Linear Systems Theory

October 2022

Gjerrit Meinsma

Department of Applied Mathematics
University of Twente

Contents

What is Systems Theory about?

Words such as *system, systems theory*, and *systems engineering* have become ingrained in various fields such as data processing, electrical engineering, economics, management, biology, theoretical computer science, and mathematics; consequently, the exact meaning of these terms is no longer clear-cut. We will therefore first attempt to describe more precisely what we will be studying in this course.

The word *system* refers to an object, device, phenomenon, or part of the environment that causes certain measurable quantities in that environment to be interrelated. We call the measurable quantities variables. In this course, we will mainly concern ourselves with *dynamical* systems. These are systems in which the variables evolve over time. In this case, the variables are often called *signals*. The variables are mostly real valued—the position of a mass in a mechanical system, the current through a wire in an electrical circuit, the height of the interest rate in a model of a national economy, etc.—or discrete—the position of a switch, a symbol equal to 0 or 1, corresponding to "on" or "off".

In order to reproduce and analyze the dynamical behavior of a system, we will consider a *mathematical model* of the system, which, to some extent, shows how the different variables in the system evolve in relation to one another. In many cases, one and the same system can be connected to different mathematical models, corresponding to different compromises between the precision or descriptive quality of the model and its simplicity. The choice of the mathematical model can also depend on which problem pertaining to the system we wish to study.

Since mathematical models are themselves also systems (of a more abstract nature), it is common to use the word "system" for both the (physical) object of study and its mathematical model. Although systems theory also deals with formulating mathematical models for specific systems, a system in this course will always be an (idealized) mathematical system.

It is essential in systems theory that some of the variables describe the relation between the system and the *environment* of the system, or between the system and other systems. These variables are called the interconnection or *external variables*, and any other variables are called *internal variables*. Think, for example, of a watch, where the external signals could include the position of the watch hands, and the internal variables could include the state of the cogs and whatever else is inside the watch. A useful representation of a system is a box with lines to the environment, where the internal variables are associated with the box and the external variables are associated with the lines; see Figure 1.

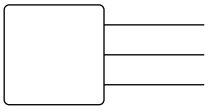

FIGURE 1: System.

In a so-called *black box* approach to the system, we cannot or do not wish to describe what happens in the box (i.e., the black box), and the description of the system will only concern the evolution of the external variables (think of the watch). A more detailed description of the box may also be available, for example that it is made up of a number of *subsystems* that are linked through their external variables; see Figure 2.

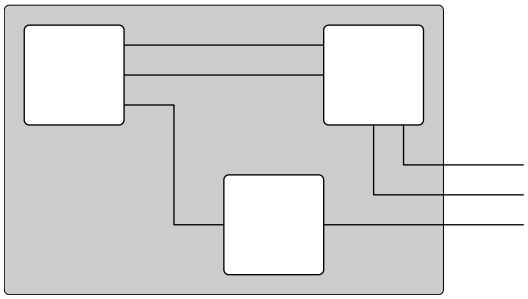

FIGURE 2: Complex system.

In many cases, it is useful to separate the external variables into *inputs* and *outputs*. Input variables can be set arbitrarily by the environment of the system—like the voltage across a voltage source in an electrical circuit or whether or not keys are pressed—while, on the other hand, the output variables are set by the input variables and internal variables—like the current through the voltage source or the symbols that appear on the monitor, respectively. Input/output systems are depicted as in Figure 3; the arrows toward the box indicate inputs, while the arrows leaving the box indicate outputs.

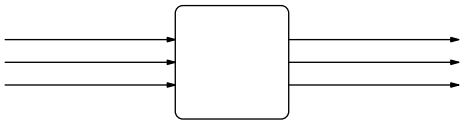

FIGURE 3: System with inputs and outputs.

In this course we will mainly deal with input/output systems because of their great importance for applications, and because they are easier to analyze than more general systems.

Using a number of examples, we will now briefly present typical problems and questions from systems theory. These examples serve as motivation, and we will encounter many of them again further on in the syllabus. Some problems are too ambitious to solve in an introductory course such as this one, some first need to be idealized because they are otherwise too complex, while others can be solved completely using the techniques that will be introduced in this course.

- *Steering and observing.* What is necessary to steer a car? On the one hand, the driver must be able to *control* the behavior of the car to some extent (which is why there is a steering wheel!), and on the other hand, the driver must *observe* the surroundings well enough to control successfully. What does the driver need to observe, and how aggressively may we pull on the steering wheel to keep the car on the road?

  The panacea here is *feedback*. Feedback is omnipresent, for example in biomedical systems such as the regulation of your blood sugar levels. We will discuss feedback extensively.

- *Tracking.* How can we, while walking, follow a moving object perfectly with our eyes? If we can describe this with a mathematical model, then we might also be able to let

cameras follow moving objects. Other similar problems: How can a control tower track the trajectories of planes? How can we ensure that the laser beam in the CD player follows the track on the CD?

- *Filtering & signal processing.* How do equalizers and noise suppressors work? How can we remove undesirable properties of signals? The "device" that does this is often called a filter. What are convolution filters (popular in, for example, video-image processing)? Why are jpeg files so small, and why are jpeg images often more fuzzy than the original picture?

- *Robust control.* How can we design a cruise control system that ensures that the car maintains a constant speed regardless of road slope or wind conditions? Similarly, how can we ensure that the laser beam in the CD player continues to follow the track despite perturbations?

- *Uncertain models.* Adjusting the temperature in an unfamiliar shower can lead to painful situations. This is because we do not know the shower, in other words, because we have a defective model of the operation of the shower. In general, it is difficult to draw conclusions using defective models. How can we bypass the defects? For example, we could take more time to adjust the temperature, or is there a more advance solution? A well-known example of successfully controlling uncertain models is *Black's negative feedback amplifier*.

- *Divide and conquer.* To model complex systems, it is helpful to view a complex system as an interconnection of subsystems. When are the interconnections set correctly? How can we simulate complex systems by combining simulations of the subsystems?

- *System identification.* How can we improve a mathematical model of, say, a car by experimenting with the car? Which experiments are necessary to obtain as much information as possible? We will not discuss this in this course.

- *Stochastic models.* The influence of noise, wind gusts, measurement errors, etc. are difficult to model. In such cases, it may be helpful to see these as realizations of stochastic processes. Systems with stochastic components form an important subfield of systems theory. We will not discuss it here.

In this course we analyse systems described by ordinary differential equations. This being a first course on systems theory we limit attention almost exclusively to *linear* and *time-invariant* systems, such as those described by ordinary differential equations with constant coefficients.

Chapter 1

Linear Differential Equations with Constant Coefficients

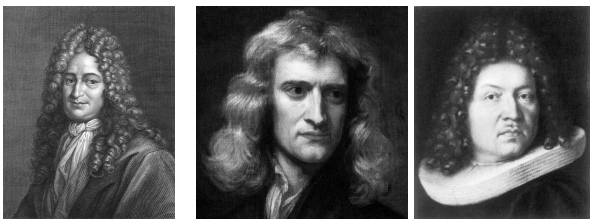

FIGURE 1.1: Gottfried Wilhelm Leibniz (1646-1716), Isaac Newton (1643-1727) and Jakob Bernoulli (1654–1705) developed the foundations of calculus and differential equations.

Differential equations — or DE's for short — are tremendously useful as models of dynamical phenomena. An advantage, and at the same time disadvantage, is that DE's do not directly provide explicit solutions of the phenomena. They merely relate variables and how they change (their derivatives). Quite often DE's are be derived from physical laws such as Newton's laws and mass balances. This is called modelling from "first principles". But differential equations are common to many other fields. For instance several DE's have been proposed that model the spread of a virus such as COVID. Weather forecasts are based on DE's, and control of drones works because we have effective models of drones described by DE's. The list goes on.

In this chapter we analyze *linear* DE's with *constant coefficients* but to get an idea of the topic we start with an overview of a couple of other types of DE's. As we will see, the theory of linear DE's has close connections with linear algebra. This is particularly true for the state representations that we consider in later chapters.

1.1 Examples and Classification of DE's

Example 1.1.1 (Pandemics — Logistic DE — nonlinear DE). How to model the spread of a virus in, say, the Netherlands? This is a complicated problem, but most models around are generalizations of an idea coined by the Belgium mathematician Pierre-François Verhulst (1804–1849). His idea was to split the population into two seperate groups: *infected* and *uninfected*, and the assumption is that the number of contacts between infected individuals and uninfected individual is linear in both the number of infected individuals and number of uni-

fected individuals. This can be explained: if, say, we have twice as many infected individuals then, on average, there will be twice as many contacts with infected individuals over, say, one day. Likewise, if we have twice as many uninfected individuals then over the course of one day on average twice as many contacts with uninfected individuals takes place.

Now assume that the size of the population does not change, so no immigration et cetera. Let $y(t)$ denote the fraction of infected individuals (hence $y(t) \in [0,1]$ with $y(t) = 1$ meaning everyone is infected). Then $1 - y(t)$ is the fraction of uninfected individuals. The number of contacts between infected and uninfected individuals by the above assumption is proportional to both infected $y(t)$ and uninfected $1 - y(t)$, so it is proportional to the *product* $y(t)(1 - y(t))$. With each contact between an infected and uninfected individual there is a certain probability that the uninfected contracts the disease. This means that, over the entire population, the increase of the number of infected individuals is proportional to this product. That gives us the DE[1]

$$\dot{y}(t) = \beta y(t)(1 - y(t)). \tag{1.1}$$

Here β is a positive number that models how infectuous the desease is. See Fig. 1.2. This DE is the famous logistic DE and its solution $y(t)$ is known as the logistic curve. It is good to realize that we assumed quite a lot here (in particular that infected individuals never recover), and clearly this model is just a first step. But it is an important first step, and many models used to this day are generalizations of this one model. □

A bit of terminology. Equation (1.1) is an example of a first-order DE because its highest derivative is of order 1. It is also a nonlinear DE because the highest derivative, $\dot{y}$, depends nonlinearly on y.

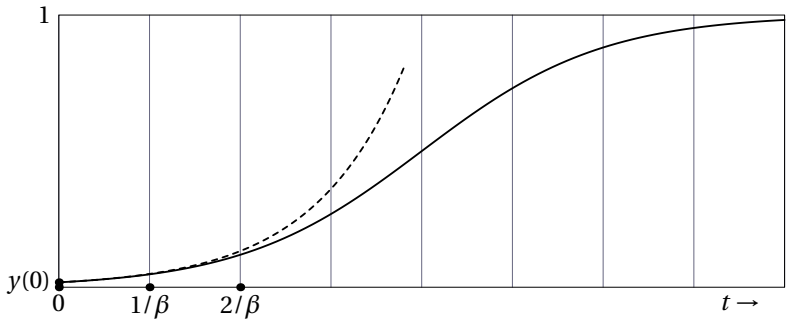

FIGURE 1.2: The solid graph is the solution $y(t)$ of the nonlinear DE (1.1) for some small initial condition $y(0)$. The solution is known as the logistic curve. The dashed graph is that of the exponential solution $y(0)e^{\beta t}$ as derived in Example 1.1.2 from a linear DE. For the given $y(0)$ the exponential solution resembles the logistic curve up to about $t = 2/\beta$.

Example 1.1.2 (Initial Pandemics — exponential growth). We continue with the pandemic example, but we simplify it a bit. At the beginning of a pandemic the fraction $y(t)$ of infected individuals is very small, so then $1 - y(t) \approx 1$. If we use this approximation then the logistic DE (1.1) simplifies to

$$\dot{y}(t) = \beta y(t).$$

[1]There are many ways to denote derivatives. The 1st derivative of $y(t)$ is denoted $\frac{\mathrm{d}}{\mathrm{d}t}y(t)$ as well as $y'(t)$ and $\dot{y}(t)$ and $y^{(1)}(t)$. Likewise the 2nd derivative is often denoted as y'' and as $\ddot{y}$ and $y^{(2)}$. We probably use them all!

This is a first-order linear DE because the highest-order derivative (here $\dot{y}$) is linear in y. Linearity simplifies the analysis a lot. In fact we will soon see that this DE has a simple and complete solution:

$$y(t) = y(0)e^{\beta t}.$$

Thus, at the beginning of a pandemic a virus spreads *exponentially* fast. (It is interesting to compare this exponential solution with that of the "true" solution $y(t)$ of the nonlinear logistic DE (1.1), see Figure 1.2.) □

In the two examples presented so far we only had one function: $y(t)$. Quite often there is also "driving term" or "input". Such functions we normally denote by $u(t)$. Here are two examples.

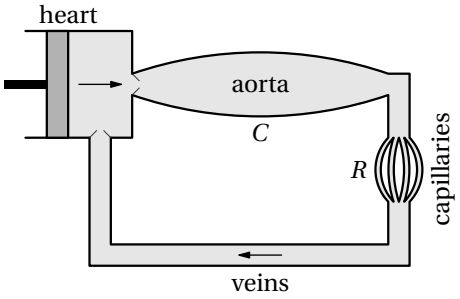

FIGURE 1.3: Blood circulation

Example 1.1.3 (Blood circulation — a DE with constant coefficients and with an input). The "windketelmodel" of blood circulation is a fairly simple model. In this model the circulatory system is considered as an interconnection of three subsystems: heart, aorta and the combination of capillaries and veins, see Fig. 1.3. The heart is modelled as a pump that produces a flow of blood into the aorta. As a result the aorta expands and the pressure, relative to the veins, builds up. This makes the blood flow from the aorta trough the capillaries into the veins and, eventually, back into the heart. The specific assumptions are:

- The flow of blood of $u(t)$ liters per second is given (probably some almost periodic signal) and does not depend on the condition of the aorta.
- Let $y(t)$ be the volume (liter) of blood in the aorta, relative to some volume of equilibrium. The flow from aorta to the veins is proportional to the the pressure difference between aorta and veins. That is, $\dot{y}(t) = -p(t)/R$ for some positive constant R, and with $p(t)$ the pressure. (The parameter R is a *resistance*.)
- The pressure difference $p(t)$ is proportional to the volume $y(t)$ in the aorta. Hence $p(t) = y(t)/C$ for some positive constante C. (This parameter is a *capacity* of the aorta.)

Combination of in and out flow gives $\dot{y}(t) = -y(t)/(RC) + u(t)$, or

$$\dot{y}(t) + \alpha y(t) = u(t), \qquad \text{for } \alpha := 1/(RC).$$

This is a 1st-order linear DE with constant coefficients. The latter refers to the fact that the parameters involved — here $\alpha := 1/(RC)$ — do not depend on t. □

It is an interesting fact that fundamentally different applications may lead to the same or similar differential equations. The differential equation in the next example has a lot in common with that of the blood ciculation example.

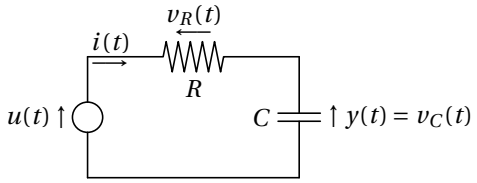

FIGURE 1.4: An RC circuit. (See Example 1.1.4.)

Example 1.1.4 (RC-circuit). Figure 1.4 depicts an electrical RC circuit. De voltage $u(t)$ over the source we assume given, and the voltage $y(t)$ over the capacitor we need to find. This voltage $y(t)$ is proportional to the charge $q(t)$ on the capacitor,

$$y(t) = \frac{q(t)}{C}. \tag{1.2}$$

Differentiating once with respect to time we find that

$$C\dot{y}(t) = i(t),$$

with $i(t)$ the current trough the capacitor. This current equals the current through the resistor, and the standard model is that this current is proportional to the voltage over the resistor,

$$v_R(t) = Ri(t) = RC\dot{y}(t).$$

The total voltage drop $v_R(t) + y(t)$ over resistor and capacitor has to add up to the supplied voltage $u(t)$, so

$$RC\dot{y}(t) + y(t) = u(t).$$

Divide by RC and we arrive at the DE

$$\dot{y}(t) + \alpha y(t) = \alpha u(t), \tag{1.3}$$

in which $\alpha = \frac{1}{RC}$. This DE is, again, a 1st-order linear DE with constant coefficients. □

All examples so far were 1st-order DE's. Often, though, we need higher order derivatives. Here is a 2nd-order DE:

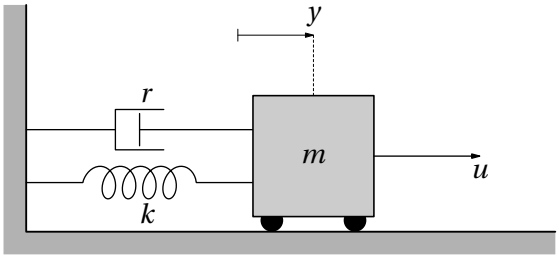

FIGURE 1.5: A mass-spring-damper system: a car connected to a wall. (See Example 1.1.5.)

Example 1.1.5 (Mass-spring-damper system). We have a car of mass m that can move in the horizontal direction, see Fig. 1.6. The mass is connected to a wall via a spring with spring constant k and damper with damping coefficient r. We may pull on the mass with a force u. Let y denote horizontal position relative to its equilibrium (meaning that $y = 0$ is zero where

the spring exerts no force). Newton's second law, combined with Hooke's law for springs, states that $m\ddot{y} = u - ky - r\dot{y}$, that is

$$m\ddot{y}(t) + r\dot{y}(t) + ky(t) = u(t). \tag{1.4}$$

This is a 2nd-order linear DE with constant coefficients. □

In the rest of this chapter we focus exclusively on linear DE's of finite order and with constant coefficients. These are DE's of the form

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = u(t). \tag{1.5}$$

Here $u : \mathbb{R} \to \mathbb{R}$ is assumed given, and we are after one or more solutions $y : \mathbb{R} \to \mathbb{R}$ of the equation. The positive integer n is the order of the DE. We normally refer to t as "time" but this is only for ease of reference; in certain applications it might mean "position" or something else, see Exercise 1.1.

1.2 Homogeneous equation

If the given function $u(t)$ is the zero function then the DE (1.5) reduces to what is called the homogeneous equation

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = 0. \tag{1.6}$$

This is a very important special case and we analyse it in detail in a series of sections. Its solutions $y(t)$ are called homogeneous solutions. In applications such solutions are sometimes referred to as *natural responses* since these are the functions that we get if we do "nothing" with the input $u(t)$ (meaning when $u(t) = 0$). In applications $y(t)$ is typically real-valued, but for the development of the theory it is better to also allow complex-valued solutions, so $y : \mathbb{R} \to \mathbb{C}$. Later we specialise it to the real-valued case.

Lemma 1.2.1 (Homogeneous subspace). *The set of all solutions $y : \mathbb{R} \to \mathbb{C}$ of a homogeneous equation* (1.6) *is a (linear) subspace of the complex vector space $\{y : \mathbb{R} \to \mathbb{C}\}$. Hence if $y_1, \ldots, y_k$ are solutions then so is any element of* $\operatorname{span}\{y_1, \ldots, y_k\}$.

Proof. We need to verify the three properties of subspace:

- the zero function is in the set
- if y_1, y_2 are in the set then so is its sum $y_1 + y_2$
- if y_1 is in the set then cy_1 is in the set for every scalar c.

The zero function $y(t) = 0$ is differentiable infinitely often, and it satisfies the homogeneous equation. Now suppose y_1 and y_2 are two solutions of the homogeneous equation (in particular they are n times differentiable.) Then $y := y_1 + y_2$ is n times differentiable, and

$$\begin{aligned} y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_0 y &= (y_1 + y_2)^{(n)} + p_{n-1}(y_1 + y_2)^{(n-1)} + \cdots + p_0(y_1 + y_2) \\ &= \left(y_1^{(n)} + p_{n-1}y_1^{(n-1)} + \cdots + p_0 y_1\right) + \left(y_2^{(n)} + p_{n-1}y_2^{(n-1)} + \cdots + p_0 y_2\right) \\ &= 0 + 0 = 0. \end{aligned}$$

Hence $y := y_1 + y_2$ satisfies the homogeneous equation as well, i.e. is in the solution set. Finally, suppose that y_1 satisfies the homogeneous equation and let $y = cy_1$ for some scalar c. Clearly if y_1 is n times differentiable then $y := cy_1$ is too, and

$$\begin{aligned} y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_0 y &= (cy_1)^{(n)} + p_{n-1}(cy_1)^{(n-1)} + \cdots + p_0(cy_1) \\ &= c(y_1^{(n)} + p_{n-1}y_1^{(n-1)} + \cdots + p_0 y_1) \\ &= c0 = 0. \end{aligned}$$

So $y := cy_1$ then satisfies the equation as well. ∎

1.3 First-order homogeneous DE

We first have a look at the case that $n = 1$. Then (1.6) is nothing but

$$\dot{y}(t) + p_0 y(t) = 0. \tag{1.7}$$

In other words, $\dot{y}(t) = -p_0 y(t)$. It is not hard to see that

$$y(t) = ce^{-p_0 t} \tag{1.8}$$

is a solution for every $c \in \mathbb{C}$. Indeed, we then have $\dot{y}(t) = -cp_0 e^{-p_0 t} = -p_0 y(t)$. The question now is: are there other solutions? The answer is no! To prove this we use a method called *variation of parameters* also known as *variation of constants*. It is a method we will use more often. It starts by writing the candidate solution $y(t)$ as in (1.8) but now with c depending on time,

$$y(t) = c(t)e^{-p_0 t}. \tag{1.9}$$

So now $c : \mathbb{R} \to \mathbb{C}$. It is important to realize that every $y(t)$ can be written in this form because $e^{-p_0 t}$ is invertible for every t. Also, it will be clear that $y(t)$ is differentiable iff so is $c(t)$, and then the product rule applies and says that

$$\dot{y}(t) = \dot{c}(t)e^{-p_0 t} - p_0 c(t)e^{-p_0 t}.$$

This $y(t) := c(t)e^{-p_0 t}$ satisfies the DE (1.7) iff

$$\begin{aligned} \dot{y}(t) + p_0 y(t) = 0 &\iff \left(\dot{c}(t)e^{-p_0 t} - p_0 c(t)e^{-p_0 t}\right) + p_0 c(t)e^{p_0 t} = 0 \\ &\iff \dot{c}(t)e^{-p_0 t} = 0 \\ &\iff \dot{c}(t) = 0 \\ &\iff c \text{ is constant.} \end{aligned}$$

Thus we proved:

Lemma 1.3.1 (All solutions of a 1st-order homogeneous DE). *Let $p_0 \in \mathbb{R}$. A differentiable function $y : \mathbb{R} \to \mathbb{C}$ satisfies $\dot{y}(t) + p_0 y(t) = 0$ for all $t \in \mathbb{R}$ iff there is a $c \in \mathbb{C}$ such that $y(t) = ce^{-p_0 t}$.* □

To put it differently, the subspace $\mathbb{Y}$ of all homogeneous solutions is spanned by the function $e^{-p_0 t}$. Notice that $c = y(0)$, so we can also write the general solution as

$$y(t) = y(0)e^{-p_0 t}, \qquad y(0) \in \mathbb{C}.$$

In applications $y(0)$ is a real number and then $y(t)$ is real-valued for all time. The number $y(0)$ is known as the initial condition (of the first-order DE), and, as the above shows, the initial condition $y(0)$ determines the solution $y(t)$ of first-order DE's uniquely for the rest of time.

Example 1.3.2 (Newton's law of cooling). Let $y(t)$ be the temperature at time t of some object sitting in a medium of constant zero degrees. Clearly, the hotter the object the faster the cooling of the object. *Newton's law of cooling* asserts that the rate of cooling is proportional to the temperature of the object, that is

$$\dot{y}(t) = -ky(t),$$

for some constant $k > 0$. This constant is sometimes called "heat transfer coefficient". The above DE is of course the same as

$$\dot{y}(t) + ky(t) = 0,$$

and so by Lemma 1.3.1 we have that

$$y(t) = y(0)\mathrm{e}^{-kt}.$$

Since $k > 0$ we see that all possible solutions converge to zero, and the larger k the faster the convergence. □

Example 1.3.3 (*RC* network). Consider the RC-circuit of Example 1.1.4. In the example we argued that the voltage $y(t)$ accross the capacitor satisfies the DE

$$\dot{y}(t) + \alpha y(t) = \alpha u(t), \qquad \alpha := \frac{1}{RC} > 0,$$

and where $u(t)$ is a given voltage. For zero voltage $u(t)$ we end up with a homogeneous equation $\dot{y}(t) + \alpha y(t) = 0$, and hence the general homogeneous solution is

$$y(t) = y(0)\mathrm{e}^{-\alpha t} = y(0)\mathrm{e}^{-t/(RC)}.$$

The larger the resistance R and/or capacitance C, the slower the convergence of the voltage accross the capacitor to zero. □

1.4 Second-order homogeneous DE

Consider next the 2nd-order DE

$$\ddot{y}(t) + p_1\dot{y}(t) + p_0 y(t) = 0. \tag{1.10}$$

Inspired by the 1st-order DE we try a solution of the exponential form,

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

with λ a number that we do not yet know. Also now it is convenient to momentarily allow *complex* solutions $y : \mathbb{R} \to \mathbb{C}$ and complex numbers $\lambda \in \mathbb{C}$. (Later, once we solved the complex case, we will switch back to the real solutions.) The first and second derivatives of $y(t) := \mathrm{e}^{\lambda t}$ are $\dot{y}(t) = \lambda\mathrm{e}^{\lambda t} = \lambda y(t)$ and $\ddot{y}(t) = \lambda^2\mathrm{e}^{\lambda t} = \lambda^2 y(t)$. Plugging this into DE (1.10) gives

$$(\lambda^2 + p_1\lambda + p_0)\mathrm{e}^{\lambda t} = 0.$$

Clearly this holds iff the number λ satisfies

$$\lambda^2 + p_1\lambda + p_0 = 0. \tag{1.11}$$

This final equation no longer involves t. Equation (1.11) is known as the characteristic equation (of the 2nd order DE), and the polynomial on the left-hand side of the equality is known as the characteristic polynomial[2] (of the 2nd-order DE). Over the complex numbers this quadratic characteristic equation has solutions

$$\lambda_{1,2} = \frac{-p_1 \pm \sqrt{p_1^2 - 4p_0}}{2} = -p_1/2 \pm \sqrt{(p_1/2)^2 - p_0}. \tag{1.12}$$

Thus each of these gives us an exponential solution of the DE:

$$y_1(t) := \mathrm{e}^{\lambda_1 t}, \qquad y_2(t) := \mathrm{e}^{\lambda_2 t}. \tag{1.13}$$

There are usually two such solutions, but it can happen that $\lambda_1 = \lambda_2$. Again the question now is: are there other solutions? Clearly yes, because as $y_1(t) := \mathrm{e}^{\lambda_1 t}$ and $y_2(t) := \mathrm{e}^{\lambda_2 t}$ are two solutions, the subspace property (Lemma 1.2.1) immediately gives that all linear combinations

$$y(t) = c_1 \mathrm{e}^{\lambda_1 t} + c_2 \mathrm{e}^{\lambda_2 t}, \qquad c_1, c_2 \in \mathbb{C}$$

are homogeneous solutions too. The lemma presented next says that these *often* constitute all solutions, but in some cases not! The proof of the lemma again exploits the method of *variations of constants*.

