathbb{C}_-$ and apply the pole placement theorem (Theorem 5.1).

Otherwise, for (A,B) not controllable and $B \neq 0$, we can use Theorem 4.11 and find a similarity transformation T such that

$$TAT^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad TB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{5.21}$$

with (A_{11}, B_1) controllable. As the eigenvalues of A_{22} are uncontrollable (see Theorem 4.17), they satisfy $\sigma(A_{22}) \subset \mathbb{C}_-$ by assumption. As the matrix pair (A_{11}, B_1) controllable, we can apply the pole placement theorem and find a matrix F_1 such that $\sigma(A_{11} + B_1F_1) \subset \mathbb{C}_-$. Then, after choosing $\bar{F} = \begin{bmatrix} F_1 & 0 \end{bmatrix}$, we obtain

$$TAT^{-1} + TB\bar{F} = \begin{bmatrix} A_{11} + B_1F_1 & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{5.22}$$

whose eigenvalues satisfy

$$\sigma(A+BF) = \sigma\big(T(A+BF)T^{-1}\big) = \sigma(A_{11} + B_1F_1) \cup \sigma(A_{22}) \subset \mathbb{C}_-. \tag{5.23}$$

Here, we used the notation $F = \bar{F}T$. □

We note that the proofs of Theorems 5.1 and 5.3 are constructive, i.e., they provide a procedure for finding a matrix F that solves Problem 5.1.

From Theorem 5.3, it is clear that only the unstable eigenvalues of the system matrix A are required to be controllable. Intuitively, this means that such eigenvalues can be "moved" in the complex plane to achieve stabilization by feedback.

The stabilization problem by state feedback (in Problem 5.1) is important enough to motivate the following definition.

Definition 5.1. *The system $\mathbf{\Sigma}(A,B)$ as in (5.1) is called stabilizable if there exists a feedback $F \in \mathbb{R}^{m\times n}$ such that $\sigma(A+BF) \subset \mathbb{C}_-$.*

Then, the result of Theorem 5.3 can be rephrased as follows.

Corollary 5.4. *The system $\mathbf{\Sigma}(A,B)$ in (5.1) is stabilizable if and only if every unstable eigenvalue of A is (A,B)-controllable, i.e., if and only if*

$$\operatorname{rank}\begin{bmatrix} A - \lambda I & B \end{bmatrix} = n, \qquad \forall\, \lambda \in \sigma(A) \text{ s.t. } \operatorname{Re}(\lambda) \geq 0. \tag{5.24}$$

The condition (5.24) is known as the Hautus test (or Popov-Belevich-Hautus test) for stabilizability.

Remark 5.1. It is very insightful to compare the Hautus test for controllability in (4.102) with the Hautus test for stabilizability in (5.24). These conditions make very clear that controllability implies stabilizability, as we have already concluded from the pole placement theorem. ◁

At this point, it is good to recall that we have restricted attention to state feedback controllers of the form (5.2) in an attempt to solve the stabilization problem. This feedback is linear, constant, and static. This raises the question if a more general class of controllers might achieve stabilization under less restrictive conditions as those in Theorem 5.3. For example, one could consider the time-varying feedback

$$u(t) = F(t)x(t), \tag{5.25}$$

or nonlinear feedback

$$u(t) = f\big(t, x(t)\big). \tag{5.26}$$

Moreover, in Section 5.3 we will consider feedback controllers that are themselves described as a linear system.

However, the following result shows that stabilizability is a necessary condition for achieving stabilization, regardless of the input u that is designed (i.e., regardless of the class of controllers).

Theorem 5.5. *Consider the system $\mathbf{\Sigma}(A, B)$ in (5.1) and assume that, for every initial condition $x_0 \in \mathbb{R}^n$, there exists an input function $u : [0, \infty) \to \mathbb{R}^m$ such that*

$$\lim_{t\to\infty} x(t; x_0, u) = 0. \tag{5.27}$$

Then, $\mathbf{\Sigma}(A, B)$ is stabilizable.

Proof. We prove by contraposition. Thus, assume that $\mathbf{\Sigma}(A, B)$ is not stabilizable, i.e., there exists $\lambda \in \sigma(A)$ such that $\mathrm{Re}(\lambda) \geq 0$ and, in addition,

$$v^{\mathrm{T}} \begin{bmatrix} \lambda I - A & B \end{bmatrix} = 0 \tag{5.28}$$

for some nonzero vector $v \in \mathbb{C}^n$, see Corollary 5.4. Now, choose the intial condition x_0 such that $v^{\mathrm{T}} x_0 \neq 0$. Note that (5.28) implies that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big\{v^{\mathrm{T}} x(t)\big\} = \lambda v^{\mathrm{T}} x(t), \tag{5.29}$$

such that the solution $x(\,\cdot\,; x_0, u)$ satisfies $v^{\mathrm{T}} x(t; x_0, u) = e^{\lambda t} v^{\mathrm{T}} x_0$ for any input function u. However, as $\mathrm{Re}(\lambda) \geq 0$, this means that (5.27) is violated, proving the theorem. □

5.2 State observers

In the previous section, it was assumed that the full state x could be measured and is therefore available for feedback. This is not often the case in practical

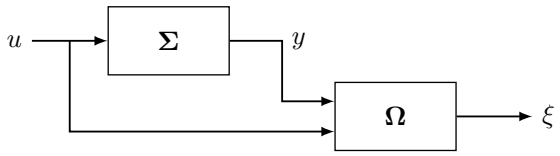


Figure 5.2: A state observer $\mathbf{\Omega}$ as in (5.31) interconnected with a linear system $\mathbf{\Sigma}$.

control problems. In this section, we therefore aim at obtaining an *estimate* of the state on the basis of knowledge of inputs and outputs of the system.

Specifically, given the linear system

$$\mathbf{\Sigma} : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{cases} \tag{5.30}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ as usual, the aim is to construct a system that generates an estimate ξ of x on the basis of measurements of u and y. To this end, introduce the general system

$$\mathbf{\Omega} : \begin{cases} \dot{w}(t) = Pw(t) + Qu(t) + Ry(t), \\ \xi(t) = Sw(t) \end{cases} \tag{5.31}$$

with state $w(t) \in \mathbb{R}^{n_w}$. Note that this system takes u and y as inputs, and outputs ξ as an estimate of x. This is illustrated in Figure 5.2.

In order to construct the matrices P, Q, R, and S in (5.31), we first need to precisely define the desired properties of $\mathbf{\Sigma}$, i.e., we need to define what is meant by an *estimate* of the state x.

To do so, we introduce the estimation error

$$e(t) = \xi(t) - x(t), \tag{5.32}$$

and consider the interconnection in Figure 5.2. Then, the time derivative of the estimation error reads

$$\begin{aligned} \dot{e}(t) &= SPw(t) + SQu(t) + SRy(t) - Ax(t) - Bu(t), \\ &= SPw(t) + SQu(t) + \big(SRC - A\big)x(t) - Bu(t), \end{aligned} \tag{5.33}$$

where the output equation $y = Cx$ is used to obtain the latter. Then, the substitution of $x = \xi - e = Sw - e$ in (5.33) leads to

$$\dot{e}(t) = \big(SP + SRCS - AS\big)w(t) + \big(A - SRC\big)e(t) + \big(SQ - B\big)u(t). \tag{5.34}$$

Now, we can use the definition of the estimation error to state the desirable properties of an estimator of the state x, which will be referred to as a state observer.

Definition 5.2. *A system $\mathbf{\Omega}$ as in (5.31) is called a state observer for $\mathbf{\Sigma}$ if, for any pair of initial conditions x_0, w_0 satisfying $e(0) = Sw_0 - x_0 = 0$ and any input function u, we have that $e(t) = 0$ for all $t \geq 0$.*

Thus, once the estimate $\xi(t)$ of the state $x(t)$ is perfect at some time t, Definition 5.2 asks for this estimate to remain perfect, irrespective of the input.

We can further strengthen the desired properties of an estimator as follows.

Definition 5.3. *A state observer $\mathbf{\Omega}$ is called stable if for each pair of initial values w_0, x_0 and any input function u, we have that*

$$\lim_{t\to\infty} e(t) = 0. \tag{5.35}$$

At this point, we return attention to the general system (5.31). For this to be a state observer as in Definition 5.2, we need that the error dynamics (5.34) is independent of both u and w. Namely, otherwise, $e(0) = 0$ would not imply that $e(t) = 0$ for $t \geq 0$. Thus, this leads to

$$SP = AS - SRCS, \qquad SQ = B \tag{5.36}$$

as necessary and sufficient conditions for $\mathbf{\Omega}$ to be a state observer. In this case, the error dynamics (5.34) reduces to

$$\dot{e}(t) = \big(A - SRC\big)e(t), \tag{5.37}$$

which is indeed easily seen to have the property that $e(0) = 0$ implies $e(t) = 0$ for all $t \geq 0$.

Under the conditions (5.36) the dynamics of $\mathbf{\Omega}$ in (5.31) leads to

$$\begin{aligned}\dot{\xi}(t) = S\dot{w}(t) &= SPw(t) + SQu(t) + SRy(t),\\ &= \big(A - SRC\big)\xi(t) + Bu(t) + SRy(t),\end{aligned} \tag{5.38}$$

where we have used $\xi(t) \in \mathbb{R}^n$ as a state variable rather than w. In this formulation, it can be seen that the matrices S and R only appear as the product SR. Therefore, we introduce $G = SR$.

The above considerations can be summarized in the following theorem.

Theorem 5.6. *Consider the system $\mathbf{\Sigma}$ in (5.30). The general form of a state observer for $\mathbf{\Sigma}$ is*

$$\dot{\xi}(t) = (A - GC)\xi(t) + Bu(t) + Gy(t). \tag{5.39}$$

with $\xi(t) \in \mathbb{R}^n$ and $G \in \mathbb{R}^{n\times p}$. Then, the estimation error $e(t) = \xi(t) - x(t)$ satisfies the dynamics

$$\dot{e}(t) = (A - GC)e(t), \tag{5.40}$$

such that the state observer (5.39) is stable if and only if $\sigma(A - GC) \subset \mathbb{C}_-$.

Remark 5.2. The general form of a state observer (as in Definition 5.2) given in (5.39) can alternatively be written as

$$\dot{\xi}(t) = A\xi(t) + Bu(t) + G\big(y(t) - C\xi(t)\big), \tag{5.41}$$

which gives the structure of the state observer as a *copy* of the system $\mathbf{\Sigma}$ with the additional term $G(y - C\xi)$. This term is sometimes referred to as the *output injection* term and provides a comparison of the actual output y with the output $C\xi$ predicted by the state observer. Thus, in case these outputs match exactly, the dynamics of the state observer reduce to that of the original system, from which it is easily concluded that Definition 5.2 is satisfied. The form (5.41) is illustrated in Figure 5.3. ◁

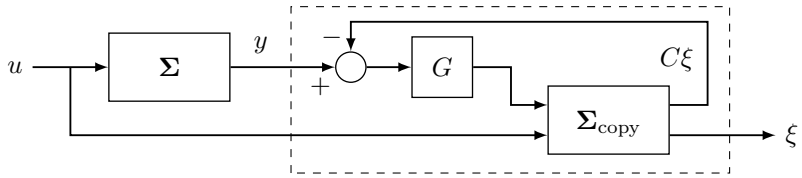


Figure 5.3: Illustration of a state observer $\mathbf{\Omega}$ as a copy of the system $\mathbf{\Sigma}$ with output injection as in (5.41). The dashed box represents $\mathbf{\Omega}$.

It is clear from Theorem 5.6 that a state observer always exists and that such observers are parameterized by the matrix G. A *stable* state observer exists if and only if G can be chosen such that $\sigma(A - GC) \subset \mathbb{C}_-$, which motivates the following definition.

Definition 5.4. *The system $\mathbf{\Sigma}$ as in (5.30) is called detectable if there exists a matrix $G \in \mathbb{R}^{n\times p}$ such that $\sigma(A - GC) \subset \mathbb{C}_-$.*

As before, we also say that the *matrix pair* (A, C) is detectable if the above condition holds.

By comparing the definition of detectability above to that of stabilizability in Definition 5.1, it is clear that these concepts are dual. We make this explicit in the following lemma, where, for clarity, we refer to stabilizability and detectability as properties of *matrix pairs* rather than systems.

Lemma 5.7. *The matrix pair (A, C) is detectable if and only if the matrix pair $(A^{\mathrm{T}}, C^{\mathrm{T}})$ is stabilizable.*

Proof. Let the matrix pair (A, C) be detectable, i.e., there exists a matrix G such that $\sigma(A - GC) \subset \mathbb{C}_-$. Note that, by taking the transpose, we also have

$$\sigma(A - GC) = \sigma\big(A^{\mathrm{T}} - C^{\mathrm{T}}G^{\mathrm{T}}\big) \subset \mathbb{C}_-, \tag{5.42}$$

such that $F = -G^{\mathrm{T}}$ is a matrix that satisfies Definition 5.1 for the matrix pair $(A^{\mathrm{T}}, C^{\mathrm{T}})$, i.e., that matrix pair is stabilizable.

The reasoning can easily be reversed to show the converse as well. □

This then leads directly to the following condition for detectability as a counterpart of Corollary 5.4.

Theorem 5.8. *The system $\mathbf{\Sigma}$ in (5.30) is detectable if and only if every unstable eigenvalue of A is (A, C)-observable, i.e., if and only if*

$$\operatorname{rank}\begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n, \qquad \forall\, \lambda \in \sigma(A) \text{ s.t. } \operatorname{Re}(\lambda) \geq 0. \tag{5.43}$$

Proof. This is an immediate consequence of Corollary 5.4 and Lemma 5.7. □

More importantly, the above developments also directly give a solution to finding a stable observer problem as in Definition 5.3.

Corollary 5.9. *Consider the system $\mathbf{\Sigma}$ in (5.30). Then, the following statements are equivalent:*

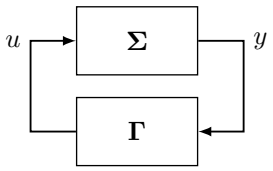


Figure 5.4: Interconnection of the system $\mathbf{\Sigma}$ in (5.44) and the dynamic output feedback controller $\mathbf{\Gamma}$ in (5.45). Compare this figure to Figure 5.1 on state feedback.

1. *there exists a stable observer for* $\mathbf{\Sigma}$*;*
2. *the matrix pair* (A, C) *is detectable;*
3. *every eigenvalue* $\lambda \in \sigma(A)$ *satisfying* $\lambda \notin \mathbb{C}_-$ *is observable.*

5.3 Stabilization by dynamic output feedback

In Section 5.1, the problem of stabilization by state feedback was considered, as formalized as Problem 5.1. The underlying assumption in this setting is that the full state x can be measured and is thus available for feedback. However, in many applications, this is not the case and only measurements of the output are available. The current section discusses the stabilization problem in this case.

Therefore, consider linear systems with outputs of the form

$$\mathbf{\Sigma} : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{cases} \tag{5.44}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$, see also (5.30). Here, it is assumed that the output y can be measured and is thus available for control.

Thus, we would like to find a controller that takes y as input and prescribes u. Whereas a *static* controller was considered in Section 5.1, a more general class of controllers are considered here. Specifically, we allow for a controller that is itself a linear system of the form

$$\mathbf{\Gamma} : \begin{cases} \dot{w}(t) = Kw(t) + Ly(t), \\ u(t) = Mw(t) + Ny(t). \end{cases} \tag{5.45}$$

Here, $w(t) \in \mathbb{R}^{n_w}$ is the state of the controller, which takes measurements of the system output $y(t) \in \mathbb{R}^p$ and prescribes the system input $u(t) \in \mathbb{R}^m$. As $\mathbf{\Gamma}$ is described as a linear system, it is often referred to as a *dynamic output feedback* controller.

The interconnection of the system $\mathbf{\Sigma}$ in (5.44) and controller $\mathbf{\Gamma}$ in (5.45) is depicted in Figure 5.4. The dynamics of this *closed-loop* system reads

$$\begin{bmatrix} \dot{x}(t) \\ \dot{w}(t) \end{bmatrix} = \begin{bmatrix} A + BNC & BM \\ LC & K \end{bmatrix} \begin{bmatrix} x(t) \\ w(t) \end{bmatrix}, \tag{5.46}$$

as is easily derived by substitution of the output equations of (5.44) and (5.45) in (5.45) and (5.44), respectively. We note that the closed-loop system (5.46) is

a homogeneous linear system with system matrix

$$A_{\mathrm{cl}} = \begin{bmatrix} A + BNC & BM \\ LC & K \end{bmatrix}, \tag{5.47}$$

such that stability of the interconnection (5.46) is determined by the eigenvalues of A_{cl} (recall Theorem 3.3).

This immediately leads to the following problem statement.

Problem 5.2. *Given matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times m}$, find an integer $n_w > 0$ and matrices*

$$K \in \mathbb{R}^{n_w \times n_w}, \quad L \in \mathbb{R}^{n_w \times p}, \quad M \in \mathbb{R}^{m\times n_w}, \quad N \in \mathbb{R}^{m\times p}, \tag{5.48}$$

such that the matrix A_{cl} in (5.47) satisfies $\sigma(A_{\mathrm{cl}}) \subset \mathbb{C}_-$.

A controller $\mathbf{\Gamma}$ for which the matrices (5.48) satisfy the above condition is referred to as an *internally stabilizing* controller for $\mathbf{\Sigma}$. We stress that such internally stabilizing controller does not only achieve $\lim_{t\to\infty} x(t) = 0$ for the closed-loop system, but also $\lim_{t\to\infty} w(t) = 0$.

Remark 5.3. It is insightful to compare the stabilization problem by (static) state feedback in Problem 5.1 to the stabilization problem by (dynamic) output feedback in Problem 5.2. In the latter, the controller $\mathbf{\Gamma}$ in (5.45) replaces the static feedback (5.2), whereas the closed-loop dynamics (5.46) is the counterpart of (5.3). ◁

A full characterization of the matrices (5.48) solving Problem 5.2 is not known (mostly because the state-space dimension of the controller n_w is free). However, in the remainder of this section, we will provide a specific solution to the stabilization problem by state feedback.

In particular, the concepts of state feedback and state observers in Sections 5.1 and 5.2, respectively, will be combined. Namely, we have seen how to construct a stabilizing state feedback $u(t) = Fx(t)$ when measurements of the state are available. However, as these measurements are not available, we will use an *estimate* of the state x as provided by a stable state observer.

Thus, consider the observer $\mathbf{\Omega}$ given in (5.39) (with state $\xi(t) \in \mathbb{R}^n$) and combine with the feedback

$$u(t) = F\xi(t), \tag{5.49}$$

using the estimate ξ instead of the actual state x. The resulting controller is given as

$$\mathbf{\Gamma} : \begin{cases} \dot{\xi}(t) &= \big(A - GC + BF\big)\xi(t) + Gy(t), \\ u(t) &= F\xi(t), \end{cases} \tag{5.50}$$

which is of the form (5.45) with $n_w = n$ and

$$K = A - GC + BF, \quad L = G, \quad M = F, \quad N = 0. \tag{5.51}$$

An illustration of the observer-based dynamic output feedback controller (5.50) is given in Figure 5.5.

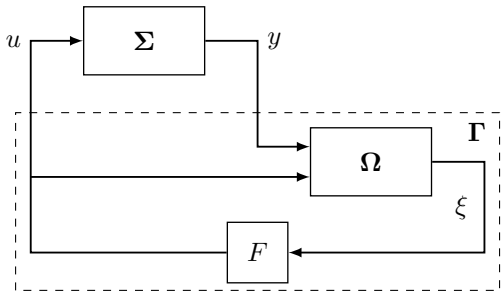


Figure 5.5: The controller $\mathbf{\Gamma}$ in (5.50), comprising a state observer $\mathbf{\Omega}$ as in (5.31) and static feedback (5.49), interconnected with a linear system $\mathbf{\Sigma}$ as in (5.44).

At this point, we have selected a specific structure for the controller $\mathbf{\Gamma}$ by combining a state observer and state feedback controller. Even though this is an intuitive choice, we still need to show that such structure can be used to solve the stabilization problem using dynamic output feedback given in Problem 5.2.

The following result shows that this can indeed be achieved.

Lemma 5.10. *Consider the system $\mathbf{\Sigma}$ in (5.44). Let $\mathbf{\Gamma}$ as in (5.50) be a dynamic output feedback controller on the basis of a stable state observer $\mathbf{\Omega}$ and a feedback F that solves the stabilization problem by state feedback. Then, $\mathbf{\Gamma}$ solves Problem 5.2.*

Proof. For the system $\mathbf{\Sigma}$ in (5.44) and controller $\mathbf{\Gamma}$ in (5.50), the closed-loop system matrix A_{cl} in (5.47) reads

$$A_{\mathrm{cl}} = \begin{bmatrix} A & BF \\ GC & A - GC + BF \end{bmatrix}. \tag{5.52}$$

Recall that $\sigma(A_{\mathrm{cl}}) = \sigma(TA_{\mathrm{cl}}T^{-1})$ for any nonsingular matrix $T \in \mathbb{R}^{2n\times 2n}$. Then, after choosing

$$T = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}, \tag{5.53}$$

a direct computation gives

$$TA_{\mathrm{cl}}T^{-1} = \begin{bmatrix} A + BF & BF \\ 0 & A - GC \end{bmatrix}. \tag{5.54}$$

As a result, due to the block upper-triangular structure of (5.54), we obtain

$$\sigma(A_{\mathrm{cl}}) = \sigma\big(A + BF\big) \cup \sigma\big(A - GC\big). \tag{5.55}$$

However, $\sigma(A + BF) \subset \mathbb{C}_-$ as F solves the stabilization problem by state feedback, see Problem 5.1. Similarly, $\sigma(A - GC) \subset \mathbb{C}_-$ as $\mathbf{\Omega}$ is a stable state observer, see Theorem 5.6. Consequently, $\sigma(A_{\mathrm{cl}}) \subset \mathbb{C}_-$, proving the lemma. □

Remark 5.4. The transformation (5.53) in the proof of Lemma 5.10 allows for an insightful interpretation. Namely, after recalling that the estimation error e is defined as $e(t) = \xi(t) - x(t)$, see (5.32), it can be observed that

$$\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = T \begin{bmatrix} x(t) \\ \xi(t) \end{bmatrix}, \tag{5.56}$$

such that (5.54) gives the corresponding closed-loop dynamics

$$\begin{bmatrix} \dot{x}(t) \\ \dot{e}(t) \end{bmatrix} = \begin{bmatrix} A + BF & BF \\ 0 & A - GC \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} \tag{5.57}$$

in terms of the system state x and estimation error e. ◁

As we have from Definition 5.1 and Corollary 5.9 that stabilizability and detectability guarantee the existence of a stabilizing static state feedback controller and stable observer, respectively, it is clear that these conditions are sufficient for a controller solving Problem 5.2 to exists. Stabilizability and detectability are however also necessary, as stated next.

Theorem 5.11. *Consider the system $\mathbf{\Sigma}$ in (5.44). There exists a stabilizing dynamic output feedback controller $\mathbf{\Gamma}$ of the form (5.45) if and only if the matrix pair (A, B) is stabilizable and the matrix pair (A, C) is detectable.*

Proof. if. As the matrix pair (A, B) is stabilizable, there exists a feedback $F \in \mathbb{R}^{m \times n}$ such that $\sigma(A+BF) \subset \mathbb{C}_-$, see Definition 5.1. Similarly, detectability of the matrix pair (A, C) in Definition 5.4 guarantees the existence of a matrix $G \in \mathbb{R}^{n \times p}$ such that $\sigma(A - GC) \subset \mathbb{C}_-$. Then, the *observer-based* controller $\mathbf{\Gamma}$ in (5.50) is a stabilizing dynamic output feedback controller for $\mathbf{\Sigma}$ by Lemma 5.10.

only if. Let $\mathbf{\Gamma}$ be a stabilizing dynamic output feedback controller for $\mathbf{\Sigma}$. This means that, for any initial state x_0 of the system $\mathbf{\Sigma}$, there exists an input function u (namely, the one generated by $\mathbf{\Gamma}$) such that $\lim_{t\to\infty} x(t; x_0, u) = 0$. Thus, by Theorem 5.5, the matrix pair (A, B) is stabilizable.

To show detectability of the matrix pair (A, C), let $\lambda \in \sigma(A)$ be an unobservable eigenvalue, i.e., there exists $p \in \mathbb{C}$, $p \neq 0$, such that

$$\begin{bmatrix} A - \lambda I \\ C \end{bmatrix} p = 0, \tag{5.58}$$

see Definition 4.9. We now consider two different trajectories of the closed-loop system:

1. let $\mathbf{\Sigma}$ have initial condition $x_0 = 0$ and denote the arbitrary initial condition of $\mathbf{\Gamma}$ by w_0. Then, the trajectory of the system, denoted by $x(\cdot)$ satisfies $\lim_{t\to\infty} x(t) = 0$ as $\mathbf{\Gamma}$ is a stabilizing controller. We use $u(\cdot)$ to denote the corresponding input function.