Lemma 1.4.1 (All complex solutions of a 2nd-order DE). *Let $p_0, p_1 \in \mathbb{R}$, and define $\lambda_{1,2} \in \mathbb{C}$ as in* (1.12)*. A twice differentiable function $y : \mathbb{R} \to \mathbb{C}$ satisfies*

$$\ddot{y}(t) + p_1 \dot{y}(t) + p_0 y(t) = 0 \qquad \forall\, t \in \mathbb{R}$$

iff there are $c_1, c_2 \in \mathbb{C}$ such that

$$y(t) = \begin{cases} c_1 \mathrm{e}^{\lambda_1 t} + c_2 \mathrm{e}^{\lambda_2 t} & \text{if } \lambda_1 \neq \lambda_2 \\ (c_1 + c_2 t)\mathrm{e}^{\lambda_1 t} & \text{if } \lambda_1 = \lambda_2 \end{cases}.$$

Proof. We write the candidate solution $y(t)$ as[3]

$$y(t) = c(t)\mathrm{e}^{\lambda_1 t} \quad \text{for some function } c(t).$$

Every $y(t)$ can be written in this form because $\mathrm{e}^{\lambda_1 t}$ is invertible for every t. So this form is without loss of generality. Also, $y(t)$ is twice differentiable iff so is $c(t)$.

Now it will be very convenient to think of the expression $\ddot{y}(t) + p_1 \dot{y}(t) + p_0 y(t)$ as the result of a "differential operator" acting on $y(t)$, that is:

$$\ddot{y}(t) + p_1 \dot{y}(t) + p_0 y(t) = (\tfrac{\mathrm{d}^2}{\mathrm{d}t^2} + p_1 \tfrac{\mathrm{d}}{\mathrm{d}t} + p_0) y(t),$$

and then to use that $(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + p_1 \frac{\mathrm{d}}{\mathrm{d}t} + p_0) = (\frac{\mathrm{d}}{\mathrm{d}t} - \lambda_2)(\frac{\mathrm{d}}{\mathrm{d}t} - \lambda_1)$. Exploiting this we see that $y(t) := y(t) = c(t)\mathrm{e}^{\lambda_1 t}$ satisfies the DE iff

$$\begin{aligned} & 0 = \ddot{y}(t) + p_1 \dot{y}(t) + p_0 y(t) \\ \iff & 0 = (\tfrac{\mathrm{d}^2}{\mathrm{d}t^2} + p_1 \tfrac{\mathrm{d}}{\mathrm{d}t} + p_0)(c(t)\mathrm{e}^{\lambda_1 t}) \\ \iff & 0 = (\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_2)(\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_1)\big(c(t)\mathrm{e}^{\lambda_1 t}\big) \\ \iff & 0 = (\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_2)(\dot{c}(t)\mathrm{e}^{\lambda_1 t}). \end{aligned}$$

[2] Later we see that it is very much related to the characteristic polynomial of square matrices / linear mappings.

[3] We could as well have taken $c(t)\mathrm{e}^{\lambda_2 t}$; the final result is the same.

That is, $v(t) := \dot{c}(t)e^{\lambda_1 t}$ is a solution of the *first-order* DE $\dot{v}(t) - \lambda_2 v(t) = 0$. According to Lemma 1.3.1 the general solution is $v(t) = de^{\lambda_2 t}, d \in \mathbb{C}$. So we can continue:

$$\begin{aligned}
&\iff \dot{c}(t)e^{\lambda_1 t} = de^{\lambda_2 t}, \quad d \in \mathbb{C} \\
&\iff \dot{c}(t) = de^{(\lambda_2-\lambda_1)t}, \quad d \in \mathbb{C} \\
&\iff c(t) \text{ is an antiderivative of } de^{(\lambda_2-\lambda_1)t} \text{ for some } d \in \mathbb{C}.
\end{aligned}$$

If $\lambda_1 = \lambda_2$ then it says that $\dot{c}(t) = d$ so the general solution $c(t)$ is $c_1 + dt$, and then the result follows. If $\lambda_1 \neq \lambda_2$ then the general antiderivative of $de^{(\lambda_2-\lambda_1)t}$ is $c(t) = c_1 + \frac{d}{\lambda_2-\lambda_1}e^{(\lambda_2-\lambda_1)t}$. Then $y(t) := c(t)e^{\lambda_1 t}$ equals $c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$ for $c_2 = d/(\lambda_2 - \lambda_1)$. Clearly d ranges over all scalars iff c_2 ranges over all scalars. ∎

Example 1.4.2 (Polynomial solution). The characteric polynomial of the DE $\ddot{y}(t) = 0$ is $\lambda^2 = 0$. According to Lemma 1.4.1 the general solution is $y(t) = (c_1 + c_2 t)e^{0t} = c_1 + c_2 t$. Indeed a 2nd derivative is zero iff the function is a polynomial of degree 1 or less. □

Example 1.4.3 (Complex characteristic roots). Suppose

$$\ddot{y}(t) + 4y(t) = 0.$$

The characteristic zeros in this case are complex

$$\lambda_{1,2} = \pm i2.$$

The complete solution set $\mathbb{Y}$ of the homogeneous equation therefore is the subspace

$$\mathbb{Y} = \text{span}\{e^{i2t}, e^{-i2t}\}. \tag{1.14}$$

These solutions appear to be complex. Normally we are only interested in *real* solutions $y(t)$. Linear Algebra tells us that $\text{span}\{e^{i2t}, e^{-i2t}\}$ equals

$$\mathbb{Y} = \text{span}\{\cos(2t), \sin(2t)\},$$

see Exercise 1.8. That is, $y(t)$ is a solution iff it has the form

$$y(t) = c_1 \cos(2t) + c_2 \sin(2t), \qquad c_1, c_2 \in \mathbb{C}.$$

In this form it is not hard to see that $y(t)$ is a *real*-valued function iff both c_1 and c_2 are *real*. Indeed if $y(t)$ is real then necessarily $c_1 = y(0) \in \mathbb{R}$ and $c_2 = y(\pi/4) \in \mathbb{R}$, and conversely if $c_1, c_2 \in \mathbb{R}$ then $y(t)$ clearly is real-valued for all t. So all *real* solutions are

$$y(t) = c_1 \cos(2t) + c_2 \sin(2t), \qquad c_1, c_2 \in \mathbb{R}.$$

□

The procedure explained in the example readily generalizes: if λ_1 is a complex root (as in "not real") of the characteristic equation then its complex conjugate $\lambda_2 = \overline{\lambda}_1$ is a root as well, so they form a pair,

$$\lambda_{1,2} = \mu \pm i\omega, \qquad \omega \neq 0.$$

Then the complex functions of the form

$$\hat{c}_1 e^{(\mu+i\omega)t} + \hat{c}_2 e^{(\mu-i\omega)t}, \qquad \hat{c}_1, \hat{c}_2 \in \mathbb{C}$$

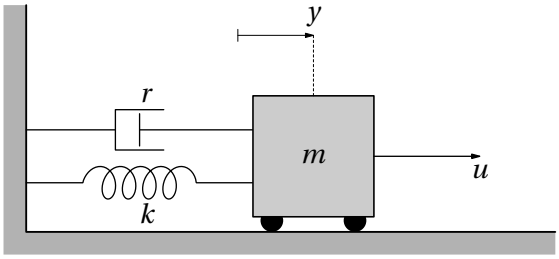

FIGURE 1.6: A car connected to a wall. (See Example 1.4.4.)

form the same subspace as

$$e^{\mu t}(c_1\cos(\omega t)+c_2\sin(\omega t)), \qquad c_1,c_2\in\mathbb{C}.$$

But the latter has the advantage that the function is real-valued iff both c_1 and c_2 are real numbers. Also notice that the *real* part of the root $\lambda_{1,2}=\mu\pm i\omega$ enters the *exponential* function while the *imaginary* part enters the two *sinusoids*.

Example 1.4.4 (Car connected to a wall). Consider again Example 1.1.5. The horizontal position $y(t)$ satisfies the DE

$$m\ddot{y}(t)+r\dot{y}(t)+ky(t)=u(t). \tag{1.15}$$

The characteristic equation is

$$m\lambda^2+r\lambda+k=0.$$

Suppose for now that $r=0$ (no damping). The characteristic equation then is

$$m\lambda^2+k=0$$

and its roots are

$$\lambda_1=i\sqrt{k/m}, \qquad \lambda_2=-i\sqrt{k/m}.$$

Therefore the solution set of the homogeneous equation is the subspace spanned by $e^{i\sqrt{k/m}t}$ and $e^{-i\sqrt{k/m}t}$ and the real solutions within this subspace are the functions of the form

$$y(t)=c_1\cos(\sqrt{k/m}\,t)+c_2\sin(\sqrt{k/m}\,t), \qquad c_1,c_2\in\mathbb{R}.$$

These are the harmonic functions with period $T=2\pi\sqrt{m/k}$. We conclude that the mass keeps on oscillation around zero in this case. This can be atttributed to the lack of damping ($r=0$).

Next suppose there is damping: we take $m=1$, $k=9.25$ and $r=1>0$. The characteristic equation then is

$$\lambda^2+\lambda+9.25=0$$

and its roots are

$$\lambda_{1,2}=\frac{-1\pm\sqrt{-36}}{2}=-0.5\pm i3.$$

The complete complex solution set therefore is $\operatorname{span}\{e^{(-0.5+i3)t},e^{(-0.5-i3)t}\}$ and the real solutions thus are

$$y(t)=e^{-0.5t}\big(c_1\cos(3t)+c_2\sin(3t)\big), \qquad c_1,c_2\in\mathbb{R}.$$

It is a damped oscillation. Every possible homogeneous solution now converges to zero as $t\to\infty$. The mass eventually comes to halt. □

1.5 Homogeneous DE

Having solved the homogeneous DE of orders $n = 1$ and $n = 2$ now gives us enough ideas to solve the homogeneous DE of arbitrary order n,

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = 0.$$

Again we contemplate a solution of exponential form

$$y(t) = \mathrm{e}^{\lambda t}.$$

Plugging this into the DE gives

$$\begin{aligned}
& 0 = y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) \\
\iff\ & 0 = \lambda^n \mathrm{e}^{\lambda t} + p_{n-1}\lambda^{n-1}\mathrm{e}^{\lambda t} + \cdots + p_1 \lambda \mathrm{e}^{\lambda t} + p_0 \mathrm{e}^{\lambda t} \\
\iff\ & 0 = \big(\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0\big)\mathrm{e}^{\lambda t} \\
\iff\ & 0 = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0.
\end{aligned}$$

The final equality does not contain t! It is a polynomial equation in λ. As before we call this the characteristic equation of the DE, and the polynomial

$$\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0 \qquad (\lambda \in \mathbb{C}) \tag{1.16}$$

is called the characteristic polynomial. The fundamental theorem of algebra says that the characteristic polynomial has precisely n zeros (aka *roots*[4]) over the complex numbers. That is, n numbers $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{C}$ exist such that

$$\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0 = \prod_{i=1}^{n}(\lambda - \lambda_i).$$

The above shows that each zero λ_i gives rise to a solution of the homogeneous DE,

$$y(t) = \mathrm{e}^{\lambda_i t}.$$

and then by the subspace property all linear combinations of $\mathrm{e}^{\lambda_1 t}, \ldots, \mathrm{e}^{\lambda_n t}$ are homogeneous solutions.

Example 1.5.1 (Third order system). Consider the 3rd-order DE

$$y^{(3)}(t) - 3y^{(2)}(t) + 2y^{(1)}(t) = 0.$$

Its characteristic polynomial is

$$\lambda^3 - 3\lambda^2 + 2\lambda = \lambda(\lambda - 1)(\lambda - 2).$$

Its three zeros are

$$\lambda_1 = 0, \quad \lambda_2 = 1, \quad \lambda_3 = 2.$$

Now the three functions $\mathrm{e}^{0t} = 1$ and e^t and e^{2t} are solutions of the homogeneous equation and so, by the subspace property, every linear combination

$$c_1 + c_2\mathrm{e}^t + c_3\mathrm{e}^{2t}$$

is a solution of the DE. Soon we will see that these are *all* solutions of the DE. □

[4] Nowadays "zeros" and "roots" seem to be synonymous, but there is also a school that says that *roots* are solutions of *equations*, while *zeros* are properties of *functions*: the zeros x of a function f are solutions of $f(x) = 0$.

Example 1.5.2 (Characteristic root of higher multiplicity). The characteristic equation of

$$y^{(3)}(t) = 0$$

is $\lambda^3 = 0$. This has one zero, $\lambda_1 = 0$, of multiplicity 3. Based on the preceding discussion we have the real solution

$$y(t) = ce^{0t} = c, \qquad c \in \mathbb{R}.$$

Indeed every constant $y(t) = c$ satisfies the DE, however there are other solutions such as $y(t) = t$ and $y(t) = t^2$. In fact it is not hard to verify that $y^{(3)} = 0$ iff $y(t)$ is a polynomial of degree 2 or less. □

The last example shows that not all solutions are exponential, and so the method presented so far does not in every case generate all solutions. The general solution can be found though, and this is the central result:

Theorem 1.5.3 (General solution of homogeneous equation). *Factorize the characteristic polynomial* (1.16) *as*

$$(\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_k)^{m_k},$$

with all λ_i distinct ($\lambda_i \neq \lambda_j \ \forall i \neq j$). Then $y(t)$ solves the homogeneous equation (1.6) *iff it has the form*

$$y(t) = \sum_{i=1}^{k} \left(c_{i,0} + c_{i,1}\, t + \cdots + c_{i,m_i-1}\, t^{m_i-1}\right) e^{\lambda_i t}$$

for some $c_{i,j} \in \mathbb{C}$.

Proof. See Appendix A.6. The proof is a bit tricky. ■

The degrees of freedom per exponential function $e^{\lambda_i t}$ equals the multiplicity m_i of the root λ_i. In this theorem we allowed for complex $c_{i,j}$ and the subspace of solutions $y(t)$ consists of all complex-valued solutions. In the examples presented next we explain how to switch back to all real-valued solutions.

Example 1.5.4. The characteristic equation of

$$y^{(n)}(t) = 0$$

is

$$\lambda^n = 0.$$

Its root is $\lambda_1 = 0$ with multiplicity $m_1 = n$. According to Thm. 1.5.3 the general solution is

$$y(t) = \left(c_{1,0} + c_{1,1}\, t + \cdots + c_{1,n-1}\, t^{n-1}\right) e^{0t} = c_{1,0} + c_{1,1}\, t + \cdots + c_{1,n-1}\, t^{n-1}.$$

That is indeed correct: the nth derivative of a function $y(t)$ is zero iff $y(t)$ is a polynomial of degree $n-1$ or less. □

Example 1.5.5. The characteristic equation of

$$y^{(3)}(t) - 4y^{(2)}(t) + 5y^{(1)}(t) - 2y(t) = 0$$

is

$$\lambda^3 - 4\lambda^2 + 5\lambda - 2 = 0.$$

It can be verified that

$$\lambda^3 - 4\lambda^2 + 5\lambda - 2 = (\lambda - 1)^2(\lambda - 2),$$

and so general *real* solution $y(t)$ is

$$(c_1 + c_2 t)\mathrm{e}^t + c_3\mathrm{e}^{2t}, \qquad c_1, c_2, c_3 \in \mathbb{R}.$$

□

1.6 Inhomogeneous equation — particular solution

Now we turn to inhomogeneous DE's,

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_0 y(t) = u(t), \tag{1.17}$$

so with $u(t)$ possibly not the zero function. In the homogeneous case we found that there are infinitely many solutions, hence we can expect infinitely many solutions here as well. However, the following lemma clarifies that we need only *one* particular solution $y(t)$ in order to generate them *all*:

Lemma 1.6.1 (Particular + homogeneous = all). *Let $u : \mathbb{R} \to \mathbb{C}$. Suppose $y_{part} : \mathbb{R} \to \mathbb{C}$ is a particular solution of* (1.17)*. Then $y(t)$ satisfies* (1.17) *iff*

$$y(t) = y_{part}(t) + y_{hom}(t)$$

for some solution $y_{hom}(t)$ of the corresponding homogeneous equation.

Proof. Let y_{part} be a solution of the inhomogeneous equation (1.17). Since $y_{\text{part}}^{(n)} + p_{n-1}y_{\text{part}}^{(n-1)} + \cdots + p_0 y_{\text{part}} = u$ the inhomogeneous equation (1.17) is equivalent to

$$y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_0 y = y_{\text{part}}^{(n)} + p_{n-1}y_{\text{part}}^{(n-1)} + \cdots + p_0 y_0.$$

This, in turn, is equivalent to

$$[y^{(n)} - y_{\text{part}}^{(n)}] + p_{n-1}[y^{(n-1)} - y_{\text{part}}^{(n-1)}] + \cdots + p_0[y - y_{\text{part}}] = 0.$$

That is, y satisfies the DE (1.17) iff $y - y_{\text{part}}$ is a solution of the homogeneous equation. ■

Thus the set of all solutions y is of the form $y_{\text{part}} + \mathbb{Y}$ with $\mathbb{Y}$ denoting the subspace of all homogeneous solutions. In Linear Algebra sets of the form $y_{\text{part}} + \mathbb{Y}$ are sometimes called *affine subspaces*. To determine a particular solution the method of variations of constants can again be used for first-order DE's (see Exercise 1.9), but for some classes of functions $u(t)$ we can also make an educated guess. This works for exponential, harmonic and polynomial $u(t)$. Table 1.1 summarizes this method. If the suggested form of the particular solution $y(t)$ happens to be a homogeneous solution then try as particular solution $ty(t)$, and if that does not work then try $t^2y(t)$ et cetera. Eventually it will work!

TABLE 1.1: Possible particular solutions $y(t)$ depending on given $u(t)$. If the suggested particular solution $y(t)$ is a homogeneous solution then try as particular solution $ty(t)$ and if that does not work then try $t^2y(t)$ et cetera.

right-hand side $u(t)$	possible particular solution $y_{\text{part}}(t)$
polynomial of degree k	polynomial of degree k
$de^{\alpha t}$	$ce^{\alpha t}$
$d_1\cos(\beta t)+d_2\sin(\beta t)$	$a\cos(\beta t)+b\sin(\beta t)$
$e^{\alpha t}(d_1\cos(\beta t)+d_2\sin(\beta t)$	$e^{\alpha t}(a\cos(\beta t)+b\sin(\beta t))$

Example 1.6.2 (Newton's law of cooling — constant medium temperature). Let $y(t)$ be the temperature of some object, and suppose the medium it sits in has constant temperature m (not necessarily zero). Newton's law of cooling says that the rate of change of temperature is proportional to the temperature *difference* between object and medium, $y(t)-m$, that is,

$$\dot{y}(t) = -k(y(t)-m), \qquad k>0, \tag{1.18}$$

or,

$$\dot{y}(t) + ky(t) = km.$$

As the right-hand side is constant (i.e. a polynomial of degree zero) Table 1.1 suggests to look for a particular solution that is constant as well,

$$y_{\text{part}}(t) = a.$$

Plugging this into the DE (1.18) gives the trivial equation $ka = km$. Clearly $a = m$ is the answer, and then the general solution follows by adding all homogeneous solutions,

$$y(t) = m + ce^{-kt}, \qquad c\in\mathbb{R}.$$

We see that temperature of the object converges exponentially fast to the temperature m of the medium. □

Example 1.6.3 (Newton's law of cooling — increasing medium temperature). Consider again Newton's law of cooling but now for an object sitting in a medium that we keep on heating up. More explicitly, suppose the temperature of the medium equals $1+t$. So we have

$$\dot{y}(t) = -k(y(t)-(1+t)), \qquad k>0, \tag{1.19}$$

that is,

$$\dot{y}(t) + ky(t) = k(1+t).$$

Table 1.1 suggests to look for a particular of the form

$$y_{\text{part}}(t) = a+bt.$$

Plugging this into the DE (1.19) gives

$$b + k(a+bt) = k(1+t).$$

The solution is $b = 1$ and $a = 1 - 1/k$, and so the general solution is the particular solution $1 - 1/k + t$ with the homogeneous solutions added to it,

$$y(t) = 1 - 1/k + t + ce^{-kt}, \qquad c \in \mathbb{R}.$$

As $t \to \infty$ the exponential term converges to zero and so the temperature $y(t)$ converges to $1 - 1/k + t$. This equals the temperature of the medium minus $1/k$. The object, so to say, tries to keep up with the increase of the temperature of the medium but it lags behind a bit, see Fig. 1.7. □

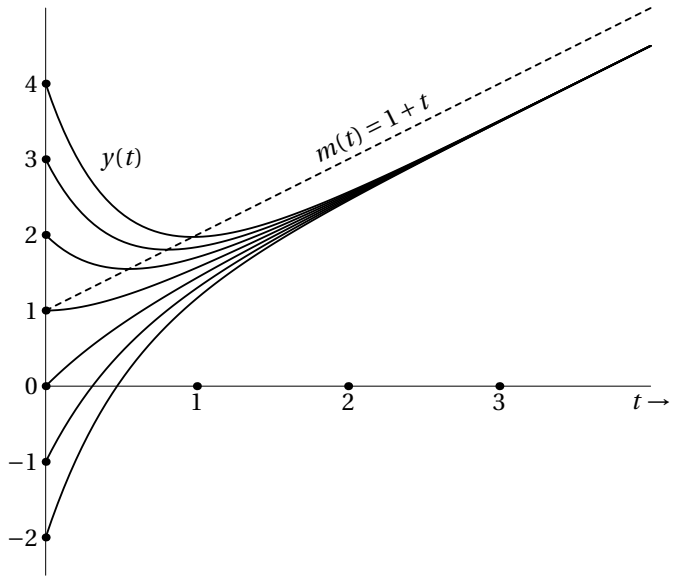

FIGURE 1.7: Several temperatures of $y(t)$ of an object sitting in a medium that heats up. The initial conditions of the seven temperatures are $y(0) = -2, \ldots, y(0) = 4$. The temperature of the medium is $m(t) = 1 + t$. We assumed that $k = 2$ (see Example 1.6.3 for details).

Example 1.6.4 (2nd-order inhomogeneous DE). Suppose

$$\ddot{y}(t) - 5\dot{y}(t) + 6y(t) = 3t + 2. \tag{1.20}$$

The characteristic equation is $\lambda^2 - 5\lambda + 6 = 0$ and its roots are $\lambda_1 = 2$ and $\lambda_2 = 3$. The homogeneous solutions hence are spanned by e^{2t} and e^{3t}. The right-hand side of (1.20) is a degree-1 polynomial, so as particular solution we try a degree-1 polynomial as well

$$y(t) = at + b.$$

Substitution of this in (1.20) gives

$$-5a + 6at + 6b = 3t + 2.$$

As this to hold for every t we must have that $a = \frac{1}{2}$ and $b = \frac{3}{4}$. A particular solution then is

$$y(t) = \frac{1}{2}t + \frac{3}{4}$$

and the general real solution is

$$y(t) = \frac{1}{2}t + \frac{3}{4} + c_1 e^{2t} + c_2 e^{3t}, \qquad c_1, c_2 \in \mathbb{R}.$$

□

Example 1.6.5 (Harmonic functions). Consider

$$\ddot{y}(t) - 4\dot{y}(t) + 4y(t) = 5\cos(t). \tag{1.21}$$

The characteristic equation is $\lambda^2 - 4\lambda + 4 = 0$ and its roots are $\lambda = 2$ (double). The homogenous solutions are the linear combinations of e^{2t} and te^{2t}. As particular solution try a function of the form

$$y(t) = a\cos(t) + b\sin(t).$$

Substitution into (1.21) gives

$$-a\cos(t) - b\sin(t) + 4a\sin(t) - 4b\cos(t) + 4a\cos(t) + 4b\sin(t) = 5\cos(t).$$

That is,

$$(3a - 4b)\cos(t) + (3b + 4a)\sin(t) = 5\cos(t).$$

Since sin and cos are independent functions the above holds iff

$$3a - 4b = 5 \quad \text{and} \quad 3b + 4a = 0.$$

These are two linear equations in the two unknowns a, b. The solution is $a = \frac{3}{5}$ and $b = -\frac{4}{5}$, and so we found the particular solution

$$y_{\text{part}}(t) = \frac{3}{5}\cos(t) - \frac{4}{5}\sin(t).$$

The general solution finally follows as

$$y(t) = c_1 e^{2t} + c_2 te^{2t} + \frac{3}{5}\cos(t) - \frac{4}{5}\sin(t).$$

□

Example 1.6.6 (Exponential functions & initial conditions). We determine the solution of

$$\ddot{y}(t) - 5\dot{y}(t) + 6\dot{y}(t) = e^{2t}, \tag{1.22}$$

that satisfies the initial conditions

$$y(0) = 1, \qquad \dot{y}(0) = 2. \tag{1.23}$$

In Example 1.6.4 we determined the general homogenous solution $y(t) = c_1 e^{2t} + c_2 e^{3t}$. As for a particular solution, based on the right-hand side one might guess a particular solution of the form $y(t) = ae^{2t}$. *However this form is already part of the homogeneous solution. In cases like these multiply the initial candidate with t*: try instead a particular solution of the form

$$y(t) = ate^{2t}$$

for some yet to be determined constant a. Substitution into (1.22) gives

$$ae^{2t}\underbrace{(4 + 4t - 5 - 10t + 6t)}_{-1} = e^{2t}.$$

This holds iff $a = -1$ and the particular hence is $y(t) = -te^{2t}$. The general solution of the inhomogeneous equation thus is

$$y(t) = c_1 e^{2t} + c_2 e^{3t} - te^{2t}.$$

What remains are the initial conditions (1.23). Using the derivative of the above $y(t)$ we find that

$$\dot{y}(t) = 2c_1 e^{2t} + 3c_2 e^{3t} - 2te^{2t} - e^{2t}.$$

This way the initial conditions (1.23) translate to conditions on c_1, c_2:

$$1 = y(0) = c_1 + c_2 - 0$$
$$2 = \dot{y}(0) = 2c_1 + 3c_2 - 0 - 1.$$

This determines the constants as $c_1 = 0, c_2 = 1$, and the solution of (1.23) subject to initial conditions (1.23) therefore is unique,

$$y(t) = e^{3t} - te^{2t}.$$

□

1.7 Asymptotic Stability & Time Constants

In many applications we are interested in how fast solutions converge to some "desired" solution. Examples are abundant. For example, how fast does the speed of a car converge to 100 km/hour. Or, how fast does the room temperature converge to 20 degrees Celcius. Also in many biological processes the result (or is it "task") is to steer certain variables in the direction of something desired, such as body temperature and glucose level.

Before analysing how fast it converges, we introduce what is called "asymptotic stability" of DE's. Asymptotic stability roughly speaking means the property of solutions returning to some given "desired" solution. It comes in various forms. For *nonlinear* DE's (not covered in this course) this is a tricky concept but for our type of linear DE's it is simple (both the definition and the lemma that characterizes it):

Definition 1.7.1 (Asymptotic stability). A DE

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = u(t), \qquad (p_i \in \mathbb{R}) \tag{1.24}$$

is asymptotically stable if $\lim_{t\to\infty} y(t) = 0$ for all homogeneous solutions of the DE. □

Let us immediately solve this problem:

Lemma 1.7.2 (Asymptotic stability). *A DE* (1.24) *is asymptotically stable if and only if all its characteristic roots have negative real part:* $\mathrm{Re}(\lambda_i) < 0$.

Proof. To each characteristic root λ_i there corresponds a real homogeneous solution

$$y(t) = \mathrm{Re}(e^{\lambda_i t}) = e^{\mathrm{Re}(\lambda_i)t}(\cos(\mathrm{Im}(\lambda_i)t) + i\sin(\mathrm{Im}(\lambda_i)t).$$

If there is a λ_i with $\mathrm{Re}(\lambda_i) \geq 0$ then clearly the above $y(t)$ does not converge to zero as $t \to \infty$. If, on the other hand, all characteristic roots satisfy $\mathrm{Re}(\lambda_i) < 0$ then Theorem 1.5.3 shows that every possible homogeneous solution converges to zero. ■

Even though we defined asymptotic stability for the homogeneous part, it is also useful for the analysis of inhomogeneous DE's. We have seen some examples already. For instance, in Example 1.6.2 on cooling of an object we found that the temperature of the object $y(t)$ obeyed the DE

$$\dot{y}(t) + ky(t) = km, \qquad k > 0.$$

Clearly $\lambda_1 = -k < 0$ so this DE is asymptotically stable and hence all homogeneous solutions converge to zero. So all solutions $y(t)$ converges to any one *particular* solution, say $y_{\text{part}}(t) = m$.

Lemma 1.7.2 says that asymptotic stability can be determined from the zeros of the characteristic polynomial. For this reason one normally defines:

Definition 1.7.3 (Asymptotically stable polynomial). A polynomial is said to be asymptotically stable if all its zeros have negative real part. □

Zeros with negative real part are referred to as stable zeros, and zeros with nonnegative real part are unstable zeros (this includes zeros on the imaginary axis).

The results might give the impression that we need to compute the zeros of the characteristic polynomial in order to figure out stability. However we only need to know if the real part of the zeros are all less than zero or not. That is an easier problem:

Lemma 1.7.4 (Asymptotic Stability of DE's of order ≤ 3). *Let $p_0, p_1, p_2 \in \mathbb{R}$. Then*

$$\begin{aligned} \lambda + p_0 \text{ is asymptotically stable} &\iff p_0 > 0 \\ \lambda^2 + p_1\lambda + p_0 \text{ is asymptotically stable} &\iff p_1, p_0 > 0 \\ \lambda^3 + p_2\lambda^2 + p_1\lambda + p_0 \text{ is asymptotically stable} &\iff p_2, p_1, p_0 > 0 \text{ and } p_2 p_1 > p_0. \end{aligned}$$

Proof. For the polynomial of degree one, $\lambda + p_0$, this is trival for then $\lambda_1 = -p_0$ which is negative iff $p_0 > 0$. The other cases follow from the Routh-Hurwitz test as explained in § 1.8, see Exercise 1.13. ■

Example 1.7.5 (Mechanical system). In Example 1.1.5 we found that the position $y(t)$ of the car satifies the 2nd-order DE

$$m\ddot{y}(t) + r\dot{y}(t) + ky(t) = u(t).$$

Here m is the mass of the car, r is a damping coefficient, and k is a spring constant. Clearly the mass is positive so we can divide the DE by m, giving the equivalent

$$\ddot{y}(t) + (r/m)\dot{y}(t) + (k/m)y(t) = (1/m)u(t).$$

According to Lemma 1.7.4 this DE is asymptotically stable iff both r/m and k/m are positive, i.e. iff both k and r are positive. The interpretation is that the spring is needed to drive the car in the direction of the origin, and damping is needed to reduce its energy to zero. □

Time constants & pole plots

Now a bit of engineering. If the DE is asymptotically stable we know that all homogeneous solutions converge to zero. But how fast do they converge? For asymptotically stable 1st-order DE's

$$\dot{y}(t) + p_0 y(t) = 0, \qquad p_0 > 0$$

this is measured by the time constant, τ, defined as

$$\tau = 1/p_0 > 0.$$

This way the general homogeneous solution can be written as[5]

$$y(t) = ce^{-t/\tau}.$$

[5] If t has dimension "time" as it usually does on our examples, then τ also has dimension time.

Figure 1.8 explains graphically what the time constant τ measures: it is the time needed for $|y(t)|$ to reduce by a factor $\mathrm{e} \approx 2.7183$. It is good to realize that this holds for *every* homogenous solution $y(t) = c\mathrm{e}^{-t/\tau}$ and at every moment in time, that is, this reduction by a factor e does not depend on the initial condition nor on the time we call "zero". The time constant is related to the "half life" which is the time need to reduce by a factor 2 (see Exercise 1.14).

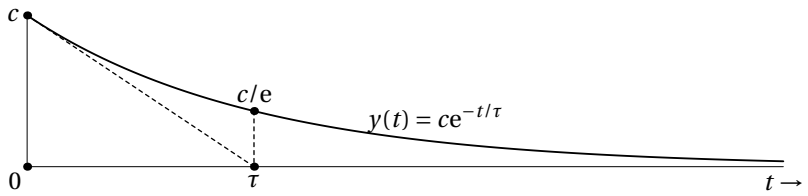

FIGURE 1.8: It takes $t = \tau$ for a function $y(t) = c\mathrm{e}^{-t/\tau}$ to reduce by a factor $\mathrm{e} = 2.718\ldots$.

Another, but equivalent, definition of the time constant for 1st-order DE's is that it is the inverse of the distance between $\lambda_1 = -p_0$ and zero, see Fig. 1.9(a). Clearly this, again, gives $\tau = 1/p_0$. The advantage of this definition is that it more easily extends to higher order DE's. If the order is, say, $n = 3$ and the DE is asymptotically stable then its "pole plot" might look like the one in Fig. 1.9(b). Now with each of the λ_i's there corresponds an exponential homogeneous solution $\mathrm{e}^{\lambda_i t}$ (if real) or $\mathrm{e}^{-\mathrm{Re}(\lambda_i)t}\cos(\mathrm{Im}(\lambda_i)t + \phi)$ (if complex) and their "time constants" $\tau_i := -1/\mathrm{Re}(\lambda_i)$ determine how fast they converge to zero. The homogeneous solution is a superposition of these exponential solutions and then typically it is the *slowest* of these exponential solutions that determines the speed of convergence. So in the case of the three "time constants" of Fig. 1.9(b) the time constant that we notice in the homogeneous solution is $\tau_2 = \tau_3$ as these are the largest (slowest) time constants. In short: it is the *smallest* distance between the characteristic roots and the imaginary axis that determines the speed of convergence.

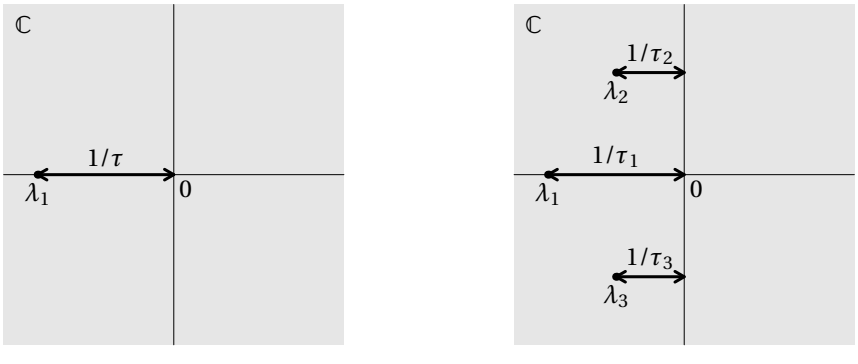

FIGURE 1.9: Two *pole plots*. In (a) there is one stable characteristic root and one time constant τ. In (b) there are three stable characteristic roots $\lambda_1, \lambda_2, \lambda_2$ each with its own "time constant" τ_1, τ_2, τ_3. The speed of convergence of the corresponding homogeneous solution usually is dominated by the *slowest* mode, i.e. by the *largest* time constant, i.e. by the characteristic root that is closest to the imaginary axis: λ_2 (or λ_3).

Example 1.7.6. In Example 1.4.4 we analyzed a car connected to a wall via a spring and damper. For certain parameter values we got the stable characteristic roots

$$\lambda_{1,2} = -1/2 \pm \mathrm{i}3.$$

Both characteristic roots are equally close to the imaginary axis (distance of 1/2), and the corresponding time constant hence is $\tau = 1/\mathrm{Re}\,(-\lambda_{1,2}) = 2$. A typical homogeneous solution $y(t)$ is shown in Fig. 1.10 together with an exponential function whose time constant equals this $\tau = 2$. The time constant $\tau = 2$ gives us an indication how fast $y(t)$ converges to zero, even though $y(t)$ oscillates. □

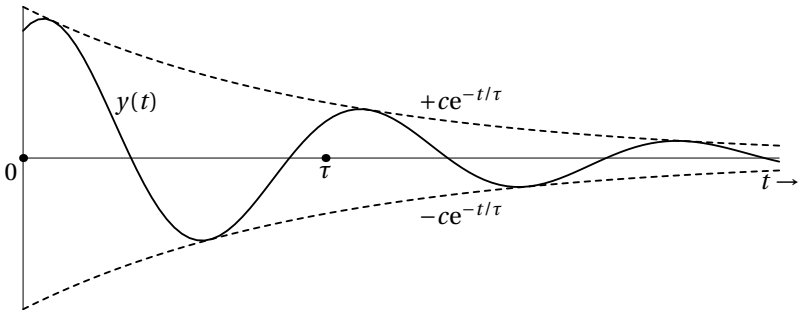

FIGURE 1.10: A possible solution $y(t)$ of $\ddot{y}(t) + \dot{y}(t) + 9.25y(t) = 0$ (solid) and $ce^{-t/\tau}$ (dashed) for $\tau = 2$. See Example 1.7.6.

1.8 Routh-Hurwitz test

We have seen that the asymptotic stability of a DE can be decided by the sign of the real parts of the zeros of the characteristic polynomial,

$$\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0.$$

Now, for general polynomials of degree greater than four, there is no finite expression for the zeros. However, to test stability we do not need to know the exact values of the zeros of the characteristic polynomial. We only need to figure out if all its zeros have negative real part, and for that finite tests do exist! This is an awesome result, developed independently by Edward Routh (1831–1907) and Adolf Hurwitz (1859–1919). Here is the famous result:

Theorem 1.8.1 (Routh–Hurwitz test). *A polynomial*

$$p_n\lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0, \qquad p_n \neq 0$$

is asymptotically stable if and only if all $n+1$ numbers in the first column of the Routh table exist and have the same sign. The Routh table is of the form

$$\begin{array}{cccccc} p_n & p_{n-2} & p_{n-4} & p_{n-6} & \cdots \\ p_{n-1} & p_{n-3} & p_{n-5} & p_{n-7} & \cdots \\ b_0 & b_2 & b_4 & \cdots & \\ b_1 & b_3 & b_5 & \cdots & \\ \cdots & & & & \end{array}$$

It has $n+1$ rows. The first two rows follow from the polynomial. The third row is constructed from the two rows above it (rows 1 and 2) by

$$\begin{bmatrix} b_0 & b_2 & \cdots \end{bmatrix} = \begin{bmatrix} p_{n-2} & p_{n-4} & p_{n-6} & \cdots \end{bmatrix} - \frac{p_n}{p_{n-1}} \begin{bmatrix} p_{n-3} & p_{n-5} & p_{n-7} & \cdots \end{bmatrix}.$$

Every following row k is constructed in the same manner from the rows directly above it (rows $k-2$ and $k-1$.) If a division by zero occurs in this construction then the polynomial is not asymptotically stable. □

A proof of this theorem is given in Appendix A.2.

Example 1.8.2 (Routh–Hurwitz). Consider the degree 3 polynomial

$$\lambda^3 + \lambda^2 + \lambda + c$$

depending on $c \in \mathbb{R}$. The Routh table then consists of $n+1 = 4$ rows:

$$\begin{array}{ll} 1 & 1 \\ 1 & c \\ 1-c & \\ c & \end{array}$$

All four numbers in the first column have the same sign exactly when $0 < c < 1$, hence the polynomial is asymptotically stable if and only $0 < c < 1$. (If $c = 0$ it has a zero at $\lambda = 0$, if $c = 1$ it has two imaginary zeros $\lambda_{1,2} = \pm\mathrm{i}$.) □

Example 1.8.3 (Degree five). Consider the degree-5 polynomial

$$2\lambda^5 + 1\lambda^4 + 4\lambda^3 + 3\lambda^2 + 6\lambda + 5.$$

The Routh table is now

$$\begin{array}{lll} 2 & 4 & 6 \\ 1 & 3 & 5 \\ -2 & -4 & \\ 1 & 5 & \\ 6 & & \\ 5 & & \end{array}$$

The six elements of the first column do not all have the same sign, so the polynomial is not asymptotically stable.

(An extension of the Routh-Hurwitz test says that there are as many unstable zeros as there are sign changes in the ordered sequence of the first column $(2, 1, -2, 1, 6, 5)$. So here that is 2. The zeros, obtained numerically, are $0.65763 \pm 1.21259\mathrm{i}$, $-0.51573 \pm 1.18754\mathrm{i}$ and -0.78379, and indeed two of them are unstable.) □

Example 1.8.4 (Division by zero). The Routh table of $\lambda^2 + 4$ should have 3 rows, but its construction breaks down because of division by zero,

$$\begin{array}{ll} 1 & 4 \\ 0 & \\ ? & \end{array}$$

The polynomial is therefore not asymptotically stable. □

With the Routh–Hurwitz test, we can establish the three claims of Lemma 1.7.4, see Exercise 1.13.

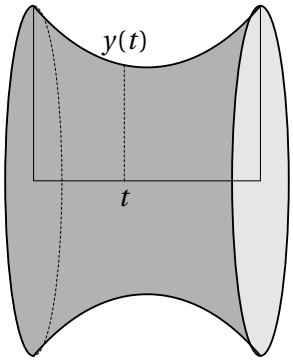

FIGURE 1.11: The shape of a soap film can be described by a DE. (See Exercise 1.1.)

1.9 Exercises

1.1 *The shape of a soap film.* Consider a soap film in between two given rings of equal radius, see Fig. 1.11. With some fancy modelling technique it can be shown that the radius $y(t)$ of the surface of revolution as a function of *position* t — see Fig. 1.11 — is a solution of the DE

$$y^2(t) = a\big(1 + \dot{y}^2(t)\big)$$

for some $a \geq 0$. (Here t is a *position* and not a *time*, but this is not relevant.)

(a) What is the order of the DE?

(b) Is the DE linear or nonlinear?

The notion of "constant coefficient DE" is a bit tricky in this case. Better is to use "t-invariance". We say a DE (in t) is "t-invariant" if $y(t)$ is a solution of the DE iff $y(t - t_0)$ is a solution of the DE for every $t_0 \in \mathbb{R}$.

(c) Is the above DE t-invariant?

(d) Is the DE (1.6) t-invariant?

(e) Suppose one or more p_i are not constant as a function of t. Argue that DE (1.6) is then not t-invariant.

1.2 Determine the general solution of the homogeneous equation

(a) $y^{(1)}(t) + 3y(t) = 0$.

(b) $2y^{(1)}(t) + 3y(t) = 0$.

(c) $y^{(2)}(t) + 2y^{(1)}(t) + y(t) = 0$.

(d) $y^{(2)}(t) + 4y(t) = 0$.

(e) $y^{(2)}(t) - 4y(t) = 0$.

(f) $y^{(4)}(t) = 0$.

1.3 Which 2nd-order DE $\ddot{y}(t) + p_1\dot{y}(t) + p_0 y(t) = 0$ has general solution

(a) $y(t) = ce^{-t} + de^{-2t}$.

(b) $y(t) = ce^{-2t} + dte^{-2t}$.

(c) $y(t) = ce^{-t}\cos(3t) + de^{-t}\sin(3t)$.

That is: for each case find the appropriate p_0, p_1.

1.4 Figure 1.12(left) shows 6 solutions $y(t)$ of different 2nd-order DEs

$$\ddot{y}(t) + b\dot{y}(t) + cy(t) = 0.$$

Figure 1.12(right) shows six configurations of the two characteristic roots $\lambda_1, \lambda_2 \in \mathbb{C}$ of $\lambda^2 + b\lambda + c = 0$. (For instance part (c) means $\lambda = -1$ and $\lambda = -2$.) Determine for each of the six solutions $y(t)$ a possible characteristic root configuration.

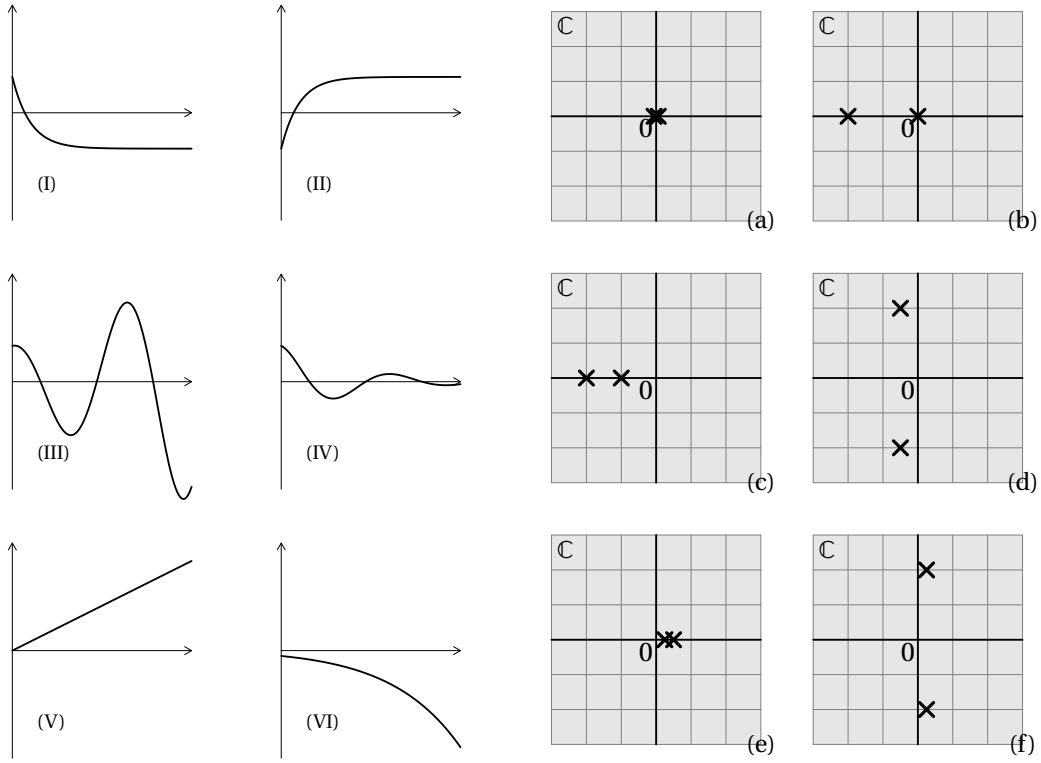

FIGURE 1.12: Six solutions $y(t)$ (left) and six root configurations λ_1, λ_2 in $\mathbb{C}$ (right).

1.5 Determine the general solution of the homogeneous equation of

(a) $y^{(2)}(t) + 2y^{(1)}(t) + 2y(t) = 0.$

(b) $y^{(2)}(t) - 4y^{(1)}(t) + 4y(t) = 0.$

(c) $y^{(2)}(t) + 7y^{(1)}(t) + 12y(t) = 0.$

(d) $y^{(4)}(t) = 0.$

(e) $y^{(1)}(t) + \beta y(t) = 0.$

(f) $y^{(3)}(t) + 2y^{(2)}(t) + 2y^{(1)}(t) = 0$

1.6 Determine the general solution of

(a) $y^{(3)}(t) + y^{(2)}(t) = 0.$

(b) $y^{(3)}(t) + y^{(1)}(t) = 0.$

(c) $y^{(3)}(t) + 2y^{(2)}(t) - 80y^{(1)}(t) = 0.$

1.7 Determine the general solution of

(a) $y^{(1)}(t) + 3y(t) = 1$.

(b) $y^{(1)}(t) + 3y(t) = \mathrm{e}^{-t}$.

(c) $y^{(1)}(t) + 3y(t) = \mathrm{e}^{-t} + 5$.

(d) $y^{(2)}(t) + 4y(t) = t^2$.

(e) $y^{(2)}(t) - 4y(t) = \mathrm{e}^{2t}$.

(f) $y^{(4)}(t) = 5$.

1.8 *Span of harmonic functions.* Consider the complex vector space of functions $f : \mathbb{R} \to \mathbb{C}$.

(a) Show that both $\cos(2t)$ and $\sin(2t)$ are linear combinations of $\mathrm{e}^{\mathrm{i}2t}, \mathrm{e}^{-\mathrm{i}2t}$.

(b) Show that both $\mathrm{e}^{2\mathrm{i}t}$ and $\mathrm{e}^{-2\mathrm{i}t}$ are linear combinations of $\cos(2t), \sin(2t)$.

(c) Argue that $\mathrm{span}\{\mathrm{e}^{2\mathrm{i}t}, \mathrm{e}^{-2\mathrm{i}t}\} = \mathrm{span}\{\cos(2t), \sin(2t)\}$.

1.9 The point of this exercise is to demonstrate that the method of variation of constants can be used to completely solve certain inhomogeneous DE's. Consider

$$\dot{y}(t) + 3y(t) = \mathrm{e}^{-t}, \tag{1.25}$$

and write $y(t)$ as $y(t) = c(t)\mathrm{e}^{-3t}$.

(a) Express DE (1.25) as a DE in $c(t)$.

(b) Determine all solutions $c(t)$ of this DE in $c(t)$.

(c) Determine all solutions $y(t)$ of (1.25).

1.10 Which of the DEs of Exercise 1.5 are asymptotically stable?

1.11 Which of these DEs is asymptotically stable?

(a) $y^{(1)}(t) + 3y(t) = 0$.

(b) $2y^{(1)}(t) + 3y(t) = 0$.