2. let $\mathbf{\Sigma}$ now have a the initial condition $x_0 = p$ and assume that the initial condition of $\mathbf{\Gamma}$ is w_0 as before. Denote this trajectory by $x'(\cdot)$ and the corresponding input function by $u'(\cdot)$.

We claim that

$$x'(t) = e^{\lambda t} p + x(t). \tag{5.59}$$

First, it can be shown that x' defined above satisfies the system dynamics. Namely, we have from (5.58) that

$$\dot{x}'(t) = \lambda e^{\lambda t} p + \dot{x}(t) = A e^{\lambda t} p + A x(t) + B u(t) = A x'(t) + B u(t), \tag{5.60}$$

where u is the input function determined by the controller for the trajectory x. Second, the use of (5.58) leads to

$$C x'(t) = C e^{\lambda} p + C x(t) = C x(t), \tag{5.61}$$

such that the trajectories x and x' lead to the same output trajectories. As the initial conditions for the controller $\mathbf{\Gamma}$ are the same for the two trajectories, we thus obtain that $u(t) = u'(t)$ for all t. Together with (5.60), this proves that the claim (5.59) holds.

However, as $\mathbf{\Gamma}$ is a stabilizing controller, we have that $\lim_{t\to\infty} x'(t) = 0$. Together with the same property for the trajectory x (see above), this implies that $\lambda \in \mathbb{C}_-$. Thus, any unobservable eigenvalue $\lambda \in \sigma(A)$ is stable, i.e., the matrix pair (A, C) is detectable. □

5.4 Exercises

Exercise 5.1. Consider the system

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} -1 & 2 & 0 & -3 \\ 0 & -2 & 0 & 0 \\ 2 & 1 & -3 & -3 \\ 0 & 2 & 0 & -4 \end{bmatrix} x(t) + \begin{bmatrix} 2 \\ 1 \\ 1 \\ 1 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} 0 & 1 & 1 & -1 \end{bmatrix} x(t). \end{aligned}$$

a. Is the system controllable? Give a basis for the reachable subspace.

b. Is the system observable? Give a basis for the unobservable subspace.

c. Is the system stabilizable?

d. Is the system detectable?

Exercise 5.2. Consider the system

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} -1 & 0 & 2 \\ 0 & -3 & 0 \\ 1 & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(t). \end{aligned}$$

a. Show that the system is not (internally) asymptotically stable.

b. Show that the system is stabilizable.

c. Put the pair (A, B) in the form of Theorem 4.13.

d. Find a matrix $F \in \mathbb{R}^{1\times 3}$ such that the state feedback $u(t) = Fx(t)$ guarantees that the closed-loop system has its eigenvalues at -1, -2, and -3.

e. Show that the system is not observable.

f. Is the system detectable?

Exercise 5.3. Let α, $\beta \in \mathbb{R}$ and consider the matrices

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 0 & 0 & \alpha \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \\ \beta \end{bmatrix}.$$

Determine all values of α and β for which the matrix pair (A, B) is stabilizable.

Exercise 5.4. Find a matrix F such that applying the static state feedback $u(t) = Fx(t)$ to the system

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} u(t)$$

yields the closed-loop characteristic polynomial

$$\Delta_{A+BF}(s) = s^4 + 3s^3 + 4s^2 + 3s + 1.$$

Exercise 5.5. Consider the linear system $\mathbf{\Sigma}$ as in (5.44) with

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

a. Show that the system is not (internally) asymptotically stable.

b. Show that the system is both controllable and observable.

c. Find matrices $F \in \mathbb{R}^{1\times 2}$ and $G \in \mathbb{R}^{2\times 1}$ such that $\sigma(A + BF) \subset \mathbb{C}_-$ and $\sigma(A - GC) \subset \mathbb{C}_-$.

d. Find matrices (K, L, M, N) such that the feedback controller

$$\begin{aligned} \dot{w}(t) &= Kw(t) + Ly(t), \\ u(t) &= Mw(t) + Ny(t), \end{aligned}$$

is internally stabilizing for $\mathbf{\Sigma}$.

Exercise 5.6. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^n$ (note that B is a vector, i.e., $m = 1$) such that the matrix pair (A, B) is controllable. Let

$$p(s) = s^n + p_{n-1}s^{n-1} + \ldots + p_1 s + p_0,$$

be a given monic polynomial with $p_i \in \mathbb{R}$. Next, let $F \in \mathbb{R}^{1\times n}$ be such that

$$\Delta_{A+BF}(s) = p(s).$$

a. Show that there exists a unique solution $\eta \in \mathbb{R}^n$ to the equations

$$\begin{aligned} \eta^{\mathrm{T}} A^k B &= 0, \qquad k = 0, 1, \ldots, n-2, \\ \eta^{\mathrm{T}} A^{n-1} B &= 1. \end{aligned}$$

b. Show that $F = -\eta^{\mathrm{T}} p(A)$, where $p(A) = A^n + p_{n-1}A^{n-1} + \ldots + p_1 A + p_0 I$.
Hint. Use that $p(A + BF) = 0$ (as a result of Cayley-Hamilton).

Exercise 5.7. Consider the linear system $\mathbf{\Sigma}$ as in (5.30) with

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

Design a stable observer for $\mathbf{\Sigma}$ such that the error dynamics (5.40) has its eigenvalues at -2, -2, and -3, i.e., $\sigma(A - GC) = \{-2, -2, -3\}$.

Exercise 5.8. Instead of the linear system $\mathbf{\Sigma}$ as in (5.30), consider the system

$$\mathbf{\Sigma} : \begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{cases} \tag{5.62}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ as usual and the direct feedthrough $Du(t)$ added.

a. Show that

$$\dot{\xi}(t) = (A - GC)\xi(t) + (B - GD)u(t) + Gy(t), \tag{5.63}$$

is a state observer for (5.62), i.e., it satisfies Definition 5.2. What is the associated dynamics of the estimation error $e(t) = \xi(t) - x(t)$?

b. Give an interpretation of the observer (5.63) as a copy of the system (5.62) with output injection as in Remark 5.2.

Exercise 5.9. Consider the linear system $\mathbf{\Sigma}$ as in (5.30). Assume that there exist matrices $S \in \mathbb{R}^{\bar{n} \times n}$, $\bar{A} \in \mathbb{R}^{\bar{n} \times \bar{n}}$, and $\bar{G} \in \mathbb{R}^{\bar{n} \times p}$ such that

$$SA - \bar{A}S = \bar{G}C. \tag{5.64}$$

a. Define $\bar{B} = SB$ and consider the system

$$\dot{w}(t) = \bar{A}w(t) + \bar{B}u(t) + \bar{G}y(t). \tag{5.65}$$

Define the error $e(t) = w(t) - Sx(t)$. Show that (5.65) is an observer (for $Sx(t)$) in the sense that $e(0) = 0$ implies $e(t) = 0$ for all $t \geq 0$, regardless of the input function u.

b. Show that the error dynamics corresponding to (5.65) is stable if and only if $\sigma(\bar{A}) \subset \mathbb{C}_-$.

c. Show that, if $n = \bar{n}$ and S is nonsingular, the observer (5.65) is similar to the general form (5.39).

In the above, we note, first, that the observer provides an estimate of Sx rather than x, and, second, that the state-space dimension $\bar{n}$ of the observer is not necessarily the same as that of the system (which is n). This raises the question of whether one can design an observer (for x) with a lower state-space dimension, which is referred to as a *reduced observer*.

In the remainder of this exercise, we will design such reduced observer for observable linear systems with a single output, i.e., $p = 1$. Without loss of generality, we assume that the matrices A and C take the observability canonical form (4.89) from Theorem 4.14.

We will show that there exists an observer for $\bar{n} = n - 1$ and where $\bar{A}$ has the form

$$\bar{A} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & -\bar{a}_0 \\ 1 & 0 & & & 0 & -\bar{a}_1 \\ 0 & 1 & \ddots & & 0 & -\bar{a}_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \ddots & 1 & 0 & -\bar{a}_{n-3} \\ 0 & 0 & \cdots & 0 & 1 & -\bar{a}_{n-2} \end{bmatrix},$$

where $\bar{a}_i \in \mathbb{R}$, $i = 0, 1, \ldots, n-2$.

d. Show that, for (A, C) in observability canonical form and $\bar{A}$ as above, there exist matrices S and $\bar{G}$ satisfying (5.64). What is the corresponding expression for $\bar{G}$?

Hint. Choose

$$S = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\bar{a}_0 \\ 0 & 1 & \ddots & \vdots & -\bar{a}_1 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -\bar{a}_{n-2} \end{bmatrix}.$$

e. Show that the eigenvalues of the error dynamics from *a.* can be placed arbitrarily, which implies that a *stable* observer of dimension $\bar{n} = n - 1$ always exists.

f. Note that the observer (5.65) only provides an estimate for $Sx(t)$, which, as $\bar{n} < n$, is not the entire state $x(t)$. How can the state $x(t)$ be reconstructed?

Hint. Note that the output measurement $y(t)$ is assumed to be available.

Chapter 6

Input-output properties

In this chapter, the input-output behavior of linear systems is studied. To this end, the impulse response matrix is introduced in Section 6.1 as a full characterization of this input-output behavior. An equivalent characterization is given by the transfer function matrix of Section 6.2, which is further studied for single-input single-output systems in Section 6.3. Finally, a notion of stability for input-output behavior is discussed in Section 6.4.

6.1 The impulse response matrix

In this section, we consider linear systems of the form

$$\mathbf{\Sigma} : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{6.1}$$

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}^m$, and output $y(t) \in \mathbb{R}^p$. To explicitly denote the matrices associated to a system $\mathbf{\Sigma}$, we will sometimes write $\mathbf{\Sigma}(A, B, C, D)$ for (6.1).

In Chapter 2 it was shown that the output trajectory of (6.1) for initial condition $x(0) = x_0$ and input function $u : \mathbb{R}_+ \to \mathbb{R}^m$ is given as

$$y(t; x_0, u) = Ce^{At}x_0 + \int_0^t Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t), \tag{6.2}$$

see Remark 2.3 for details. Here, we stress that we take the initial time t_0 to satisfy $t_0 = 0$ (and thus write $y(\,\cdot\,; x_0, u)$ instead of $y(\,\cdot\,; t_0, x_0, u)$ to denote the output trajectory), which is motivated by the result on time-invariance in Theorem 2.13.

It is noted that, for $x_0 = 0$, the output solution (6.2) does not explicitly contain the state x. To further analyze the effect of the state on the output trajectory, we introduce the change of coordinates

$$\bar{x}(t) = Tx(t) \tag{6.3}$$

with $T \in \mathbb{R}^{n\times n}$ a nonsingular matrix. In these coordinates, it is easy to see that

the system (6.1) can be written as

$$\boldsymbol{\Sigma}: \begin{cases} \dot{\bar{x}}(t) = TAT^{-1}\bar{x}(t) + TBu(t), \\ y(t) = CT^{-1}\bar{x}(t) + Du(t). \end{cases} \tag{6.4}$$

This motivates the following definition.

Definition 6.1. *Two systems $\boldsymbol{\Sigma}(A, B, C, D)$ and $\boldsymbol{\Sigma}(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ of the form (6.1) are called similar if there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that*

$$\bar{A} = TAT^{-1}, \quad \bar{B} = TB, \quad \bar{C} = CT^{-1}, \quad \bar{D} = D. \tag{6.5}$$

Remark 6.1. Changes of coordinates were also considered in the analysis of controllability and observability properties in Chapter 4, see in particular Section 4.3. There, it was shown that controllability and observability are invariant under such changes of coordinates. ◁

Given (6.4), the following result is perhaps not surprising.

Theorem 6.1. *Let the systems $\boldsymbol{\Sigma}(A, B, C, D)$ and $\boldsymbol{\Sigma}(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ be similar and let $y(\,\cdot\,; 0, u)$ and $\bar{y}(\,\cdot\,; 0, u)$ denote their output trajectories for zero initial condition and common input function $u : J \to \mathbb{R}^m$. Then,*

$$y(t; 0, u) = \bar{y}(t; 0, u), \tag{6.6}$$

for all $t \in J$.

Proof. As $\boldsymbol{\Sigma}(A, B, C, D)$ and $\boldsymbol{\Sigma}(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ are similar, there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that (6.5) holds. As a result, we have that

$$\bar{C}e^{\bar{A}t}\bar{B} = CT^{-1}e^{TAT^{-1}t}TB = CT^{-1}\big(Te^{At}T^{-1}\big)TB = Ce^{At}B, \tag{6.7}$$

for all $t \in \mathbb{R}$. Here, we have used Lemma 2.6 on the matrix exponential to obtain the second equality.

Now, considering the output trajectory of $\boldsymbol{\Sigma}(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ gives

$$\begin{aligned} \bar{y}(t; 0, u) &= \int_0^t \bar{C}e^{\bar{A}(t-\tau)}\bar{B}u(\tau)\,\mathrm{d}\tau + \bar{D}u(t), \\ &= \int_0^t Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t) = y(t, 0, u), \end{aligned} \tag{6.8}$$

where (6.7) as well as $\bar{D} = D$ from (6.5) are used to obtain (6.8). This proves the desired result (6.6). □

Even though the result of Theorem 6.1 is relatively simple, it has important consequences. Namely, it implies that the choice of coordinates does not influence the so-called *input-output behavior* of a linear system (for zero initial conditions). We will frequently exploit this observation in the remainder of this chapter.

To further characterize the input-output behavior given through the general form (6.2), we will consider the output trajectory for specific input functions.

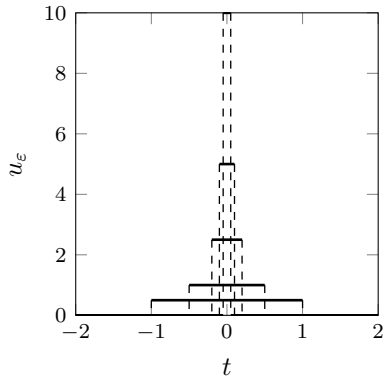


Figure 6.1: The function u_ε in (6.9) for $\varepsilon = 1, 0.5, 0.2, 0.1, 0.05$.

To this end, consider the function $u_\varepsilon : \mathbb{R} \to \mathbb{R}$, characterized by a parameter $\varepsilon > 0$, as

$$u_\varepsilon(t) = \begin{cases} \frac{1}{2\varepsilon} & \text{if } -\varepsilon \leq t \leq \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{6.9}$$

For some values of ε, this function is depicted in Figure 6.1, from which it is also easy to see that $\int_{-\infty}^{\infty} u_\varepsilon(t)\,\mathrm{d}t = 1$.

Using (6.9) as an input to the linear system (6.1) for $x_0 = 0$, we obtain the following result.

Lemma 6.2. *Consider the system (6.1) with $D = 0$ and define*

$$y_\varepsilon(t) = \int_{-\varepsilon}^{t} Ce^{A(t-\tau)}Bu_\varepsilon(\tau)\,\mathrm{d}\tau. \tag{6.10}$$

Then, we have that

$$\lim_{\varepsilon \to 0^+} y_\varepsilon(t) = \begin{cases} Ce^{At}B, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{6.11}$$

Proof. First, for $t \leq -\varepsilon$, it is clear that we have $y_\varepsilon(t) = 0$ for any $\varepsilon > 0$. Next, using the property $e^{A(t-\tau)} = e^{At}e^{-A\tau}$ (see Lemma 2.8) and the definition of the matrix exponential, we can write, for $t \geq \varepsilon$,

$$y_\varepsilon(t) = \int_{-\varepsilon}^{t} Ce^{A(t-\tau)}Bu_\varepsilon(\tau)\,\mathrm{d}\tau, \tag{6.12}$$

$$= \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} Ce^{At}\left(I - A\tau + \frac{A^2\tau^2}{2!} - \frac{A^3\tau^3}{3!} + \dots\right)B\,\mathrm{d}\tau \tag{6.13}$$

Since the sum for the matrix exponential converges uniformly and absolutely on $[-\varepsilon, \varepsilon]$, the integral can be distributed over the sum to obtain

$$y_\varepsilon(t) = \frac{1}{2\varepsilon}Ce^{At}\left(I(2\varepsilon) + \frac{A^2(2\varepsilon)^3}{3!} + \dots\right)B, \tag{6.14}$$

$$= Ce^{At}\left(I + \frac{A^2(2\varepsilon)^2}{3!} + \dots\right)B, \tag{6.15}$$

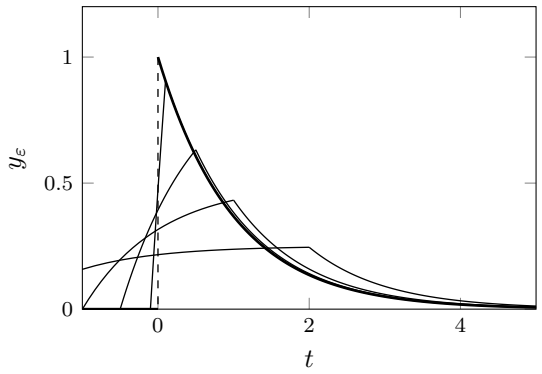


Figure 6.2: Output y_ε for the example (6.16) with input functions u_ε as in (6.9) for $\varepsilon = 2, 1, 0.5, 0.1$ (thin lines) as well as its limit e^{-t}, $t \geq 0$, and zero otherwise (thick line).

after which it can be observed that taking the limit $\varepsilon \to 0^+$ leads to (6.11). □

The inputs u_ε in (6.9) can be regarded as short impulses applied to the system. Lemma 6.2 shows that, when these impulses become increasingly short but of higher intensity, the output of the system (6.1) converges to the right-hand-side of (6.11). Therefore, we will refer to the right-hand-side of (6.11) as the *impulse response* (a formal definition is postponed to Definition 6.2).

Example 6.1. To illustrate the outputs to impulsive inputs of the form (6.9), let $A = -1$, $B = C = 1$, such that

$$Ce^{At}B = e^{-t}. \tag{6.16}$$

The outputs y_ε are given in Figure 6.2. ◇

We would like to regard the impulse response as the output for a given input function (rather than as a limit). Motivated by the result of Lemma 6.2, we could intuitively define this input function as the "limit" of a sequence of inputs of the form (6.9), even though the result is not a function in the ordinary sense. Instead, we consider the *Dirac delta function* δ.

Remark 6.2. The Dirac delta function (also referred to as the Dirac distribution), denoted δ, is not a function in the ordinary sense. However, as mentioned above, it can intuitively be defined as the "limit" of functions of the form (6.9), see also Figure 6.1. Alternatively, it can be treated as a function $\delta : \mathbb{R} \to \mathbb{R}$ with the defining properties that

1. $\delta(t) = 0$ for all $t \neq 0$;

2. for any continuous function $\phi : \mathbb{R} \to \mathbb{R}$, we have for any $t \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \phi(t - \tau)\delta(\tau)\, \mathrm{d}\tau = \phi(t). \tag{6.17}$$

The relation of the Dirac delta function with (6.9) can be made more explicit by noting that we also have

$$\lim_{\varepsilon \to 0^+} \int_{-\infty}^{\infty} \phi(t-\tau) u_\varepsilon(\tau) \, \mathrm{d}\tau = \phi(t), \tag{6.18}$$

for any continuous function $\phi : \mathbb{R} \to \mathbb{R}$. As a result, we can write

$$\int_{-\infty}^{\infty} \phi(t-\tau)\delta(\tau) \, \mathrm{d}\tau = \lim_{\varepsilon \to 0^+} \int_{-\infty}^{\infty} \phi(t-\tau) u_\varepsilon(\tau) \, \mathrm{d}\tau. \tag{6.19}$$

A rigorous treatment of the Dirac delta function would involve the study of so-called distributions, see, e.g., [9] for a discussion from a system theoretic perspective. ◁

From now on, we will treat the Dirac delta function as a function with the defining properties of Remark 6.2. Thus, we can take the input

$$u(t) = e_i \delta(t), \tag{6.20}$$

where $e_i \in \mathbb{R}^m$ is a vector of all zeros except for a one in element i. Thus, the input (6.20) can be regarded as an impulsive input in the i-th input. The resulting output (again for zero initial conditions) is given by

$$\begin{aligned} y(t; 0, e_i\delta) &= \int_0^t C e^{A(t-\tau)} B e_i \delta(\tau) \, \mathrm{d}\tau + D e_i \delta(t), \\ &= \big(C e^{At} B + D\delta(t) \big) e_i, \end{aligned} \tag{6.21}$$

for $t \geq 0$, as can be concluded after using the property (6.17).

This motivates the following definition.

Definition 6.2 (Impulse response matrix)**.** *For the system $\mathbf{\Sigma}$ in (6.1), its impulse response matrix is the matrix-valued function $H : \mathbb{R} \to \mathbb{R}^{p \times m}$ defined as*

$$H(t) = \begin{cases} C e^{At} B + D\delta(t), & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{6.22}$$

Thus, the impulse response matrix collects the impulse responses for the system (6.1) in the sense that the element H_{ji} is the response in output j to an impulse (in the form of a Dirac delta function) in input i.

The impulse response matrix has the important property that it can be used to fully characterize the output trajectory of (6.1) *for any* input function, i.e., not only for (6.20). This is stated next.

Theorem 6.3. *Consider the system $\mathbf{\Sigma}$ with impulse response matrix (6.22) and let the input function $u : \mathbb{R}_+ \to \mathbb{R}^m$ be given. Then,*

$$y(t; 0, u) = \int_0^t H(t-\tau) u(\tau) \, \mathrm{d}\tau, \tag{6.23}$$

for $t \geq 0$.

Proof. From the general output trajectory (6.2), we have for $x_0 = 0$ that

$$\begin{aligned} y(t;0,u) &= \int_0^t Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t), \\ &= \int_0^t Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + \int_0^t D\delta(t-\tau)u(\tau)\,\mathrm{d}\tau, \end{aligned} \tag{6.24}$$

where the property (6.17) of the Dirac delta function is obtained to obtain the latter. Then, collecting the integrals and comparing the result with (6.22) leads to (6.23). □

A second important property of the impulse response matrix is that it is independent of the choice of coordinates, as stated in the following theorem.

Theorem 6.4. *Let the systems $\mathbf{\Sigma}(A,B,C,D)$ and $\mathbf{\Sigma}(\bar{A},\bar{B},\bar{C},\bar{D})$ be similar and denote their impulse response matrices by H and $\bar{H}$, respectively. Then,*

$$H(t) = \bar{H}(t), \tag{6.25}$$

for all $t \in \mathbb{R}$.

Proof. This is a direct consequence of (6.7) in the proof of Theorem 6.1. □

The above two theorems illustrate the importance of the impulse response matrix. Namely, it completely captures the input-output behavior of a linear system (for zero initial condition) and is independent on the choice of coordinates in the state-space representation (6.1). For these reasons, the form (6.23) is sometimes said to characterize the *external behavior* of a linear system.

Example 6.2. Recall the mass-spring-damper system from Example 1.1, which is a linear system with matrices

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad D = 0, \tag{6.26}$$

when the vector of outputs contains both the position and velocity of the mass. We take $m = 1$ in this example, after which the following cases can be distinguished:

1. Taking $c = 3$ and $k = 2$, then the impulse response matrix (for $t \geq 0$), depicted in Figure 6.3, reads

$$H(t) = \begin{bmatrix} e^{-t} - e^{-2t} \\ 2e^{2t} - e^{-t} \end{bmatrix}, \quad t \geq 0. \tag{6.27}$$

   This corresponds to the so-called overdamped case in which A has two real eigenvalues.

2. For $c = 2$ and $k = 1$, we obtain

$$H(t) = \begin{bmatrix} te^{-t} \\ e^{-t}(1-t) \end{bmatrix}, \quad t \geq 0, \tag{6.28}$$

   corresponding to the case when A has one repeated eigenvalue. This is referred to as the critically damped case, which is depicted in Figure 6.4.

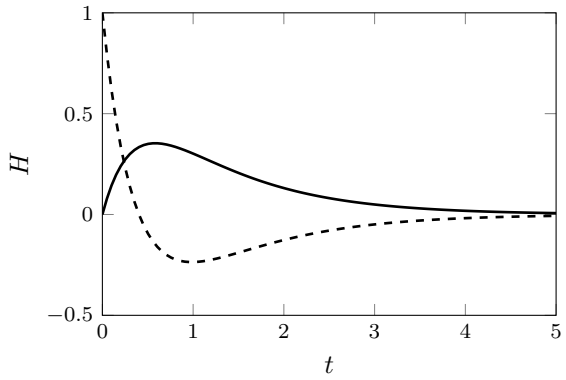


Figure 6.3: Impulse response matrix H in (6.27) for the "overdamped" mass-spring-damper system: position (H_1, solid) and velocity (H_2, dashed).