(c) $y^{(2)}(t) + 2y^{(1)}(t) + y(t) = 0$.

(d) $y^{(2)}(t) + 4y(t) = 0$.

(e) $y^{(2)}(t) - 4y(t) = 0$.

(f) $y^{(4)}(t) = 0$.

1.12 Is $P(\lambda) = \lambda^5 + 2\lambda^4 + 3\lambda^2 + 4\lambda + 5$ asymptotically stable?

An extension of Thm. 1.8.1 is as follows: *if* the first column of the Routh table does not contain zeros, then the polynomial does not have imaginary zeros, and the number of sign changes in the first column of the table – as you go from the first entry down to the last entry – is equal to the number of unstable zeros. (A zero $\lambda \in \mathbb{C}$ is *unstable* if $\mathrm{Re}(\lambda) \geq 0$).

How many unstable zeros does $\lambda^5 + 2\lambda^4 + 3\lambda^2 + 4\lambda + 5$ have?

1.13 *Routh–Hurwitz test for degree 2 and 3 polynomials.*

(a) Prove that a $\lambda^2 + p_1\lambda + p_0$ is asymptotically stable iff $p_1, p_0 > 0$.

(b) Show that $\lambda^3 + p_2\lambda^2 + p_1\lambda + p_0$ is asymptotically stable iff $p_2, p_1, p_0 > 0$ and $p_2 p_1 > p_0$.

1.14 *Half life.* The half life of an exponential function $f(t)$ can be defined as the time κ needed to reduce by a factor two, that is $f(t+\kappa) = \frac{1}{2} f(t)$ for all t.

(a) Show that the half life is well defined and positive and follows uniquely from $f(t)$ if $f(t) = ce^{-p_0 t}$ for some $c \neq 0$ and $p_0 > 0$. (Well defined means that κ does not depend on t.)

(b) Express the half life κ in terms of the time constant τ assuming $f(t) = ce^{-p_0 t}$ for some $c \neq 0, p_0 > 0$.

Tougher Exercises

1.15 *Schur–Cohn–Jury criterion.* In this exercise, $p(s)$ is a real polynomial. We know that a system $p(\frac{\mathrm{d}}{\mathrm{d}t})y = 0$ is asymptotically stable if and only if all zeros of $p(\lambda) = 0$ have negative real part. This seems to imply that asymptotic stability can only be determined by computing the zeros. That is not the case. A simple test suffices:

(a) Show that the coefficients of an asymptotically stable polynomial $p(s)$ all have the same sign.

(b) Show that a nonconstant polynomial $p(\lambda)$ is asymptotically stable only when $|p(-1)/p(1)| < 1$. [*Hint:* Use part (a).]

(c) Prove that a nonconstant polynomial $p(\lambda)$ is asymptotically stable if and only if $|p(-1)/p(1)| < 1$ and

$$q(\lambda) := p(\lambda) - \frac{p(-1)}{p(1)} p(-\lambda)$$

is asymptotically stable.

[*Hint:* Define the family of polynomials $r_\eta(\lambda) := (1-\eta)p(\lambda) + \eta q(\lambda)$ and show that for all $\eta \in [0,1]$ they have the same degree and the same imaginary zeros if $|p(-1)/p(1)| < 1$.]

The good news is that $q(-1) = 0$, so the problem can be reduced to that of the polynomial $q(\lambda)/(\lambda+1)$ (of lower degree). Continuing this way, the problem is solved in n steps (with $n = \deg(p)$).

1.16 *Time-invariance.* We know already that the set $\mathbb{Y}$ of all solutions y of a homogeneous DE (1.6) forms a (linear) subspace (see Lemma 1.2.1). But it has another structural property: time invariance. A subset $\mathbb{V}$ of functions $f : \mathbb{R} \to \mathbb{C}$ is called *time invariant* if $f \in \mathbb{V}$ iff for every $t_0 \in \mathbb{R}$ also $g \in \mathbb{V}$ for $g(t) := f(t - t_0)$.

(a) Show that the set of solutions $\mathbb{Y}$ of (1.6) is time invariant.

(b) Show that the set of polynomials $\mathbb{P}_n$ of degree n are less is time invariant.

(c) Argue that a set $\mathbb{V}$ of differentiable functions is a time invariant subspace of dimension 1 iff $\mathbb{V} = \mathrm{span}\{e^{\lambda t}\}$ for some $\lambda \in \mathbb{C}$.

1.17 *Associative property of differential operators.* In the proof of Lemma 1.4.1 we used, without proof, the associative rule

$$(\tfrac{\mathrm{d}}{\mathrm{d}t} - a)\Big((\tfrac{\mathrm{d}}{\mathrm{d}t} - b)y(t)\Big) = \Big((\tfrac{\mathrm{d}}{\mathrm{d}t} - a)(\tfrac{\mathrm{d}}{\mathrm{d}t} - b)\Big)y(t).$$

That is that $(\frac{\mathrm{d}}{\mathrm{d}t} - a)(\dot{y}(t) - by(t))$ equals $(\frac{\mathrm{d}^2}{\mathrm{d}t^2} - (a+b)\frac{\mathrm{d}}{\mathrm{d}t} + ab)y(t)$.

(a) Verify the above associative property.

(b) Verify that $(\frac{\mathrm{d}}{\mathrm{d}t} - a)\big((\frac{\mathrm{d}}{\mathrm{d}t} - b)y(t)\big) = (\frac{\mathrm{d}}{\mathrm{d}t} - b)\big((\frac{\mathrm{d}}{\mathrm{d}t} - a)y(t)\big)$.

(c) Let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, and take $m \in \mathbb{N}, m \leq n$. Prove that

$$\big((\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_1)\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_m)\big)\Big(\big((\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_{m+1})\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t} - \lambda_n)\big)y(t)\Big)$$

does not depend on m. Hence we can simply write $(\frac{\mathrm{d}}{\mathrm{d}t} - \lambda_1)\cdots(\frac{\mathrm{d}}{\mathrm{d}t} - \lambda_n)y(t)$ without putting more brackets, and, also, re-ordering the λ's does not change the result.

(d) Prove that $(\frac{\mathrm{d}}{\mathrm{d}t} - \lambda)(\mathrm{e}^{st}c(t)) = \mathrm{e}^{st}((\frac{\mathrm{d}}{\mathrm{d}t} - (\lambda - s))c(t))$.

(e) Use the previous part to determine all solutions of $(\frac{\mathrm{d}}{\mathrm{d}t} - \lambda)^n y(t) = 0$.

Chapter 2

State Representations

2.1 Introduction

An nth order DE

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = u(t) \tag{2.1}$$

can also be expressed as a *first*-order DE, but then in *several* components. This is called a *system* of DE's. The procedure is simple: define the vector of functions $x(t)$ as

$$x(t) := \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{n-1}(t) \\ x_n(t) \end{bmatrix} := \begin{bmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(n-2)}(t) \\ y^{(n-1)}(t) \end{bmatrix}.$$

Its derivative $\dot{x}(t)$ satisfies

$$\dot{x}(t) = \begin{bmatrix} y^{(1)}(t) \\ y^{(2)}(t) \\ \vdots \\ y^{(n-1)}(t) \\ y^{(n)}(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \\ -p_{n-1}x_{n-1}(t) - \cdots - p_1 x_2(t) - p_0 x_1(t) + u(t) \end{bmatrix}.$$

In the last equality we used (2.1). In matrix form this is

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(t) \\ y(t) &= \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} x(t). \end{aligned} \tag{2.2}$$

This is a so-called *state representation* of the DE. It is important to understand that this representation is *equivalent* to DE (2.1). Indeed, if all we know is that x satisfies (2.2) then it follows from the final part that $y = x_1$, and then the first row of the matrix equation gives us that $\dot{y} = \dot{x}_1 = x_2$, and then the 2nd row that $\ddot{y} = \ddot{x}_1 = \dot{x}_2 = x_3$, et cetera: $x = (y, \dot{y}, \ddot{y}, \ldots, y^{(n-1)})$, and then the final row of the matrix equation is the DE (2.1).

We will soon see that state representations are extremely convenient for analysis and simulation[1], and also for *control* (the topic of the following three chapters). We introduced here state representations as an alternative for high-order DE's, but in modelling of complex systems quite often we immediately work towards state representations. Here is an example.

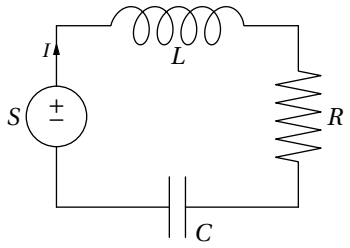

FIGURE 2.1: *RLC* network.

Example 2.1.1 (*RLC* electrical circuit). Consider the electrical circuit of Figure 2.1. This consists of a capacitor C, an inductor (coil) L, a resistor R, and a voltage source S. Let V_C, V_L, V_R, and V be the voltages across C, L, R, and S, respectively, and let I_C, I_L, I_R, and I be the currents through these elements. Kirchoff's voltage and current laws lead to the following balance equations:

$$V = V_L + V_C + V_R, \qquad I_L = I_R = I_C = I. \tag{2.3}$$

The constitutive equations of a linear capacitor, inductor, and resistor are respectively given by

$$\begin{cases} V_C(t) = \frac{1}{C} q(t) \\ \dot{q}(t) = I_C(t) \end{cases}, \qquad \begin{cases} I_L(t) = \frac{1}{L} \phi(t) \\ \dot{\phi}(t) = V_L(t) \end{cases}, \qquad V_R(t) = R I_R(t) \tag{2.4}$$

for constants C, L, and R. Here q is the charge on the capacitor and ϕ is the magnetic flux of the inductor. By choosing

$$x(t) = \begin{bmatrix} q(t) \\ \phi(t) \end{bmatrix},$$

we can rewrite (2.3) and (2.4) as

$$\begin{bmatrix} \dot{q}(t) \\ \dot{\phi}(t) \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{L} \\ -\frac{1}{C} & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} q(t) \\ \phi(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} V(t). \tag{2.5}$$

This is a DE with input $u(t) := V(t)$, the voltage across the voltage source. As output $y(t)$ we can for example take the current $I(t)$ through the voltage source, in which case

$$y(t) = \begin{bmatrix} 0 & \frac{1}{L} \end{bmatrix} \begin{bmatrix} q(t) \\ \phi(t) \end{bmatrix}, \tag{2.6}$$

or the charge on the capacitor, in which case

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} q(t) \\ \phi(t) \end{bmatrix}. \tag{2.7}$$

□

[1] Simulation of DE's means solving the DE (approximately) on a computer.

In this chapter, we consider systems with several inputs,

$$u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_{n_u}(t) \end{bmatrix},$$

and several outputs,

$$y(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_{n_y}(t) \end{bmatrix}$$

and a state with several components

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

that satisfy a system of 1st-order DE's of the form

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{cases} \tag{2.8}$$

The number of components of x is always denoted by n. The first equation in (2.8),

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

is called the state equation. Here A is an $n \times n$ matrix and B is an $n \times n_u$ matrix. The second equation in (2.8) is

$$y(t) = Cx(t) + Du(t).$$

This is called the output equation of the system. Here C is an $n_y \times n$ matrix and D is an $n_y \times n_u$ matrix. In most applications the matrix D is actually zero, so then the output is just a linear combination of the entries of the state, $y(t) = Cx(t)$.

For ease of notation, we sometimes leave out the time and simply write $\dot{x} = Ax + Bu$ and $y = Cx + Du$.

2.2 Solutions of State Equations

In this section, we determine the general solution $x(t), y(t)$ of

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

for a given $u : \mathbb{R} \to \mathbb{R}^{n_u}$. Here, as always, $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n_u}$, $C \in \mathbb{R}^{n_y \times n}$, $D \in \mathbb{R}^{n_y \times n_u}$, and $x(t)$ an n-dimensional signal. As soon as we have $x(t)$, we also have the output $y(t)$, because $y(t) = Cx(t) + Du(t)$. In other words, the only difficulty lies in the state equation $\dot{x}(t) = Ax(t) + Bu(t)$.

Example 2.2.1 (Variation of constants). In this example, we determine $x(t)$ in the case $n = n_u = 1$. That is, A and B are scalars, $A = a$, $B = b$, and the state has only one component, $x(t) = x_1(t)$. Thus

$$\dot{x}_1(t) = ax_1(t) + bu(t). \tag{2.9}$$

For $u(t) \equiv 0$, this equation reduces to the homogeneous equation $\dot{x}_1(t) = ax_1(t)$. The solution of this homogeneous equation is known to be

$$x_1(t) = z\mathrm{e}^{at}, \quad z \in \mathbb{C}$$

for an arbitrary constant z. The method of *variation of constants* consists in writing a candidate solution $x_1(t)$ of (2.9) as

$$x_1(t) = z(t)\mathrm{e}^{at},$$

where $z(t)$ is now a function of t. Every $x_1(t)$ can be written as $x_1(t) = z(t)\mathrm{e}^{at}$ because e^{at} is invertible. We have

$$\begin{aligned}
\dot{x}_1(t) = ax_1(t) + bu(t) &\iff \dot{z}(t)\mathrm{e}^{at} + az(t)\mathrm{e}^{at} = az(t)\mathrm{e}^{at} + bu(t) \\
&\iff \dot{z}(t)\mathrm{e}^{at} = bu(t) \\
&\iff \dot{z}(t) = \mathrm{e}^{-at}bu(t) \\
&\iff z(t) = z_0 + \int_0^t \mathrm{e}^{-a\tau}bu(\tau)\,\mathrm{d}\tau \qquad (z_0 \in \mathbb{C}).
\end{aligned}$$

The general solution $x_1(t) = z(t)\mathrm{e}^{at}$ is therefore

$$x_1(t) = \mathrm{e}^{at}\Big(z_0 + \int_0^t \mathrm{e}^{-a\tau}bu(\tau)\,\mathrm{d}\tau\Big) = \mathrm{e}^{at}z_0 + \int_0^t \mathrm{e}^{a(t-\tau)}bu(\tau)\,\mathrm{d}\tau.$$

□

We will see that the method of variation of constants also works for states with more than one component ($n > 1$). In the example above, we used the exponential function e^{at}. In the general n-dimensional case, its role is taken over by the matrix exponential e^{At}, with $A \in \mathbb{R}^{n\times n}$. In analogy with the Taylor series expansion of e^a,

$$\mathrm{e}^a = \sum_{k=0}^{\infty} \frac{1}{k!}a^k = 1 + a + \frac{1}{2!}a^2 + \frac{1}{3!}a^3 + \cdots$$

we define the following.

Definition 2.2.2 (Matrix exponential). The matrix exponential e^A of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as

$$\mathrm{e}^A = \sum_{k=0}^{\infty} \frac{1}{k!}A^k = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots. \tag{2.10}$$

□

For every square matrix A, this series is convergent.

Lemma 2.2.3 (Matrix exponential). *Let $A, F \in \mathbb{R}^{n\times n}$. Four characteristic properties of the matrix exponential are:*

1. $\mathrm{e}^0 = I$ *for the zero matrix* $0 \in \mathbb{R}^{n\times n}$.

2. *If $AF = FA$, then $e^A e^F = e^{A+F}$.*

3. *e^A is invertible and $(e^A)^{-1} = e^{-A}$.*

4. *Let $t \in \mathbb{R}$; then $\frac{d}{dt} e^{At} = Ae^{At} = e^{At}A$.*

Proof. The proof of parts 2 and 4 use some results about convergence of series. Certain technical details are discussed in another course (Analysis).

1. Follows immediately from the definition (formula (2.10)).

2. Written out, the product $e^A e^F$ is

$$e^A e^F = \Big(\sum_{k=0}^{\infty} \frac{1}{k!} A^k\Big)\Big(\sum_{m=0}^{\infty} \frac{1}{m!} F^m\Big) = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty} \frac{1}{k!\,m!} A^k F^m.$$

   We see that the coefficient of $A^k F^m$ is equal to $\frac{1}{k!m!}$. Written out, e^{A+F} gives

$$e^{A+F} = \sum_{n=0}^{\infty} \frac{1}{n!}(A+F)^n = \sum_{n=0}^{\infty} \left(\frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} A^k F^{n-k} \right) = \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{1}{k!(n-k)!} A^k F^{n-k}.$$

   Here too, the coefficient of the factor $A^k F^m$ is equal to $\frac{1}{k!m!}$. So $e^A e^F = e^{A+F}$. (So, where did we use that $AF = FA$? See Exercise 2.6.)

3. Apply part 2 with $F = -A$.

4. The derivative of the series $\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k$ equals the series of the derivatives[2] (that is, summation and differentiation may be interchanged):

$$\frac{d}{dt} e^{At} = \frac{d}{dt}\Big(\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k\Big) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k \frac{d}{dt} t^k = \sum_{k=1}^{\infty} \frac{1}{(k-1)!} A^k t^{k-1} = \sum_{m=0}^{\infty} \frac{1}{m!} A^{m+1} t^m = e^{At} A.$$

   Since A commutes with itself, we also have $e^{At} A = \sum_{n=0}^{\infty} \frac{1}{n!} t^n A^{n+1} = Ae^{At}$.

■

With these properties of the matrix exponential, we can redo the method of variation of constants of Example 2.2.1. We write the candidate solution $x(t)$ of $\dot{x}(t) = Ax(t) + Bu(t)$ as

$$x(t) = e^{At} z(t)$$

with $z(t) \in \mathbb{R}^n$ a yet to be determined function of time. Every $x(t)$ can be written this way, because e^{At} is invertible. We have

$$\begin{aligned}
\dot{x}(t) = Ax(t) + Bu(t) &\iff Ae^{At} z(t) + e^{At} \dot{z}(t) = Ae^{At} z(t) + Bu(t) \\
&\iff e^{At} \dot{z}(t) = Bu(t) \\
&\iff \dot{z}(t) = e^{-At} Bu(t) \\
&\iff z(t) = z(t_0) + \int_{t_0}^{t} e^{-A\tau} Bu(\tau)\, d\tau \qquad (z(t_0) \in \mathbb{R}^n) \\
&\iff x(t) = e^{At}\Big(e^{-At_0} x(t_0) + \int_{t_0}^{t} e^{-A\tau} Bu(\tau)\, d\tau\Big) \\
&\iff x(t) = e^{A(t-t_0)} x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)} Bu(\tau)\, d\tau.
\end{aligned}$$

In summary, we have the following result.

[2] The proper statement (proved in *Analysis*) is this: since the series $\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k$ converges at $t = 0$, and the series of derivatives $\sum_k \frac{d}{dt}(\frac{1}{k!} A^k t^k)$ converges *uniformly* on any bounded interval, we have that $\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k$ converges and is differentiable everywhere, and that $\frac{d}{dt}(\sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k)$ equals $\sum_{k=0}^{\infty} \frac{d}{dt}(\frac{1}{k!} A^k t^k)$.

Theorem 2.2.4 (Solution of the state equations). *Let $t_0 \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n_u}$. The state equation $\dot{x}(t) = Ax(t) + Bu(t)$ for any given piecewise continuous $u(t)$ has a unique solution $x(t)$, given by*

$$x(t) = \mathrm{e}^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} \mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau. \tag{2.11}$$

The output $y(t) = Cx(t) + Du(t)$ follows uniquely, and is

$$y(t) = C\mathrm{e}^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} C\mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t). \tag{2.12}$$

Proof. See the discussion preceding this theorem. There is one subtlety: we have until now tacitly assumed that $\dot{x}$ exists, but if u is piecewise scontinuous, then the x defined in (2.11) is well defined and continuous, but not necessarily differentiable in the classic sense. What does $\dot{x} = Ax + Bu$ mean in this case? Appendix A.1 gives more details on this, but it only scratches the surface, and it is not really what this course is about! ∎

Note that (2.12) is of the form

$$y(t) = \mathscr{H}(x(t_0), u(\tau)|_{\tau\in[t_0,t]}) \quad \forall\, t \geq t_0.$$

If we think of t_0 as "the present" and t as some time in the future, $t > t_0$, then the above says that the present $x(t_0)$ together with the future of the input, $\{u(\tau) \mid \tau \in [t_0, t]\}$ is sufficient to determine the future of the output, $y(t)$. In other words,

> *the only information we need from the past, $t < t_0$, in order to continue into the future, $t > t_0$, is the present $x(t_0)$.*

This is known as the "state property" and for this reason we call x a state of the system. Our state x at each moment in time is an element of a finite-dimensional space, $\mathbb{R}^n$, and such systems are called finite-dimensional systems. The state property is exploited in simulation:

Example 2.2.5 (Simulation using state representations). To simulate (2.8), we discretize the state equation $\dot{x}(t) = Ax(t) + Bu(t)$. Let $h > 0$ be a (small) step size. Since $(x(t+h) - x(t))/h \approx \dot{x}(t)$ we have

$$\begin{aligned} x(t+h) &\approx x(t) + h\dot{x}(t) \\ &= x(t) + h(Ax(t) + Bu(t)). \end{aligned}$$

So if the initial state $x(0)$ is known (together with the input) then we also approximately know

$$x(h) \approx x(0) + h(Ax(0) + Bu(0)).$$

Now that we have $x(h)$ we also approximately know $x(2h)$ because, by the same rule,

$$x(2h) \approx x(h) + h(Ax(h) + Bu(h)).$$

Likewise $x(3h)$ follows as $x(3h) \approx x(2h) + h(Ax(2h) + Bu(2h))$, et cetera. This method of simulation is called *Euler's method*. More sophisticated methods will be explored in a course on numerical methods. For simulation we only need to keep track of the current state $x(kh)$, that is, all values of the state prior to that may be discarded. The following simple PYTHON script illustrates this point.

```
from math import sin
u = lambda t: sin(t)     # let's say our input is u(t) = sin(t)
a=-1                     # consider state model ẋ(t) = ax(t)
b=1                      #                            +bu(t)

h = 0.01                 # some small stepsize >0
t = 0                    # whatever initial time
x = 0                    # whatever initial state (at initial time)
tend = 10                # end time of simulation
while t<tend:            # simulate the state model:
  x+=h*(a*x+b*u(t))      # WE ONLY NEED TO KEEP TRACK OF THE CURRENT STATE
  t+=h                   # x ≈ x(t)

print(f'x({t})={x}\n')   # x(10.009999999999831)=0.1427120062412903
```

□

If the input is the zero function, $u(t) = 0$, then the dynamics are fully determined by $x(t_0)$ at whatever t_0. Indeed, the state then satisfies $\dot{x}(t) = Ax(t)$, and so according to (2.11), its solution is $x(t) = \mathrm{e}^{A(t-t_0)}x(t_0)$. For $t_0 = 0$ this obviously means that

$$x(t) = \mathrm{e}^{At}x(0).$$

The vector $x(0)$ is known as the initial state or the vector of "initial conditions".

Example 2.2.6 (Initial conditions of a DE). Consider once more the homogeneous DE

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = 0.$$

As explained on page 31 we can equivalently describe the DE as $\dot{x} = Ax, y = Cx$ with $x = (y, y^{(1)}, \ldots, y^{(n-1)})$ and certain matrices A, C. The general homogeneous solution according to Thm. 2.2.4 is

$$y(t) = C\mathrm{e}^{At}x(0).$$

It shows that the vector of initial conditions

$$x(0) = (y(0), y^{(1)}(0), \ldots, y^{(n-1)}(0))$$

uniquely determines the solution of the homogeneous equation for all time. In the previous chapter we saw this for 1st-order DE's. Now we know it holds for all nth-order DE's. □

2.2.1 The Entries of e^{At}

To better understand the behavior of the system, it is useful to write the matrix exponential e^{At} in a more explicit form. Only in a few cases can this be done directly using the power series (2.10). An example where the power series is handy, is a diagonal matrix. If Λ is diagonal,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \tag{2.13}$$

then the corresponding matrix exponential $e^{\Lambda t}$ is nothing else than the diagonal matrix of the exponentials,

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_n t} \end{bmatrix}.$$

This follows from (2.10),

$$\begin{aligned} e^{\Lambda t} &= \begin{bmatrix} 1 & 0 \\ & \ddots \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \lambda_1 t & 0 \\ & \ddots \\ 0 & \lambda_n t \end{bmatrix} + \frac{1}{2!} \begin{bmatrix} \lambda_1^2 t^2 & 0 \\ & \ddots \\ 0 & \lambda_n^2 t^2 \end{bmatrix} + \cdots \\ &= \begin{bmatrix} 1 + \lambda_1 t + \frac{1}{2!}\lambda_1^2 t^2 + \cdots & & 0 \\ & \ddots & \\ 0 & & 1 + \lambda_n t + \frac{1}{2!}\lambda_n^2 t^2 + \cdots \end{bmatrix} = \begin{bmatrix} e^{\lambda_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n t} \end{bmatrix}. \end{aligned}$$

This result serves as a basis for the determination of the elements of e^{At} for general A. If, for example, A has an eigendecomposition, that is, if there exists an invertible matrix T such that

$$A = T\Lambda T^{-1}$$

for some diagonal matrix Λ, then

$$e^{At} = Te^{\Lambda t}T^{-1}$$

and because the elements of $e^{\Lambda t}$ are easy to compute, the elements of e^{At} follow. We state this result for general matrices $\Lambda = M$.

Lemma 2.2.7. *Let M and T be square matrices of the same dimension, and assume that T is invertible. Then*

$$e^{TMT^{-1}} = Te^{M}T^{-1}.$$

Proof.

$$\begin{aligned} e^{TMT^{-1}} &= I + TMT^{-1} + \frac{1}{2!}(TMT^{-1})(TMT^{-1}) + \cdots \\ &= I + TMT^{-1} + \frac{1}{2!}TM^2T^{-1} + \cdots \\ &= T\big(I + M + \frac{1}{2!}M^2 + \cdots\big)T^{-1} = Te^{M}T^{-1}. \end{aligned}$$

■

The decomposition $A = T\Lambda T^{-1}$, with Λ diagonal, is called an eigendecomposition because the columns of T are then eigenvectors, and the corresponding diagonal elements of Λ are the eigenvalues. Indeed, if we multiply $A = T\Lambda T^{-1}$ on the right by T, we get

$$AT = T\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix},$$

hence the kth column T_k of T satisfies $AT_k = \lambda_k T_k$. This relation allows us to determine T and Λ (provided that they exist; see further on), and we conclude that the computation of e^{At} can be reduced to the computation of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A and the corresponding linearly independent eigenvectors $v_1, \ldots, v_n$. The eigenvalues λ_i of A are the zeros of the characteristic polynomial $\chi_A(\lambda)$ of A, defined as[3]

$$\chi_A(\lambda) = \det(\lambda I - A).$$

By the fundamental theorem of algebra, the characteristic polynomial has exactly n zeros $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ (of which some can coincide, in which case those zeros are called *multiple*).

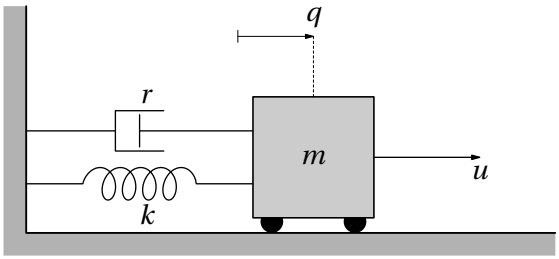

FIGURE 2.2: A car connected to a wall. (See examples 2.2.8 and 2.2.9.)

Example 2.2.8 (Matrix exponential for a mass-damper system). Consider once again the mass-spring-damper system of Example 1.1.5, see Fig. 2.2. We found that $m\ddot{q}(t) + r\dot{q}(t) + kq(t) = u(t)$, where $q(t)$ is the position of the mass (the car), and $u(t)$ is an external force acting on the car. With $x := \left[\begin{smallmatrix} q \\ \dot{q} \end{smallmatrix}\right]$ this becomes the state model

$$\begin{bmatrix} \dot{q}(t) \\ \ddot{q}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -r/m \end{bmatrix} \begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1/m \end{bmatrix} u(t).$$

In this example we focus on the computation of e^{At} so with

$$A = \begin{bmatrix} 0 & 1 \\ -k/m & -r/m \end{bmatrix}.$$

In fact, in this example we consider the case that $k = 0$ (no spring). Then

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -r/m \end{bmatrix}. \tag{2.14}$$

Its characteristic polynomial is

$$\chi_A(\lambda) = \det(\begin{bmatrix} \lambda & -1 \\ 0 & \lambda + \frac{r}{m} \end{bmatrix}) = \lambda(\lambda + \tfrac{r}{m}).$$

This has zeros $\lambda_1 = 0$, $\lambda_2 = -\frac{r}{m}$. The corresponding eigenvectors v_1 and v_2 can be determined using the equations

$$0 = (\lambda_1 I - A)v_1 = \begin{bmatrix} 0 & -1 \\ 0 & \frac{r}{m} \end{bmatrix} v_1, \qquad 0 = (\lambda_2 I - A)v_2 = \begin{bmatrix} -\frac{r}{m} & -1 \\ 0 & 0 \end{bmatrix} v_2.$$

[3] Sometimes the characteristic polynomial is defined as $\det(A - \lambda I)$. This equals $(-1)^n \det(\lambda I - A)$. The choice $\det(\lambda I - A)$ has the advantage that its leading coefficient is always +1, that is, $\det(\lambda I - A) = \lambda^n + \cdots$.

This gives, for example,

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 1 \\ -\frac{r}{m} \end{bmatrix}.$$

We can therefore take T and Λ to be

$$T = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & -\frac{r}{m} \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{r}{m} \end{bmatrix},$$

and, hence,

$$\mathrm{e}^{At} = T\mathrm{e}^{\Lambda t}T^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & -\frac{r}{m} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{e}^{-\frac{r}{m}t} \end{bmatrix} \begin{bmatrix} 1 & \frac{m}{r} \\ 0 & -\frac{m}{r} \end{bmatrix} = \begin{bmatrix} 1 & \frac{m}{r}(1-\mathrm{e}^{-\frac{r}{m}t}) \\ 0 & \mathrm{e}^{-\frac{r}{m}t} \end{bmatrix}.$$

If we do not exert any force u on the mass, then the state $x = \left[\begin{smallmatrix} q \\ \dot{q} \end{smallmatrix}\right]$ is given by $x(t) = \mathrm{e}^{At}x(0)$, that is,

$$\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} 1 & \frac{m}{r}(1-\mathrm{e}^{-\frac{r}{m}t}) \\ 0 & \mathrm{e}^{-\frac{r}{m}t} \end{bmatrix} \begin{bmatrix} q(0) \\ \dot{q}(0) \end{bmatrix}.$$

This seems quite reasonable, because if we write it out, we get

$$\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} q(0) + \frac{m}{r}(1-\mathrm{e}^{-\frac{r}{m}t})\dot{q}(0) \\ \mathrm{e}^{-\frac{r}{m}t}\dot{q}(0) \end{bmatrix},$$

and we see that the initial velocity $\dot{q}(0)$ decreases exponentially. In the end (as $t \to \infty$), the velocity becomes zero, and the mass comes to rest at $q(\infty) = q(0) + \frac{m}{r}\dot{q}(0)$. We now also see that heavier masses take longer to come to a standstill, and that for them, the final position lies further away from the initial position. □

Example 2.2.9 (Matrix exponential for a mass-spring system). We continue with the previous example, but now we include a spring, $k > 0$, and leave out the damper, $r = 0$. The A-matrix is then equal to

$$A = \begin{bmatrix} 0 & 1 \\ -k/m & 0 \end{bmatrix}.$$

This has characteristic polynomial $\chi_A(\lambda) = \lambda^2 + \frac{k}{m}$, which has imaginary zeros,

$$\lambda_1 = \mathrm{i}\omega, \qquad \lambda_2 = -\mathrm{i}\omega \qquad \text{for } \omega := \sqrt{k/m}.$$

The corresponding eigenvectors have complex components (verify this yourself),

$$v_1 = \begin{bmatrix} -\mathrm{i} \\ \omega \end{bmatrix}, \quad v_2 = \begin{bmatrix} \mathrm{i} \\ \omega \end{bmatrix},$$

but the matrix exponential does not:

$$\begin{aligned} \mathrm{e}^{At} &= \begin{bmatrix} -\mathrm{i} & \mathrm{i} \\ \omega & \omega \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\mathrm{i}\omega t} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}\omega t} \end{bmatrix} \begin{bmatrix} -\mathrm{i} & \mathrm{i} \\ \omega & \omega \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \dfrac{\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}}{2} & \dfrac{\mathrm{e}^{\mathrm{i}\omega t} - \mathrm{e}^{-\mathrm{i}\omega t}}{2\mathrm{i}\omega} \\ -\omega\dfrac{\mathrm{e}^{\mathrm{i}\omega t} - \mathrm{e}^{-\mathrm{i}\omega t}}{2\mathrm{i}} & \dfrac{\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{-\mathrm{i}\omega t}}{2} \end{bmatrix} = \begin{bmatrix} \cos(\omega t) & \frac{\sin(\omega t)}{\omega} \\ -\omega\sin(\omega t) & \cos(\omega t) \end{bmatrix}. \end{aligned}$$

In the final identity we used Euler's formula, $e^{\pm i\omega t} = \cos(\omega t) \pm i\sin(\omega t)$. If we do not exert any force u on the mass, then the state $x = \left[\begin{smallmatrix} q \\ \dot{q} \end{smallmatrix}\right]$ is given by $x(t) = e^{At}x(0)$, which means that

$$\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} \cos(\omega t) & \frac{\sin(\omega t)}{\omega} \\ -\omega\sin(\omega t) & \cos(\omega t) \end{bmatrix} \begin{bmatrix} q(0) \\ \dot{q}(0) \end{bmatrix}.$$

For instance if the initial conditions are $q(0) = 1, \dot{q}(0) = 0$ then

$$\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} \cos(\omega t) \\ -\omega\sin(\omega t) \end{bmatrix}.$$

The mass keeps swinging back and forth about zero. The period of the motion is $2\pi/\omega = 2\pi\sqrt{m/k}$. Large masses m (with respect to k) need more time to complete a period. □

If $\det(\lambda I - A)$ has multiple zeros, then it may happen that A does not have n linearly independent eigenvectors. The matrix A is then not diagonalizable, and the procedure of finding $A = T\Lambda T^{-1}$ described above breaks down. If the matrix A is not diagonalizable, we can use the "Jordan normal form" of A. Every square matrix has a Jordan normal form. This is a decomposition

$$A = TJT^{-1}$$

with T an invertible matrix and J the Jordan normal form of A, which is a matrix of a block-diagonal structure

$$J = \begin{bmatrix} J_1 & 0 & \cdots \\ 0 & J_2 & \ddots \\ \vdots & \ddots & \ddots \end{bmatrix}$$

with J_k ($k = 1, 2, \ldots$) the so-called Jordan blocks

$$J_k = \begin{bmatrix} \lambda_k & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_k \end{bmatrix}.$$

A special case is that where all J_k are of dimension 1×1, that is, $J_k = \lambda_k \in \mathbb{C}$. In this case, the decomposition is an eigendecomposition. In general, however, the Jordan blocks J_k have a higher dimension, and then ones appear on the superdiagonal. For every Jordan block J_k, the matrix exponential can be determined using the definition (2.10). This gives (see Exercise 2.8) the upper triangular matrix

$$e^{J_k t} = e^{\lambda_k t} \begin{bmatrix} 1 & t & \frac{1}{2!}t^2 & \cdots & \frac{1}{(m-1)!}t^{m-1} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{2!}t^2 \\ \vdots & \ddots & \ddots & \ddots & t \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}. \tag{2.15}$$

It now follows that e^{At} equals

$$e^{At} = T \begin{bmatrix} e^{J_1 t} & 0 & \cdots \\ 0 & e^{J_2 t} & \ddots \\ \vdots & \ddots & \ddots \end{bmatrix} T^{-1}. \tag{2.16}$$

Note that the elements of $e^{J_k t}$ are of the form $\alpha t^m e^{\lambda t}$ for some $m \in \mathbb{N}$ and $\lambda, \alpha \in \mathbb{C}$, and because *every* square matrix A has a Jordan normal form, all elements of *every* matrix exponential are linear combinations of functions of the form $t^m e^{\lambda t}$. For complex λ, the function $e^{\lambda t}$ is usually split into a real part and an imaginary part,

$$\lambda = \mu + i\omega \implies e^{\lambda t} = e^{\mu t} e^{i\omega t} = e^{\mu t}\big(\cos(\omega t) + i\sin(\omega t)\big).$$

Hence, as we already saw in Example 2.2.9, there can also be terms with cosines and sines in the elements of e^{At}.

2.2.2 Fundamental set of solutions

The fact that the solution of $\dot{x}(t) = Ax(t)$ is uniquely determined by $x(0)$ can sometimes be exploited to quickly determine e^{At} (and, hence, all solutions of $\dot{x}(t) = Ax(t)$). This is possible if we know "sufficiently many" specific solutions. As an example, suppose

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

Say, we guess these two solutions:

$$X_1(t) := \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}, \qquad X_2(t) := := \begin{bmatrix} -\sin(t) \\ \cos(t) \end{bmatrix}$$

It is easy to verify that both indeed satisfy $\dot{X}_i(t) = AX_i(t)$. Now the good news: by the linearity property then also the linear combination

$$x(t) := \begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix} x_0 = \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix} x_0$$

satisfies $\dot{x}(t) = Ax(t)$. This solution has initial state $x(0) = \left[\begin{smallmatrix} \cos(0) & -\sin(0) \\ \sin(0) & \cos(0) \end{smallmatrix}\right] x_0 = x_0$, but since initial states determine the solution completely — $x(t) = e^{At} x_0$ — we must have that $\left[\begin{smallmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{smallmatrix}\right]$ in fact equals e^{At}. A bit more general:

Lemma 2.2.10 (Fundamental set of solutions). *Let $A \in \mathbb{R}^{n\times n}$. Suppose $X_1, \ldots, X_n : \mathbb{R} \to \mathbb{R}^n$ are n solutions of $\dot{x}(t) = Ax(t)$. If $X : \mathbb{R} \to \mathbb{R}^{n\times n}$ defined as $X(t) = \begin{bmatrix} X_1(t) & X_2(t) & \cdots & X_n(t) \end{bmatrix}$ is nonsingular at $t = 0$, then $e^{At} = X(t)X^{-1}(0)$.*

Proof. Every linear combination x of solutions $X_1, \ldots, X_n$ is a solution of $\dot{x} = Ax$. In particular

$$x(t) := X(t)X^{-1}(0)x_0.$$

Clearly then $x(0) = X(0)X^{-1}(0)x_0 = x_0$ so we can also express $x(t)$ as $x(t) = e^{At}x_0$. Since this holds for every x_0 it must be that $e^{At} = X(t)X^{-1}(0)$. ∎

if $X(0)$ is nonsingular we say that $\{X_1, \ldots, X_n\}$ is a fundamental set of solutions of $\dot{x}(t) = Ax(t)$. A generalization is discussed in Exercise 2.9.

Example 2.2.11. Consider

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t)$$

Clearly $\dot{x}_2(t) = 0$ so $x_2(t)$ is constant, $x_2(t) = c$. The equation $\dot{x}_1(t) = c$ then means $x_1(t) = ct + d$. Thus we have at least these two solutions (for $c = 1$ and $d = 0$ and $d = 1$):

$$X_1(t) = \begin{bmatrix} t \\ 1 \end{bmatrix}, \qquad X_2(t) = \begin{bmatrix} t+1 \\ 1 \end{bmatrix},$$

Let

$$X(t) = \begin{bmatrix} t & t+1 \\ 1 & 1 \end{bmatrix}.$$

Clearly $X(0)$ is nonsingular, so

$$e^{At} = X(t)X^{-1}(0) = \begin{bmatrix} t & t+1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

□

2.2.3 Coordinate Transformations

Another way to look at the computation of e^{At} in Lemma 2.2.7 is as follows. The choice of state x is not unique. For instance, in the mass-spring-damper example (Example 2.2.8) we took $x := \left[\begin{smallmatrix} q \\ \dot{q} \end{smallmatrix}\right]$ but we might as well have chosen[4] $x := \left[\begin{smallmatrix} q \\ m\dot{q} \end{smallmatrix}\right]$. More generally, if we have a state x, then we can choose to change to the transformed state z defined by

$$z(t) = T^{-1}x(t) \quad \text{for some invertible } T \in \mathbb{R}^{n\times n}. \tag{2.17}$$

Such a transformation is called a state transformation. Because $x = Tz$, the entries z_k are the coordinates with respect to the basis formed by the columns $T_k \in \mathbb{R}^n$ of T, while the x_k are the coordinates with respect to the *standard basis* of $\mathbb{R}^n$.

Substituting $z = T^{-1}x$ and $x = Tz$ in the equations

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

gives

$$\begin{aligned} \dot{z} &= T^{-1}\dot{x} = T^{-1}(Ax + Bu) = T^{-1}ATz + T^{-1}Bu \\ y &= CTz + Du. \end{aligned}$$

In other words, a coordinate transformation $x \to z = T^{-1}x$ corresponds to the following transformation of system matrices:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \to \begin{bmatrix} T^{-1}AT & T^{-1}B \\ CT & D \end{bmatrix}. \tag{2.18}$$

We say that two state representations (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are isomorphic if they can be changed into each other through a state transformation. Note that a state transformation does not change the relation between u and y.

This transformation changes the homogeneous equation (that is, the equation for $u \equiv 0$)

$$\dot{x} = Ax$$

[4] $m\dot{q}$ is known as "impulse" of the mass. This is a popular choice, also because the product of the two state components, $qm\dot{q}$ is then a "power".

into

$$\dot{z} = \Lambda z, \qquad \Lambda := T^{-1}AT.$$

This holds for every $\Lambda = T^{-1}AT$, but is particularly interesting when Λ, as before, is diagonal,

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

In this case, $\dot{z} = \Lambda z$ is nothing else than n uncoupled equations in the components of z,

$$\dot{z}_k = \lambda_k z_k \qquad \forall k \in \{1, \ldots, n\}.$$

As we know, the solution of this is $z_k(t) = \mathrm{e}^{\lambda_k t} z_k(0)$, so we obtain once again that

$$x(t) = Tz(t) = T\mathrm{e}^{\Lambda t} z(0) = T\mathrm{e}^{\Lambda t} T^{-1} x(0).$$

2.2.4 Geometric Interpretation of Solutions of $\dot{x} = Ax$

Because the transformation matrix T does not depend on time, $x(t)$ and $z(t) = T^{-1}x(t)$ exhibit a quantitatively similar dynamical behavior. Because $\dot{z} = \Lambda z$ is uncoupled (provided that Λ is diagonal), it is easier to first analyze the behavior in the z-domain and later transfer the analysis to the x-domain. This is particularly helpful in providing insight into second-order systems. These are systems whose state has two components ($n = 2$).

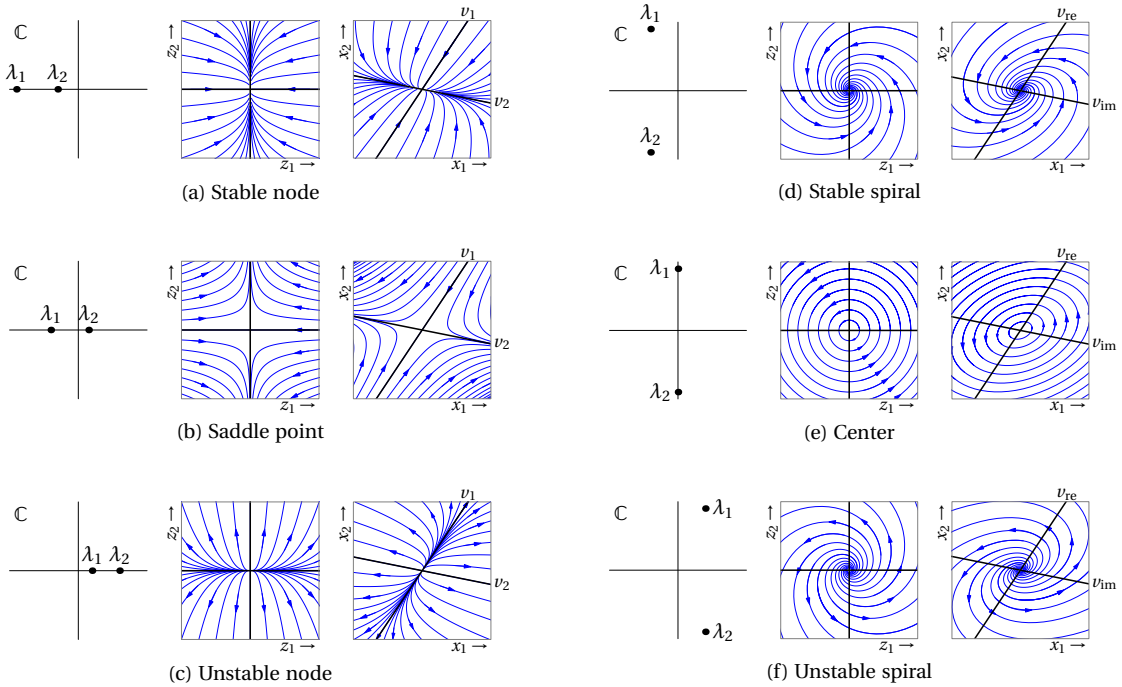

FIGURE 2.3: Phase portraits for diagonalizable second-order systems $\dot{x}(t) = Ax(t)$ for different types of eigenvalues $\lambda_{1,2} \in \mathbb{C}$. In the phase portraits in the (x_1, x_2)-plane, the eigenvectors v_1, v_2 are also shown.

Example 2.2.12 (Phase portraits of second-order systems). Consider $\dot{x}(t) = Ax(t)$ and assume that $A \in \mathbb{R}^{2\times 2}$ and that the matrix is diagonalizable.

1. Stable node. If A has two different negative eigenvalues, then after the transformation $z = T^{-1}x$, the system equation is given by

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \qquad (\lambda_1 < \lambda_2 < 0).$$

   Both states converge to zero, but z_1 does so more quickly because $\lambda_1 < \lambda_2 < 0$. The trajectories in the (z_1, z_2)-plane therefore converge more quickly to the z_2-axis than to the z_1-axis. See Figure 2.3(a). Such plots, where z_2 is set out against z_1, are called *phase portraits*. Figure 2.3(a) also shows the phase portrait of $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] = T\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ (on the right). With each column v_1, v_2 of T corresponds a characteristic trajectory,

$$x(t) = v_1 z_1(t), \qquad x(t) = v_2 z_2(t).$$

   These are the trajectories along the eigenvectors v_1 and v_2; we can recognize them as the straight lines in the phase portrait.

2. Saddle point. Figure 2.3(b) shows phase portraits in the case where A has one real negative eigenvalue and one real positive eigenvalue. To the negative eigenvalue corresponds a characteristic trajectory $x(t) = v_1 z_1(t)$ that converges to the origin. To the positive eigenvalue corresponds a characteristic trajectory $x(t) = v_2 z_2(t)$ that moves away from the origin. All other trajectories also diverge.

3. Unstable node. If both eigenvalues of A are positive (and real), both components of z (and therefore of $x = Tz$) diverge; see Figure 2.3(c).

4. Stable spiral. If the eigenvalues of A are not real, they are necessarily each other's complex conjugates, $\lambda_1 = \mu + \mathrm{i}\omega$, $\lambda_2 = \mu - \mathrm{i}\omega$. The corresponding eigenvectors v_1 and v_2 can then also be chosen as complex conjugates (verify): $v_1 = v_{\mathrm{re}} + \mathrm{i}v_{\mathrm{im}}$ en $v_2 = v_{\mathrm{re}} - \mathrm{i}v_{\mathrm{im}}$. After the transformation $z = \begin{bmatrix} v_1 & v_2 \end{bmatrix}^{-1} x$, we get

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} \mu + \mathrm{i}\omega & 0 \\ 0 & \mu - \mathrm{i}\omega \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

   but this does not give much insight because $z(t)$ is then complex valued. Instead, we use the transformation $z = \begin{bmatrix} v_{\mathrm{re}} & -v_{\mathrm{im}} \end{bmatrix}^{-1} x$. This gives

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \underbrace{\begin{bmatrix} \mu & -\omega \\ \omega & \mu \end{bmatrix}}_{\Lambda} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

   from which we can deduce (see Exercise 2.11) that

$$\mathrm{e}^{\Lambda t} = \mathrm{e}^{\mu t} \begin{bmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{bmatrix}.$$

   Figure 2.3(d) shows the situation when $\mu < 0$. The solutions spiral around the origin and converge to it.

5. Center. Consider the stable focus, but this time with $\mu = 0$. In this case, the phase portrait consists of concentric circles around the origin (in the (z_1, z_2)-plane) and ellipses around the origin (in the (x_1, x_2)-plane). See Figure 2.3(e).

6. Unstable spiral. If the eigenvalues are complex, namely $\lambda_1 = \mu + i\omega$ and $\lambda_2 = \mu - i\omega$ with $\mu > 0$, then x_1 and x_2 diverge. This looks the same as the stable spiral, except that it spirals away from the origin; see Figure 2.3(f).

If the eigenvalues coincide, $\lambda_1 = \lambda_2$, then the A-matrix may not be diagonalizable. We do not consider these cases. They will be studied in the course *Ordinary Differential Equations*. □

Example 2.2.13 (Classification of the damping for second-order systems). Consider, once more, the mass-spring-damper system from Example 2.2.8 with $k > 0$ and $r > 0$, and assume that we are not exerting any external force u, i.e. $u = 0$. We then have

$$\dot{x} = Ax \quad \text{with} \quad A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{r}{m} \end{bmatrix}.$$

The eigenvalues of A are

$$\lambda_{1,2} = -\frac{r}{2m} \pm \sqrt{\frac{r^2}{4m^2} - \frac{k}{m}}.$$

Depending on the values of k and r, we can distinguish three cases:

Underdamped. If $\frac{r^2}{4m^2} - \frac{k}{m} < 0$, then λ_1 and λ_2 are complex eigenvalues, with $\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2 = -\frac{r}{2m} < 0$. This is called the underdamped situation and occurs when the damping r is weak. The phase portrait is as in Figure 2.3(d). Both x_1 and x_2 converge to zero, but they do so oscillating about zero. The mass m therefore continuously oscillates about the equilibrium, and only comes to a halt in the limit.

Overdamped. If $\frac{r^2}{4m} - \frac{k}{m} > 0$, then λ_1 and λ_2 are both real, and we have $\lambda_{1,2} < 0$. This is called the overdamped situation and occurs when r is sufficiently large. The phase portrait is as in Figure 2.3(a). In this case, x_1 and x_2 can at most have a local extremum, after which they converge monotonically to zero. The mass m therefore shoots through the origin at most once, after which it converges monotonically to it.