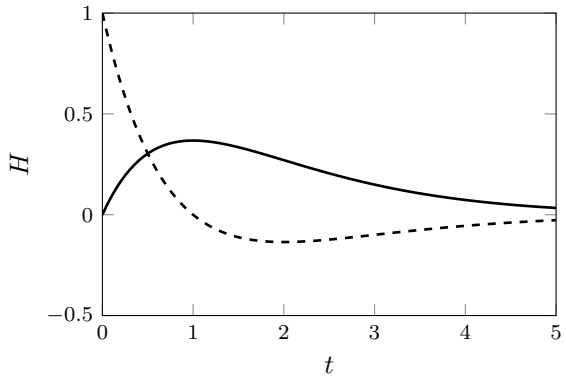


Figure 6.4: Impulse response matrix H in (6.28) for the "critically damped" mass-spring-damper system: position (H_1, solid) and velocity (H_2, dashed).

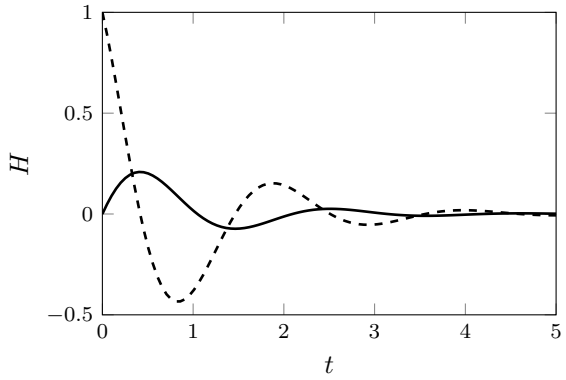


Figure 6.5: Impulse response matrix H in (6.29) for the "underdamped" mass-spring-damper system: position (H_1, solid) and velocity (H_2, dashed).

3. In case $c = 2$ and $k = 10$, it can be computed that

$$H(t) = \begin{bmatrix} \frac{1}{3}e^{-t}\sin 3t \\ e^{-t}\big(\cos 3t - \frac{1}{3}\sin 3t\big) \end{bmatrix}, \quad t \geq 0. \tag{6.29}$$

It can be seen that oscillations occur in the impulse response for the position and velocity, which corresponds to the so-called underdamped case. Then, A has two complex conjugate eigenvalues. An illustration is given in Figure 6.5.

A fourth case, where $c = 0$ and A has two eigenvalues on the imaginary axis, is omitted. ◇

The observation that the impulse response is independent of the choice of coordinates allows for a further characterization, as stated next.

Theorem 6.5. *Consider the system $\mathbf{\Sigma}$ in (6.1) and its impulse response matrix (6.22). Then, the following statements hold:*

1. *assume that the matrices A, B, and C are structured as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}, \tag{6.30}$$

then the impulse response matrix satisfies

$$H(t) = \begin{cases} C_1 e^{A_{11}t}B_1 + D\delta(t), & t \geq 0, \\ 0, & t < 0; \end{cases} \tag{6.31}$$

2. *assume that the matrices A, B, and C are structured as*

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \qquad C = \begin{bmatrix} C_1 & 0 \end{bmatrix}, \tag{6.32}$$

then the impulse response matrix satisfies

$$H(t) = \begin{cases} C_1 e^{A_{11}t}B_1 + D\delta(t), & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{6.33}$$

Proof. We only prove statement *1* as statement *2* follows similarly.

Given the structure (6.30) and the definition of the matrix exponential (recall Definition 2.1), it can be verified that

$$Ce^{At}B = C_1 e^{A_{11}t}B_1 \tag{6.34}$$

for all $t \in \mathbb{R}$. The result (6.31) now follows immediately. □

The importance of this result can be recognized after comparing the partitioning (6.30) with that in Theorem 4.11, which shows that such structure can always be obtained if $\mathbf{\Sigma}$ is not controllable (potentially after a change of coordinates). In fact, the matrices A_{11} and B_1 can be chosen to form a controllable pair, such that statement *1* of Theorem 6.5 essentially states that the impulse response of a system is only dependent on its controllable subsystem. Similarly, statement *2* states that the impulse response is independent of the unobservable subsystem (see Theorem 4.12).

The above result can intuitively be understood by recalling that the impulse response fully characterizes the input-output behavior of the linear system (6.1) and does not involve the state x explicitly. Consequently, states that cannot be influenced by the input or states that cannot be observed from the output do not affect the input-output behavior.

Remark 6.3. The input-output perspective on linear systems allows for a more general definition of input-output systems. Namely, for a given function $H : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{m \times p}$, we can define a linear input-output system through the integral representation

$$y(t) = \int_{-\infty}^{\infty} H(t, \tau) u(\tau) \, \mathrm{d}\tau, \tag{6.35}$$

that maps input functions $u : \mathbb{R} \to \mathbb{R}^m$ to output functions $y : \mathbb{R} \to \mathbb{R}^p$. In this case, the function H is sometimes referred to as the *kernel* of the integral representation (6.35). It is clear that such input-output system is linear. In addition, the system (6.35) is

- *causal* (or *non-anticipating*) if $H(t, \tau) = 0$ for all $\tau > t$. In this case, the output at time t only depends on past input values (i.e., on $u(s)$ for $s \leq t$) and not on future input values.

- *time invariant* if $H(t + s, \tau + s) = H(t, \tau)$ for all (t, τ) and all $s \in \mathbb{R}$, i.e., $H(t, \tau) = H(t - \tau, 0)$ for all (t, τ). In this case, there is no explicit time dependence in the linear system in the sense that a time-shifted input function leads to a time-shifted output function.

For a time-invariant system, we sometimes write $H(t, \tau)$ (with some abuse of notation) as a function of one variable, i.e., $H(t - \tau)$. Then, (6.35) leads to

$$y(t) = \int_{-\infty}^{\infty} H(t - \tau) u(\tau) \, \mathrm{d}\tau, \tag{6.36}$$

Such time-invariant system is causal if $H(t) = 0$ for $t < 0$, in which case the integration limits can be changed to

$$y(t) = \int_{-\infty}^{t} H(t - \tau) u(\tau) \, \mathrm{d}\tau. \tag{6.37}$$

With this terminology in mind, we see that the impulse response matrix (6.22) for a linear state-space system (6.1) gives rise to a time-invariant and causal integral representation. Here, the lower integration limit can be changed from $-\infty$ to 0 in case input functions are taken that satisfy $u(t) = 0$ for all $t < 0$.

Nonetheless, we stress that the integral representation (6.35) of the input-output system is more general than the input-output behavior generated by a linear state-space system (6.1). As an example, consider the moving average system

$$y(t) = \frac{1}{T} \int_{t-T}^{t} u(\tau) \, \mathrm{d}\tau, \tag{6.38}$$

which is a linear time-invariant causal system of the form (6.35) with

$$H(t,\tau) = \begin{cases} \frac{1}{T}, & 0 \le t-\tau \le \frac{1}{T}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.39}$$

However, there does *not* exists a state-space system of the form (6.1) that generates the same input-output behavior. ◁

Remark 6.4. It is insightful to consider linear *time-invariant* input-output systems as in (6.36) in Remark 6.3 under *periodic* inputs. Specifically, assume that $m = p = 1$ (the system has a single input and single output) and let $u : \mathbb{R} \to \mathbb{R}$ be periodic with period $T > 0$, i.e.,

$$u(t+T) = u(t) \tag{6.40}$$

for all $t \in \mathbb{R}$. Then, after a change of integration variable in (6.36), we obtain

$$y(t) = \int_{-\infty}^{\infty} H(t-\tau)u(\tau)\,\mathrm{d}\tau = \int_{-\infty}^{\infty} H(\tau)u(t-\tau)\,\mathrm{d}\tau, \tag{6.41}$$

from which it is readily concluded that

$$y(t+T) = \int_{-\infty}^{\infty} H(\tau)u(t+T-\tau)\,\mathrm{d}\tau = \int_{-\infty}^{\infty} H(\tau)u(t-\tau)\,\mathrm{d}\tau = y(t) \tag{6.42}$$

for all $t \in \mathbb{R}$, i.e., the output y is periodic with the *same* period as the input u.

We can strengthen this observation for input functions of the form $u(t) = \sin \omega t$ for some frequency $\omega \in \mathbb{R}$, which, by Euler's formula, can be written as

$$u(t) = \frac{e^{i\omega t} - e^{-i\omega t}}{2i}, \tag{6.43}$$

with i satisfying $i^2 = -1$ the imaginary unit. The substitution of (6.43) in (6.41) gives

$$y(t) = \int_{-\infty}^{\infty} H(\tau)\frac{e^{i\omega(t-\tau)} - e^{-i\omega(t-\tau)}}{2i}\,\mathrm{d}\tau, \tag{6.44}$$

$$= \frac{1}{2i}\int_{-\infty}^{\infty} H(\tau)e^{-i\omega\tau}\,\mathrm{d}\tau\, e^{i\omega t} - \frac{1}{2i}\int_{-\infty}^{\infty} H(\tau)e^{i\omega\tau}\,\mathrm{d}\tau\, e^{-i\omega t}, \tag{6.45}$$

$$= \frac{1}{2i}T(i\omega)e^{i\omega t} - \frac{1}{2i}T(-i\omega)e^{-i\omega t}. \tag{6.46}$$

Here, we have defined

$$T(i\omega) = \int_{-\infty}^{\infty} H(\tau)e^{-i\omega\tau}\,\mathrm{d}\tau, \tag{6.47}$$

which is known as the *Fourier transform* of H. Assuming that H is such that the integral in (6.47) is well-defined, it is easily observed (recall that H is real-valued) that $T(-i\omega) = \overline{T(i\omega)}$, where $\bar{c}$ is the complex conjugate of a complex number c. Then, again using Euler's formula, (6.46) can be written as

$$y(t) = \frac{T(i\omega) - \overline{T(i\omega)}}{2i}\cos \omega t + \frac{T(i\omega) + \overline{T(i\omega)}}{2}\sin \omega t \tag{6.48}$$

$$= \mathrm{Im}(T(i\omega))\cos \omega t + \mathrm{Re}(T(i\omega))\sin \omega t, \tag{6.49}$$

$$= |T(i\omega)|\sin\big(\omega t + \arg(T(i\omega))\big), \tag{6.50}$$

where the final equality is obtained by simple trigonometric identities. Here, $|T(i\omega)|$ and $\arg(T(i\omega))$ are the magnitude (or modulus) and phase (or argument) of the complex number $T(i\omega)$.

Hence, for a sinusoidal input $u(t) = \sin \omega t$, the output of a linear time-invariant input-output system is again a sinusoid with the *same* frequency, but with different amplitude and phase, namely $|T(i\omega)|$ and $\arg(T(i\omega))$. As the function (6.47) completely characterizes the response to periodic inputs, it is sometimes referred to as the *frequency response function*.

We will also derive this frequency response function as a special case of the transfer function matrix introduced in the next section, which will also provide insights in when the integral in (6.47) is well-defined. ◁

6.2 The transfer function matrix

Whereas we considered impulses as the inputs to the linear system (6.1) in the previous section, we will start this section by considering the output response of (6.1) to input signals of the form

$$u(t) = \hat{u}e^{\sigma t}, \tag{6.51}$$

where $\sigma \in \mathbb{R}$ and $\hat{u} \in \mathbb{R}^m$. We assume that σ is chosen such that the matrix $(\sigma I - A)$ is nonsingular, i.e., σ is not an eigenvalue of A. Then, we can show that there exists unique solutions for the state and output of the form

$$x(t) = \hat{x}e^{\sigma t}, \qquad y(t) = \hat{y}e^{\sigma t}, \tag{6.52}$$

with $\hat{x} \in \mathbb{R}^n$ and $\hat{y} \in \mathbb{R}^p$. Namely, a substitution of (6.51) and (6.52) in the dynamic equation for the system (6.1) leads to

$$\sigma \hat{x}e^{\sigma t} = A\hat{x}e^{\sigma t} + B\hat{u}e^{\sigma t}, \tag{6.53}$$

which can be seen (note that $e^{\sigma t} > 0$ for all $t \in \mathbb{R}$) to have the unique solution

$$\hat{x} = (\sigma I - A)^{-1}B\hat{u}. \tag{6.54}$$

Similarly, the substitution of (6.51) and (6.52) in the output equation for the system (6.1) yields

$$\hat{y}e^{\sigma t} = C\hat{x}e^{\sigma t} + D\hat{u}e^{\sigma t}, \tag{6.55}$$

after which the use of (6.54) shows that $\hat{y}$ can be obtained directly from $\hat{u}$ as

$$\hat{y} = \big(C(\sigma I - A)^{-1}B + D\big)\hat{u}. \tag{6.56}$$

Thus, for input functions of the form (6.51), there exists a unique output of the form (6.52) and, more importantly, this output can be obtained without explicitly solving the differential equation (6.1). Instead, only the evaluation of the matrix function in (6.56) is needed and the relation between the input function (characterized through $\hat{u}$) and the output function is obtained by a simple multiplication. We will soon see that this matrix function plays an important role in the analysis of linear systems.

Remark 6.5. In the above discussion, we note that there is no mention of the initial condition for the differential equation (6.1). Instead, we can think of the functions (6.51) and (6.52) as being defined for all $t \in \mathbb{R}$. Therefore, the solutions (6.52) are sometimes referred to as the *steady state* response to the input (6.51). ◁

We will generalize the above procedure to more general input functions through the so-called Laplace transform. Before giving its definition, we note that a function[1] $f : \mathbb{R}_+ \to \mathbb{R}$ is *exponentially bounded* if there exists M, $\alpha > 0$ such that

$$|f(t)| \leq Me^{\alpha t} \tag{6.57}$$

for all $t \geq 0$. Here, we refer to α as the bounding exponent.

Now, the Laplace transform can be defined as follows.

Definition 6.3. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be exponentially bounded with bounding exponent α. Then, the Laplace transform of f, denoted $\mathcal{L}(f)$, is defined as*

$$\mathcal{L}(f)(s) = \int_0^\infty f(t)e^{-st}\,\mathrm{d}t, \tag{6.58}$$

where $s \in \mathbb{C}$ with $\mathrm{Re}(s) > \alpha$.

We stress that $\mathcal{L}(f)$ is itself a function, but of the so-called Laplace variable s rather than time t. Moreover, the Laplace transform can be defined for functions that are vector-valued or matrix-valued by a straightforward extension of Definition 6.3.

Remark 6.6. Using the assumption that f is exponentially bounded, it can be shown that the integral (6.58) converges uniformly in any domain (in the complex plane) of the form $\mathrm{Re}(s) \geq \beta$ if $\beta > \alpha$. In this domain, the Laplace transform is an analytic function.

For values of $s \in \mathbb{C}$ that are not in this domain, the integral (6.58) does not necessarily converge. Nonetheless, it follows from the theory of complex functions that there exists a unique so-called *analytic continuation* of $\mathcal{L}(f)$ that extends $\mathcal{L}(f)$ to a function that is defined and analytic on the entire complex plane, with the exception of a number of isolated singularities. Since we will not distinguish between a $\mathcal{L}(f)$ and its analytic continuation, we will not pay much attention to the region of convergence of the Laplace transform (6.58) in the remainder of this chapter. ◁

Example 6.3. Consider the function

$$f(t) = e^{\sigma t}, \tag{6.59}$$

with $\sigma \in \mathbb{R}$. The function is clearly exponentially bounded, with any bounding exponent $\alpha > \sigma$. Using the definition, its Laplace transform is given as

$$\mathcal{L}(f)(s) = \int_0^\infty e^{\sigma t}e^{-st}\,\mathrm{d}t = \int_0^\infty e^{(\sigma-s)t}\,\mathrm{d}t = \frac{1}{s-\sigma}\left(1 - \lim_{t\to\infty} e^{(\sigma-s)t}\right). \tag{6.60}$$

[1] We use the notation $\mathbb{R}_+ = [0, \infty)$.

The integral converges if $\mathrm{Re}(s) > \sigma$, in which case

$$\mathcal{L}(f)(s) = \frac{1}{s-\sigma}. \tag{6.61}$$

However, following Remark 6.6, we can ignore the issue of convergence and simply associate the rational function (6.61) with the transform of (6.59). ◇

The following properties of the Laplace transform hold.

Theorem 6.6. *Consider the Laplace transform (6.58). Then, the following properties hold:*

1. *the Laplace transform is linear, i.e.,*

$$\mathcal{L}(f+f') = \mathcal{L}(f) + \mathcal{L}(f'), \qquad \mathcal{L}(\alpha f) = \alpha\mathcal{L}(f), \tag{6.62}$$

   for any exponentially bounded functions f, $f' : \mathbb{R}_+ \to \mathbb{R}$ and $\alpha \in \mathbb{C}$;

2. *if $f : \mathbb{R}_+ \to \mathbb{R}$ is continuously differentiable and $\dot{f}$ is exponentially bounded, then f is exponentially bounded and*

$$\mathcal{L}(\dot{f}) = s\mathcal{L}(f) - f(0); \tag{6.63}$$

3. *if u, $h : \mathbb{R}_+ \to \mathbb{R}$ are exponentially bounded, then*

$$y(t) = \int_0^t h(t-\tau)u(\tau)\,\mathrm{d}\tau \tag{6.64}$$

   is exponentially bounded and, moreover,

$$\mathcal{L}(y) = \mathcal{L}(h)\mathcal{L}(u). \tag{6.65}$$

Proof. This is Exercise 6.4. □

Now, we return to solutions of the linear system (6.1), taking initial condition $x_0 = 0$ and some input function $u : \mathbb{R}_+ \to \mathbb{R}^m$. However, rather than considering the input function u and the resulting state and output trajectories as functions of time, we consider their Laplace transforms as

$$\hat{x}(s) = \mathcal{L}(x)(s), \qquad \hat{u}(s) = \mathcal{L}(u)(s), \qquad \hat{y}(s) = \mathcal{L}(y)(s). \tag{6.66}$$

Then, taking the Laplace transformation of the dynamics in (6.1) leads to

$$\mathcal{L}(\dot{x}) = \mathcal{L}(Ax + Bu) = A\mathcal{L}(x) + B\mathcal{L}(u), \tag{6.67}$$

where linearity of the Laplace transform (see statement *1* of Theorem 6.6) is used to obtain the final equality. Subsequently, the use of property (6.63) together with the definitions (6.66) yields

$$s\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s), \tag{6.68}$$

after recalling that $x(0) = 0$. Solving for $\hat{x}$ now gives

$$\hat{x}(s) = (sI - A)^{-1}B\hat{u}(s). \tag{6.69}$$

A similar derivation can be performed by taking the Laplace transform of the output equation in (6.1), after which substitution of (6.69) can be shown to give

$$\hat{y}(s) = \big(C(sI - A)^{-1}B + D\big)\hat{u}(s). \tag{6.70}$$

The result (6.70) gives a direct relation between the Laplace transform of the input function and the Laplace transform of the output trajectory, where the latter can be obtained after a simple multiplication.

This result is so important that this multiplication factor warrants the following definition.

Definition 6.4 (Transfer function matrix)**.** *Consider the system* $\mathbf{\Sigma}$ *in (6.1). Then, the function*

$$T(s) = C(sI - A)^{-1}B + D \tag{6.71}$$

in indeterminate s *is called the transfer function matrix of* $\mathbf{\Sigma}$*.*

Thus, by relating the Laplace transforms of the input and output, the transfer function matrix gives a full characterization of the input-output behavior (or external behavior) of a linear system $\mathbf{\Sigma}$ of the form (6.1), albeit in the so-called Laplace domain (i.e., in the variable s).

Next, after recalling that the input-output behavior in the time domain can be given through the convolution form (6.23) and by observing the property (6.65) of the Laplace transform, it follows immediately that the transfer function matrix in Definition 6.4 is related to the impulse response matrix in Definition 6.2.

This is made precise in the following theorem.

Theorem 6.7. *Consider the system* $\mathbf{\Sigma}$ *in (6.1), its impulse response matrix* H *in (6.22), and transfer function matrix* T *in (6.71). Then,*

$$T(s) = \mathcal{L}(H)(s), \tag{6.72}$$

for all s *such that* $\mathrm{Re}(s) > \Lambda(A)$*, with* $\Lambda(A)$ *the spectral abscis*[2] *of* A*.*

Proof. As a first step, we will consider the Laplace transform of e^{At}. A direct use of the definition (6.58) (but extended for matrix-valued functions) gives

$$\mathcal{L}(e^{At})(s) = \int_0^\infty e^{At}e^{-st}\,\mathrm{d}t = \int_0^\infty e^{-(sI-A)t}\,\mathrm{d}t, \tag{6.73}$$

where Lemma 2.8 is used. To evaluate this integral, note that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big\{-(sI - A)^{-1}e^{-(sI-A)t}\Big\} = e^{-(sI-A)t} \tag{6.74}$$

by Lemma 2.4, after which substitution in (6.73) yields

$$\mathcal{L}(e^{At})(s) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\Big\{-(sI - A)^{-1}e^{-(sI-A)t}\Big\}\,\mathrm{d}t, \tag{6.75}$$

$$= (sI - A)^{-1}\left(I - \lim_{t\to\infty} e^{-(sI-A)t}\right). \tag{6.76}$$

[2]Recall that $\Lambda(A) = \max\{\mathrm{Re}(\lambda) \mid \lambda \in \sigma(A)\}$, see Remark 3.3.

For $s \in \mathbb{C}$ such that $\mathrm{Re}(s) > \Lambda(A)$, we have that the eigenvalues of $A - sI$ have strictly negative real parts, i.e., $\sigma(A - sI) \subset \mathbb{C}_-$. As a result of Theorem 3.3,

$$\lim_{t\to\infty} e^{-(sI-A)t} = 0 \tag{6.77}$$

for such s, after which it follows that

$$\mathcal{L}(e^{At})(s) = (sI - A)^{-1}. \tag{6.78}$$

Next, we note that, by definition of the Dirac delta function,

$$\mathcal{L}(\delta)(s) = \int_0^\infty \delta(t)e^{-st}\,\mathrm{d}t = 1 \tag{6.79}$$

for all $s \in \mathbb{C}$.

Now, combining results and using the definition of the impulse response matrix in (6.22), we obtain

$$\begin{aligned}\mathcal{L}(H)(s) = \mathcal{L}\big(Ce^{At}B + D\delta(t)\big) &= C\mathcal{L}(e^{At})B + D\mathcal{L}(\delta),\\ &= C(sI - A)^{-1}B + D\end{aligned} \tag{6.80}$$

for $s \in \mathbb{C}$ such that $\mathrm{Re}(s) > \Lambda(A)$. Here, we have used the linearity of the Laplace transform, see Theorem 6.6. This proves the desired result after recalling the definition of the transfer function matrix in (6.71). □

Remark 6.7. In the derivation of the result (6.79), we have implicitly assumed that the contribution of the impulse at $t = 0$ is taken into account in the integral, even though this equals the lower integration limit. A more rigorous definition of the Laplace transform would therefore be

$$\mathcal{L}^-(f)(s) = \lim_{\varepsilon\to 0^-} \int_\varepsilon^\infty f(s)e^{-st}\,\mathrm{d}t = \int_{0^-}^\infty f(s)e^{-st}\,\mathrm{d}t. \tag{6.81}$$

In this chapter, whenever the Dirac delta function is involved, we assume the Laplace transform to be defined as (6.81). ◁

Remark 6.8. Assuming that $\Lambda(A) < 0$, we can evaluate (6.72) at the imaginary axis $s = i\omega$ to obtain

$$T(i\omega) = \mathcal{L}(H)(i\omega) = \int_0^\infty H(t)e^{-i\omega t}\,\mathrm{d}t, \tag{6.82}$$

where Definition 6.2 is used. Now, after recalling that $H(t) = 0$ for $t < 0$, it is clear that (6.82) is exactly the definition of the Fourier transform in (6.47) in Remark 6.4. From this remark it is clear that the evaluation of the transfer function on the imaginary axis gives the frequency response function, which has the nice interpretation of relating sinusoidal input and output signals. ◁

Given the fact that the impulse response matrix is independent of the choice of coordinates in which a linear system (6.1) is expressed, the relation also implies that the transfer function matrix is independent of this choice as well. This is stated next without proof (a direct proof on the basis of Definition 6.4 is also easy to state).