Critically damped. The boundary case between underdamped and overdamped is when $\frac{r^2}{4m^2} - \frac{k}{m} = 0$. In this case, λ_1 and λ_2 are real and equal, $\lambda_1 = \lambda_2 = -\frac{r}{2m} < 0$. This is called the critically damped situation. The mass just barely refrains from oscillating, and the system has in common with the overdamped situation that the mass passes through the equilibrium at most once.

In the critically damped situation, the damper is just strong enough to prevent oscillations, but not strong enough to make the system excessively sluggish; see Figure 2.4. □

2.3 Stability of Equilibrium Points

In the previous example, we saw that in certain cases of

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

the solution $x(t)$ converges to zero as $t \to \infty$. In other cases, $x(t)$ might move around the origin or diverge away from the origin. In all cases, the solution $x(t)$ remains at zero if the initial condition is $x(0) = 0$. We therefore call the origin an equilibrium (point). However, a small disturbance can move $x(t)$ away from this equilibrium point, and then it depends on the situation whether $x(t)$ returns to the equilibrium point or not.

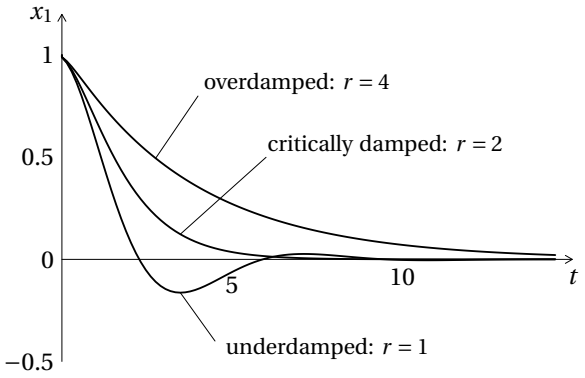

FIGURE 2.4: Position $x_1(t)$ in an overdamped, critically damped, and underdamped mass-spring-damper system (Example 2.2.13 with $m = 1$ and $k = 1$ and $x_1(0) = 1, \dot{x}_1(0) = 0$.)

Definition 2.3.1 (Equilibrium point). An $\bar{x} \in \mathbb{R}^n$ is an equilibrium (point) of $\dot{x}(t) = Ax(t)$ if $x(t) := \bar{x}$ is a constant solution of the differential equation. □

Clearly constant is equivalent to $\dot{x}(t)$ being zero, so the equilibriun points $\bar{x}$ of $\dot{x}(t) = Ax(t)$ are precisely those that lie in the null space of A:

$$0 = A\bar{x}.$$

The origin $0 \in \mathbb{R}^n$ is an equilibrium point of every $\dot{x}(t) = Ax(t)$. If A is invertible, then the origin is the *only* equilibrium point of $\dot{x}(t) = Ax(t)$, see Exercise 2.1. Asymptotic stability is the property of solutions returning back to the equilibrium. As already mentioned in Chapter 1 stability is an important and tricky concept for *nonlinear* DE's, but for our type of linear systems we simply define:

Definition 2.3.2 (Asymptotic stability). A state equation $\dot{x}(t) = Ax(t) + Bu(t)$ is asymptotically stable if for $u = 0$ all solutions $x(t)$ converge to 0 as $t \to \infty$. □

Lemma 2.3.3 (Asymptotic stability). *Consider the system $\dot{x}(t) = Ax(t)$ with equilibrium point $\bar{x} = 0$. The following four statements are equivalent:*

1. *the system is asymptotically stable.*
2. $\lim_{t\to\infty} x(t) = 0$ *for all* $x(0) \in \mathbb{R}^n$.
3. *All eigenvalues of A have negative real part.*
4. *The characteristic polynomial* $\det(\lambda I - A)$ *is an asymptotically stable polynomial.*

Proof. 1. $\Longleftrightarrow$ 2.: $x(t) = e^{At}x(0)$, that is, the solution is completely determined by $x(0)$. Now the result follows immediately from the definition of asymptotic stability.

2. $\Longleftrightarrow$ 3.: If all eigenvalues λ_k have real part < 0 then all entries of (2.15) converge to zero, and then so do all entries of e^{At} as given in (2.16). Consequently $x(t) = e^{At}x(0) \to 0$ as $t \to \infty$. Conversely, if there is an eigenvalue with $\mathrm{Re}\,\lambda \geq 0$, then $x(t) := ve^{\lambda t}$ is a solution if v is a corresponding eigenvector. It does not converge to zero for $t \to \infty$. (If λ is not real, take $x(t) = \mathrm{Re}\,(ve^{\lambda t})$, and verify that this signal is not the zero function.)

3. $\Longleftrightarrow$ 4. standard linear algebra: the eigenvalues of A are the zeros of its characteristic polynomial $\det(\lambda I - A)$. ■

Example 2.3.4 (Mass-spring-damper). The mass-spring-damper system has A-matrix given by

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{r}{m} \end{bmatrix}.$$

Suppose that $m > 0$. The system is asymptotically stable iff its characteristic polynomial is asymptotically stable (i.e. all eigenvalues of A have negative real part). The characteristic polynomial is

$$\chi_A(\lambda) = \det(\lambda I - A) = \lambda(\lambda + \tfrac{r}{m}) + \tfrac{k}{m} = \lambda^2 + \tfrac{r}{m}\lambda + \tfrac{k}{m}.$$

In Lemma 1.7.4 we found that this polynomial is asymptotically stable iff both $k > 0$ and $r > 0$. The characteristic polynomial of our matrix A is identical to characteristic polynomial of the 2nd-order DE of Example 1.7.5. Of course this is not a surprise, in fact this always holds, see Exercise 2.7. □

2.4 A final note on higher-order DE's

We conclude this chapter with a remark about higher-order DE's. At the beginning of this chapter we saw that nth-order DE's

$$y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) = u(t)$$

can be turned into an equivalent state representation using the substitution

$$x := \begin{bmatrix} y \\ y^{(1)} \\ \vdots \\ y^{(n-2)} \\ y^{(n-1)} \end{bmatrix}.$$

This gives the nth order state representation

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix} x + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} u \\ y &= \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} x. \end{aligned} \tag{2.19}$$

However, this method does not automatically generalize to the case where the DE also contains one or more derivatives of u, as in the following nth order DE:

$$\begin{aligned} &y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1 y^{(1)}(t) + p_0 y(t) \\ &\quad = \quad q_n u^{(n)}(t) + q_{n-1}u^{(n-1)}(t) + \cdots + q_1 u^{(1)}(t) + q_0 u(t). \end{aligned} \tag{2.20}$$

(We did *not* consider such DE's in Chapter 1.) This type of DE also admits a state representation, but its construction is different. The following example shows how.

Example 2.4.1 (Simulation diagram). Consider the DE

$$\ddot{y} + 5\dot{y} + 6y = 7\dot{u} + 8u. \tag{2.21}$$

To derive the state representation, we bring all terms except $\ddot{y}$ to the right-hand side of the equation,

$$\ddot{y} = -5\dot{y} - 6y + 7\dot{u} + 8u.$$

Next, we integrate the equation as often as necessary to get rid of the derivatives

$$\begin{aligned} y &= \iint \left[-5\dot{y} - 6y + 7\dot{u} + 8u\right] \\ &= \int \Big[-5y + 7u + \int \left[-6y + 8u\right]\Big]. \end{aligned}$$

As a last step, we assign a state component x_k, with $k = 1, 2$, to each of the antiderivatives,

$$y = \underbrace{\int \Big[-5y + 7u + \underbrace{\int \left[-6y + 8u\right]}_{x_1}\Big]}_{x_2}.$$

The so defined state components satisfy

$$\begin{cases} \dot{x}_1 = -6y + 8u, \\ \dot{x}_2 = -5y + 7u + x_1, \\ \ \ y = x_2. \end{cases}$$

We can now read off the state representation directly:

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & -6 \\ 1 & -5 \end{bmatrix} x + \begin{bmatrix} 8 \\ 7 \end{bmatrix} u, \\ y &= \begin{bmatrix} 0 & 1 \end{bmatrix} x. \end{aligned}$$

Ready. □

The general process to construct a state representation from an ordinary nth order DE

$$y^{(n)} + p_{n-1} y^{(n-1)} + \cdots + p_0 y = q_n u^{(n)} + q_{n-1} u^{(n-1)} + \cdots + q_0 u \tag{2.22}$$

is not more complicated than the example we just gave. For typographical reasons, we only present the method for third-order DE's. We first bring all terms except $y^{(n=3)}$ to the right of the equal sign, and group the terms with the same order of derivative,

$$y^{(3)} = q_3 u^{(3)} + [q_2 u^{(2)} - p_2 y^{(2)}] + [q_1 u^{(1)} - p_1 y^{(1)}] + [q_0 u - p_0 y].$$

Then we integrate $n = 3$ times,

$$y = q_3 u + \int \Big[q_2 u - p_2 y + \int \left[q_1 u - p_1 y + \int [q_0 u - p_0 y]\right]\Big]$$

and assign a state component $x_1, \ldots, x_{n=3}$ to each of the antiderivatives,

$$y = q_3 u + \underbrace{\int \Big[q_2 u - p_2 y + \underbrace{\int [q_1 u - p_1 y + \underbrace{\int [q_0 u - p_0 y]}_{x_1}]}_{x_2}\Big]}_{x_3}. \tag{2.23}$$

This way, the state components satisfy

$$\begin{aligned}\dot{x}_1 &= q_0 u - p_0 y,\\ \dot{x}_2 &= q_1 u - p_1 y + x_1,\\ \dot{x}_3 &= q_2 u - p_2 y + x_2,\\ y &= q_3 u + x_3,\end{aligned}$$

or, in matrix form,

$$\dot{x} = \begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix} x + \begin{bmatrix}q_0 & -p_0\\ q_1 & -p_1\\ q_2 & -p_2\end{bmatrix}\begin{bmatrix}u\\ y\end{bmatrix}, \tag{2.24}$$

$$y = \begin{bmatrix}0 & 0 & 1\end{bmatrix} x + q_3 u. \tag{2.25}$$

In particular, we see that $y = x_3 + q_3 u$. This allows us to eliminate y in (2.24), and the result is a state representation that is called the observer canonical form (the reasons for this name will become clear in § 3.5):

$$\begin{aligned}\dot{x} &= \begin{bmatrix}0 & 0 & -p_0\\ 1 & 0 & -p_1\\ 0 & 1 & -p_2\end{bmatrix} x + \begin{bmatrix}q_0 - p_0 q_3\\ q_1 - p_1 q_3\\ q_2 - p_2 q_3\end{bmatrix} u,\\ y &= \begin{bmatrix}0 & 0 & 1\end{bmatrix} x + q_3 u.\end{aligned}$$

For general n we have the following result.

Lemma 2.4.2 (Observer canonical form). *A pair $u, y : \mathbb{R} \to \mathbb{R}$ is a solution of*

$$y^{(n)} + p_{n-1} y^{(n-1)} + \cdots + p_0 y = q_n u^{(n)} + \cdots + q_0 u$$

if and only if there exists an $x : \mathbb{R} \to \mathbb{R}^n$ such that

$$\begin{aligned}\dot{x} &= \begin{bmatrix}0 & \cdots & \cdots & 0 & -p_0\\ 1 & \ddots & & \vdots & -p_1\\ 0 & \ddots & \ddots & \vdots & \vdots\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & 1 & -p_{n-1}\end{bmatrix} x + \begin{bmatrix}q_0 - p_0 q_n\\ q_1 - p_1 q_n\\ \vdots\\ \vdots\\ q_{n-1} - p_{n-1} q_n\end{bmatrix} u,\\ y &= \begin{bmatrix}0 & \cdots & \cdots & 0 & 1\end{bmatrix} x + q_n u.\end{aligned} \tag{2.26}$$

Proof (idea only). For simplicity, we again take $n = 3$. It follows from the above that if (u, y) satisfies (2.22), then we can take x_i as constructed in (2.23). These x_i satisfy (2.24,2.25) by construction. Conversely, if (2.24,2.25) hold, then substitution shows that x satisfies (2.23). Differentiate (2.23) n times, and we obtain (2.22). ∎

Note that the A -matrix of the observer canonical form is the transpose of the A-matrix of (2.19).

Example 2.4.3 (Observer canonical form). As we know, a pair (u, y) satisfies

$$\dot{y} = \dot{u}$$

if and only if $y = u + c$ for some constant $c \in \mathbb{R}$. The observer canonical form of $\dot{y} = \dot{u}$ is (verify this yourself)

$$\begin{aligned}\dot{x} &= 0,\\ y &= x + u.\end{aligned}$$

Since $\dot{x}$ is zero, we have that x is constant. The state x takes over the role of c in $y = u + c$. □

The state x in the observer canonical form usually does not have a physical interpretation such as the current or voltage in Example 2.1.1.

About the degee of differentiability of the input. A comment about the degree of differentiability of the input is in order. For DE (2.20) to be well defined one would normally insist on having an input that is differentiable n times. However, in the equivalent state model (2.26) we merely need u to be integrable! The difference already shows up in the last example: $\dot{y} = \dot{u}$. Here one would want u to be differentiable, while in the equivalent $y = u + c$ anything goes. For a proper discussion of this problem we would have to introduce the concept of "weak solutions" of DE's. We will not do that, but we do want to emphasise that with state models we can allow more inputs, such as discontinuous inputs, and this, in part, explains why state models are popular when we want to simulate the DE based on samples of the input.

2.5 Exercises

2.1 Comprehension questions (on the whole chapter). Prove or give a counterexample.

(a) There exists an $A \in \mathbb{R}^{n\times n}$ such that

$$e^{At} = \begin{bmatrix} e^{t} & e^{t} \\ 0 & e^{-t} \end{bmatrix}.$$

(b) There exists an $A \in \mathbb{R}^{n\times n}$ such that

$$e^{At} = \begin{bmatrix} e^{t} & te^{t} \\ 0 & e^{-t} \end{bmatrix}.$$

(c) If $A \in \mathbb{R}^{n\times n}$ is invertible, then 0 is the only equilibrium point of $\dot{x} = Ax$.

(d) If $A \in \mathbb{R}^{n\times n}$ is singular, then $\dot{x} = Ax$ has infinitely many equilibrium points.

(e) If $A \in \mathbb{R}^{n\times n}$ is singular, then no equilibrium point of $\dot{x} = Ax$ is stable.

2.2 Let

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

(a) Determine e^{At} using the definition.

(b) Determine the general solution $(x_1(t), x_2(t), x_3(t))$ of

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = 0$$

using part (a) and verify your answer.

2.3 Determine e^{At} for the following matrices:

(a) $A = \begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix}$.

(b) $A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$.

(c) $A = \begin{bmatrix} -8 & 8 \\ -15 & 14 \end{bmatrix}$.

(d) $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}$.

(e) $A = \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}$. [Hint: easier than/when you think!]

(f) $A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$

(g) $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ [Hint: use the definition of e^{At}]

2.4 Determine e^{At} for $A = \begin{bmatrix} 4 & -2 \\ 3 & -1 \end{bmatrix}$ and verify that $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{At} = A\mathrm{e}^{At}$.

2.5 Determine e^{At} for the antidiagonal $n \times n$ matrix with ones on the antidiagonal,

$$A = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

[Hint: Use the definition of the matrix exponential, and realize that $\cosh(t) := \frac{1}{2}(\mathrm{e}^t + \mathrm{e}^{-t})$ and $\sinh(t) := \frac{1}{2}(\mathrm{e}^t - \mathrm{e}^{-t})$.]

2.6 *Commuting matrices.* In which step of the proof of Lemma 2.2.32 do we use the assumption that $AF = FA$?

2.7 *Companion matrix.* Consider the *(right) companion matrix*

$$A = \begin{bmatrix} 0 & \cdots & \cdots & 0 & -p_0 \\ 1 & \ddots & & \vdots & -p_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -p_{n-1} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

This is a square matrix with a one under each diagonal element and zeros elsewhere, except in the last column.

(a) Show that the characteristic polynomial $\chi_A(\lambda)$ is equal to

$$\lambda^n + p_{n-1}s^{n-1} + \cdots p_1\lambda + p_0.$$

(b) Let λ be an eigenvalue of A. Show that $v \in \mathbb{C}^{1\times n}$ is a corresponding *left* eigenvector if-and-only-if

$$v = c\begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{n-1} \end{bmatrix}$$

for some nonzero constant c.

(c) Let A be the companion matrix

$$A = \begin{bmatrix} 0 & 0 & -6 \\ 1 & 0 & -11 \\ 0 & 1 & -6 \end{bmatrix}.$$

Determine e^{At} using parts (a) and (b).

2.8 *Matrix exponential of a Jordan block.* Prove Equation (2.15).

2.9 *Fundamental set of solutions.* Prove the following generalization of Lemma 2.2.10:

Let $A \in \mathbb{R}^{n\times n}$. Suppose $X_1, \dots, X_n : \mathbb{R} \to \mathbb{R}^n$ are n solutions of $\dot{x}(t) = Ax(t)$. If $X : \mathbb{R} \to \mathbb{R}^{n\times n}$ defined as $X(t) = \begin{bmatrix} X_1(t) & X_2(t) & \cdots & X_n(t) \end{bmatrix}$ is nonsingular for some $t \in \mathbb{R}$, then $X(t_0)$ is invertible for every $t_0 \in \mathbb{R}$, and $e^{A(t-t_0)} = X(t)X^{-1}(t_0)$.

2.10 Consider the problem of Example 2.2.11. We already figured out that $x_1(t) = ct + d, x_2(t) = c$ are all the possible solutions. Thus fundamental sets $\{X_1, X_2\}$ are necessarily of the form

$$X(t) := \begin{bmatrix} X_1(t) & X_2(t) \end{bmatrix} := \begin{bmatrix} c_1 t + d_1 & c_2 t + d_2 \\ c_1 & c_2 \end{bmatrix}.$$

(a) Under what conditions on c_1, c_2, d_1, d_2 do they form a fundamental set?

(b) Verify that singularity of $X(t)$ does not depend on t (this is in line with the claim of Exercise 2.9).

(c) Suppose $\{X_1, X_2\}$ is a fundamental set. Determine $X(t)X^{-1}(0)$ and explain why you are not surprised that it does not depend on c_1, c_2, d_1, d_2.

2.11 *Stable spiral.* Consider part 4 of Example 2.2.12.

(a) Show that after the transformation $z = \begin{bmatrix} v_{\text{re}} & -v_{\text{im}} \end{bmatrix}^{-1} x$, the system $\dot{x} = Ax$ is given by $\dot{z} = \Lambda z$ with

$$\Lambda = \begin{bmatrix} \mu & -\omega \\ \omega & \mu \end{bmatrix}.$$

(b) Determine $e^{\Lambda t}$ for Λ as above. (Hint: Use that $\left[\begin{smallmatrix} \mu & 0 \\ 0 & \mu \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 & -\omega \\ \omega & 0 \end{smallmatrix}\right]$ commute.)

(c) Show that $\|e^{\Lambda t} z_0\| = e^{\mu t}\|z_0\|$, where $\|\cdot\|$ is the usual Euclidean norm. (This shows that for $\mu < 0$, the solutions z of $\dot{z} = \Lambda z$ are all strictly decreasing in the sense that $\|z(t)\|$ is strictly monotonically decreasing.)

2.12 *Stability.* Study the asymptotic stability of the system $\dot{x} = Ax$ for the following matrices:

(a) $A = \begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix}$.

(b) $A = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$.

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

(d) $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

(e) $A = \begin{bmatrix} 0 & A_{12} \\ A_{21} & 0 \end{bmatrix}$ with $A_{ij} \in \mathbb{R}^{n_i \times n_j}$.

(f) $A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$ with $A_{ij} \in \mathbb{R}^{n_i \times n_j}$.

2.13 *Stationary solution.* Consider (2.8) and assume that it is asymptotically stable. If $u(t)$ is a constant signal, $u(t) = u_\infty$, then $y(t)$ also converges to a constant signal $y_\infty = \lim_{t\to\infty} y(t)$. Express y_∞ in terms of A, B, C, D, u_∞ (without using integrals).

2.14 *Differentiated and integrated outputs.* Consider (2.8) with $D = 0$.

(a) Determine a state representation for the system with input u and output $z := \dot{y}$.

(b) Determine a state representation for the system with input u and output z, where z is an arbitrary signal satisfying $\dot{z} = y$.

2.15 *State representations.* Give state representations of $\dot{y} + y = z$, $\dot{z} + z = v$, $\dot{v} + v = u$ with input u and output y.

2.16 Determine state representations of the following systems:

(a) $y^{(1)} + 4y = 2u$

(b) $y^{(1)} + 4y = u^{(1)} + 2u$

(c) $\dot{y} + 2y = \dot{u} - 3u$

(d) $3y^{(3)} + 2y^{(2)} + y^{(1)} = u^{(2)}$

(e) $y^{(4)} = u$

(f) $y^{(3)} = u^{(3)}$

2.17 Consider

$$\begin{cases} \dot{x} &= Ax + Bu, \\ y &= Cx + u. \end{cases}$$

Show that y can also be chosen as input (and u as output) and determine the corresponding state representation

$$\begin{cases} \dot{x} &= \cdots x + \cdots y, \\ u &= \cdots x + \cdots y. \end{cases}.$$

Tougher Exercises

2.18 *Differential equations that do not admit a state representation.*

(a) Explain that $y(t) = \dot{u}(t)$ does not admit a state representation with input u and output y [Hint: Find a continuous u for which y is not continuous, and explain that this is not possible in state representations.]

(b) Explain that

$$p_n y^{(n)}(t) + p_{n-1} y^{(n-1)}(t) + \cdots + p_0 y(t) = q_n u^{(n)}(t) + q_{n-1} u^{(n-1)}(t) + \cdots + q_0 u(t)$$

does not admit a state representation (with input u and output y) if $p_n = 0$ and $q_n \neq 0$.

2.19 *Monotonically decreasing state.*

(a) Prove that $\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|^2 < 0$ for all solutions $x(t) \neq 0$ of $\dot{x} = Ax$ if and only if the matrix $A + A^{\mathrm{T}}$ has only negative eigenvalues.

(b) Prove that all eigenvalues of $A \in \mathbb{R}^{n\times n}$ have negative real part if $A + A^{\mathrm{T}}$ has only negative eigenvalues.

2.20 *Smoothness of the output.* Consider (2.8).

(a) Under what conditions on A, B, C, D is y continuous for every bounded u?

(b) Under what conditions on A, B, C, D is y continuously differentiable for every bounded u?

(c) Let $k \in \mathbb{N}$. Under what conditions on A, B, C, D is y at least k times continuously differentiable for every bounded u?

2.21 *Alternating system.* Assume that we have two matrices $A_1, A_2 \in \mathbb{R}^{n\times n}$, and consider the alternating system $\dot{x}(t) = A(t)x(t)$ with $A(t) \in \mathbb{R}^{n\times n}$ a time-varying matrix that switches between A_1 and A_2, that is,

$$\text{for all } t, \text{ either } A(t) = A_1 \text{ or } A(t) = A_2. \tag{2.27}$$

(a) Give an example of two asymptotically stable systems $\dot{x} = A_1 x$ and $\dot{x} = A_2 x$ for which $\dot{x}(t) = A(t)x(t)$ is *not* stable. [Hint: Take $n = 2$ and let the points where the system switches depend on $x(t)$.]

(b) Assume that A_1 and A_2 commute. Show that $\dot{x} = A(t)x$ is asymptotically stable for all $A(t)$ of the form (2.27) if and only if both $\dot{x} = A_1 x$ and $\dot{x} = A_2 x$ are asymptotically stable.

2.22 Let A be a 2×2 matrix with distinct eigenvalues $\lambda_1 \neq \lambda_2$. We can write A as

$$A = \lambda_2 P_1 + \lambda_1 P_2,$$

for

$$P_1 := \frac{1}{\lambda_2 - \lambda_1}(A - \lambda_1 I),$$
$$P_2 := \frac{1}{\lambda_1 - \lambda_2}(A - \lambda_2 I).$$

(a) Show that $P_1 P_2 = P_2 P_1$.

(b) Show that $P_1^2 = P_1$ en $P_2^2 = P_2$.

(c) Show that

$$e^{At} = e^{\lambda_2 t} P_1 + e^{\lambda_1 t} P_2.$$

(d) Find a similar method to determine e^{At} when the eigenvalues coincide.

2.23 *Discretized systems.* We are given the continuous-time system (2.8). Assume that the input u is piecewise constant of the form

$$u(t) = u(kT) \quad \text{for } kT \leq t < (k+1)T, \quad k \in \mathbb{Z}.$$

When we are only interested in the values of the output y at times kT, that is,

$$\tilde{y}[k] := y(kT), \qquad k \in \mathbb{Z},$$

it suffices to study the discrete-time system

$$\begin{aligned} \tilde{x}[k+1] &= F\tilde{x}[k] + G\tilde{u}[k], \\ \tilde{y}[k] &= C\tilde{x}[k] + D\tilde{u}[k]. \end{aligned}$$

(a) Show this and express F and G in terms of A and B.

(b) Determine F and G for $A = \left[\begin{smallmatrix} -2 & 0 \\ 0 & -3 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$.

Chapter 3

Controllability and Observability

In Chapter 2, we spent much time analyzing the autonomous system $\dot{x} = Ax$, in other words, the system $\dot{x} = Ax + Bu$ with input equal to zero, $u \equiv 0$. However, in applications we can often choose u any way we want (within reasonable limits), and thereby direct the behavior of x to some degree. This is called *controllability*. For example, it is because of the driver that a car does not go about its business autonomously but rather follows a trajectory determined by the driver.

Another fundamental notion from systems theory, which is closely connected to controllability, is *observability*. In order to steer the car successfully, we must of course keep our eyes and ears open. The question is then what we must watch, and what we must listen to. Is what we see and hear even sufficient to keep the car on the road? The systems theory abstraction of this idea begins with the question whether we can reconstruct, or *observe*, the internal variables (the state x) based on only the external variables (u, y).

3.1 Reachability

Consider the initially-at-rest system

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = 0 \tag{3.1}$$

and recall that the state $x(t)$ can be expressed explicitly as

$$x(t) = \int_0^t \mathrm{e}^{A(t-\tau)} Bu(\tau)\,\mathrm{d}\tau.$$

By reachability, we mean the possibility to reach, from the resting position $x(0) = 0$, any arbitrary state using a well-chosen input signal.

Definition 3.1.1 (Reachability). A system $\dot{x} = Ax + Bu$ is reachable if for every $x_1 \in \mathbb{R}^n$ and $x(0) = 0$, there is a $t_1 > 0$ and an input $u : [0, t_1] \to \mathbb{R}^{n_u}$ such that $x(t_1) = x_1$. □

We also say that "the pair (A, B)" is reachable. Reachability says that *eventually*, we can reach any desired state from $x(0) = 0$. It does not say that we should be able to do so during a previously imposed time horizon t_1, just that such a finite horizon t_1 exists (possibly depending on x_1).

Example 3.1.2 (Reachability). A simple example of an unreachable system is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u. \tag{3.2}$$

Because the A-matrix is diagonal and u can only influence the second state component x_2, it will be clear that x_1 cannot be influenced (controlled) by u. This system is therefore not reachable. Reachability is more difficult to analyze for the system

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}\frac{1}{2} & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix} u. \tag{3.3}$$

Here too, u cannot influence the first state component x_1 *directly*, but the difference with the previous system is that x_2 can influence the first component x_1, and because x_2 in turn can be influenced by u, it might still be possible to sent x_1 to a desired value. But if that is possible, can it also be done in such a way that, at the same time, x_2 is sent to a (different) desired value? □

We first study which states $x(t_1)$ can be reached at a given $t_1 > 0$. We denote the set of possible states $x(t_1)$ by $\mathbb{X}(t_1)$:

$$\mathbb{X}(t_1) = \left\{ \int_0^{t_1} \mathrm{e}^{A(t_1-\tau)} B u(\tau)\,\mathrm{d}\tau \mid u : [0,t_1] \to \mathbb{R}^{n_u} \right\}. \tag{3.4}$$

It is a subspace of $\mathbb{R}^n$ (see Exercise 3.2).

Lemma 3.1.3 (Unreachable states). *Let $t_1 > 0$ and $\eta \in \mathbb{R}^n$. The following four statements are equivalent:*

1. *$\eta \perp \mathbb{X}(t_1)$; that is, $\eta^{\mathrm{T}} x(t_1) = 0$ for all possible $x(t_1)$.*
2. *$\eta^{\mathrm{T}} \mathrm{e}^{At} B = 0$ for all $t \in [0, t_1]$.*
3. *$\eta^{\mathrm{T}} A^k B = 0$ for all $k = 0, 1, \ldots$.*
4. *$\eta^{\mathrm{T}} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0$.*

Proof. We prove (1) $\implies$ (2) $\implies$ (3) $\implies$ (4) $\implies$ (1).

(1) $\implies$ (2): We have $\eta^{\mathrm{T}} x(t_1) = 0$ for all u. This holds, in particular, for $u(\tau) = (\eta^{\mathrm{T}} \mathrm{e}^{A(t_1-\tau)} B)^{\mathrm{T}}$. For this input, we have

$$0 = \eta^{\mathrm{T}} x(t_1) = \int_0^{t_1} \eta^{\mathrm{T}} \mathrm{e}^{A(t_1-\tau)} B u(\tau)\,\mathrm{d}\tau = \int_0^{t_1} \|\eta^{\mathrm{T}} \mathrm{e}^{A(t_1-\tau)} B\|^2\,\mathrm{d}\tau.$$

This implies that $\eta^{\mathrm{T}} \mathrm{e}^{At} B = 0$ for all $t = t_1 - \tau \in [0, t_1]$.

(2) $\implies$ (3). Differentiating the equality $\eta^{\mathrm{T}} \mathrm{e}^{At} B = 0$ a number of times gives $\eta^{\mathrm{T}} A^k \mathrm{e}^{At} B = 0 \forall\, k = 0, 1, 2, \ldots$. For $t = 0$, this says that $\eta^{\mathrm{T}} A^k B = 0$.

(3) $\implies$ (4). Trivial.

(4) $\implies$ (1). We use the Cayley–Hamilton theorem. This theorem says that a matrix A satisfies its own characteristic equation. That is, if

$$\chi_A(\lambda) := \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_0 := \det(\lambda I - A),$$

then $\chi_A(A) = 0$, that is,

$$A^n = -(p_{n-1}A^{n-1} + p_{n-2}A^{n-2} + \cdots + p_1 A + p_0 I).$$

So A^n is a linear combination of lower powers of A. But then A^{n+1} is also a linear combination of $I, A, \ldots, A^{n-1}$ because

$$\begin{aligned} A^{n+1} = AA^n &= -A\big(p_{n-1}A^{n-1} + p_{n-2}A^{n-2} + \cdots + p_1 A + p_0 I\big) \\ &= -\big(p_{n-1}A^n + p_{n-2}A^{n-1} + \cdots + p_1 A^2 + p_0 A\big) \\ &= -p_{n-1}A^n + \text{linear combination of } A^{n-1}, \ldots, A \\ &= \text{linear combination of } A^{n-1}, \ldots, A, I. \end{aligned}$$

Continuing this way, we see that every A^k is a linear combination of $I, A, A^2, \ldots, A^{n-1}$. Consequently, every $A^k B$ is a linear combination of $B, AB, A^2B, \ldots, A^{n-1}B$.

Now, if $\eta^{\mathrm{T}} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0$, then the Cayley–Hamilton theorem implies that $\eta^{\mathrm{T}} A^k B = 0$ for all $k \geq 0$, and consequently that for every input, we have

$$\eta^{\mathrm{T}} x(t_1) = \int_0^{t_1} \eta^{\mathrm{T}} \mathrm{e}^{A(t_1-\tau)} B u(\tau)\,\mathrm{d}\tau = \int_0^{t_1} \sum_{k=0}^{\infty} \eta^{\mathrm{T}} A^k B \frac{(t_1-\tau)^k}{k!} u(\tau)\,\mathrm{d}\tau = 0.$$

In other words, η is orthogonal to every element of $\mathbb{X}(t_1)$. ∎

The set $\mathbb{X}(t_1)$ and the column space of the matrix $\begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix}$ therefore have the same orthogonal complement. But as $\mathbb{X}(t_1)$ is a subspace of $\mathbb{R}^n$, this means that $\mathbb{X}(t_1)$ is *equal* to this column space:

$$\mathbb{X}(t_1) = \mathrm{im}(\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}) := \{Bu_0 + ABu_1 + \cdots + A^{n-1}Bu_{n-1} \mid u_i \in \mathbb{R}^{n_u}\}.$$

Also, because this column space does not depend on t_1, the space $\mathbb{X}(t_1)$ is also independent of t_1 (provided $t_1 > 0$). Apparently, the states that are at all reachable, are reachable for every positive $t_1 > 0$, regardless of how small it is. The matrix

$$\mathscr{C} := \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \tag{3.5}$$

is called the controllability matrix. It follows from the above that the set of reachable states $x(t)$ for $t > 0$ is equal to the column space

$$\mathrm{im}(\mathscr{C}).$$

This is a subspace of $\mathbb{R}^n$ and is called the reachable subspace. The system is thus reachable if and only if $\mathrm{im}(\mathscr{C}) = \mathbb{R}^n$. This is the case if and only if $\mathscr{C}$ has full row rank (i.e. rank n).

The input u often consists of one element. In this case, B is a matrix with one column, and the controllability matrix $\mathscr{C}$ is therefore square. For square matrices, full row rank is equivalent to invertibility. Reachability can now be tested easily.

Example 3.1.4. The system (3.2) has state dimension $n = 2$ and input dimension $n_u = 1$. The controllability matrix $\mathscr{C}$ is then a 2×2 matrix, namely

$$\mathscr{C} = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}.$$

This matrix is not invertible, hence does not have full row rank. The system (3.2) is therefore not reachable (as we had already deduced in Example 3.1.2). The controllability matrix of the system (3.3) is equal to

$$\mathscr{C} = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

This matrix is invertible (and therefore has full row rank), and this system is therefore reachable. □

We summarize this reachability test in the following theorem. In this theorem, we also explicitly give an input that realizes the control objective.

Theorem 3.1.5 (Reachability). *Let $\mathscr{C}$ be the controllability matrix as in* (3.5)*. The following five statements are equivalent:*

1. (A, B) *is reachable.*

2. $\operatorname{im}(\mathscr{C}) = \mathbb{R}^n$.

3. $\mathscr{C}$ *has full row rank.*

4. *The controllability Gramian* $P(t) := \int_0^t \mathrm{e}^{A\tau} B B^{\mathrm{T}} \mathrm{e}^{A^{\mathrm{T}}\tau}\,\mathrm{d}\tau$, *is invertible for all* $t > 0$.

5. *The* controllability Gramian $P(t)$ *is invertible for some* $t > 0$.

If these hold, then every state x_1 *is reachable for every positive horizon* $t_1 > 0$ *and*

$$u_*(t) := B^{\mathrm{T}} \mathrm{e}^{A^{\mathrm{T}}(t_1 - t)} P^{-1}(t_1) x_1 \tag{3.6}$$

is one of the many inputs that achieves $x(t_1) = x_1$. *Moreover, the (squared) norm of this* u_* *is*

$$\|u_*\|^2 := \int_0^{t_1} u_*^{\mathrm{T}}(t) u_*(t)\,\mathrm{d}t = x_1^{\mathrm{T}} P^{-1}(t_1) x_1,$$

and no other input that achieves $x(t_1) = x_1$ *has a smaller norm.*

Proof. We prove (1) $\implies$ (2) $\implies$ (3) $\implies$ (4) $\implies$ (5) $\implies$ (1).

(1) $\implies$ (2) is trivial.

(2) $\implies$ (3): Proof by contradiction: If $\mathscr{C}$ has less than full row rank, then there exists an $\eta \in \mathbb{R}^n$ with $\eta \neq 0$ such that $\eta^{\mathrm{T}}\mathscr{C} = 0$. But then by Lemma 3.1.3, the nonzero η is orthogonal to the reachable subspace $\mathbb{R}^n$. This is impossible, so $\mathscr{C}$ has full row rank.

(3) $\implies$ (4): Let $t > 0$. We must prove that $P(t)\eta = 0$ implies $\eta = 0$. Suppose $P(t)\eta = 0$. We then also have $\eta^{\mathrm{T}} P(t)\eta = 0$ and therefore

$$0 = \eta^{\mathrm{T}} P(t)\eta = \int_0^t \eta^{\mathrm{T}} \mathrm{e}^{A\tau} B B^{\mathrm{T}} \mathrm{e}^{A^{\mathrm{T}}\tau}\eta\,\mathrm{d}\tau = \int_0^t \|\eta^{\mathrm{T}} \mathrm{e}^{A\tau} B\|^2\,\mathrm{d}\tau.$$

This is only possible if $\eta^{\mathrm{T}} \mathrm{e}^{A\tau} B$ is identical to the zero function (on $[0, t]$). As in Lemma 3.1.3, it follows that $\eta^{\mathrm{T}} A^k B = 0$ for all $k \geq 0$ and therefore, by the same lemma, that $\eta^{\mathrm{T}}\mathscr{C} = 0$. Since $\mathscr{C}$ has full row rank, this can only happen if $\eta = 0$.

(4) $\implies$ (5) is trivial.

(5) $\implies$ (1) follows by verifying that (3.6) holds (see Exercise 3.6).

In Exercise 3.6, you must show that $\|u_*\|^2 = x_1^{\mathrm{T}} P^{-1}(t_1) x_1$. It remains to prove that u_* has optimal norm. Suppose that u_1 is one of the inputs that achieve $x(t_1) = x_1$. It then follows by linearity that

$$\int_0^{t_1} \mathrm{e}^{A(t_1 - \tau)} B [u_1(\tau) - u_*(\tau)]\,\mathrm{d}\tau = x_1 - x_1 = 0.$$

Consequently,

$$\begin{aligned}
\int_0^{t_1} u_*^{\mathrm{T}}(\tau)[u_1(\tau) - u_*(\tau)]\,\mathrm{d}\tau &= \int_0^{t_1} x_1^{\mathrm{T}} P^{-1}(t_1) \mathrm{e}^{A(t_1-\tau)} B [u_1(\tau) - u_*(\tau)]\,\mathrm{d}\tau \\
&= x_1^{\mathrm{T}} P(t_1)^{-1} \int_0^{t_1} \mathrm{e}^{A(t_1-\tau)} B [u_1(\tau) - u_*(\tau)]\,\mathrm{d}\tau \\
&= 0.
\end{aligned}$$

(in the course *Linear Structures*, we would say that u_* and $u_1 - u_*$ are "orthogonal" to each other in a suitable inner product.) It thus follows that the norm of u_1 is at least that of u_*:

$$\begin{aligned}\|u_1\|^2 = \|u_* + (u_1 - u_*)\|^2 &= \int_0^{t_1} (u_* + (u_1 - u_*))^{\mathrm{T}}(u_* + (u_1 - u_*))\,\mathrm{d}\tau \\ &= \int_0^{t_1} u_*^{\mathrm{T}} u_* + 2\underbrace{u_*^{\mathrm{T}}(u_1 - u_*)}_{\text{integrates to 0}} + (u_1 - u_*)^{\mathrm{T}}(u_1 - u_*)\,\mathrm{d}\tau \\ &= \|u_*\|^2 + \|u_1 - u_*\|^2 \geq \|u_*\|^2.\end{aligned}$$

■

3.2 Controllability

Reachability is defined for systems with $x(0) = 0$. For systems with an arbitrary initial state $x(0) = x_0$ we have the following analogous definition.

Definition 3.2.1 (Controllability). A system $\dot{x} = Ax + Bu$ is controllable if for every pair of states $x_0, x_1 \in \mathbb{R}^n$, and $x(0) = x_0$, there is a $t_1 \geq 0$ and an input $u : [0, t_1] \to \mathbb{R}^{n_u}$ such that $x(t_1) = x_1$. □

Controllability obviously implies reachability, but for our type of system, $\dot{x} = Ax + Bu$, reachability also implies controllability. Indeed, if the system is reachable, then there also exists an input u_* that sends $x(0) = 0$ to $x(t_1) = x_1 - \mathrm{e}^{At_1}x_0$. This u_* sends $x(0) = x_0$ to $x(t_1) = x_1$ because

$$x(t_1) = \mathrm{e}^{At_1}x_0 + \underbrace{\int_0^{t_1} \mathrm{e}^{A(t_1-\tau)}Bu_*(\tau)\,\mathrm{d}\tau}_{x_1 - \mathrm{e}^{At_1}x_0} = x_1.$$

Hence controllability and reachability are equivalent. From here on, we will usually call it controllability.

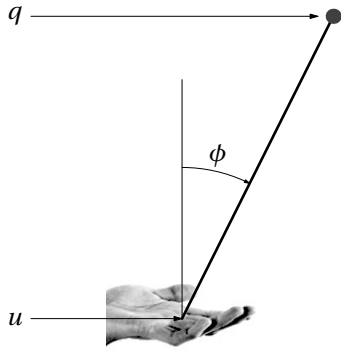

FIGURE 3.1: Inverted pendulum. See Example 3.2.2.

Example 3.2.2 (Juggler). Consider an ideal pendulum consisting of a mass m attached to a massless rigid stick of length ℓ_1, which can rotate freely in its pivot point (the palm of the hand); see Figure 3.1. Assume that we can set the position of the hand freely in one horizontal direction; this is the input variable u. We indicate the angle of the pendulum with respect to the vertical position by ϕ. Newton's second law gives the *nonlinear* DE

$$\cos(\phi)\ddot{u} + \ell\ddot{\phi} = g\sin(\phi). \tag{3.7}$$

The mass m does not play a role in the model (this DE is derived in Appendix A.3.) For $\phi \approx 0$ we have $\cos(\phi) \approx 1$ and $\sin(\phi) \approx \phi$. With these approximations the DE becomes a *linear*[1] DE,

$$\ddot{u} + \ell\ddot{\phi} = g\phi. \tag{3.8}$$

As state variables, we choose $q = u + \ell\phi$ and $v := \dot{q} = \dot{u} + \ell\dot{\phi}$. This q is the horizontal displacement of the top of the pendulum, and v is its velocity. With these choices, we can write (3.8) as the state representation

$$\begin{bmatrix} \dot{q} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{g}{\ell} & 0 \end{bmatrix} \begin{bmatrix} q \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{g}{\ell} \end{bmatrix} u. \tag{3.9}$$

The controllability matrix $\mathscr{C}$ is now given by

$$\mathscr{C} = \begin{bmatrix} 0 & -\frac{g}{\ell} \\ -\frac{g}{\ell} & 0 \end{bmatrix}. \tag{3.10}$$

Since $\det(\mathscr{C}) = g^2/\ell^2 \neq 0$, we have that (3.9) is controllable. In the neighborhood of the vertical position, we can therefore control both the *position* and the *velocity* of the pendulum by moving the pivot point. Not bad!

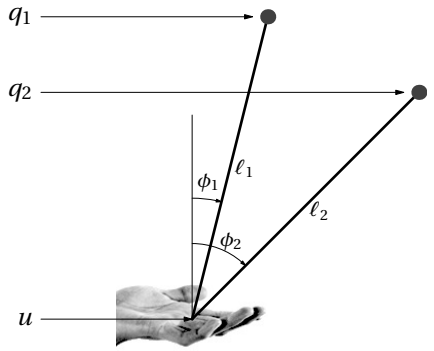

FIGURE 3.2: Two inverted pendula. See Example 3.2.2.

Now, assume that our juggler has *two* ideal pendula on his hand; see Figure 3.2. As above, we obtain linearized equations

$$\begin{aligned} \ddot{u} + \ell_1\ddot{\phi}_1 &= g\phi_1, \quad q_1 := u + \ell_1\phi_1, \\ \ddot{u} + \ell_2\ddot{\phi}_2 &= g\phi_2, \quad q_2 := u + \ell_2\phi_2, \end{aligned} \tag{3.11}$$

and the state representation

$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{v}_1 \\ \dot{v}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{g}{\ell_1} & 0 & 0 & 0 \\ 0 & \frac{g}{\ell_2} & 0 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ v_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -\frac{g}{\ell_1} \\ -\frac{g}{\ell_2} \end{bmatrix} u. \tag{3.12}$$

Let $\alpha := -\frac{g}{\ell_1}, \beta := -\frac{g}{\ell_2}$. Then the controllability matrix is

$$\mathscr{C} = \begin{bmatrix} 0 & \alpha & 0 & -\alpha^2 \\ 0 & \beta & 0 & -\beta^2 \\ \alpha & 0 & -\alpha^2 & 0 \\ \beta & 0 & -\beta^2 & 0 \end{bmatrix}. \tag{3.13}$$

[1] This is known as "linearization" of the DE. In the course *Ordinary Differential Equations* you will learn more about it.

(You may want to verify this.) This controllability matrix has rank 4 if and only if the matrix $\left[\begin{smallmatrix} \alpha & -\alpha^2 \\ \beta & -\beta^2 \end{smallmatrix}\right]$ has rank 2, or, equivalently, $\alpha\beta^2 - \alpha^2\beta \neq 0$, that is, $\alpha \neq \beta$. Hence the system is controllable if $\ell_1 \neq \ell_2$ and uncontrollable if $\ell_1 = \ell_2$. It may not be surprising that the system is uncontrollable if the pendula have the same length. That the system *is* controllable if the lengths differ is less intuitive, in fact it is quite spectacular! □

3.3 Kalman Controllability Decomposition & the Hautus Test

Consider a system in state z (not x) and suppose it has the following structure,

$$\begin{bmatrix} \dot{z}_c \\ \dot{z}_{uc} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} z_c \\ z_{uc} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u \tag{3.14}$$

with $z_c \in \mathbb{R}^q$ and $z_{uc} \in \mathbb{R}^{n-q}$, for some $q \leq n$. Because of the zero blocks in the lower left corner of the A-matrix and in the lower part of the B-matrix, it is intuitively clear that we cannot reach z_{uc} using u. Indeed, if $z_{uc}(0) = 0$, then it follows from $\dot{z}_{uc} = A_{22} z_{uc}$ that $z_{uc}(t) = 0$ for all t, regardless of the choice of u. This also follows from the reachable subspace, $\mathrm{im}(\mathscr{C}_z)$, because its controllability matrix equals

$$\mathscr{C}_z = \begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n-1}B_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{3.15}$$

(verify this yourself), and the reachable subspace $\mathrm{im}(\mathscr{C}_z)$ therefore satisfies

$$\mathrm{im}(\mathscr{C}_z) \subseteq \begin{bmatrix} \mathbb{R}^q \\ 0 \end{bmatrix}.$$

If z_{uc} begins in the origin, $z_{uc}(0) = 0$, then it stays in the origin, $z_{uc}(t) = 0 \forall\, t$, and then the system reduces to

$$\dot{z}_c = A_{11} z_c + B_1 u.$$

It is clear that the system (3.14) is uncontrollable if z_{uc} has at least one component (if $n - q > 0$). We will now show that using a state transformation $z = T^{-1}x$, *every* system can be written in the form above, and such that the subsystem $\dot{z}_c = A_{11} z_c + B_1 u$ *is* controllable.

Lemma 3.3.1 (Kalman controllability decomposition). *For every system $\dot{x} = Ax + Bu$, there exists a state transformation, $z = T^{-1}x$, such that in the new coordinates, we have*

$$\begin{bmatrix} \dot{z}_c \\ \dot{z}_{uc} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} z_c \\ z_{uc} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u \tag{3.16}$$

with (A_{11}, B_1) controllable. The form is known as the Kalman controllability decomposition.

Proof. If $\mathscr{C}$ has full row rank, then we do not need a transformation: $z = x$, $A_{11} = A$, and A_{22} is the "empty" matrix.

Now, suppose that $\mathscr{C}$ does not have full row rank, $q := \mathrm{rank}(\mathscr{C}) < n$. In this case, the reachable subspace is a q-dimensional subspace of $\mathbb{R}^n$. Take a basis $\{v_1, \ldots, v_q\}$ of this subspace, and extend it to a basis of $\mathbb{R}^n$,

$$(v_1, \ldots, v_q, \tilde{v}_{q+1}, \ldots, \tilde{v}_n).$$

Let z be the coordinate vector of x with respect to this new basis, that is,

$$x = Tz, \qquad T := \begin{bmatrix} v_1 & v_2 & \cdots & \tilde{v}_n \end{bmatrix}.$$

So $z = T^{-1}x$. By construction, x is in the reachable subspace iff z is of the form $z = (z_1, \ldots, z_q, 0, \ldots, 0)$. Hence the reachable subspace in terms of z is $\left[\begin{smallmatrix}\mathbb{R}^q\\0\end{smallmatrix}\right]$. The state transformation transforms the system $\dot{x} = Ax + Bu$ into $\dot{z} = T^{-1}ATz + T^{-1}Bu$, and the controllability matrix into

$$\mathscr{C}_z := \begin{bmatrix} T^{-1}B & T^{-1}AT\,T^{-1}B & \cdots & T^{-1}A^{n-1}B \end{bmatrix} = T^{-1}\mathscr{C}. \tag{3.17}$$

Since the reachable subspace for z is $\left[\begin{smallmatrix}\mathbb{R}^q\\0\end{smallmatrix}\right]$ and all columns of the new matrix $T^{-1}B$ are part of the reachable subspace, the new matrix $T^{-1}B$ must be of the form $\left[\begin{smallmatrix}B_1\\0\end{smallmatrix}\right]$. Consequently, the transformed system is of the form

$$\begin{bmatrix} \dot{z}_{\mathrm{c}} \\ \dot{z}_{\mathrm{uc}} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_{\mathrm{c}} \\ z_{\mathrm{uc}} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u.$$

If we would now have $A_{21} \neq 0$, then, because of the reachability of z_{c}, we could send $z_{\mathrm{c}}(t)$ (for some t) to a vector satisfying $A_{21}z_{\mathrm{c}}(t) \neq 0$. But then, we would have $\dot{z}_{\mathrm{uc}}(t) \neq 0$, which contradicts the fact that the reachable subspace is equal to $\left[\begin{smallmatrix}\mathbb{R}^q\\0\end{smallmatrix}\right]$. A contradiction, so $A_{21} = 0$. ∎

Example 3.3.2. Consider the system

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

The controllability matrix is then

$$\mathscr{C} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 2 & 5 \\ 0 & 1 & 2 \end{bmatrix}.$$

This has rank 2 and is therefore not invertible. The first two columns of $\mathscr{C}$ span the reachable subspace. As new basis $\{v_1, v_2, v_3\}$, we choose the first two columns of $\mathscr{C}$ and an arbitrary column vector v_3 that is independent of the first two, for example $v_3 = (0,0,1)^{\mathrm{T}}$. So

$$T = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

The transformed system $\dot{z} = T^{-1}ATz + T^{-1}Bu$ then becomes (verify this yourself)

$$\dot{z} = \left[\begin{array}{cc|c} 0 & 1 & -1 \\ 1 & 2 & 2 \\ \hline 0 & 0 & 1 \end{array}\right] z + \left[\begin{array}{c} 1 \\ 0 \\ \hline 0 \end{array}\right] u.$$

And the transformed controllability matrix becomes

$$\mathscr{C}_z = T^{-1}\mathscr{C} = \left[\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & 2 \\ \hline 0 & 0 & 0 \end{array}\right].$$

It confirms that the reachable subspace for z is $\left[\begin{smallmatrix}\mathbb{R}^2\\0\end{smallmatrix}\right]$. □

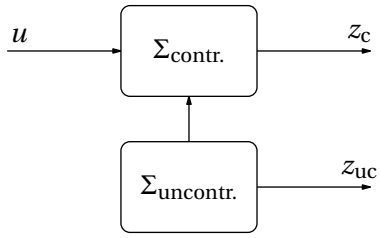

FIGURE 3.3: Kalman controllability decomposition.