Theorem 6.8. *Let the systems* $\mathbf{\Sigma}(A, B, C, D)$ *and* $\mathbf{\Sigma}(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ *be similar denote their transfer function matrices by* T *and* $\bar{T}$*, respectively. Then,*

$$T(s) = \bar{T}(s), \tag{6.83}$$

for all $s \in \mathbb{C}$*.*

We have now established the transfer function matrix as a full characterization of the input-output behavior of a linear system. In the remainder of this section, we take a closer look at the structure of such transfer function matrix.

To do so, note that *Cramer's rule* states that

$$(sI - A)^{-1} = \frac{1}{\Delta_A(s)} \operatorname{adj}(sI - A), \tag{6.84}$$

where $\Delta_A(s) = \det(sI - A)$ is the characteristic polynomial of the matrix A and $\operatorname{adj}(sI - A)$ is the *adjugate* of the matrix $sI - A$.

Remark 6.9. The adjugate $\operatorname{adj} M$ of a matrix $M \in \mathbb{R}^{n\times n}$ is the transpose of the matrix of cofactors. This means that the element (j, i) in $\operatorname{adj} M$ is given by cofactor (i, j), i.e.,

$$(\operatorname{adj} M)_{ji} = (-1)^{i+j} m_{ij}, \tag{6.85}$$

with m_{ij} the determinant of the matrix obtained by removing row i and column j from M. ◁

Using (6.84), the transfer function can be written as

$$C(sI - A)^{-1}B + D = \frac{1}{\Delta_A(s)} C \operatorname{adj}(sI - A)B + D. \tag{6.86}$$

As $\operatorname{adj}(sI - A)$ consists of determinants of submatrices of $sI - A$, it is a matrix of polynomials of degree at most $n-1$. Therefore, as $C \operatorname{adj}(sI - A)B$ comprises linear combinations of these polynomials, the matrix $C \operatorname{adj}(sI - A)B$ also contains polynomials of degree at most $n - 1$. Note also that $C \operatorname{adj}(sI - A)B \in \mathbb{C}^{p\times m}$ for each $s \in \mathbb{C}$ such that $sI - A$ is invertible. Thus,

$$\frac{1}{\Delta_A(s)} C \operatorname{adj}(sI - A)B \tag{6.87}$$

is a matrix of rational functions (recall that Δ_A is the characteristic polynomial of A and is of degree n).

The following example illustrates these observations.

Example 6.4. Consider again the mass-spring-damper system in Example 1.1 and recall that it is a linear system (6.1) with matrices

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad D = 0, \tag{6.88}$$

when the position of the mass is taken as the output.

As a first step in obtaining the transfer function matrix, the computation of $\operatorname{adj}(sI - A)$ leads to

$$\operatorname{adj}(sI - A) = \begin{bmatrix} s + \frac{c}{m} & 1 \\ -\frac{k}{m} & s \end{bmatrix}, \tag{6.89}$$

as can be concluded from the definition of the adjugate. Then, as a result of Cramer's rule in (6.84),

$$(sI - A)^{-1} = \frac{1}{s^2 + \frac{c}{m}s + \frac{k}{m}} \begin{bmatrix} s + \frac{c}{m} & 1 \\ -\frac{k}{m} & s \end{bmatrix}. \tag{6.90}$$

Then, it follows directly that

$$T(s) = C(sI - A)^{-1}B + D = \frac{\frac{1}{m}}{s^2 + \frac{c}{m}s + \frac{k}{m}} = \frac{1}{ms^2 + cs + k}. \tag{6.91}$$

It is interesting to note that, if we choose the velocity of the mass as the output (instead of its position), the output matrix C needs to be replaced by

$$C' = \begin{bmatrix} 0 & 1 \end{bmatrix}. \tag{6.92}$$

In this case, the transfer function becomes

$$T'(s) = C'(sI - A)^{-1}B + D = \frac{\frac{s}{m}}{s^2 + \frac{c}{m}s + \frac{k}{m}} = \frac{s}{ms^2 + cs + k}. \tag{6.93}$$

We observe that $T'(s) = sT(s)$ which, after recalling that the velocity is the time-derivative of the position, can be seen to correspond with the property (6.63) of the Laplace transform.

Finally, assuming that measurements of both the position and velocity of the mass are available, the output matrix C'' reads $C'' = I$. In this case, the corresponding transfer function *matrix* reads

$$T''(s) = \begin{bmatrix} T(s) \\ T'(s) \end{bmatrix} = \begin{bmatrix} \frac{1}{ms^2 + cs + k} \\ \frac{s}{ms^2 + cs + k} \end{bmatrix} = \frac{1}{ms^2 + cs + k} \begin{bmatrix} 1 \\ s \end{bmatrix}, \tag{6.94}$$

where common factors have been collected to obtain the final result. ◇

Remark 6.10 (Bode plot). Transfer functions (or transfer function matrices) are typically depicted using a so-called *Bode plot*. A Bode plot of an individual transfer function T (i.e., one element from the transfer function matrix, corresponding to a single input and a single output) can be generated after evaluating T for $s = i\omega$, with i the imaginary unit and $\omega \in \mathbb{R}$. For a given ω, $T(i\omega)$ is a complex number. It can be written as

$$T(i\omega) = |T(i\omega)|e^{i\arg(T(i\omega))}, \tag{6.95}$$

with $|T(i\omega)|$ the magnitude (or modulus) and $\arg(T(i\omega))$ the phase (or argument) of $T(i\omega)$. Now, a Bode plot contains two graphs which depict the magnitude and phase of $T(i\omega)$, respectively, both as a function of ω. It is customary to use a logarithmic scale for the ω-axis. We recall that the function $\omega \mapsto T(i\omega)$ is sometimes referred to as the *frequency response function*, see Remarks 6.4 and 6.8.

The Bode plot of a transfer function of a system is a convenient way to evaluate the dynamic behavior of such system as it easier to interpret than the matrices describing a state-space realization $\mathbf{\Sigma}(A, B, C, D)$. This is especially apparent after recalling that these matrices can differ even for similar systems, whereas the transfer function is invariant under such coordinate transformations. The Bode plot is therefore an important tool for engineers. ◁

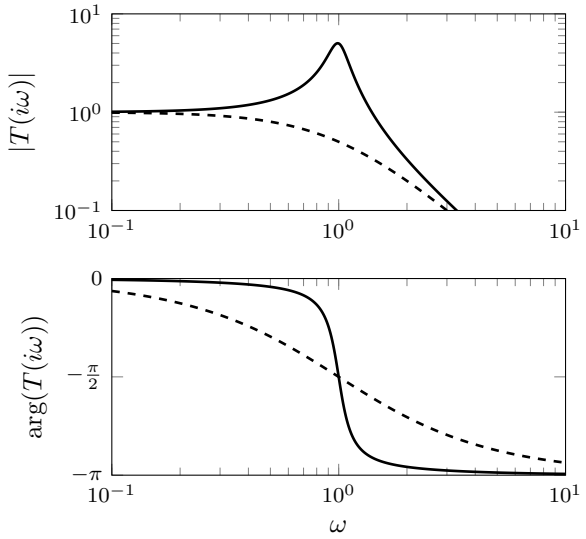


Figure 6.6: Bode plot for the transfer function (6.91) for $m = 1$, $k = 1$, and $c = 0.2$ (solid) or $c = 1$ (dashed).

Example 6.5. Consider the transfer function (6.91) for the mass-spring-damper system (when the position is measured). Note that

$$T(i\omega) = \frac{1}{-m\omega^2 + ci\omega + k}, \tag{6.96}$$

after which the Bode plot is given in Figure 6.6. ◇

6.3 Transfer functions for SISO systems

Whereas transfer function matrices for general linear systems were introduced in the previous section, this section considers linear systems with a single input and a single output. For such single-input single-output (SISO) systems, the transfer function matrix reduces to a scalar transfer function. In this section, properties of these transfer functions are considered.

Thus, we consider single-input single-output linear systems in the standard form

$$\mathbf{\Sigma}_{\text{SISO}} : \begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t), \end{cases} \tag{6.97}$$

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}$, and output $y(t) \in \mathbb{R}$ (note that this is a system as in (6.1) with $m = 1$ and $p = 1$). For simplicity, we omit the direct feedthrough term ($D = 0$).

As the transfer function of a system (6.97) is a rational function, we will consider rational functions (with indeterminate s) as

$$T(s) = \frac{p_m s^m + p_{m-1} s^{m-1} + \ldots + p_1 s + p_0}{q_n s^n + q_{n-1} s^{n-1} + \ldots + q_1 s + q_0}, \tag{6.98}$$

with $p_0, \ldots, p_m \in \mathbb{R}$, $p_m \neq 0$ and $q_0, \ldots, q_n \in \mathbb{R}$, $q_n \neq 0$. Such function is said to be *proper* if $n \geq m$ and *strictly proper* if $n > m$. We stress that, in this section, m denotes the degree of the numerator polynomial rather than the number of inputs to the linear system (6.97).

The following result relates the coefficients of the transfer function (6.98) to the controllability canonical form of a single-input system of Theorem 4.13.

Theorem 6.9. *Consider the SISO system $\mathbf{\Sigma}_{\text{SISO}}$ in (6.97) with matrices*

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad C^{\mathrm{T}} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-3} \\ c_{n-2} \\ c_{n-1} \end{bmatrix}. \tag{6.99}$$

Then, its transfer function reads

$$T(s) = \frac{c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \ldots + c_1 s + c_0}{s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0}. \tag{6.100}$$

Proof. A direct but lengthy computation shows that $\operatorname{adj}(sI - A)$ is of the form

$$\operatorname{adj}(sI - A) = \begin{bmatrix} \star & \cdots & \star & 1 \\ \star & \cdots & \star & s \\ \vdots & \ddots & \vdots & \vdots \\ \star & \cdots & \star & s^{n-1} \end{bmatrix}, \tag{6.101}$$

where the stars denote entries that are not of interest. As a result, we have

$$\operatorname{adj}(sI - A)B = \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix}, \tag{6.102}$$

after which it follows immediately that

$$C \operatorname{adj}(sI - A)B = c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \ldots + c_1 s + c_0. \tag{6.103}$$

Moreover, after noting that the matrix A is in companion form, we have that its characteristic polynomial reads

$$\Delta_A(s) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0, \tag{6.104}$$

see Exercise 4.13. Then, combining the results (6.103) and (6.104) using Cramer's rule in (6.86), we obtain the desired result (6.100). □

The result of Theorem 6.9 provides a clear relation between the system matrices of a linear system and its transfer function, when the system is in the controllability canonical form (6.99). This is not only convenient when computing the transfer function for a given linear system, but is mostly relevant in addressing the *converse* question of finding system matrices for a given transfer function.

Specifically, assume that the transfer function (6.98) is given and that $m \leq n-1$, i.e., the transfer function is strictly proper. Then, after scaling numerator and denominator by q_n to make the denominator monic, the transfer function is of the form (6.100) and corresponding system matrices are given by (6.99). The linear system $\mathbf{\Sigma}_{\text{SISO}}$ is in this case called a *realization* of the transfer function. We note that such realization is not unique as a result of Theorem 6.8 (see Theorem 6.11 for a realization based on the observability canonical form).

Example 6.6. Consider the transfer function given by the rational function

$$T(s) = \frac{s^2 + 2s + 1}{s^3 + 4s^2 + s - 6}. \tag{6.105}$$

Using Theorem 6.9, it is immediate that the SISO system (6.97) with

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 6 & -1 & -4 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}, \tag{6.106}$$

has (6.105) as its transfer function, i.e., is a realization of (6.105). Moreover, by construction, the matrix pair (A, B) is controllable. ◇

Remark 6.11. The above discussion is easily extended when $m = n$ in the transfer function (6.98). Namely, after assuming $q_n = 1$ for simplicity, it is immediate that

$$\begin{aligned} &\frac{p_n s^n + p_{n-1} s^{n-1} + \ldots + p_1 s + p_0}{s^n + q_{n-1} s^{n-1} + \ldots + q_1 s + q_0} \\ &\qquad = p_n + \frac{(p_{n-1} - p_n q_{n-1}) s^{n-1} + \ldots + (p_1 - p_n q_1) s + (p_0 - p_n q_0)}{s^n + q_{n-1} s^{n-1} + \ldots + q_1 s + q_0}. \end{aligned} \tag{6.107}$$

Here, the second term at the right-hand-side of (6.107) is a strictly proper transfer function for which a realization is given by Theorem 6.9 for the choice

$$a_i = q_i, \qquad c_i = p_i - p_n q_i \tag{6.108}$$

for $i = 0, 1, \ldots, n-1$. The addition of a direct feedthrough matrix $D = p_n$ then leads to

$$C(sI - A)^{-1} B + D = \frac{p_n s^n + p_{n-1} s^{n-1} + \ldots + p_1 s + p_0}{s^n + q_{n-1} s^{n-1} + \ldots + q_1 s + q_0}. \tag{6.109}$$

Thus, for any proper rational transfer function, there exists a realization of the form (6.97), although a direct feedthrough term needs to be added in case the transfer function is not strictly proper.

The topic of *realization theory* provides a more thorough investigation of the linear systems that generate a certain given transfer function, also for systems that are not single-input single-output. ◁

Remark 6.12. Transfer functions (6.98) also allow for an insightful interpretation in the time domain. To this end, recall from (6.70) that the transfer function related the (Laplace transforms of the) input and output as

$$\frac{\hat{y}(s)}{\hat{u}(s)} = T(s), \tag{6.110}$$

which can be written as

$$\begin{aligned}&\big(q_n s^n + q_{n-1}s^{n-1} + \ldots + q_1 s + q_0\big)\hat{y}(s)\\ &\qquad = \big(p_m s^m + p_{m-1}s^{m-1} + \ldots + p_1 s + p_0\big)\hat{u}(s).\end{aligned} \tag{6.111}$$

After recalling from Theorem 6.6 that $s\mathcal{L}(y)$ (for $y(0) = 0$) is the Laplace transform of $\dot{y}$, it follows that (6.111) is the Laplace transform of

$$\begin{aligned}&q_n y^{(n)}(t) + q_{n-1}y^{(n-1)}(t) + \ldots + q_1\dot{y}(t) + q_0 y(t)\\ &\qquad = p_m u^{(m)}(t) + p_{m-1}u^{(m-1)}(t) + \ldots + p_1\dot{u}(t) + p_0 u(t),\end{aligned} \tag{6.112}$$

where we have used the notation

$$y^{(k)}(t) = \frac{\mathrm{d}^k y}{\mathrm{d}t^k}(t) \tag{6.113}$$

(and similar for u) to denote higher-order time derivatives.

The time-domain input-output description (6.112) can also immediately be placed in the state-space form (6.97) if $m \leq n-1$. Namely, define x_1 to satisfy

$$q_n x_1^{(n)}(t) + q_{n-1}x_1^{(n-1)}(t) + \ldots + q_1\dot{x}_1(t) + q_0 x_1(t) = u(t), \tag{6.114}$$

such that $y = p_m x_1^{(m)} + \ldots + p_1\dot{x}_1 + p_0 x_1$ satisfies (6.112). Then, after letting $x_2, \ldots, x_n$ be such that

$$x_i = \frac{\mathrm{d}^{i-1}x_1}{\mathrm{d}t^{i-1}}, \quad i = 2, 3, \ldots, n, \tag{6.115}$$

it follows that the dynamics (6.112) can be written as (6.97) with the matrices (6.99) with $a_i = q_i$ and $c_i = p_i$. In this latter step, we have again assumed that $q_n = 1$ for simplicity. ◁

From Theorem 4.13 (see also (4.87)), it is clear that the system (6.97) with matrices (6.99) is controllable, i.e., it gives a controllable realization of the transfer function (6.100). The following result gives, for controllable systems, a relation between the transfer function and observability. Here, we recall that two polynomials p and q are *coprime* if they do not have common factors, i.e., they do not have common roots.

Theorem 6.10. *Consider the SISO system $\mathbf{\Sigma}_{\text{SISO}}$ in (6.97) and let the matrix pair (A, B) be controllable. Then, the polynomials*

$$p(s) = C\,\mathrm{adj}(sI - A)B, \qquad q(s) = \Delta_A(s) \tag{6.116}$$

are coprime if and only if the matrix pair (A, C) is observable.

Proof. First, we note that the system (6.97) can be brought in a form in which its matrices have the structure (6.99). Namely, this follows from the assumption that the matrix pair (A, B) is controllable and Theorem 4.13. As the transfer function remains unchanged after such change of coordinates (see Theorem 6.8), we may assume without loss of generality that $\mathbf{\Sigma}_{\text{SISO}}$ is characterized by the matrices (6.99). Now, we prove sufficiency and necessity separately.

only if. To prove this statement by contraposition, let the matrix pair (A, C) be not observable. Then, by Theorem 4.15, there exists an eigenvalue $\lambda \in \sigma(A)$

that is not (A, C)-observable, i.e., there exists a nonzero vector $v \in \mathbb{C}^n$ such that

$$\begin{bmatrix} A - \lambda I \\ C \end{bmatrix} v = 0. \tag{6.117}$$

To further characterize this vector v, denote

$$v = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^{\mathrm{T}}. \tag{6.118}$$

From (6.117), it can be concluded that v is an eigenvector of A for the eigenvalue λ. Specifically, the eigenvalue equation $Av = \lambda v$ leads to

$$\begin{aligned} v_2 &= \lambda v_1, \\ v_3 &= \lambda v_2, \\ &\vdots \\ v_n &= \lambda v_{n-1}, \\ -a_0 v_1 - a_1 v_2 - \ldots - a_{n-1} v_n &= \lambda v_n. \end{aligned} \tag{6.119}$$

Here, the first $n-1$ equations in (6.119) show that $v_k = \lambda^{k-1} v_1$ for $k = 1, 2, \ldots, n$, such that the eigenvalue v is of the form

$$v = v_1 \begin{bmatrix} 1 & \lambda & \cdots & \lambda^{n-1} \end{bmatrix}^{\mathrm{T}}. \tag{6.120}$$

The substitution of this result in the final equation of (6.119) gives, after rearranging terms,

$$0 = \big(\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0\big) v_1 = \Delta_A(\lambda) v_1. \tag{6.121}$$

As $v_1 \neq 0$ (otherwise, the vector v would be zero), we have thus recovered the known result that each eigenvalue of A is a root of the characteristic equation.

However, we also have from (6.117) that $Cv = 0$, which can be evaluated as

$$\begin{aligned} 0 = Cv &= \big(c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \ldots c_1\lambda + c_0\big) v_1, \\ &= C \operatorname{adj}(\lambda I - A) B v_1. \end{aligned} \tag{6.122}$$

Here, the final equality follows from (6.103) in the proof of Theorem 6.9. Thus, we have that λ is a root of both Δ_A and $C \operatorname{adj}(sI - A)B$, meaning that these polynomials are not coprime and finalizing the proof by contraposition.

if. Using again contraposition, let the polynomials (6.116) be not coprime. Then, they have a common root, i.e., there exists $\lambda \in \mathbb{C}$ such that $\Delta_A(\lambda) = 0$ and $C \operatorname{adj}(\lambda I - A)B = 0$. By the first equation, λ is an eigenvalue of A. By following the same steps as in the proof for the *only if* above, the corresponding eigenvalue v is necessarily of the form (6.120) (note that this eigenvalue is unique up to scalar multiplication). Following again a similar reasoning as above, it can be shown that (6.117) holds for this eigenpair (λ, v), such that λ is an unobservable eigenvalue of A. □

Example 6.7. Consider again the transfer function (6.105) from Example 6.6, for which

$$\begin{aligned} p(s) &= s^2 + 2s + 1 &&= (s+1)^2, \\ q(s) &= s^3 + 4s^2 + s - 6 &&= (s-1)(s+2)(s+3). \end{aligned} \tag{6.123}$$

It is clear that p and q have no common factors (i.e., are coprime), such that the matrix pair (A, C) in (6.106) is observable. Recall that the matrix pair (A, B) is controllable by construction. ◇

Example 6.8. Consider the system $\mathbf{\Sigma}_{\text{SISO}}$ with

$$A = \begin{bmatrix} 0 & 1 \\ 1 & \varepsilon \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad C = \begin{bmatrix} -1 & 1 \end{bmatrix}, \tag{6.124}$$

which is easily seen to be controllable as (A, B) is in controllability canonical form. To verify observability, compute

$$\begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & -1+\varepsilon \end{bmatrix}, \tag{6.125}$$

such that the matrix pair (A, C) is observable if and only if $\varepsilon \neq 0$. Moreover, a direct computation gives

$$\Delta_A(s) = s^2 - \varepsilon s - 1, \qquad C \operatorname{adj}(sI - A)B = s - 1, \tag{6.126}$$

such that the transfer function reads

$$T(s) = \frac{s-1}{s^2 - \varepsilon s - 1}. \tag{6.127}$$

Note that in the case $\varepsilon = 0$ (i.e., when (A, C) is not observable), we obtain

$$T(s) = \frac{s-1}{s^2 - 1} = \frac{s-1}{(s-1)(s+1)} = \frac{1}{s+1}, \tag{6.128}$$

such that the transfer function can be simplified. The cancellation of the factor $s - 1$ is referred to as a *pole-zero cancellation.* ◇

The above results rely on controllability of the single-input single-output system (6.97). Given the duality between controllability and observability, analogous results can be obtained when focusing on observability properties.

The first of these results is the following, which is the counterpart of Theorem 6.9 and is stated without proof.

Theorem 6.11. *Consider the SISO system* $\mathbf{\Sigma}_{\text{SISO}}$ *in (6.97) with matrices*

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & -a_0 \\ 1 & 0 & & & 0 & -a_1 \\ 0 & 1 & \ddots & & 0 & -a_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & & \ddots & 1 & 0 & -a_{n-2} \\ 0 & 0 & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} b_0 \\ b_1 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{bmatrix}, \quad C^{\mathrm{T}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{6.129}$$

Then, its transfer function reads

$$T(s) = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \ldots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0}. \tag{6.130}$$

Here, we note that the pair (A, C) given by (6.129) is observable and that the matrices (6.129) define the so-called observability canonical form (see Theorem 4.14).

Then, the counterpart of Theorem 6.10 is given as follows.

Theorem 6.12. *Consider the SISO system* $\mathbf{\Sigma}_{\text{SISO}}$ *in (6.97) and let the matrix pair* (A, C) *be observable. Then, the polynomials*

$$p(s) = C \operatorname{adj}(sI - A)B, \qquad q(s) = \Delta_A(s) \tag{6.131}$$

are coprime if and only if the matrix pair (A, B) *is controllable.*

6.4 Input-output stability

In Chapter 3, stability of autonomous systems is defined by ignoring the role of inputs and outputs. In this section, we consider a notion of stability for systems with inputs and outputs.

However, before defining such stability notions, we take a closer look at the transfer function matrix for a linear system $\mathbf{\Sigma}$ as in (6.1). Specifically, we recall that, when this system is single-input single-output (SISO), its transfer function is a rational function

$$T(s) = \frac{p(s)}{q(s)} = \frac{p'(s)}{q'(s)}, \tag{6.132}$$

where p and q are polynomials. As p and q are not necessarily coprime, common factors might be canceled and we denote p' and q' as the numerator and denominator polynomials of T after cancellation of these common factors (i.e., p' and q' are coprime).

For systems that are not SISO, the transfer function matrix (see Definition 6.4) takes the general form

$$T(s) = \frac{1}{q(s)} P(s), \tag{6.133}$$

with q a polymomial and P a matrix of polynomials. Specifically, we have seen that $q(s) = \Delta_A(s)$ and $P(s) = C \operatorname{adj}(sI - A)B$.

Now, we are in a position to define *poles* of a transfer function (matrix).

Definition 6.5. *A complex number $\lambda \in \mathbb{C}$ is called a pole of the scalar transfer function (6.132) if it is a root of q'. It is a pole of the transfer function matrix (6.133) if it is a pole of at least one of its elements.*

Thus, poles of a transfer function give the location of its singularities. As $q(s) = \Delta_A(s)$ in (6.132) and (6.133), it is clear that any pole is an eigenvalue of A. However, the converse is not necessarily true as cancellations between the numerator and denominator polynomial might occur.

The following theorem, which is stated without proof, gives a full characterization of the relation between poles of a transfer function and the eigenvalues of A.

Theorem 6.13. *Consider a linear system $\mathbf{\Sigma}$ as in (6.1) and its transfer function matrix T. Then, the following hold:*

1. *if $\lambda \in \mathbb{C}$ is a pole of T, then it is an eigenvalue of A, i.e., $\lambda \in \sigma(A)$;*
2. *if $\lambda \in \sigma(A)$, the matrix pair (A, B) is controllable, and the matrix pair (A, C) is observable, then λ is a pole of T.*

Remark 6.13. Even though no proof is given of statement *2* above, some intuition can be given. Namely, from Theorem 6.5, it follows that the impulse response matrix of a linear system is only dependent on its controllable and observable part. As the impulse response matrix and transfer function matrix give an equivalent representation of a linear system (through the Laplace transform, see Theorem 6.7 for one side of this implication), the transfer function only depends on the controllable and observable part of $\mathbf{\Sigma}$. ◁

Now, we define a notion of stability on the input-output behavior of a linear system.

Definition 6.6. *The system $\mathbf{\Sigma}$ in (6.1) is called externally stable (or bounded-input bounded-output stable) if there exists a scalar $\gamma > 0$ such that, for any bounded input function $u : \mathbb{R}_+ \to \mathbb{R}^m$,*

$$\sup_{t\in\mathbb{R}_+} |y(t;0,u)| \leq \gamma \sup_{t\in\mathbb{R}_+} |u(t)|. \tag{6.134}$$

Thus, a system is externally stable if, for any bounded input function u, the resulting output function y is bounded as well. Here, it is stressed that the initial condition is chosen to be zero, which is the standard choice in considering input-output behavior (see also Section 6.1). Moreover, the parameter γ in (6.134) can be regarded as an upper bound on the amplification from the input signal to the output. The smallest γ for which (6.134) holds is therefore sometimes called the *gain* of the system $\mathbf{\Sigma}$.

We also note that, due to linearity, the definition of external stability is equivalent to asking for the implication

$$|u(t)| \leq 1 \quad \forall t \in \mathbb{R}_+ \quad \Longrightarrow \quad |y(t;0,u)| \leq \gamma \quad \forall t \in \mathbb{R}_+ \tag{6.135}$$

to hold.

Definition 6.6 is given for arbitrary linear systems $\mathbf{\Sigma}$ of the form (6.1), but the following result shows that the direct feedthrough term (given by $Du(t)$) does not influence external stability.

Lemma 6.14. *A system $\mathbf{\Sigma}(A, B, C, D)$ as in (6.1) is externally stable if and only if $\mathbf{\Sigma}(A, B, C, 0)$ is externally stable.*

Proof. To prove the result, let $y(\,\cdot\,;0,u)$ and $\bar{y}(\,\cdot\,;0,u)$ denote the output solutions for $\mathbf{\Sigma}(A, B, C, D)$ and $\mathbf{\Sigma}(A, B, C, 0)$, respectively, for zero initial condition and a common input function $u : \mathbb{R}_+ \to \mathbb{R}^m$. By the general form of this solution (see (6.2)), we have, for all $t \geq 0$,

$$y(t;0,u) = \bar{y}(t;0,u) + Du(t). \tag{6.136}$$

Let $\mathbf{\Sigma}(A, B, C, 0)$ be externally stable with gain $\bar{\gamma} > 0$. Then, for any $t \geq 0$,

$$|y(t;0,u)| \leq |\bar{y}(t;0,u)| + |Du(t)| \leq \bar{\gamma} \sup_{t\in\mathbb{R}_+} |u(t)| + \|D\| \sup_{t\in\mathbb{R}_+} |u(t)|, \tag{6.137}$$

where $\|D\| = \sup\{|Du| \mid |u| = 1\}$ is the standard matrix norm. Thus, we have that $\mathbf{\Sigma}(A, B, C, D)$ is externally stable and its gain γ can be chosen to satisfy

$$\gamma \leq \bar{\gamma} + \|D\|. \tag{6.138}$$

The proof of the converse follows similarly. □

Now, the following result on external stability can be stated.

Theorem 6.15. *Consider the system $\mathbf{\Sigma}$ in (6.1) and its transfer function matrix T. Then, the following statements are equivalent:*

1. *$\mathbf{\Sigma}$ is externally stable;*

2. $\displaystyle\int_0^\infty \left\|Ce^{At}B\right\| \mathrm{d}t < \infty$;

3. $\lim_{t\to\infty} Ce^{At}B = 0$;

4. *all poles of* T *are in* $\mathbb{C}_-$.

Proof. In all statements below, we assume that $D = 0$, which does not pose any restrictions due to Lemma 6.14.

2 $\Rightarrow$ *1.* By Theorem 6.3 (for $D = 0$), we can write

$$\left|y(t;0,u)\right| = \left|\int_0^t Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau\right| \leq \int_0^t \left|Ce^{A(t-\tau)}Bu(\tau)\right|\mathrm{d}\tau \tag{6.139}$$

for any $t \geq 0$. Using the definition of the matrix norm (recall that, for a matrix M, $\|M\| = \sup\{|Mx| \mid |x| = 1\}$, see (2.9)), this can be further bounded as

$$\begin{aligned}\left|y(t;0,u)\right| &\leq \int_0^t \left\|Ce^{A(t-\tau)}B\right\||u(\tau)|\,\mathrm{d}\tau,\\ &\leq \left(\int_0^t \left\|Ce^{A(t-\tau)}B\right\|\mathrm{d}\tau\right)\sup_{t\in\mathbb{R}_+}|u(t)|,\\ &\leq \left(\int_0^\infty \left\|Ce^{At}B\right\|\mathrm{d}t\right)\sup_{t\in\mathbb{R}_+}|u(t)|,\end{aligned} \tag{6.140}$$

for all $t \geq 0$. Here, we have used that $|u(t)| \leq \sup_{t\in\mathbb{R}_+}|u(t)|$ for all $t \geq 0$ and that the supremum exists as u is assumed to be bounded. It is clear that (6.140) implies (6.134), such that $\boldsymbol{\Sigma}$ is externally stable. In fact, we have shown that the parameter γ in (6.134) can be bounded as

$$\gamma \leq \int_0^\infty \left\|Ce^{At}B\right\|\mathrm{d}t. \tag{6.141}$$

1 $\Rightarrow$ *2.* To show this implication, let (j,i) be a pair of indices that characterize the output j and input i, respectively. Now, fix any $T > 0$ and define the input function

$$u(t) = e_i\tilde{u}(t) \tag{6.142}$$

with e_i the i-th column of the identity matrix and $\hat{u} : \mathbb{R}_+ \to \mathbb{R}_+$ the scalar function

$$\tilde{u}(t) = \begin{cases}\operatorname{sign}\left(Ce^{A(T-t)}B\right)_{ji}, & 0 \leq t \leq T,\\ 0, & t > T,\end{cases} \tag{6.143}$$

where $(Ce^{A(T-t)}B)_{ji}$ denotes element (j,i) of the matrix $Ce^{A(T-t)}B$. For this input, we have

$$\begin{aligned}y_j(T;0,u) &= \int_0^T \left(Ce^{A(T-\tau)}B\right)_{ji}\tilde{u}(\tau)\,\mathrm{d}\tau\\ &= \int_0^T \left|\left(Ce^{A(T-\tau)}B\right)_{ji}\right|\mathrm{d}\tau,\end{aligned} \tag{6.144}$$

where y_j denotes component j of the output y. However, due to the definition of external stability in Definition 6.6 and after noting that $\sup_{t\in\mathbb{R}_+} |u(t)| = 1$, it holds that

$$\left|y_j(T, 0, u)\right| \leq \gamma, \tag{6.145}$$

such that

$$\int_0^T \left|\left(Ce^{At}B\right)_{ji}\right| \mathrm{d}t \leq \gamma, \tag{6.146}$$

for any $T > 0$ and any pair of indices (j, i). This implies statement *2*.

2 $\Leftrightarrow$ *3*. This follows immediately.

2 $\Rightarrow$ *4*. By statement *2*, it follows that the integral

$$\int_0^\infty Ce^{At}Be^{-st} \mathrm{d}t \tag{6.147}$$

converges for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) \geq 0$. After noting that this integral is nothing more than the Laplace transform, Theorem 6.7 shows that the transfer function matrix T has no singularities in this domain. Thus, all poles of T are in $\mathbb{C}_- = \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}$.

4 $\Rightarrow$ *2*. This can be shown using the inverse Laplace transform, but this is not within the scope of these notes. □

Theorem 6.15 provides a relation between external stability, the impulse response matrix (through $Ce^{At}B$, recall Definition 6.2), and the transfer function matrix. This is perhaps not surprising as the impulse response matrix and transfer function matrix provide a full characterization of the input-output behavior of a linear system.

In Chapter 3, we have defined (asymptotic) stability of autonomous systems as a property of *state* trajectories without considering inputs and outputs. This notion of stability is sometimes referred to as internal stability (see also Remark 3.1). We formalize this in the following definition.

Definition 6.7. *The system $\boldsymbol{\Sigma}$ in (6.1) is called internally stable if the system*

$$\dot{x}(t) = Ax(t) \tag{6.148}$$

is asymptotically stable (see Definition 3.1).

Now, the notions of internal and external stability can be related as follows.

Theorem 6.16. *Consider the system $\boldsymbol{\Sigma}$ in (6.1). Then, the following statements hold:*

1. *if $\boldsymbol{\Sigma}$ is internally stable, then it is externally stable;*
2. *if $\boldsymbol{\Sigma}$ is externally stable, the matrix pair (A, B) is controllable, and the matrix pair (A, C) is observable, then $\boldsymbol{\Sigma}$ is internally stable.*

Proof. 1. By internal stability, $\lim_{t\to\infty} e^{At} = 0$. It is clear that this implies $\lim_{t\to\infty} Ce^{At}B = 0$, which is equivalent to external stability by Theorem 6.15.

2. This follows immediately after recalling from Theorem 6.13 that the poles of the transfer function (determining external stability) are exactly the eigenvalues of A (determining internal stability). Nonetheless, we give an independent proof here.

By external stability, we have that

$$\lim_{t\to\infty} Ce^{At}B = 0. \tag{6.149}$$

Note that each element in $Ce^{At}B$ is the sum of terms of the form $t^k e^{\lambda t}$ with $\lambda \in \sigma(A)$ and k a nonnegative integer. It is clear that time differentiation does not lead to terms with exponents that are different from the exponents in $Ce^{At}B$. This implies

$$\lim_{t\to\infty} CAe^{At}B = \lim_{t\to\infty} Ce^{At}AB = 0, \tag{6.150}$$

where we have used that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} Ce^{At}B = CAe^{At}B = Ce^{At}AB, \tag{6.151}$$

see Lemma 2.4. Repeating this reasoning, it is easy to show that

$$\lim_{t\to\infty} CA^k e^{At} A^\ell B = 0, \tag{6.152}$$

for any nonnegative integers k, ℓ. This then implies

$$\lim_{t\to\infty} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} e^{At} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0. \tag{6.153}$$

Now, as the matrix pair (A, B) is controllable, we have that there exists a right inverse for $[B \;\; AB \;\; \cdots \;\; A^{n-1}B]$, i.e.,

$$\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \Pi_c = I, \tag{6.154}$$

for some matrix $\Pi_r \in \mathbb{R}^{nm\times n}$. Specifically, after denoting

$$\mathcal{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}, \tag{6.155}$$

it can be seen that the choice

$$\Pi_c = \mathcal{C}^{\mathrm{T}}(\mathcal{C}\mathcal{C}^{\mathrm{T}})^{-1} \tag{6.156}$$

satisfies (6.154). Similarly, by observability of the matrix pair (A, C), there exists a matrix $\Pi_o \in \mathbb{R}^{n\times np}$ satisfying

$$\Pi_o \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = I, \tag{6.157}$$

after which it follows that

$$0 = \lim_{t\to\infty} \Pi_o \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} e^{At} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \Pi_r = \lim_{t\to\infty} e^{At}, \tag{6.158}$$

showing that $\mathbf{\Sigma}$ is internally stable and finalizing the proof. □

Remark 6.14. Statement *2* of Theorem 6.16 shows that controllability and observability are sufficient conditions for external stability to imply internal stability. In fact the result can be strengthened to:

- if $\mathbf{\Sigma}$ is externally stable, the matrix pair (A, B) is *stabilizable*, and the matrix pair (A, C) is *detectable*, then $\mathbf{\Sigma}$ is internally stable.

A proof of this statement can follow a similar reasoning as the proof of Theorem 6.16 after using the canonical form for uncontrollable systems in (4.49) in Theorem 4.11 and noting that $\sigma(A_{22}) \subset \mathbb{C}_-$ due to stabilizability (see Theorem 4.17. $\lhd$

6.5 Exercises

Exercise 6.1. Consider three linear systems $\mathbf{\Sigma}(A_i, B_i, C_i, 0)$ with

$$A_1 = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad C_1 = \begin{bmatrix} 1 & 0 \end{bmatrix},$$
$$A_2 = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 0 & 1 \end{bmatrix},$$
$$A_3 = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad B_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad C_3 = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

a. Show that all three systems have the same impulse response (matrix).

b. Show that $\mathbf{\Sigma}(A_1, B_1, C_1, 0)$ and $\mathbf{\Sigma}(A_2, B_2, C_2, 0)$ are similar.

c. Is $\mathbf{\Sigma}(A_3, B_3, C_3, 0)$ similar to the other systems?

Exercise 6.2. Consider the linear system $\mathbf{\Sigma}(A, B, C, 0)$ with

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}, \qquad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

for $\sigma \in \mathbb{R}$ and $\omega > 0$. Determine the impulse response (matrix) and plot it for various values of σ.

Exercise 6.3. Using Theorem 6.6, find the Laplace transforms of

$$f_1(t) = \sin(t), \qquad f_2(t) = \cos(t), \qquad f_3(t) = t^k e^{\lambda t},$$

where k is a nonnegative integer.

Exercise 6.4. Give a proof of Theorem 6.6. Use integration by parts for statement *2* and a change in the order of integration for statement *3*.

Exercise 6.5. Consider again the system in Exercise 6.2. Compute its transfer function (matrix).

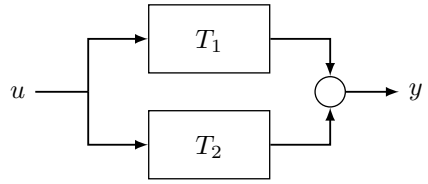


Figure 6.7: Parallel interconnection of the transfer function (matrices) T_1 and T_2. Note that the circle indicates a summer, i.e., the incoming signals are summed to form the outgoing signal.

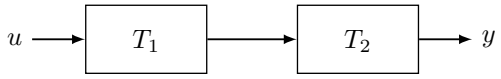


Figure 6.8: Series interconnection of the transfer function (matrices) T_1 and T_2.

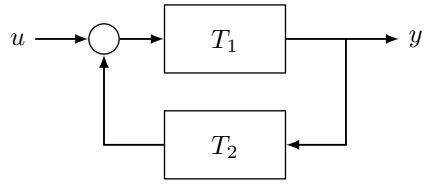


Figure 6.9: Feedback interconnection of the transfer function (matrices) T_1 and T_2.

Exercise 6.6. Find the transfer function matrix of the system

$$\dot{x}(t) = \begin{bmatrix} a & b & 0 \\ -b & a & 0 \\ 0 & 0 & c \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} u(t), \qquad y(t) = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} x(t),$$

where a, b, $c \in \mathbb{R}$ are nonzero.

Exercise 6.7. Consider two systems $\mathbf{\Sigma}(A_i, B_i, C_i, D_i)$, $i = 1, 2$, and let T_i denote their corresponding transfer function matrices. For each of the following interconnection structures, *1)* find the transfer function from input u to output y, and, *2)* find a state-space representation of this transfer function.

a. The parallel interconnection of Figure 6.7.

b. The series interconnection of Figure 6.8.

c. The feedback interconnection of Figure 6.9 for $D_1 = 0$ and $D_2 = 0$, i.e., the transfer function matrices are strictly proper.

d. The feedback interconnection of Figure 6.9 for general feedthrough matrices D_i. Is the feedback interconnection well-defined for all matrices D_i, $i = 1, 2$?

Exercise 6.8. Consider the system $\mathbf{\Sigma}(A, B, C, D)$ with transfer function T.

a. Assume that the matrix pair (A, B) is not controllable, such that, by Theorem 4.11, there exists coordinates in which

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix},$$

with (A_{11}, B_1) controllable. Show by a direct computation that

$$T(s) = C_1(sI - A_{11})^{-1}B_1 + D.$$

b. Formulate the counterpart of the above result in case that the matrix pair (A, C) is not observable.

Exercise 6.9. Consider the single-input single-output system (6.97) and let (A, B) be not controllable. Prove that, for any C, the polynomials

$$p(s) = C\,\mathrm{adj}(sI - A)B, \qquad q(s) = \Delta_A(s)$$

are not coprime. *Hint.* Use Theorem 4.11 and the result of Exercise 6.8.

Exercise 6.10. Consider the polynomials

$$p(s) = s^2 + 1, \qquad q(s) = s^3 + s^2 + s + 1,$$

and note that p and q are not coprime.

a. Construct two systems $\mathbf{\Sigma}(A_i, B_i, C_i, 0)$, $i = 1, 2$, both with state-space dimension 3, such that

- $\mathbf{\Sigma}(A_1, B_1, C_1, 0)$ is controllable but not observable;
- $\mathbf{\Sigma}(A_2, B_2, C_2, 0)$ is observable but not controllable;
- both systems have the transfer function $T(s) = \frac{p(s)}{q(s)}$.

b. Does there exists a system $\mathbf{\Sigma}(A_3, B_3, C_3, 0)$ with state-space dimension 3 that is both controllable and observable and for which $T(s) = \frac{p(s)}{q(s)}$?

Exercise 6.11. Consider two systems $\mathbf{\Sigma}(A_i, B_i, C_i, 0)$, $i = 1, 2$, and let T_i denote their corresponding transfer function matrices.

a. Show that the linear system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}, \\ y(t) &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \end{aligned}$$

has the transfer function matrix $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$.

b. Use the above observation to determine a state-space representation of the transfer function

$$T(s) = \begin{bmatrix} \dfrac{1}{(s+2)(s+3)(s+4)} & \dfrac{1}{(s+2)(s+3)(s+5)} \end{bmatrix}.$$

What is the corresponding state-space dimension?

c. Note that, after collecting common factors, the above transfer function can be written as

$$T(s) = \frac{1}{(s+2)(s+3)} \begin{bmatrix} \dfrac{1}{(s+4)} & \dfrac{1}{(s+5)} \end{bmatrix}.$$

Use this to find an alternative state-space representation of the transfer function of *lower* state-space dimension.

Hint. Use the results of Exercise 6.7.

Exercise 6.12. Consider the linear system according to the time-domain input-output description (see Remark 6.12)

$$\ddot{y}(t) + \dot{y}(t) + y(t) = \dot{u}(t) + p_0 u(t), \tag{6.159}$$

where $p_0 \neq 0$.

a. Find a state-space representation of (6.159). Note that such *realization* is not unique.

b. Compute the transfer function of (6.159) using the result from *a.*

c. Compute the transfer function of (6.159) using the Laplace transform properties of Theorem 6.6, without using the state-space representation.

Exercise 6.13. As a first step in the direction of Remark 6.14, prove the following statement:

- if $\mathbf{\Sigma}$ is externally stable, the matrix pair (A, B) is *stabilizable*, and the matrix pair (A, C) is *observable*, then $\mathbf{\Sigma}$ is internally stable.

Hint. Use Theorem 4.11 and the ideas in the proof of Theorem 6.16.

Appendix A

Ordinary differential equations

In this appendix, we review some basics on (systems of) ordinary differential equations. This appendix is mostly based on [10], see [4, 2] for other sources on differential equations.

A.1 Scalar differential equations

In this section, we consider scalar ordinary differential equations of the form

$$\dot{x}(t) = f\big(t, x(t)\big), \tag{A.1}$$

where t is the so-called independent variable and x is an unknown function (of t). We also sometimes refer to x as the dependent variable. The dot in $\dot{x}$ denotes the derivative of x with respect to t, i.e.,

$$\dot{x}(t) = \frac{\mathrm{d}x}{\mathrm{d}t}(t), \tag{A.2}$$

and we call (A.1) a scalar ordinary differential equation if $f : D \to \mathbb{R}$ for some domain $D \subset \mathbb{R} \times \mathbb{R}$. Finally, we will often omit the argument t in (A.1) and write

$$\dot{x} = f(t, x) \tag{A.3}$$

instead.

Now, solutions to (A.1) are defined as follows.

Definition A.1. *Let $J \subset \mathbb{R}$ be an interval. A function $x : J \to \mathbb{R}$ is called a solution to the (scalar) differential equation (A.1) (in J) if x is differentiable, $(t, x(t)) \in D$ for all $t \in J$, and (A.1) holds for all $t \in J$.*

The scalar differential equation (A.1) has a simple geometric interpretation. Namely, if x is a solution of (A.1) that passes through (t_0, x_0) (in this case, $x_0 = x(t_0)$), then $f(t_0, x_0)$ gives the slope of the curve x at this point. This is illustrated in Figure A.1.

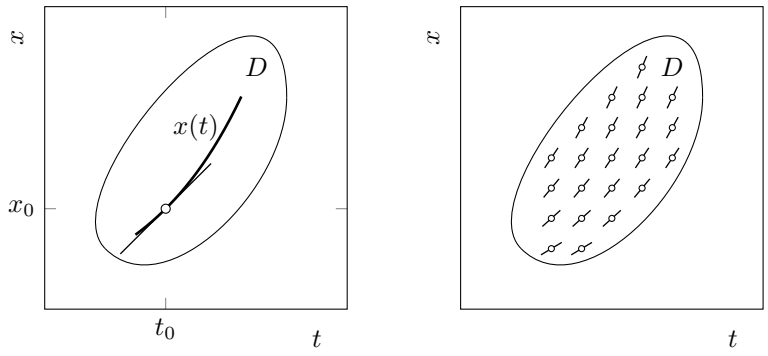


Figure A.1: A solution x with its slope $f(t_0, x_0)$ (left) and the corresponding direction field (right).

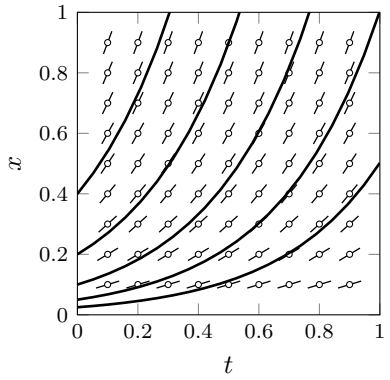


Figure A.2: Direction field and solutions for the differential equation (A.4) with $a = 3$. The solutions are drawn for $c = 0.025, 0.05, 0.1, 0.2, 0.4$.

Following this interpretation, the slope $f(t, x)$ can be indicated for each $(t, x) \in D$. This leads to a so-called *direction field* as given in the right panel of Figure A.1. A solution to the differential equation (A.1) can then be regarded as a curve that fits the direction field.

Example A.1. Consider the linear differential equation

$$\dot{x}(t) = ax(t), \tag{A.4}$$

with $x(t) \in \mathbb{R}$ and for some $a \in \mathbb{R}$ (we can thus take the domain D to be $\mathbb{R} \times \mathbb{R}$). It is easily verified that

$$x(t) = ce^{at} \tag{A.5}$$

is a solution of (A.4) for every c. This is illustrated in Figure A.2. ◇

Whereas the above example illustrates that differential equations generally have an entire class of solutions, we are often interested in a solution that passes

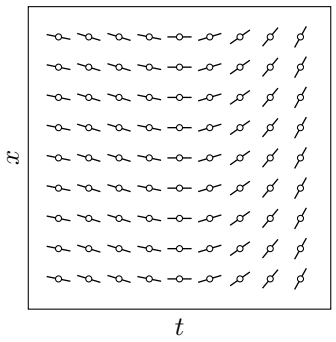


Figure A.3: Illustration of the direction field for scalar differential equations of the form $\dot{x}(t) = f(t)$ as in (A.9).

through a certain point (t_0, x_0). This leads to the so-called *initial value problem*, which can be stated as follows.

Problem A.1 (Initial value problem)**.** *Given a function $f : D \to \mathbb{R}$ for some $D \subset \mathbb{R} \times \mathbb{R}$ and a point (t_0, x_0), find a solution $x : J \to \mathbb{R}$ (as in Definition A.1 with J such that $t_0 \in J$) such that*

$$\dot{x}(t) = f\big(t, x(t)\big), \qquad x(t_0) = x_0. \tag{A.6}$$

Here, the equation $x(t_0) = x_0$ is referred to as the initial condition.

Example A.1 can be extended to include the initial condition.

Example A.2. For the differential equation (A.4) and initial condition (t_0, x_0), the initial value problem is stated as finding a solution $x : \mathbb{R} \to \mathbb{R}$ such that

$$\dot{x}(t) = ax(t), \qquad x(t_0) = x_0. \tag{A.7}$$

It is ready checked that

$$x(t; t_0, x_0) = x_0 e^{a(t-t_0)} \tag{A.8}$$

is a solution to this initial value problem. Here, we have introduced the notation $x(\,\cdot\,; t_0, x_0)$ to denote a solution to the initial value problem. Note that the initial condition thus determines the parameter c in (A.5) as $c = x_0 e^{-at_0}$. ◇

We will now consider the direction fields for three classes of scalar differential equations for which solutions x can be computed explicitly.

Equations of the form $\dot{x} = f(t)$. Consider the differential equation

$$\dot{x}(t) = f(t), \tag{A.9}$$

where f is independent of x and assumed to be continuous on an interval J. In this case, we can take the set D to equal $J \times \mathbb{R}$. It is clear that the corresponding direction field is independent of x, see Figure A.3 for an illustration. This suggest that, if $x(\cdot)$ is a solution to the differential equation, then $x(\cdot) + c$ is a solution as well for any $c \in \mathbb{R}$.

The value of c can be determined in case an accompanying initial value problem is considered, i.e., when an initial condition $x(t_0) = x_0$ is given. Namely, by the fundamental theorem of calculus, we have that

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau, \tag{A.10}$$

solves the initial value problem (A.9) with initial condition (t_0, x_0). In fact, (A.10) is the *unique* solution to the initial value problem.

Equations of the form $\dot{x} = g(x)$ (autonomous equations). Whereas the previous paragraph considers functions f that are independent of x, we will now consider functions f that are independent of t instead. Using the notation $f(t, x) = g(x)$, this leads to the differential equation

$$\dot{x}(t) = g\big(x(t)\big), \tag{A.11}$$

where it is assumed that g is continuous in an interval X (then, $D = \mathbb{R} \times X$). In this case, the direction field is independent of t, see Figure A.2 for an example. From this, we see that, if $x(t)$ is a solution to the differential equation (A.11), then $x(t + \tau)$ is a solution to (A.11) as well for any $\tau \in \mathbb{R}$.

Specifically, any solution to (A.11) can be characterized through the implicit equation

$$H(x) = t + \tau, \tag{A.12}$$

where H is such that

$$\frac{\mathrm{d}H}{\mathrm{d}x}(x) = \frac{1}{g(x)}. \tag{A.13}$$

Namely, for $g(x) \neq 0$, we have by the chain rule that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(H(x)\big) = \frac{\mathrm{d}H}{\mathrm{d}x}(x)\dot{x} = \frac{1}{g(x)}\dot{x} = 1, \tag{A.14}$$

where the latter equality follows from (A.11). It is clear (after integration) that (A.14) is equivalent to (A.12), which shows the desired result.

Differential equations of the form (A.11) are referred to as *autonomous*. As the independent variable t often represents time and solutions can be shifted in the independent variable, such differential equations are also called *time-invariant*.

Example A.3. Returning again to the linear differential equation (A.4) in Example A.1, we can use the above approach to derive the solution (A.5). To this end, recall that the anti-derivative of $\frac{1}{g(x)} = \frac{1}{ax}$ is given by

$$H(x) = \tfrac{1}{a} \ln |x|. \tag{A.15}$$

Thus, the solution to (A.4) is of the form

$$\ln |x| = a(t + \tau), \tag{A.16}$$

which is equivalent to

$$|x(t)| = e^{a\tau}e^{at} \tag{A.17}$$

and, subsequently,

$$x(t) = ce^{at}. \tag{A.18}$$

Here, $c \in \mathbb{R}$ is an arbitrary constant (such that $c = \pm e^{a\tau}$ or $c = 0$). ◇

The strategy that was to characterize solutions for differential equations of the form (A.11) can be applied to a more general class of differential equations known as *separable* differential equations.

Equations of the form $\dot{x} = f(t)g(x)$ (separable equations). Consider a differential equation of the form

$$\dot{x}(t) = f(t)g\big(x(t)\big), \tag{A.19}$$

where f and g are assumed to be continuous on intervals J and X, respectively (such that $D = J \times X$).

Following the same approach as in the previous example, let H be a function such that (A.13) holds and let F be such that

$$\frac{\mathrm{d}F}{\mathrm{d}t}(t) = f(t). \tag{A.20}$$

Then, the class of solutions of (A.19) can be written implicitly as

$$H(x) = F(t) + c, \tag{A.21}$$

with $c \in \mathbb{R}$, which leads to $x(t) = H^{-1}(F(t)+c)$ in case the inverse function H^{-1} exists. Again, the solution can be verified by differentiation with respect to t and the chain rule, leading to

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(H(x)\big) = \frac{1}{g(x)}\dot{x} = f(t), \tag{A.22}$$

which equals the original equation (A.19).

A more practical view to separable equations is obtained by writing the differential equation (A.19) as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(t)g(x), \tag{A.23}$$

which can be rewritten to obtain

$$\frac{1}{g(x)}\,\mathrm{d}x = f(t)\,\mathrm{d}t. \tag{A.24}$$

Then, integrating both sides gives

$$\int \frac{1}{g(x)}\,\mathrm{d}x = \int f(t)\,\mathrm{d}t, \tag{A.25}$$

where one can recognize the functions H and F on the left- and right-hand side, respectively. We will apply this perspective to find solutions for the following example.

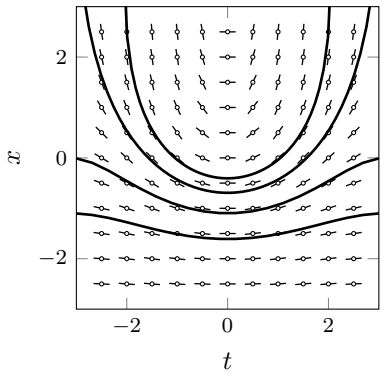


Figure A.4: The direction field and solutions (for $c = 0.5, 1, 2, 4$) for the differential equation (A.26).

Example A.4. Consider the differential equation

$$\dot{x}(t) = \sin(t)e^{x(t)}, \tag{A.26}$$

where the right-hand size is defined on $D = \mathbb{R} \times \mathbb{R}$. By separation of variables, we obtain

$$\frac{1}{e^x}\,\mathrm{d}x = \sin(t)\,\mathrm{d}t \tag{A.27}$$

and, subsequently,

$$\int e^{-x}\,\mathrm{d}x = -e^{-x} = \int \sin(t)\,\mathrm{d}t = -\cos(t) - c. \tag{A.28}$$

Note that solutions are only obtained when $\cos(t) + c > 0$, in which case

$$x(t) = -\ln\big(\cos(t) + c\big). \tag{A.29}$$

The direction field of (A.26) and the corresponding solutions (A.29) are illustrated in Figure A.4. ◇

We finish this section with some examples to show aspects of differential equations. The first example shows that solutions are not necessarily defined for all $t \in \mathbb{R}$.

Example A.5. Consider the differential equation

$$\dot{x}(t) = x^2(t), \tag{A.30}$$

whose right-hand side is defined on $D = \mathbb{R} \times \mathbb{R}$. By separation of variables, it can be shown that solutions have the form

$$x(t) = -\frac{1}{t - c}, \tag{A.31}$$

which have a vertical asymptote for $t = c$. Specifically, for the initial condition $(t_0, x_0) = (0, 1)$, the solution to the corresponding initial value problem reads

$$x(t) = -\frac{1}{t-1}, \tag{A.32}$$

which is not defined for $t = 1$. However, as the branch for $t > 1$ does not satisfy the initial condition $x(0) = 1$, the solution is only valid for $t < 1$. Hence, we say that the solution is defined only for $t < 1$. ◇

Next, solutions to initial value problems are not necessarily unique, as shown in the following example.

Example A.6. Consider the initial value problem

$$\dot{x}(t) = \sqrt{|x(t)|}, \qquad x(-2) = 0 \tag{A.33}$$

defined on $D = \mathbb{R} \times \mathbb{R}$. We immediately observe that

$$x(t) = 0, \quad t \in \mathbb{R} \tag{A.34}$$

is a solution to (A.33). Moreover, by symmetry of the function $\sqrt{|x|}$, we have that $-x(-t)$ is a solution if $x(t)$ is a solution. We can therefore limit attention to solutions for which $x(t) \geq 0$, after which separation of variables gives

$$\int \frac{1}{\sqrt{x}}\,\mathrm{d}x = 2\sqrt{x} = \int \mathrm{d}t = t + c, \tag{A.35}$$

such that solutions exist for $t \geq -c$. Combining this result for the trivial solution (A.34), we have that all functions x that are piecewise defined as

$$x(t) = \begin{cases} 0 & , \quad t < -c, \\ \frac{1}{4}(t+c)^2 & , \quad t \geq -c, \end{cases} \tag{A.36}$$

are solutions to the differential equation $\dot{x} = \sqrt{x}$. In fact, for each $c \leq 2$, the initial condition $x(-2) = 0$ is satisfied and x presents a solution to the initial value problem. Here, note that the solution (A.36) is differentiable at $t = -c$ and thus satisfies the conditions in Definition A.1. An illustration is given in Figure A.5. ◇

From the above examples, we thus see that solutions to initial value problems are not guaranteed to exist and are not necessarily unique. A brief discussion on uniqueness of solutions for general ordinary differential equations will be postponed to Section A.3, but *linear* differential equations are studied in the next section.

A.2 Linear differential equations

In this section, we consider a special class of scalar ordinary differential equations as in (A.1) known as *linear* ordinary differential equations.

A linear differential equation is an equation of the form

$$\dot{x}(t) = a(t)x(t) + b(t), \tag{A.37}$$

where $a : J \to \mathbb{R}$ and $b : J \to \mathbb{R}$ are functions defined on some interval $J \subset \mathbb{R}$. These functions are assumed to be continuous. The equation (A.37) is called *homogeneous* when $b(t) = 0$ for all t and *nonhomogeneous* (or *inhomogeneous*) otherwise.

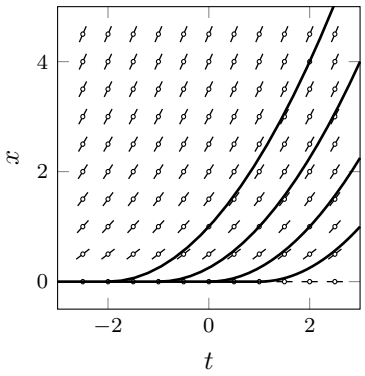


Figure A.5: The direction field and solutions (for $c = -1, 0, 1, 2$) for the differential equation (A.33).

Remark A.1. Roughly speaking, the differential equation (A.37) is called linear because the terms related to the dependent variable x (namely, both x and $\dot{x}$) only appear linearly. A more rigorous explanation can be given after defining the differential operator

$$L(x) = \dot{x} - a(t)x. \tag{A.38}$$

Note that the operator takes *functions* $x : J \to \mathbb{R}$ as an input and associates to each function another function $L(x) = \dot{x} - a(t)x$. For this to be well-defined, we take as the domain of L the functions $x : J \to \mathbb{R}$ that are differentiable.

Now, we can easily show that the operator L is linear. Namely, let x, $x' : J \to \mathbb{R}$ be two differentiable functions and α, $\alpha' \in \mathbb{R}$. Then,

$$\begin{aligned} L\big(\alpha x + \alpha' x'\big) &= \tfrac{\mathrm{d}}{\mathrm{d}t}\big\{\alpha x + \alpha' x'\big\} - a(t)\big(\alpha x + \alpha' x'\big), \\ &= \alpha\dot{x} + \alpha'\dot{x}' - a(t)\alpha x - a(t)\alpha' x', \\ &= \alpha L(x) + \alpha' L(x'). \end{aligned} \tag{A.39}$$

Note that, using the differential operator (A.38), the linear differential equation (A.37) can be written as

$$L(x) = b(t), \tag{A.40}$$

and $L(x) = 0$ in the homogeneous case. ◁

In the remainder of this section, we will study solutions (as in Definition A.1) of the linear differential equation (A.37).

The homogeneous case is considered first.

Homogeneous equations. After noting that the homogeneous differential equation

$$\dot{x}(t) = a(t)x(t) \tag{A.41}$$

is a separable equation as in (A.19), we obtain with

$$H(x) = \int \frac{1}{x}\,\mathrm{d}x = \ln|x|, \qquad F(t) = \int a(t)\,\mathrm{d}t, \tag{A.42}$$

that any solution satisfies

$$\ln|x| = F(t) + c, \tag{A.43}$$

see (A.21). This leads to the class of solutions solution

$$x(t) = Ce^{F(t)}, \tag{A.44}$$

where $C = \pm e^c$ or $C = 0$.

Remark A.2. An alternative approach to finding the class of solutions to (A.41) is by introduction of the so-called *integrating factor* $e^{-F(t)}$ with F as in (A.42). Namely, as $e^{-F(t)} \neq 0$ for all t, we have after multiplication with the integrating factor that (A.41) is equivalent to

$$e^{-F(t)}\dot{x}(t) - e^{-F(t)}a(t)x(t) = 0. \tag{A.45}$$

However, we have that

$$e^{-F(t)}\dot{x}(t) - e^{-F(t)}a(t)x(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left\{e^{-F(t)}x(t)\right\}, \tag{A.46}$$

as follows from the chain rule and the definition of F. Thus, the direct integration of (A.46), hereby using (A.45), leads to

$$e^{-F(t)}x(t) = C, \tag{A.47}$$

for some constant $C \in \mathbb{R}$, which is equivalent to the class of solutions (A.44) found before. ◁

The above result, characterizing the class of solutions for (A.41), can be used to find the solution to the intial value problem. This is formalized in the following lemma.

Lemma A.1. *Consider the intial value problem*

$$\dot{x}(t) = a(t)x(t), \qquad x(t_0) = x_0, \tag{A.48}$$

where $a : J \to \mathbb{R}$ is continuous and $t_0 \in J$. Then, the unique solution is

$$x(t; t_0, x_0) = x_0 e^{F(t)}, \qquad F(t) = \int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau, \tag{A.49}$$

for $t \in J$.

Proof. It is readily verified that (A.49) is a solution to (A.48), i.e., it satisfies both the dynamics and the initial condition.

Uniqueness follows from the observation that *any* solution to the differential equation (A.41) is of the form (A.44). However, we give an independent proof here. To this end, introduce the short-hand notation $x(t) = x(t; t_0, x_0)$ and let x' be a second solution to the initial value problem. Then, after introducing $z(t) = x(t) - x'(t)$, it is easy to see that $\dot{z}(t) = a(t)z(t)$ and, moreover,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{e^{-F(t)}z(t)\right\} = e^{-F(t)}\dot{z}(t) - e^{-F(t)}a(t)z(t) = 0. \tag{A.50}$$

Here, the chain rule and $\frac{\mathrm{d}}{\mathrm{d}t}F(t) = a(t)$ are used. Now, integration gives

$$\int_{t_0}^{t} \frac{\mathrm{d}}{\mathrm{d}\tau}\Big\{e^{-F(\tau)}z(\tau)\Big\}\,\mathrm{d}\tau = e^{-F(t)}z(t) - e^{-F(t_0)}z(t_0) = 0. \tag{A.51}$$

However, since x and x' are both assumed to be solutions to the initial value problem (A.48), we have that $z(t_0) = 0$ and, subsequently, that $e^{-F(t)}z(t) = 0$ for all $t \in J$. This implies that $z(t) = 0$, such that $x(t) = x'(t)$ for all $t \in J$, i.e., (A.49) is the unique solution. □

In the following example, a homogeneous differential equation is solved by separation of variables.

Example A.7. Consider the homogeneous differential equation

$$\dot{x}(t) = -\frac{4}{t}x(t), \quad t > 0. \tag{A.52}$$

We solve the differential equation by separation of variables. This leads to

$$\int \frac{1}{x}\,\mathrm{d}x = -\int \frac{4}{t}\,\mathrm{d}t, \tag{A.53}$$

such that

$$\ln|x| = -4\ln|t| + c = \ln t^{-4} + c, \tag{A.54}$$

for a constant $c \in \mathbb{R}$. Subsequently, we obtain the solution

$$x(t) = \frac{C}{t^4}, \tag{A.55}$$

where $C = \pm e^c$ or $C = 0$. ◇

Finally, we use the integrating factor approach to solve the same example.

Example A.8. Given (A.52), we obtain from (A.42) that

$$F(t) = \int -\frac{4}{t}\,\mathrm{d}t = -4\ln|t| = \ln t^{-4}, \tag{A.56}$$

such that the integrating factor reads $e^{-F(t)} = e^{-\ln t^{-4}} = t^4$. Then, multiplication of (A.52) with the integrating factor gives

$$0 = t^4\dot{x}(t) + 4t^3x(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}\big\{t^4x(t)\big\}, \tag{A.57}$$

such that solutions satisfy

$$t^4x(t) = C \tag{A.58}$$

for some constant $C \in \mathbb{R}$. We note that this solution equals (A.55) (recall that $t > 0$ is considered). ◇

Nonhomogeneous equations. We now turn attention to the nonhomogeneous linear differential equation (A.37), which is repeated as

$$\dot{x}(t) = a(t)x(t) + b(t). \tag{A.59}$$

The class of solutions to (A.59) can be found by exploiting the result for the homogeneous case through a method known as *variation of constants*. To this end, let

$$z(t) = e^{-F(t)}x(t), \tag{A.60}$$

where $e^{-F(t)}$ is the integrating factor from Remark A.2, with F in (A.42). Then, similar to (A.46) and again exploiting $\frac{\mathrm{d}}{\mathrm{d}t}F(t) = a(t)$, we obtain

$$\dot{z}(t) = e^{-F(t)}\dot{x}(t) - e^{-F(t)}a(t)x(t) = e^{-F(t)}b(t). \tag{A.61}$$

As the right-hand side is independent of z, the differential equation (A.61) can be solved by direct integration to obtain

$$z(t) = \int e^{-F(t)}b(t)\,\mathrm{d}t + C, \tag{A.62}$$

after which (A.60) gives $x(t) = e^{F(t)}z(t)$ (note that $e^{-F(t)} \neq 0$ for all t) and, subsequently,

$$x(t) = Ce^{F(t)} + e^{F(t)}\int e^{-F(t)}b(t)\,\mathrm{d}t. \tag{A.63}$$

Thus, any solution to (A.59) can be written as (A.63).

Remark A.3. In the derivation above, the integrating factor from Remark A.2 is exploited. However, the name *variation of constants* comes from a different perspective on the same approach. Namely, the homogeneous case gives the class of solutions $x(t) = Ce^{F(t)}$. Then, replacing the constant C by a function $z(\cdot)$ gives exactly the expression (A.60). ◁

After repeating the above procedure with the proper definite integrals rather than indefinite integrals leads to the solution of the corresponding initial value problem. This is stated in the following lemma, which is given without proof.

Lemma A.2. *Consider the intial value problem*

$$\dot{x}(t) = a(t)x(t) + b(t), \qquad x(t_0) = x_0, \tag{A.64}$$

where a, $b : J \to \mathbb{R}$ are continuous and $t_0 \in J$. Then, the unique solution is

$$x(t; t_0, x_0) = x_0e^{F(t)} + e^{F(t)}\int_{t_0}^{t} e^{-F(\tau)}b(\tau)\,\mathrm{d}\tau, \tag{A.65}$$

for $t \in J$ and with F as in (A.49).

Returning to the general solution (A.63), we observe that it consists of two parts. Specifically, the first term equals the general solution for the homogeneous linear differential equation (A.37), see (A.44). Moreover, the second term does not contain any free parameters.

These observations can be explained by considering two solutions x, $x' : J \to \mathbb{R}$ to the nonhomogeneous differential equation (A.59). Then, we have that

$$\dot{x}(t) - \dot{x}'(t) = a(t)x(t) + b(t) - a(t)x'(t) - b(t) = a(t)\big(x(t) - x'(t)\big), \quad \text{(A.66)}$$

i.e., the difference $x - x'$ satisfied the *homogeneous* equation (A.41). This implies that the general solution to (A.59) can be written in the form

$$x(t) = x_h(t) + x_p(t). \quad \text{(A.67)}$$

Here, x_h is the so-called *homogeneous solution* and is a solution to

$$\dot{x}_h(t) = a(t)x_h(t), \quad \text{(A.68)}$$

whereas x_p is referred to as the *particular solution* and satisfies

$$\dot{x}_p(t) = a(t)x_p(t) + b(t). \quad \text{(A.69)}$$

Stated differently, any solution to the nonhomogeneous equation (A.59) can be written as by fixing *one* solution to (A.69) (the particular solution) and adding a homogeneous solution. Namely, let

$$x_p(t) = C_p e^{F(t)} + e^{F(t)} \int e^{-F(t)} b(t)\, \mathrm{d}t \quad \text{(A.70)}$$

be a particular solution (see (A.63)). Then, adding a homogeneous solution $x_h(t) = C_h e^{F(t)}$ as in (A.67) gives

$$x(t) = (C_h + C_p)e^{F(t)} + e^{F(t)} \int e^{-F(t)} b(t)\, \mathrm{d}t, \quad \text{(A.71)}$$

which is indeed another solution of the form (A.59).

We use this perspective in the following example.

Example A.9. Consider the nonhomogeneous differential equation

$$\dot{x}(t) = -\frac{4}{t}x(t) + \frac{\sin t}{t^4}, \quad t > 0, \quad \text{(A.72)}$$

and note that the homogeneous equation was solved in Example A.7. Thus, we have from (A.55) that

$$x_h(t) = \frac{C}{t^4}. \quad \text{(A.73)}$$

We now use variation of constants to solve the nonhomogeneous equation (A.72). Thus, following Remark A.3, replace C in the homogeneous solution (A.73) by z to obtain $z(t) = t^4 x(t)$

$$\dot{z}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}\big\{t^4 x(t)\big\} = t^4\dot{x}(t) + 4t^3 x(t) = \sin t. \quad \text{(A.74)}$$

Here, the latter equality is obtained by substitution of (A.72). Consequently, $z(t) = -\cos t$ is a solution, which leads to

$$x_p(t) = \frac{z(t)}{t^4} = -\frac{\cos t}{t^4}. \quad \text{(A.75)}$$

Note that we are only looking for *one* particular solution so there is no harm in not considering the integration constant in solving $\dot{z} = \sin t$. Now, combining the homogeneous solution (A.73) to the particular solution (A.75) gives

$$x(t) = \frac{C}{t^4} - \frac{\cos t}{t^4} = \frac{C - \cos t}{t^4} \tag{A.76}$$

as the general solution to the nonhomogeneous equation (A.72). ◇

Similar to the homogeneous case, we use the method of the integrating factor as an alternative to solve the same nonhomogeneous equation.

Example A.10. Recall from Example A.8 that the integrating factor for (A.72) is given as by $e^{-F(t)} = t^4$. Then, multiplication of (A.72) with the integrating factor gives, after rearranging terms,

$$\sin t = t^4\dot{x}(t) + 4t^3x(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}\big\{t^4x(t)\big\}, \tag{A.77}$$

such that any solution satisfies

$$t^4x(t) = -\cos t + C. \tag{A.78}$$

Note that this gives the same solutions as in (A.76). ◇

A.3 Systems of differential equations

In Section A.1 scalar ordinary differential equation of the form (A.1) were considered. In this section, we consider a set of coupled scalar differential equations. Such set is known as a *system* of differential equations and can be written as

$$\begin{aligned} \dot{x}_1(t) &= f_1\big(t, x_1(t), x_2(t), \ldots, x_n(t)\big), \\ &\vdots \\ \dot{x}_n(t) &= f_n\big(t, x_1(t), x_2(t), \ldots, x_n(t)\big). \end{aligned} \tag{A.79}$$

Here, the n functions f_i are taken to be defined on a set $D \subset \mathbb{R} \times \mathbb{R}^n$, i.e., $f_i : D \to \mathbb{R}$. After defining the vector notation

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad f(t,x) = \begin{bmatrix} f_1\big(t, x_1, x_2, \ldots, x_n\big) \\ \vdots \\ f_n\big(t, x_1, x_2, \ldots, x_n\big) \end{bmatrix}, \tag{A.80}$$

we can write (A.79) in a more compact manner as

$$\dot{x}(t) = f\big(t, x(t)\big), \tag{A.81}$$

where $f : D \to \mathbb{R}^n$ and $x(t) \in \mathbb{R}^n$. We note that differentiation and integration of a vector x as in (A.80) are defined component-wise.

As the system of differential equations (A.81) presents a direct extension of the scalar differential equation (A.1), it is no surprise that solutions to (A.81) are defined similarly to the scalar case. Nonetheless, we state the following definition for completeness.

Definition A.2 (Solutions)**.** *Let $J \subset \mathbb{R}$ be an interval. A function $x : J \to \mathbb{R}^n$ is called a solution to the system of differential equations (A.81) (in J) if x is differentiable, $(t, x(t)) \in D$ for all $t \in J$, and (A.81) holds for all $t \in J$.*

At this point, we also give a precise statement of the initial value problem.

Problem A.2 (Initial value problem)**.** *Given a function $f : D \to \mathbb{R}^n$ for some $D \subset \mathbb{R} \times \mathbb{R}^n$ and a point $(t_0, x_0) \in D$, find a solution $x : J \to \mathbb{R}^n$ (as in Definition A.2 with J such that $t_0 \in J$) such that*

$$\dot{x}(t) = f\big(t, x(t)\big), \qquad x(t_0) = x_0. \tag{A.82}$$

Here, the equation $x(t_0) = x_0$ is referred to as the initial condition.

By integration of (A.82) and assuming that f is continuous in D, it follows that the initial value problem is equivalent to finding a solution $x : J \to \mathbb{R}^n$ to the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f\big(\tau, x(\tau)\big)\, \mathrm{d}\tau. \tag{A.83}$$

Note that the initial condition $x(t_0) = x_0$ is automatically satisfied for solutions x satisfying (A.83).

We recall from Example A.6 that solutions to an initial value problem are not necessarily unique. In the remainder of this section, we would like to state a sufficient condition for guaranteeing uniqueness of solutions.

To this end, we first define the Euclidean norm[1] $|\cdot|$ on $\mathbb{R}^n$ as

$$|x| = \sqrt{x^{\mathrm{T}} x} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}, \tag{A.84}$$

which allows us to state the following definition.

Definition A.3 (Lipschitz continuity)**.** *A function $f : D \to \mathbb{R}^n$ with $D \subset \mathbb{R} \times \mathbb{R}^n$ is called Lipschitz (in x) at a point $(t', x') \in D$ if there exist constants $L > 0$ and $r > 0$ such that*

$$\big|f(t, x) - f(t, x')\big| \leq L|x - x'| \tag{A.85}$$

for all (t, x) such that $|x - x'| < r$, $|t - t'| < r$, and $(t, x) \in D$. If f is Lipschitz (in x) for all $(t', x') \in D$, it is said to be locally Lipschitz (in x) on D.

Lipschitz continuity is a crucial condition in guaranteeing uniqueness of solutions of (systems of) ordinary differential equations. This is stated in the following important result. The proof is outside the scope of these lecture notes, but can be found in, e.g., [10].

Theorem A.3. *Let $f : D \to \mathbb{R}^n$ with domain $D \subset \mathbb{R} \times \mathbb{R}^n$ be continuous and locally Lipschitz (in x) on D. If $(t_0, x_0) \in D$, then there exists a solution to the initial value problem (A.82). This solution is unique and can be extended to the left and right up to the boundary of D.*

[1] Recall that a norm on $\mathbb{R}^n$ is a function $|\cdot| : \mathbb{R}^n \to [0, \infty)$ such that, 1) $|x| = 0$ if and only if $x = 0$, 2) $|\lambda x| = |\lambda||x|$ for each $\lambda \in \mathbb{R}$, and 3) the triangle inequality $|x + x'| \leq |x| + |x'|$ holds for all x, $x' \in \mathbb{R}^n$. The Euclidian norm is an example that satisfies these properties.

Remark A.4. The proof of Theorem A.3 makes use of the sequence $\{x^{(k)}\}$ of continuous functions $x^{(k)} : J \to \mathbb{R}^n$ given as

$$x^{(k+1)}(t) = x_0 + \int_{t_0}^{t} f\big(\tau, x^{(k)}(\tau)\big)\,\mathrm{d}\tau, \tag{A.86}$$

and shows (using the Banach fixed-point theorem) that this sequence uniformly converges to the solution of the initial value problem for any initial term $x^{(0)}$. This approach is known as the *method of successive approximations*. ◁

Remark A.5. For scalar functions $f : \mathbb{R} \to \mathbb{R}$, the Lipschitz condition (A.85) can be written as

$$\frac{|f(x) - f(x')|}{|x - x'|} = \left|\frac{f(x) - f(x')}{x - x'}\right| \leq L. \tag{A.87}$$

This implies that, in a plot of $f(x)$ as a function of x, a straight line joining the points $(x, f(x))$ and $(x', f(x'))$ can not have a slope whose absolute value is larger than L. As a result, any function that has infinite slope at some point is not Lipschitz at that point. An example is the function $f(x) = \sqrt{|x|}$ in Example A.6, which has infinite slope at $x = 0$. Indeed, we saw in this example that unique solutions cannot be guaranteed. ◁

Remark A.6. One of the motivations for studying *systems* of differential equations (A.79) is that they can be used to represent *higher-order* differential equations. To illustrate this, denote

$$y^{(k)}(t) = \frac{\mathrm{d}^k y}{\mathrm{d}t^k}(t), \tag{A.88}$$

for $k = 1, 2, \ldots$ (then, $\dot{y}(t) = y^{(1)}(t)$) and consider the n-th order scalar differential equation

$$y^{(n)}(t) = f\big(t, y(t), \dot{y}(t), \ldots, y^{(n-1)}(t)\big), \tag{A.89}$$

where $n \geq 1$. Following the general definition, a function $y : J \to \mathbb{R}$ is called a solution if it is differentiable at least n times and satisfies (A.89) for all $t \in J$. The equation (A.89) can be written as system of first-order differential equations by introducing

$$x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} y & \dot{y} & \cdots & y^{(n-1)} \end{bmatrix}^{\mathrm{T}}, \tag{A.90}$$

such that (A.89) leads to

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \dot{x}_{n-1}(t) \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \\ f\big(t, x_1(t), x_2(t), \ldots, x_{n-1}(t)\big) \end{bmatrix}. \tag{A.91}$$

The higher-order differential equation (A.89) and the system of first-order differential equations (A.91) are equivalent in the sense that a solution to (A.89) (as described above) is a solution to (A.91) (as in Definition A.2) and vice versa.

A similar approach can be used to transform a *system* of higher-order equations to a system of first-order differential equations. ◁

A.4 Exercises

Exercise A.1. Consider the differential equation

$$\dot{x}(t) = \big(a - bx(t)\big)x(t)$$

with a, $b > 0$. This equation is known as the *logistic equation* and describes the evolution of a population, where $x(t) \in \mathbb{R}$ is the population size at time t.

a. An *equilibrium* $\bar{x} \in \mathbb{R}$ is a constant solution, i.e., it satisfies $x(t; t_0, \bar{x}) = \bar{x}$ for all t_0, t. Compute the equilibria for the logistic equation. What is the interpretation of these equilibria?

b. Draw the direction field for the differential equation. Indicate the equilibria in the direction field. For simplicity, take $a = 6$ and $b = 1$.

c. Using separation of variables, compute the general solution of the differential equation. Do not take values for a and b here but leave them as parameters.

d. Show that, for any $x_0 > 0$, the solutions to the initial value problem

$$\dot{x}(t) = \big(a - bx(t)\big)x(t), \qquad x(t_0) = x_0,$$

satisfy $x(t; t_0, x_0) > 0$ for all $t \geq t_0$ and, moreover, $\lim_{t\to\infty} x(t; t_0, x_0) = \frac{a}{b}$.

e. Draw the solution to the initial value problem and the direction field in the same plot for various initial conditions x_0 (you can take $t_0 = 0$). Here, take at least one initial condition satisfying $x_0 > \frac{a}{b}$ and one satisfying $0 < x_0 < \frac{a}{b}$. Again, take $a = 6$ and $b = 1$.

Exercise A.2. Solve the following initial value problems. In each case, give the maximum interval of existence of the solution and verify the answer by differentiation.

a. $\dot{x}(t) = \dfrac{e^t + \sin t}{2x^2(t)}, \quad x(0) = 0$

b. $\dot{x}(t) = \dfrac{1}{(t+1)\cos x(t)}, \quad x(0) = 0$

c. $\dot{x}(t) = \dfrac{tx^2(t)\big(t^2 - x^2(t)\big)}{\big(x(t) + t\big)\big(x^2(t) - tx(t)\big)}, \quad x(0) = 1$

d. $\dot{x}(t) = \dfrac{e^{-x^2(t)}}{x(t)(2t + t^2)}, \quad x(2) = 0$

e. $\dot{x}(t) = \dfrac{x(t)\log x(t)}{\sin t}, \quad x(\frac{1}{2}\pi) = e^e$

f. $\dot{x}(t) = \dfrac{\cos t}{\cos^2 x(t)}, \quad x(\pi) = \frac{1}{4}\pi$

Exercise A.3. Compute the general solution for the following linear differential equations. In each case verify the answer by differentiation.

a. $t\dot{x}(t) + x(t) = e^t, \quad t > 0$

b. $e^t\dot{x}(t) + 2e^t x(t) = 1$

c. $t\dot{x}(t) + 3x(t) = \dfrac{\sin t}{t^2}, \quad t > 0$

d. $\dot{x}(t) + x(t)\tan t = \cos^2(t), \quad -\frac{1}{2}\pi < t < \frac{1}{2}\pi$

e. $t\dot{x}(t) + 2x(t) = 1 - \dfrac{1}{t}, \quad t > 0$

f. $(1+t)\dot{x}(t) + x(t) = \sqrt{t}, \quad t > 0$

g. $\dot{x}(t) + 2x(t) = e^t \sin t$

Exercise A.4. Consider the nonlinear scalar differential equation

$$\dot{x}(t) + a(t)x(t) + b(t)\big(x(t)\big)^\gamma = 0, \tag{A.92}$$

with $\gamma \in \mathbb{R}$ satisfying $\gamma > 1$ and where $a, b : \mathbb{R} \to \mathbb{R}$ are continuous functions. The differential equation (A.92) is called *Bernoulli's equation*.

a. Show that $x(t) = 0$ for all $t \in \mathbb{R}$ is a solution to (A.92).

b. Let x be a solution to (A.92) and assume that there exists $t_0 \in \mathbb{R}$ such that $x(t_0) > 0$. Motivate that $x(t) > 0$ for all t for which the solution is defined (we will call such solution a positive solution).

Hint. Think of the direction field. You can assume that, for a given initial condition $x(t_0)$, the solution to the corresponding initial value problem is unique.

c. Let x be a positive solution. Define

$$z(t) = \big(x(t)\big)^{1-\gamma}$$

and show that z is a positive solution to the *linear* differential equation

$$\dot{z}(t) + (1-\gamma)a(t)z(t) + (1-\gamma)b(t) = 0. \tag{A.93}$$

d. Using the result from *c.*, compute the general solution to the differential equation

$$\dot{x}(t) + x(t) - e^t\big(x(t)\big)^3 = 0.$$

You can restrict attention to positive solutions x.

e. Assume that γ is an even integer and let z be a negative solution to (A.93). Show that

$$x(t) = -|z(t)|^{\frac{1}{1-\gamma}}$$

is a negative solution to (A.92).

f. Compute the solution to the initial value problem

$$\dot{x}(t) + \frac{x(t)}{1+t} + (1+t)\big(x(t)\big)^4 = 0, \qquad x(0) = -1,$$

and also indicate the maximum interval of existence of this solution.

Exercise A.5. Assuming that all functions are differentiable, show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left\{\int_{a(x)}^{b(x)} f(x,t)\,\mathrm{d}t\right\} = \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t + f\big(x,b(x)\big)\frac{\mathrm{d}b}{\mathrm{d}x}(x) - f\big(x,a(x)\big)\frac{\mathrm{d}a}{\mathrm{d}x}(x). \tag{A.94}$$

Hint: define

$$F(x,y,z) = \int_y^z f(x,t)\,\mathrm{d}t \tag{A.95}$$

and use the chain rule together with the fundamental theorem of calculus. This result is known as the Leibniz integral rule.

Appendix B

The Jordan canonical form

Some aspects of the Jordan canonical form are reviewed in this appendix, focusing on the computation of this canonical form. Details can be found in texts on linear algebra or matrix analysis, see, e.g., [3, 5]. The final section presents the Cayley-Hamilton theorem, for which the Jordan canonical form gives a compact proof.

B.1 The Jordan canonical form

For a matrix $A \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ is called an eigenvalue if there exists a nonzero vector $v \in \mathbb{C}^n$ such that

$$Av = \lambda v. \tag{B.1}$$

The set of all eigenvalues of A, known as the *spectrum* of A, will be denoted as $\sigma(A)$. Moreover, it is well-known that all eigenvalues can be found as the roots of the characteristic polynomial of A, i.e.,

$$\Delta_A(\lambda) = \det(\lambda I - A) = 0 \tag{B.2}$$

if and only if $\lambda \in \sigma(A)$.

These two equivalent approaches for characterizing eigenvalues allow for two different notions of *multiplicity* of an eigenvalue, which are in general not the same. First, the characterization (B.2) leads to the definition of algebraic multiplicity.

Definition B.1. *The algebraic multiplicity of an eigenvalue $\lambda \in \sigma(A)$, denoted a_λ, is its multiplicity as a root of the characteristic polynomial Δ_A, i.e.,*

$$\Delta_A(s) = (s - \lambda)^{a_\lambda} p(s), \tag{B.3}$$

for some polynomial p satisfying $p(\lambda) \neq 0$.

Note that, due to the fundamental theorem of algebra, we have that

$$\sum_{i=1}^{\ell} a_{\lambda_i} = n, \tag{B.4}$$

where $\lambda_1, \ldots, \lambda_\ell$ are the *distinct* eigenvalues of A.

To obtain a notion of multiplicity based on the definition (B.1), we first define the so-called eigenspace for a given eigenvalue.

Definition B.2. *The eigenspace of an eigenvalue $\lambda \in \sigma(A)$, denoted E_λ, is defined as*

$$E_\lambda = \ker(A - \lambda I) = \big\{ v \in \mathbb{C}^n \;\big|\; (A - \lambda I)v = 0 \big\}. \tag{B.5}$$

It is readily verified that E_λ is a subspace of $\mathbb{C}^n$. Its dimension is known as the geometric multiplicity of the associated eigenvalue, as defined next.

Definition B.3. *The geometric multiplicity of an eigenvalue $\lambda \in \sigma(A)$, denoted g_λ, is the dimension of the corresponding eigenspace, i.e., $g_\lambda = \dim E_\lambda$.*

Stated differently, the geometric multiplicity is the number of linearly independent eigenvectors that can be found for a given eigenvalue.

The concepts of algebraic and geometric multiplicity are illustrated by means of the following simple example.

Example B.1. Consider the matrices

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \qquad A' = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}, \tag{B.6}$$

which are easily seen to have the same characteristic polynomial. Namely,

$$\Delta_A(s) = \Delta_{A'}(s) = (s-2)^2, \tag{B.7}$$

such that both A and A' have the single eigenvalue $\lambda = 2$ with algebraic multiplicity $a_\lambda = 2$.

To compute the geometric multiplicity of λ for the matrix A, consider

$$\ker(A - \lambda I) = \ker \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}, \tag{B.8}$$

leading to $g_\lambda = 2$. However, a similar computation for the matrix A' leads to

$$\ker(A - \lambda I) = \ker \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \tag{B.9}$$

and we have, with some abuse of notation, $g'_\lambda = 1$. ◇

The above example shows that the geometric multiplicity g_λ of an eigenvalue can be smaller than its algebraic multiplicity. This result can be proven in general, which is stated next.

Theorem B.1. *Let $A \in \mathbb{C}^{n\times n}$. Then, the following hold:*

1. *$g_\lambda \leq a_\lambda$ for any $\lambda \in \sigma(A)$;*
2. *the matrix A is diagonalizable, i.e., there exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ such that $T^{-1}AT$ is diagonal, if and only if $g_\lambda = a_\lambda$ for all $\lambda \in \sigma(A)$.*

The second statement of Theorem B.1 is an important result. To elaborate on this, note that it implies that the diagonal matrix $T^{-1}AT$ has the eigenvalues of A on the diagonal. Specifically, the transformation T can be chosen such that

$$T^{-1}AT = \begin{bmatrix} \lambda_1 I_{a_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{a_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_\ell I_{a_\ell} \end{bmatrix}, \tag{B.10}$$

with $\lambda_1, \ldots, \lambda_\ell$ the distinct eigenvalues of A and where we have used the short-hand notation $a_i = a_{\lambda_i}$.

Remark B.1. When taking a more geometric perspective on eigenvalues, the following two statements can be shown to hold:

1. E_λ is an A-invariant subspace of $\mathbb{C}^n$, i.e., $v \in E_\lambda$ implies $Av \in E_\lambda$;
2. $E_\lambda \cap E_{\lambda'} = \{0\}$ for any two distinct λ, $\lambda' \in \sigma(A)$.

Moreover, as the geometric multiplicity g_λ is defined as the dimension of E_λ, there exists vectors $q_\lambda^1, \ldots, q_\lambda^g \in \mathbb{C}^n$ with $g = g_\lambda$ that form a basis for E_λ, i.e.,

$$E_\lambda = \operatorname{span}\{q_\lambda^1, \ldots, q_\lambda^g\}. \tag{B.11}$$

Now, statement *2* in Theorem B.1 can be interpreted as follows. As the result $E_\lambda \cap E_{\lambda'} = \{0\}$ above implies that bases of distinct eigenvalues are linearly independent, it follows from $g_i = a_i$ and (B.4) that the collection of bases for E_{λ_i} forms a basis for $\mathbb{C}^n$ when $i = 1, 2, \ldots, \ell$ ranges over the number of distinct eigenvalues. Stated differently, the matrix

$$T = \left[\begin{array}{ccc:ccc:c:ccc} q_{\lambda_1}^1 & \cdots & q_{\lambda_1}^{g_1} & q_{\lambda_2}^1 & \cdots & q_{\lambda_2}^{g_2} & \cdots & q_{\lambda_\ell}^1 & \cdots & q_{\lambda_\ell}^{g_\ell} \end{array} \right] \in \mathbb{C}^{n \times n} \tag{B.12}$$

is nonsingular. Moreover, it follows the definition of E_λ that $Aq_{\lambda_i}^j = \lambda_i q_{\lambda_i}^j$, such that we have

$$AT = T \begin{bmatrix} \lambda_1 I_{a_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{a_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_\ell I_{a_\ell} \end{bmatrix}. \tag{B.13}$$

Thus, the matrix (B.12) achieves (B.10) and the above gives a constructive procedure for finding the transformation in statement *2* of Theorem B.1. We note that, in the case of distinct eigenvalues, the basis q_λ of E_λ is given by a single eigenvector. In the above, we have used the short-hand notation $a_i = a_{\lambda_i}$ and $g_i = g_{\lambda_i}$. ◁

We illustrate this with an example.

Example B.2. Consider the matrix

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{B.14}$$

A direct computation gives the characteristic polynomial $\Delta_A(s) = (s+1)^2(s-2)$, leading two distinct eigenvalues with algebraic multiplicities

$$\begin{aligned} \lambda_1 &= -1, \quad & a_{\lambda_1} &= 2, \\ \lambda_2 &= 2, & a_{\lambda_2} &= 1. \end{aligned} \tag{B.15}$$

Then, a computation of the eigenspace for λ_1 gives

$$E_{\lambda_1} = \ker(A - \lambda_1 I) = \ker \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \operatorname{span}\left\{ q^1_{\lambda_1}, q^2_{\lambda_1} \right\}, \tag{B.16}$$

where

$$q^1_{\lambda_1} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \qquad q^2_{\lambda_1} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}. \tag{B.17}$$

Thus, the geometric multiplicity equals $g_{\lambda_1} = 2$, which equals the algebraic multiplicity of λ_1. We stress however that, even though the geometric multiplicity is uniquely defined, the basis (B.17) for E_λ is not and other bases are possible.

Repeating this computation for λ_2 yields

$$E_{\lambda_2} = \ker(A - \lambda_2 I) = \ker \begin{bmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{bmatrix} = \operatorname{span}\{q^1_{\lambda_2}\}, \quad q^1_{\lambda_2} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{B.18}$$

i.e., the geometric multiplicity reads $g_{\lambda_2} = 1$. Note that, due to the definition of an eigenvalue (B.1) and statement *1* in Theorem B.1, we always have that $g_\lambda = 1$ for any eigenvalue λ with algebraic multiplicity one.

Following the approach of Remark B.1, collecting the bases for E_{λ_1} and E_{λ_2} leads to the definition of T as

$$T = \begin{bmatrix} q^1_{\lambda_1} & q^2_{\lambda_1} & q^1_{\lambda_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}, \tag{B.19}$$

which is indeed easily seen to be nonsingular. Moreover, a direct computation gives

$$T^{-1}AT = \left[\begin{array}{cc|c} -1 & 0 & 0 \\ 0 & -1 & 0 \\ \hline 0 & 0 & 2 \end{array}\right] = \begin{bmatrix} \lambda_1 I_{g_1} & 0 \\ 0 & \lambda_2 I_{g_2} \end{bmatrix} \tag{B.20}$$

as expected. ◇

As a consequence of Theorem B.1, the diagonalization of a matrix A is not possible if there are eigenvalues whose geometric multiplicity is strictly less than the algebraic multiplicity. In such cases, we are interested in finding a transformation that *almost* diagonalizes A.

To make this explicit, we first define a specific matrix structure, known as a Jordan block.

Definition B.4. *A Jordan block $J_k(\lambda) \in \mathbb{C}^{k\times k}$ is the matrix*

$$J_k(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 & 0 \\ 0 & & & \ddots & \lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}. \tag{B.21}$$

Using this definition, the following fundamental result can be stated, which shows that any matrix can be written in a so-called *Jordan canonical form*.

Theorem B.2. *For any matrix $A \in \mathbb{C}^{n\times n}$, there exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ such that $A = TJT^{-1}$ with*

$$J = \begin{bmatrix} J_{k_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{k_2}(\lambda_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & J_{k_r}(\lambda_r) \end{bmatrix}, \tag{B.22}$$

where $\lambda_i \in \sigma(A)$, $i = 1, 2, \ldots, r$ and $n = k_1 + k_2 + \ldots + k_r$. Conversely, let $\lambda \in \sigma(A)$ be any eigenvalue of A. Then, $\lambda = \lambda_i$ for some $i \in \{1, 2, \ldots, r\}$.

We note that the matrix J in (B.22) plays the same role as the diagonal matrix containing the eigenvalues in (B.10).

It is important to stress that the parameters $\lambda_1, \ldots, \lambda_r$ in (B.22) are eigenvalues of A, but they are not necessarily distinct. Stated differently, an eigenvalue λ of A might correspond to multiple parameters λ_i. To make this explicit, let $\lambda \in \sigma(A)$ and let $\mathcal{J} \subset \{1, 2, \ldots, r\}$ be the set of indices such that $\lambda = \lambda_j$ for all $j \in \mathcal{J}$. Thus, the indices $\mathcal{J}$ characterize the Jordan blocks corresponding to the eigenvalue λ. Then, the following statements hold:

1. the *geometric* multiplicity of λ equals the number of Jordan blocks corresponding to λ, i.e., $g_\lambda = |\mathcal{J}|$, with $|\mathcal{J}|$ the cardinality (the number of elements) of the set $\mathcal{J}$;
2. the *algebraic* multiplicity of λ equals the sum of the dimensions of the Jordan blocks corresponding to λ, i.e., $a_\lambda = \sum_{j\in\mathcal{J}} k_j$.

Remark B.2. We remark that these statements imply the result of Theorem B.1. Namely, if $a_\lambda = g_\lambda$, this implies that every Jordan block is of size one. As a result, the matrix J is diagonal and equals the matrix at the right-hand side of (B.10) (potentially after reordering the diagonal elements). Consequently, the diagonalization of Theorem B.1 is a special case of Theorem B.2. ◁

The proof of Theorem B.2 is very involved and out of the scope of these notes. Instead, the next section focuses on the computation of the Jordan canonical form. The results required for this computation also provide some insights in (parts of) the proof of Theorem B.2.

B.2 Computation of the Jordan canonical form

We recall from Remark B.1 that the eigenspaces E_λ can be used to explicitly construct a matrix T that achieves the diagonalization of a matrix A, as long as the algebraic multiplicity of each eigenvalue equals its geometric multiplicity.

In order to find a transformation T that almost achieves diagonalization (in the sense specified by the Jordan canonical form in Theorem B.2) in the general case that $g_\lambda \leq a_\lambda$, we introduce the *generalized* eigenspace of an eigenvalue.

Definition B.5. *The generalized eigenspace of an eigenvalue $\lambda \in \sigma(A)$, denoted K_λ, is defined as*

$$K_\lambda = \big\{ v \in \mathbb{C}^n \;\big|\; (A - \lambda I)^p v = 0 \text{ for some integer } p > 0 \big\}. \tag{B.23}$$

By comparing the definitions of the eigenspace E_λ (Definition B.2) and the generalized eigenspace K_λ, it is immediate that $E_\lambda \subset K_\lambda$ for each $\lambda \in \sigma(A)$. Thus, we also have that $g_\lambda = \dim E_\lambda \leq \dim K_\lambda$.

In fact, it can be shown that the dimension of the generalized eigenspace K_λ equals the *algebraic* multiplicity of the eigenvalue λ, as stated in the following theorem.

Theorem B.3. *Let $A \in \mathbb{C}^{n\times n}$, take $\lambda \in \sigma(A)$ and consider the generalized eigenspace (B.23). Then, the following hold:*

1. *K_λ is an A-invariant subspace of $\mathbb{C}^n$, i.e., $v \in K_\lambda$ implies $Av \in K_\lambda$;*
2. *$\dim K_\lambda = a_\lambda$ with a_λ the algebraic multiplicity of λ;*
3. *$K_\lambda \cap K_{\lambda'} = \{0\}$ for any two distinct λ, $\lambda' \in \sigma(A)$.*

Proof. We only prove statement *1.*

1. First, it needs to be verified that K_λ is indeed a subspace of $\mathbb{C}^n$. Clearly, $0 \in K_\lambda$. Next, let v and v' be in K_λ, such that

$$(A - \lambda I)^p v = 0, \qquad (A - \lambda I)^{p'} v' = 0. \tag{B.24}$$

for some positive integers p and p'. As a result,

$$\begin{aligned}(A - \lambda I)^{p+p'}(v + v') &= (A - \lambda I)^{p+p'} v + (A - \lambda I)^{p+p'} v' \\ &= (A - \lambda I)^{p'} (A - \lambda I)^p v + (A - \lambda I)^p (A - \lambda I)^{p'} v' \\ &= 0 + 0,\end{aligned} \tag{B.25}$$

and we have that $v + v' \in K_\lambda$. The proof that $\alpha v \in K_\lambda$ for $v \in K_\lambda$ and $\alpha \in \mathbb{C}$ is immediate.

To show A-invariance, let $v \in K_\lambda$, i.e., $(A - \lambda I)^p v = 0$ for some positive integer p. Then,

$$(A - \lambda I)^p (Av) = (A - \lambda I)^p Av = A(A - \lambda I)^p v = 0, \tag{B.26}$$

such that $Av \in K_\lambda$. In the above, we have used the fact that $(A - \lambda I)^p$ and A commute, which is easily shown by induction on p. □

The result of Theorem B.3 has important consequences as it allows for the partial diagonalization of the matrix A. Namely, statement *2* implies the existence of vectors $q_\lambda^1, \ldots, q_\lambda^a \in \mathbb{C}^n$ with $a = a_\lambda$ that form a basis for K_λ. Thus, we have

$$K_\lambda = \operatorname{span}\{q_\lambda^1, \ldots, q_\lambda^a\}. \tag{B.27}$$

with $a = a_\lambda$ the algebraic multiplicity of λ (and the dimension of K_λ). By statement *3*, the bases corresponding to the generalized eigenspaces of two distinct eigenvalues are independent, such that (B.4) implies that the matrix T collecting all basis vectors as

$$T = \left[\begin{array}{ccc|ccc|c|ccc} q_{\lambda_1}^1 & \cdots & q_{\lambda_1}^{a_1} & q_{\lambda_2}^1 & \cdots & q_{\lambda_2}^{a_2} & \cdots & q_{\lambda_\ell}^1 & \cdots & q_{\lambda_\ell}^{a_\ell} \end{array}\right] \in \mathbb{C}^{n\times n} \tag{B.28}$$

is nonsingular. Here, the shorthand notation $a_i = a_{\lambda_i}$ is used. Then, the fact that the subspaces K_λ are A-invariant (statement *1* in Theorem B.3) gives

$$AT = T \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_\ell \end{bmatrix}, \tag{B.29}$$

for some matrices $A_i \in \mathbb{C}^{a_i \times a_i}$, $i = 1, 2, \ldots, \ell$.

Remark B.3. At this point, it is insightful to compare the equations (B.27), (B.28), and (B.29), with the developments in Remark B.1. We stress that in the case $E_{\lambda_i} = K_{\lambda_i}$ (which holds when $a_{\lambda_i} = g_{\lambda_i}$), the matrix A_i in (B.29) satisfies $A_i = \lambda_i I$. ◁

As the definition of the generalized eigenspaces K_λ in Definition B.5 already achieves partial diagonalization in the sense that the form (B.29) is obtained, we now turn attention to selecting a suitable basis (B.27) for K_λ. Namely, the basis for K_{λ_i} will determine the specific structure of A_i.

As a first step in this direction, we define so-called *cycles* of generalized eigenvectors.

Definition B.6. *Let $\lambda \in \sigma(A)$ and let $v \in K_\lambda$. Suppose that k is the smallest positive integer for which $(A - \lambda I)^k v = 0$. Then, the ordered set*

$$\big((A - \lambda I)^{k-1} v,\, (A - \lambda I)^{k-2} v,\, \ldots,\, (A - \lambda I) v,\, v\big), \tag{B.30}$$

denoted as $c_\lambda(v)$, is called a cycle (of length k) of generalized eigenvectors corresponding to λ.

It is readily verified that each vector in the cycle is indeed a generalized eigenvector. Moreover, the first element (known as the *initial vector*) is an eigenvector, i.e., $(A - \lambda I)^{k-1} v \in E_\lambda$. On the other hand, if v is an eigenvector corresponding to λ, then (v) is the corresponding cycle of length 1.

The relevance of defining cycles of generalized eigenvectors can be made apparent after introducing the shorthand notation

$$v^i = (A - \lambda I)^{k-i} v \tag{B.31}$$

to denote the elements of a cycle $c_\lambda(v)$. Namely, a direct calculation then gives

$$Av^1 = (A - \lambda I)v^1 + \lambda v^1 = (A - \lambda I)^k v + \lambda v^1 = \lambda v^1, \tag{B.32}$$

where Definition B.6 is used. Similarly, for $i = 2, 3, \ldots, k$,

$$Av^i = (A - \lambda I)v^i + \lambda v^i = (A - \lambda I)^{k-i+1} v + \lambda v^i = v^{i-1} + \lambda v^i, \tag{B.33}$$

such that collecting (B.32) and (B.33) yields

$$A\begin{bmatrix} v^1 & v^2 & \cdots & v^k \end{bmatrix} = \begin{bmatrix} v^1 & v^2 & \cdots & v^k \end{bmatrix} \begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} \tag{B.34}$$

$$= \begin{bmatrix} v^1 & v^2 & \cdots & v^k \end{bmatrix} J_k(\lambda). \tag{B.35}$$

Thus, a cycle of length k corresponding to λ leads to a Jordan block of size $k \times k$ for this eigenvalue.

After defining the span of a cycle as the span of its generalized eigenvectors according to

$$\operatorname{span}\{c_\lambda(v)\} = \operatorname{span}\{v^1, \ldots, v^k\}, \tag{B.36}$$

we can formalize properties of the cycle $c_\lambda(v)$ as follows.

Theorem B.4. *Let $\lambda \in \sigma(A)$, take $v \in K_\lambda$ and consider the cycle of generalized eigenvectors (B.30). Then, the following hold:*

1. $\operatorname{span}\{c_\lambda(v)\}$ *is an A-invariant subspace of K_λ;*
2. $\dim \operatorname{span}\{c_\lambda(v)\} = k$*;*
3. $\operatorname{span}\{c_\lambda(v_1)\} \cap \operatorname{span}\{c_\lambda(v_2)\} = \{0\}$ *for any two cycles of generalized eigenvectors $c_\lambda(v_1)$, $c_\lambda(v_2)$ whose intial vectors are linearly independent.*

Note that the property of A-invariance in statement *1* is proven by (B.35), whereas the result $\operatorname{span}\{c_\lambda(v)\} \subset K_\lambda$ immediately follows from the definition of a cycle. The second statement implies that the vectors $v^1, \ldots, v^k$ are linearly independent. At this point, we stress that cycle length k above might be smaller than the dimension of K_λ (given by the algebraic multiplicity a_λ), in which case a single cycle is not sufficient to provide a basis for K_λ. However, statement *3* in Theorem B.4 suggests that multiple cycles can be combined.

The following theorem states that a basis for the generalized eigenspace can be constructed in this way.

Theorem B.5. *Let $\lambda \in \sigma(A)$. Then, the generalized eigenspace K_λ has a basis consisting of cycles.*

This result is important as it proves the existence of the Jordan canonical form in Theorem B.2. Namely, we already had from Theorem B.3 that finding bases for the generalized eigenspaces K_{λ_i} leads to a block-diagonal form as in (B.29). Now, for each eigenvalue λ_i, the result in Theorem B.5 further refines this by guaranteeing an additional structure in A_i. Namely, as a basis consisting of cycles exists for K_{λ_i}, it follows from (B.35) that A_i has itself a block-diagonal structure, in which each block is a Jordan block corresponding to the eigenvalue λ_i.

However, one important question remains. Namely, what is the number of Jordan blocks for a given eigenvalue λ and what are the corresponding sizes? Or, stated differently, what is the number of independent cycles and their corresponding lengths such that they constitute a basis for K_λ?

To answer this question, let $\lambda \in \sigma(A)$ and define the parameters

$$\begin{aligned} r_{\lambda,1} &= n - \operatorname{rank}(A - \lambda I), \\ r_{\lambda,j} &= \operatorname{rank}\big((A-\lambda I)^{j-1}\big) - \operatorname{rank}\big((A-\lambda I)^{j}\big), \quad j = 2, 3, \ldots. \end{aligned} \tag{B.37}$$

It can be shown that there exists an integer $k > 0$ such that $r_{\lambda,k} \neq 0$ and $r_{\lambda,k+i} = 0$ for all $i > 0$. This integer is sometimes referred to as the *index* of the eigenvalue λ. Then, on the basis of the parameters $r_{\lambda,1}, \ldots, r_{\lambda,k}$, we can construct a so-called *dot diagram*. This is a diagram with k rows, were row j has $r_{\lambda,j}$ dots. This is most easily understood by an example.

Example B.3. For a given eigenvalue $\lambda \in \sigma(A)$, assume that $r_{\lambda,1} = 4$, $r_{\lambda,2} = 3$, $r_{\lambda,3} = 2$. Then, the dot diagram is obtained as

$$\begin{array}{lcl} r_{\lambda,1} = 4 & & \bullet\ \bullet\ \bullet\ \bullet \\ r_{\lambda,2} = 3 & \Rightarrow & \bullet\ \bullet\ \bullet \\ r_{\lambda,3} = 2 & & \bullet\ \bullet \end{array} \tag{B.38}$$

Note that the dot diagram has $r_{\lambda,1}$ columns. ◇

The dot diagram encodes the number of Jordan blocks and their sizes, for a given eigenvalue λ. This is made precise in the following theorem.

Theorem B.6. *Let $\lambda \in \sigma(A)$ and consider the parameters $r_{\lambda,j}$, $j = 1, 2, \ldots$ as in (B.37) as well as the associated dot diagram. Then, the following hold:*

1. *the number of columns in the dot diagram gives the number of independent cycles;*
2. *the number of dots in each column gives the length of the corresponding cycle.*

Now, we return to the dot diagram in Example B.3.

Example B.4. Given the dot diagram (B.38), it follows from Theorem B.6 that a basis for K_λ exists that consists of four cycles, with lengths 3, 3, 2, and 1. Specifically, denoting these cycles as $c_\lambda(v_1)$, $c_\lambda(v_2)$, $c_\lambda(v_3)$, and $c_\lambda(v_4)$, respectively, these can be associated to the dot diagram as

$$\begin{array}{llll} \bullet\ (A-\lambda I)^2 v_1 & \bullet\ (A-\lambda I)^2 v_2 & \bullet\ (A-\lambda I) v_3 & \bullet\ v_4 \\ \bullet\ (A-\lambda I) v_1 & \bullet\ (A-\lambda I) v_2 & \bullet\ v_3 & \\ \bullet\ v_1 & \bullet\ v_2 & & \end{array} \tag{B.39}$$

under the condition that the initial vectors are linearly independent (see Theorem B.4, statement *3*). Taking a perspective in terms of Jordan blocks, we have that the dot diagram (B.38) shows that the eigenvalue λ leads to four Jordan blocks, of size 3, 3, 2, and 1, respectively. ◇

Remark B.4. The parameters r_i defined in (B.37) allow for an insightful interpretation. Namely, from the definition it is clear that

$$r_{\lambda,1} = n - \operatorname{rank}(A - \lambda I) = \dim \ker(A - \lambda I) = \dim E_\lambda = g_\lambda, \tag{B.40}$$

i.e., $r_{\lambda,1}$ is the geometric multiplicity of the eigenvalue λ. This represents the number of cycles in the Jordan canonical form (or the number of Jordan blocks for λ, see also Theorem B.2). Similarly, the parameter $r_{\lambda,j}$ can be interpreted as the number of Jordan blocks of size at least j (cycles of length at least j) for the eigenvalue λ. The dot diagram is sometimes also called a Ferrers diagram or Young diagram. ◁

Based on the above developments, we can define the following procedure for computing the Jordan canonical form as in Theorem B.2.

1. Compute the eigenvalues λ_i and their algebraic multiplicity a_{λ_i} using the characteristic polynomial (B.2).

For each eigenvalue λ_i, perform the following steps.

2. Compute $\operatorname{rank}(A - \lambda_i I)^j$, for $j = 1, 2, \ldots$ until $\operatorname{rank}(A - \lambda_i I)^j = n - a_{\lambda_i}$.

3. Determine the dot diagram for λ_i using (B.37).

4. Compute each cycle using the dot diagram. There are many possible choices here, but make sure that the initial vectors of each cycle are linearly independent.

Finally, collect the results for the eigenvalues in the final steps.

5. Use the dot diagrams to construct the matrix J.

6. Use the corresponding cycles to construct the matrix T.

If we are only interested in finding the structure of the matrix J in (B.22), the first three steps (with step five) are sufficient and step four and six can be omitted. Namely, for each eigenvalue, the dot diagram fully determines the number of Jordan blocks and their sizes.

We finish this section by applying the above procedure to an example.

Example B.5. Consider the matrix

$$A = \begin{bmatrix} 2 & -1 & 0 & 1 \\ 0 & 3 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 3 \end{bmatrix}, \tag{B.41}$$

for which the characteristic polynomial can be calculated as

$$\Delta_A(s) = \det(sI - A) = (s-2)^3(s-3). \tag{B.42}$$

Thus, we have two distinct eigenvalues with algebraic multiplicities 3 and 1, respectively, i.e.,

$$\begin{aligned} \lambda_1 &= 2, \quad a_{\lambda_1} = 3, \\ \lambda_2 &= 3, \quad a_{\lambda_2} = 1. \end{aligned} \tag{B.43}$$

Consider the eigenvalue $\lambda_1 = 2$. We recall from Theorem B.3 that its algebraic multiplicity ensures that $\dim K_{\lambda_1} = 3$. This also means that the dot diagram for λ_1 will have three dots. In order to find the diagram, compute

$$\operatorname{rank}(A - \lambda_1 I) = \operatorname{rank} \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} = 2. \tag{B.44}$$

Here, we note that the geometric multiplicity of λ_1 can be obtained from this result as $g_{\lambda_1} = n - \operatorname{rank}(A - \lambda_1 I) = 4 - 2 = 2$. We proceed by computing

$$(A - \lambda_1 I)^2 = \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}^2 = \begin{bmatrix} 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 \end{bmatrix} \tag{B.45}$$

such that

$$\operatorname{rank}(A - \lambda_1 I)^2 = \operatorname{rank} \begin{bmatrix} 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 \end{bmatrix} = 1. \tag{B.46}$$

We note that $\operatorname{rank}(A - \lambda_1 I)^2 = n - a_{\lambda_1}$, such that there is no need to compute the rank for higher powers. Thus, the above results can be used in (B.37) to obtain the parameters $r_{\lambda_1,j}$ as

$$\begin{aligned} r_{\lambda_1,1} &= n - \operatorname{rank}(A - \lambda_1 I) = 4 - 2 = 2, \\ r_{\lambda_1,2} &= \operatorname{rank}(A - \lambda_1 I) - \operatorname{rank}(A - \lambda_1 I)^2 = 2 - 1 = 1, \end{aligned} \tag{B.47}$$

leading to the dot diagram

(B.48)

By Theorem B.6, we have that there are two Jordan blocks associated to λ_1. These are of size 2×2 and 1×1, respectively. We remark that the computation of $r_{\lambda_1,2}$ in (B.47) was in fact not needed. Namely, we had already concluded that there are three dots in the dot diagram (as $\dim K_{\lambda_1} = a_{\lambda_1} = 3$), such that $r_{\lambda_1,1} = 2$ only leaves the possibility $r_{\lambda_1,2} = 1$.

For the eigenvalue $\lambda_2 = 3$, we immediately have that $\dim K_{\lambda_2} = 1$ as λ_2 has algebraic multiplicity one (in fact, $K_{\lambda_2} = E_{\lambda_2}$). Even though a direct computation would lead to $r_{\lambda_2,1} = 1$ and the trivial dot diagram

(B.49)

such computation is not needed. Thus, combining the above observations, we have that the matrix A is similar to

$$J = \left[\begin{array}{cc|c|c} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 2 & 0 \\ \hline 0 & 0 & 0 & 3 \end{array}\right], \tag{B.50}$$

as stated in Theorem B.2.

In addition, we would like to find the matrix T that achieves the desired transformation. To this end, we first consider again $\lambda_1 = 1$ and recall that the dot diagram (B.48) shows that there exists a basis for K_{λ_1} comprising two cycles, of length 2 and 1. Recalling Definition B.6 and focusing on the cycle of length 2 first, we need to find $v \in K_{\lambda_1}$ such that

$$(A - \lambda_1 I)^2 v = 0, \qquad (A - \lambda_1 I) v \neq 0. \tag{B.51}$$

In more geometric terms, we have $v \in \ker(A-\lambda_1 I)^2$ with

$$\ker(A-\lambda_1 I)^2 = \ker \begin{bmatrix} 0 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 \end{bmatrix} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \end{bmatrix} \right\} \tag{B.52}$$

where the result (B.45) is exploited. Here, we stress that any $v \in \ker(A-\lambda_1 I)^2$ also satisfies $v \in K_{\lambda_1}$ (see Definition B.5) such that there is no need to verify $v \in K_{\lambda_1}$ separately. From (B.51), the vector v should however be such that $v \notin \ker(A-\lambda_1 I)$, where

$$\ker(A-\lambda_1 I) = \ker \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\} = E_{\lambda_1}. \tag{B.53}$$

Using (B.52) and (B.53), we can observe that

$$v = v_1 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix} \tag{B.54}$$

is a vector that satisfies (B.51). This leads to the cycle

$$c_{\lambda_1}(v_1) = \big((A-\lambda_1 I)v_1, v_1\big) = \left(\begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix} \right). \tag{B.55}$$

A second cycle of length 1 remains to be found. Specifically, we are looking for $v_2 \in \ker(A-\lambda_1 I)$ that is linearly independent from the vectors in $c_{\lambda_1}(v_1)$ (by Theorem B.4 it is sufficient to check independence with respect to the initial vector only). Using (B.53), it can be shown that

$$v_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{B.56}$$

satisfies these criteria such that the second cycle corresponding to λ_1 reads

$$c_{\lambda_1}(v_2) = \big(v_2\big) = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right). \tag{B.57}$$

As the eigenvalue $\lambda_2 = 3$ has algebraic multiplicity 1 (and, hence, geometric multiplicity 1), any vector $v \in \ker(A-\lambda_2 I) = E_{\lambda_2}$ with

$$\ker(A-\lambda_2 I) = \ker \begin{bmatrix} -1 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\} \tag{B.58}$$

is a basis for $K_{\lambda_2} = E_{\lambda_2}$. After choosing

$$v = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{B.59}$$

leading to the trivial cycle $c_{\lambda_2}(v) = (v)$, we can collect the cycles (B.55), (B.57), and $c_{\lambda_2}(v)$ to obtain the transformation matrix

$$T = \begin{bmatrix} -1 & 0 & 1 & 1 \\ -1 & 1 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}. \tag{B.60}$$

Finally, a direct computation verifies $A = TJT^{-1}$ with J in (B.50). ◇

B.3 The Cayley-Hamilton theorem

We conclude this appendix with a result from linear algebra that is used frequently in the study of linear systems.

To this end, recall that the *characteristic polynomial* of a matrix $A \in \mathbb{C}^{n\times n}$, denoted Δ_A, is defined as

$$\Delta_A(s) = \det(sI - A). \tag{B.61}$$

The properties of the determinant function imply that this is indeed a polynomial, i.e., it is of the form

$$\Delta_A(s) = a_n s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0, \tag{B.62}$$

where $a_0, \ldots, a_n \in \mathbb{C}$. In fact, we have $a_n = 1$, such that (B.61) is a *monic* polynomial of degree n. Moreover, if A is a real matrix, then the coefficients $a_0, \ldots, a_n$ are real scalars.

The following result is known as the *Cayley-Hamilton theorem.*

Theorem B.7. *Let $A \in \mathbb{C}^{n\times n}$ and consider its characteristic polynomial (B.62). Then,*

$$a_n A^n + a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 I = 0. \tag{B.63}$$

Proof. This is Exercise B.4. □

Loosely speaking, the Cayley-Hamilton theorem states that any matrix is a root of its own characteristic polynomial, which is sometimes written as

$$\Delta_A(A) = 0. \tag{B.64}$$

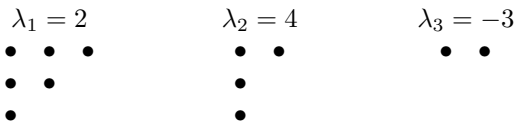
Importantly, as $a_n = 1$, the result (B.63) implies that A^n can be written as

$$A^n = -a_{n-1}A^{n-1} - a_{n-2}A^{n-2} - \ldots - a_1 A - a_0 I, \tag{B.65}$$

i.e., A^n is a linear combination of $I, A, \ldots, A^{n-1}$. Moreover, it can easily be derived from (B.65) that, for any integer $k \geq n$, A^k is a linear combination of $I, A, \ldots, A^{n-1}$. Namely, multiplication of (B.65) by A and proceeding by induction shows this result.

B.4 Exercises

Exercise B.1. Let A be a matrix with three distinct eigenvalues $\lambda_1 = 2$, $\lambda_2 = 4$, and $\lambda_3 = -3$. Assume that the dot diagrams for the eigenvalues are given as



(B.66)

Determine the Jordan canonical form J of A.

Exercise B.2. Let A be a matrix whose Jordan canonical form is given as

$$J = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}. \tag{B.67}$$

a. Determine the characteristic polynomial of A.

b. For each eigenvalue $\lambda_i \in \sigma(A)$, determine the dot diagram.

c. For which eigenvalues λ_i, if any, does $E_{\lambda_i} = K_{\lambda_i}$?

Exercise B.3. Let $\lambda \in \sigma(A)$ and consider the parameters $r_{\lambda,j}$ defined in (B.37). Show that

$$r_{\lambda,1} \geq r_{\lambda,2} \geq r_{\lambda,3} \geq \dots. \tag{B.68}$$

Exercise B.4. Use the Jordan canonical form of A to prove the Cayley-Hamilton theorem in Theorem B.7.

Bibliography

[1] S. Abbott. *Understanding Analysis*. Springer, New York, USA, second edition, 2015.

[2] V. I. Arnold. *Ordinary Differential Equations*. Springer-Verlag Berlin Heidelberg, Germany, 1992.

[3] S. H. Friedberg, A. J. Insel, and L. E. Spence. *Linear Algebra*. Pearson Education, Upper Saddle River, USA, fourth edition, 2003.

[4] J. K. Hale. *Ordinary Differential Equations*. Rober E. Krieger Publishing Company, New York, USA, 1980.

[5] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, United Kingdom, second edition, 2013.

[6] G. Meinsma. Elementary proof of the Routh-Hurwitz test. *Systems & Control Letters*, 25(4):237–242, 1995.

[7] R. J. Minnichelli, J. J. Anagnost, and C. A. Desoer. An elementary proof of Kharitonov's stability theorem with extensions. *IEEE Transactions on Automatic Control*, 34(9):995–998, 1989.

[8] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[9] H. L. Trentelman, A. A. Stoorvogel, and M. L. J. Hautus. *Control Theory for Linear Systems*. Communications and Control Engineering. Springer-Verlag, London, Great Britain, 2001.

[10] W. Walter. *Ordinary Differential Equations*. Graduate Texts in Mathematics. Springer-Verlag, New York, USA, 1998.