Figure 3.3 illustrates this decomposition. The controllable part z_c is influenced by the input u and the uncontrollable part z_{uc}, while the uncontrollable part is influenced by nothing (apart from an initial condition $z_{uc}(0)$).

It should be clear that a state transformation $z = T^{-1}x$ does not change controllability: a controllable system with state x is still controllable in state $z = T^{-1}x$, and an *un*controllable system is still uncontrollable after a transformation. The importance of the Kalman controllability decomposition is, among other things, that it translates controllability into matrix properties, which are often easier to handle. A good example is the proof of the Hautus test[2]:

Theorem 3.3.3 (Hautus test). *The system $\dot{x} = Ax + Bu$ is controllable if and only if the $n \times (n + n_u)$ matrix depending on s,*

$$\begin{bmatrix} sI - A & B \end{bmatrix},$$

has full row rank for all $s \in \mathbb{C}$.

Proof. Suppose that $\begin{bmatrix} sI - A & B \end{bmatrix}$ does not have full row rank for some s. Then there exists an $s_0 \in \mathbb{C}$ and a nonzero vector $\eta \in \mathbb{C}^n$ such that $\eta^{\mathrm{T}} \begin{bmatrix} s_0 I - A & B \end{bmatrix} = 0$. In particular, η^{T} is a left eigenvector of the A-matrix: $\eta^{\mathrm{T}} A = s_0 \eta^{\mathrm{T}}$. This η is orthogonal to the reachable subspace, because

$$\eta^{\mathrm{T}} \mathscr{C} = \eta^{\mathrm{T}} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = \begin{bmatrix} \eta^{\mathrm{T}}B & s_0\eta^{\mathrm{T}}B & \cdots & s_0^{n-1}\eta^{\mathrm{T}}B \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The controllability matrix therefore does not have full row rank, and the system is uncontrollable.

Conversely, suppose that the system is uncontrollable. Since a state transformation does not change the rank of $\begin{bmatrix} sI - A & B \end{bmatrix}$ (see Exercise 3.7), we may assume without loss of generality that the system is in the form of the Kalman controllability decomposition (3.16). The matrix A_{22} of (3.16) is nonempty because the system is uncontrollable. Let η_{uc}^{T} be a left eigenvector of A_{22} and let s_0 be its eigenvalue. (That is, $\eta_{uc}^{\mathrm{T}} A_{22} = s_0 \eta_{uc}^{\mathrm{T}}$ and $\eta_{uc}^{\mathrm{T}} \neq 0$.) We then have

$$\begin{bmatrix} 0 & \eta_{uc}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} s_0 I - A_{11} & -A_{12} & B_1 \\ 0 & s_0 I - A_{22} & 0 \end{bmatrix} = 0.$$

Hence $\begin{bmatrix} s_0 I - A & B \end{bmatrix}$ does not have full row rank. ∎

Example 3.3.4. The system

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

[2] Malo Hautus was a professor at the University of Eindhoven. The test is also called the PBH test, after Popov, Belevitch, and Hautus. They discovered the test shortly after one another in 1966, 1968, and 1969, respectively.

is uncontrollable because

$$\mathscr{C} = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix},$$

and this is not invertible. This also follows from the Hautus test: consider the matrix

$$\begin{bmatrix} sI - A & B \end{bmatrix} = \begin{bmatrix} s+2 & -1 & 1 \\ 0 & s+1 & 1 \end{bmatrix}.$$

Since we are only interested in the rank, we may carry out elementary operations on the rows and columns. By subtracting the first row from the second, we obtain

$$\begin{bmatrix} s+2 & -1 & 1 \\ -s-2 & s+2 & 0 \end{bmatrix}. \tag{3.18}$$

The last row is zero if $s = -2$, so the rank of the matrix decreases for $s = -2$. Consequently, the system is uncontrollable. □

In the above example we could easily spot the rank loss of (3.18). For other more complicated systems that may be harder. However, what always should work is this: since $sI - A$ loses rank at precisely the eigenvalues s of A, the matrix $\begin{bmatrix} sI - A & B \end{bmatrix}$ can only lose rank if s is an eigenvalue of A. In the above example the eigenvalues of A are $s_1 = -1$ and $s_2 = -2$ and so to test for controllability we need only verify the rank of *two* matrices

$$\begin{bmatrix} (-1)I - A & B \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} (-2)I - A & B \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Clearly the rank of the second is less than 2 (and it happens at $s = -2$).

3.4 Observability

The second fundamental property of systems is *observability*. Loosely speaking, observability means that we can determine the state by looking at only the external behavior (u, y). The output now also plays a role. We consider systems of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{aligned} \tag{3.19}$$

Definition 3.4.1 (Observability). A system (3.19) is observable if there exists a $t_1 > 0$ such that for every triple of solutions (u_1, x_1, y_1), (u_2, x_2, y_2) of (3.19) with the same external behavior,

$$u_1(t) = u_2(t), \quad y_1(t) = y_2(t) \qquad \forall\, t \in [0, t_1],$$

also the state is the same,

$$x_1(t) = x_2(t) \qquad \forall\, t \in [0, t_1].$$

□

The system (3.19) is therefore observable if, from the knowledge of the input and output signals over a sufficiently long time interval $[0, t_1]$, we can uniquely determine the state signal on the interval $[0, t_1]$. The following $n_y n \times n$ matrix is crucial in the characterization of observability:

$$\mathscr{O} := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}. \tag{3.20}$$

This matrix $\mathscr{O}$ is called the observability matrix (compare this with the definition of the controllability matrix $\mathscr{C}$).

We first study the case where the external signals are zero at all times, $u(t) = 0, y(t) = 0$. This reduces the system (3.19) to

$$\begin{aligned} \dot{x}(t) &= Ax(t), \\ 0 &= Cx(t) \qquad \forall\, t > 0; \end{aligned}$$

that is,

$$Ce^{At}x_0 = 0 \qquad \forall\, t > 0.$$

For observability, this last equality must hold only for $x_0 = 0$. In general, x_0 does not need to be zero, and we define the t_1-unobservable subspace $\mathbb{X}^{\mathrm{uo}}(t_1)$ as those initial states for which the output is zero over $[0, t_1]$,

$$\mathbb{X}^{\mathrm{uo}}(t_1) := \{x_0 \in \mathbb{R}^n \mid Ce^{At}x_0 = 0 \forall\, t \in [0, t_1]\}. \tag{3.21}$$

Lemma 3.4.2 (Unobservable subspace). *Let $t_1 > 0$ and $\eta \in \mathbb{R}^n$. The following four statements are equivalent:*

1. $\eta \in \mathbb{X}^{uo}(t_1)$.
2. $Ce^{At}\eta = 0$ *for all* $t \in [0, t_1]$.
3. $CA^k\eta = 0$ *for all* $k = 0, 1, 2, \ldots$
4. $\eta \in \ker(\mathscr{O})$.

Proof. We prove (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) $\Longrightarrow$ (4) $\Longrightarrow$ (1). The proofs closely resemble those of Lemma 3.1.3 on reachability.

(1) $\Longrightarrow$ (2) is by definition of $\mathbb{X}^{\mathrm{uo}}(t_1)$. (2) $\Longrightarrow$ (3) and (3) $\Longrightarrow$ (4) are exercises (Exercise 3.22). This leaves (4) $\Longrightarrow$ (1): By the Cayley–Hamilton theorem, for every $k \geq 0$, the matrix A^k is a linear combination of $I, A, \ldots, A^{n-1}$. If $\eta \in \ker(\mathscr{O})$, then by definition $C\eta = 0, CA\eta = 0, \ldots, CA^{n-1}\eta = 0$, so by the Cayley–Hamilton theorem, $CA^k\eta = 0$ for all $k \geq 0$. It follows that

$$Ce^{At}\eta = C\big(I + tA + \tfrac{t^2}{2!}A^2 + \tfrac{t^3}{3!}A^3 + \cdots\big)\eta = C\eta + tCA\eta + \tfrac{t^2}{2!}CA^2\eta + \tfrac{t^3}{3!}CA^3\eta + \cdots = 0.$$

In other words, $\eta \in \mathbb{X}^{\mathrm{uo}}(t_1)$. ∎

The t_1-unobservable subspace $\mathbb{X}^{\mathrm{uo}}(t_1)$ is therefore equal to $\ker(\mathscr{O})$, and because the latter does not depend on t_1, the subspace $\mathbb{X}^{\mathrm{uo}}(t_1)$ is apparently independent of t_1 as well (provided $t_1 > 0$). We therefore have

$$\ker(\mathscr{O}) = \{x_0 \in \mathbb{R}^n \mid Ce^{At}x_0 = 0 \forall\, t > 0\}.$$

This is called the unobservable subspace. It consists of all states for which $y(t) = 0$ for all t if $u(t) = 0$ for all t. The following theorem should not come as a surprise.

Theorem 3.4.3 (Observability). *Let $\mathcal{O}$ be the observability matrix defined in* (3.20)*. The following three statements are equivalent:*

1. *The system is observable.*
2. $\ker(\mathcal{O}) = \{0\}$.
3. *The $\mathcal{O}$ has full column rank (rank n).*

Proof. We prove (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) $\Longrightarrow$ (1).

(1) $\Longrightarrow$ (2): By contradiction: If $\ker(\mathcal{O}) \neq \{0\}$, then in addition to $\{0\}$, the unobservable subspace also contains a nonzero vector x_0. Then $x(t) = 0$ and $x(t) = \mathrm{e}^{At}x_0$ both are consistent with zero external behavior for all time $((u(t), y(t)) = (0,0) \forall\, t)$. So the system is not observable.

(2) $\Longrightarrow$ (3). This is a standard result from linear algebra.

(3) $\Longrightarrow$ (1). Let $x_1, x_2 : [0, t_1] \to \mathbb{R}^n$ be two states with the same external behavior $u : [0, t_1] \to \mathbb{R}^{n_u}$, $y : [0, t_1] \to \mathbb{R}^{n_y}$. In particular, we can express y two ways:

$$y(t) = C\mathrm{e}^{At}x_1(0) + \int_0^t C\mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}t + Du(t),$$
$$y(t) = C\mathrm{e}^{At}x_2(0) + \int_0^t C\mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}t + Du(t).$$

The difference between these two expressions of $y(t)$ shows that

$$C\mathrm{e}^{At}[x_1(0) - x_2(0)] = 0 \qquad \forall\, t \in [0, t_1].$$

So $x_1(0) - x_2(0)$ is an element of the unobservable subspace $\ker(\mathcal{O})$. Since $\mathcal{O}$ has full column rank, this subspace is $\{0\}$; that is, $x_1(0) = x_2(0)$. But then $x_1(t) = \mathrm{e}^{At}x_0 + \int_0^t \mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}t = x_2(t)$ for all t. ∎

Even though the definition of observability also considers the input u, observability apparently does not depend on the matrices B and D. We therefore often speak of the observability of the matrix *pair* (A, C).

Example 3.4.4 (Unobservable system). Consider the system

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x,$$
$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$$

For every α, the state $x = \left[\begin{smallmatrix} 0 \\ \alpha \end{smallmatrix}\right]$ is a constant solution of this system (verify this yourself) and it gives $y(t) = 0$ for all t. So the state cannot be observed on the basis of the output y. This also follows from Theorem 3.4.3 because

$$\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

and this is not invertible. The nullspace, $\ker(\mathcal{O})$, is the unobservable subspace. In this case, this is $\left[\begin{smallmatrix} 0 \\ \mathbb{R} \end{smallmatrix}\right]$, so all vectors of the form $\left[\begin{smallmatrix} 0 \\ \alpha \end{smallmatrix}\right]$. □

The output y often consists of one element. In that case C is a matrix with one row and $\mathcal{O}$ is square. Then the system is observable if and only if $\mathcal{O}$ is invertible.

Example 3.4.5 (Observability in the inverted pendulum). Consider the system with two inverted pendula in Example 3.2.2. If we can only observe one of the two angles ϕ_1 and ϕ_2, then the system is unobservable. Take, for example, the output

$$y = \phi_1 = \begin{bmatrix} \frac{1}{\ell_1} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ v_1 \\ v_2 \end{bmatrix} - \frac{1}{\ell_1} u. \tag{3.22}$$

Then

$$\mathcal{O} = \begin{bmatrix} \frac{1}{\ell_1} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\ell_1} & 0 \\ \frac{g}{\ell_1^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{g}{\ell_1^2} & 0 \end{bmatrix} \tag{3.23}$$

and so (verify this yourself) $\operatorname{rank}(\mathcal{O}) = 2 < 4$. If, on the other hand, we take for the output y the *difference* between the two angles, that is,

$$y = \phi_1 - \phi_2 \approx \begin{bmatrix} \frac{1}{\ell_1} & \frac{-1}{\ell_2} & 0 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ v_1 \\ v_2 \end{bmatrix} + (\frac{1}{\ell_2} - \frac{1}{\ell_1}) u, \tag{3.24}$$

then the observability matrix equals

$$\mathcal{O} = \begin{bmatrix} \frac{1}{\ell_1} & \frac{-1}{\ell_2} & 0 & 0 \\ 0 & 0 & \frac{1}{\ell_1} & \frac{-1}{\ell_2} \\ \frac{g}{\ell_1^2} & \frac{-g}{\ell_2^2} & 0 & 0 \\ 0 & 0 & \frac{g}{\ell_1^2} & \frac{-g}{\ell_2^2} \end{bmatrix}. \tag{3.25}$$

It is clear that the rank of this matrix decreases if $\ell_1 = \ell_2$. This is also the only way the rank can decrease (since the determinant of $\mathcal{O}$ is equal to $-g^2(\ell_1 - \ell_2)^2/(\ell_1\ell_2)^4$). Therefore, in the case $\ell_1 \neq \ell_2$, applying any input $u(t)$, $t \in [0, t_1]$, and recording the resulting output $y(t)$, $t \in [0, t_1]$ equal to the difference between the angles, allows us to uniquely determine the complete 4-dimensional state $x(t) = [q_1(t), q_2(t), v_1(t), v_2(t)]^{\mathrm{T}}$. Not bad! □

In analogy with he Kalman controllability decomposition, there is an observability decomposition.

Lemma 3.4.6 (Kalman observability decomposition). *Every system $\dot{x} = Ax + Bu$, $y = Cx + Du$ is isomorphic[3] to a system of the form*

$$\begin{aligned} \begin{bmatrix} \dot{z}_o \\ \dot{z}_{uo} \end{bmatrix} &= \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_o \\ z_{uo} \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u, \\ y &= \begin{bmatrix} C_1 & 0 \end{bmatrix} \begin{bmatrix} z_o \\ z_{uo} \end{bmatrix} + Du \end{aligned} \tag{3.26}$$

with (C_1, A_{11}) observable. This form is called the Kalman observability decomposition.

Proof. (A, C) is observable if and only if $(A^{\mathrm{T}}, C^{\mathrm{T}})$ is controllable. Apply the Kalman controllability decomposition to $(A^{\mathrm{T}}, C^{\mathrm{T}})$. ■

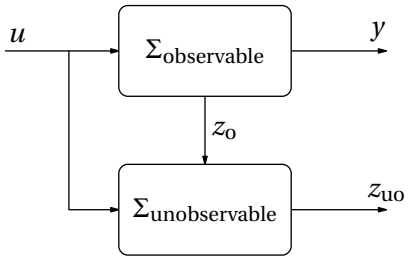

FIGURE 3.4: Kalman observability decomposition.

Figure 3.4 illustrates this decomposition. The input u and the observable state z_{o} may influence the unobservable state z_{uo}, but this unobservable z_{uo} does not influence the output y. The proof of Lemma 3.4.6 does not immediately yield an algorithm for finding this decomposition. The next example explains how to find it.

Example 3.4.7 (How to determine the Kalman observability decomposition). Consider the system without input

$$\dot{x} = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 1 & 3 & 1 \end{bmatrix} x,$$
$$y = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} x.$$

The observability matrix of this system is

$$\mathscr{O}_x = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 3 & 5 & 2 \end{bmatrix},$$

and it is not invertible. (We added a subscript x to $\mathscr{O}$ to stress that this is with respect to state x.) It is not difficult to show that $\ker(\mathscr{O}_x)$ is spanned by $(1,-1,1)^{\mathrm{T}}$. This is how we choose the *last* column of T. Next we simply choose the other columns of T in such a way that T is invertible. For example,

$$T = \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{array}\right].$$

Now, by construction, the unobservable subspace $\ker(\mathscr{O}_x)$ is spanned by the last column of T. So in the new coordinates, $z := T^{-1}x$, it is spanned by $z = (0,0,1)^{\mathrm{T}}$. In these new coordinates, we have the desired decomposition (verify this yourself)

$$\dot{z} = \left[\begin{array}{cc|c} -1 & -1 & 0 \\ 2 & 3 & 0 \\ \hline 1 & 3 & -1 \end{array}\right] z,$$
$$y = \left[\begin{array}{cc|c} 1 & 1 & 0 \end{array}\right] z$$

and the observability matrix becomes

$$\mathscr{O}_z = \mathscr{O}_x T = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 3 & 5 & 0 \end{array}\right].$$

[3] See § 2.2.3

As expected, the unobservable subspace $\ker(\mathscr{O}_z)$ in the new state z are the vectors of the form $z = (0,0,z_3)^{\mathrm{T}}$. □

The observability matrix of (3.26) is

$$\mathscr{O}_z = \begin{bmatrix} C_1 & 0 \\ C_1 A_{11} & 0 \\ \vdots & \vdots \\ C_1 A_{11}^{n-1} & 0 \end{bmatrix}$$

and the unobservable subspace, $\ker(\mathscr{O}_z)$, thus is the set of states of the form

$$z = \begin{bmatrix} 0 \\ z_{\mathrm{uo}} \end{bmatrix}.$$

To conclude, we formulate the Hautus test for observability.

Theorem 3.4.8 (Hautus test—observability). *The system $\dot{x} = Ax + Bu, y = Cx + Du$ is observable if and only if the $(n + n_y) \times n$ matrix depending on s,*

$$\begin{bmatrix} sI - A \\ C \end{bmatrix},$$

has full column rank for all $s \in \mathbb{C}$.

Proof. See Exercise 3.23. ∎

3.5 Canonical Representations

We present a number of canonical representations for systems that are either controllable or observable. We restrict ourselves to systems with a single input and a single output.

The first canonical form has a controllability matrix equal to the identity.

Lemma 3.5.1. *Suppose $n_u = 1$. For every controllable system $\dot{x} = Ax + Bu$ there is a unique state transformation $v = T^{-1}x$ for which the system takes the form*

$$\dot{v} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & -p_0 \\ 1 & \ddots & & \vdots & -p_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -p_{n-1} \end{bmatrix} v + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} u. \tag{3.27}$$

This has controllability matrix $\mathscr{C}_v = I$, and the p_i are the coefficients of the characteristic polynomial of the A-matrix: $\det(sI - A) = s^n + p_{n-1}s^{n-1} + \cdots + p_1 s + p_0$.

Proof. A state transformation $v = T^{-1}x$ transforms the controllability matrix $\mathscr{C}_x$ into $\mathscr{C}_v = T^{-1}\mathscr{C}_x$ (see (3.17)). It is easy to see that the controllability matrix of (3.27) is $\mathscr{C}_v = I$ so *if* a tranformation to (3.27) exists then the T is unique: $T = \mathscr{C}_x$.

We now show that $T = \mathscr{C}_x$ indeed does the job. Then by construction $\mathscr{C}_v = I$. Since the first column of $\mathscr{C}_v = I$ is the transformed B, this transformed B must be the column vector $(1,0,\ldots,0)$. Likewise, the second through nth columns of $\mathscr{C}_v = I$ are equal to the first through next-to-last columns of the transformed A (verify this yourself). Denote the last column of the transformed A-matrix as $(-p_0,\ldots,-p_{n-1})$. We have seen in Exercise 2.7 that these p_i are the coefficients of the characteristic polynomial of the matrix. Since state transformations do not change characteristic polynomials this is also the characteristic polynomial of A. ∎

In itself, this canonical form is of limited use, but it allows us to deduce the important *controller canonical form*, formulated next. At first glance, this closely resembles the previous form, but the A-matrix is transposed, and the B-matrix is ordered differently.

Theorem 3.5.2 (Controller canonical form). *Suppose $n_u = 1$. Every controllable system $\dot{x} = Ax + Bu$ is isomorphic to a system of the form*

$$\dot{z} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix} z + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} u. \tag{3.28}$$

This is called the controller canonical form. This form is unique: the p_i are the coefficients of the characteristic polynomial $\det(sI - A) = s^n + p_{n-1}s^{n-1} + \cdots + p_0$*, and $z = T^{-1}x$ where*

$$T = \mathscr{C}_x \mathscr{C}_z^{-1}.$$

Here $\mathscr{C}_x$ and $\mathscr{C}_z$ are the controllability matrices of $\dot{x} = Ax + Bu$ and the system (3.28)*, respectively.*

Proof. Define the p_i from the characteristic polynomial of the matrix A: $\det(sI - A) = s^n + p_{n-1}s^{n-1} + \cdots + p_0$. Since the matrix of (3.28) is a companion matrix its characteristic polynomial equals that of A.

Because the controller canonical form (3.28) is controllable (see Exercise 3.25), by Lemma 3.5.1 it is isomorphic to (3.27) through $v = \mathscr{C}_z^{-1}z$. The controllable system $\dot{x} = Ax + Bu$ is also isomorphic to (3.27) (through $v = \mathscr{C}_x^{-1}x$). So system $\dot{x} = Ax + Bu$ is isomorphic to (3.28), and $z = \mathscr{C}_z v = \mathscr{C}_z \mathscr{C}_x^{-1}x$ is the desired transformation from x to z. ∎

The computation of $T = \mathscr{C}_x \mathscr{C}_z^{-1}$ can be laborious. It is just a bit easier to determine this T using that

$$T = \begin{bmatrix} \eta \\ \eta A \\ \vdots \\ \eta A^{n-1} \end{bmatrix}^{-1} \quad \text{with} \quad \eta := \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathscr{C}_x^{-1}. \tag{3.29}$$

This allows us to avoid having to determine $\mathscr{C}_z$. (Formula (3.29) is derived in Appendix A.4.)

Example 3.5.3 (Construction of a controller canonical form). Consider

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

The controllability matrix and its inverse are

$$\mathscr{C}_x = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}, \qquad \mathscr{C}_x^{-1} = \begin{bmatrix} 0 & 1 \\ 1/4 & -1/4 \end{bmatrix}.$$

The row vector η is defined as the last row of $\mathscr{C}_x^{-1}$, so $\eta = \begin{bmatrix} 1/4 & -1/4 \end{bmatrix}$. We can now compute the inverse of T from (3.29),

$$T^{-1} = \begin{bmatrix} 1/4 & -1/4 \\ 3/4 & 1/4 \end{bmatrix} \quad \text{and therefore} \quad T = \begin{bmatrix} 1 & 1 \\ -3 & 1 \end{bmatrix}.$$

The matrices of the controller canonical form (including output) now follow:

$$\begin{bmatrix} A_z & B_z \\ C_z & 0 \end{bmatrix} = \begin{bmatrix} T^{-1}AT & T^{-1}B \\ CT & 0 \end{bmatrix}.$$

This gives

$$\begin{aligned} \dot{z} &= \begin{bmatrix} 0 & 1 \\ -8 & 3 \end{bmatrix} z + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \\ y &= \begin{bmatrix} 1 & 1 \end{bmatrix} z. \end{aligned}$$

Actually only the C_z-matrix needs to be computed here because the B_z-matrix we know to be of the form $\begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$ (as a column) and the A_z-matrix is a companion matrix whose coefficients in the bottom row are easily derived from the characteristic polynomial of the original matrix: $\det(\lambda I - A) = \det\left[\begin{smallmatrix} \lambda-1 & -3 \\ 2 & \lambda-2 \end{smallmatrix}\right] = \lambda^2 - 3\lambda + 8$. □

In analogy with the controller canonical form, we have the observer canonical form.

Lemma 3.5.4 (Observer canonical form). *Suppose $n_u = n_y = 1$. Every observable system $\dot{x} = Ax + Bu, y = Cx$ is isomorphic to a system of the form*

$$\begin{aligned} \dot{z} &= \begin{bmatrix} 0 & \cdots & \cdots & 0 & -p_0 \\ 1 & \ddots & & \vdots & -p_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -p_{n-1} \end{bmatrix} z + \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ \vdots \\ q_{n-1} \end{bmatrix} u, \\ y &= \begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} z. \end{aligned} \tag{3.30}$$

This is called the observer canonical form. This form is unique: the p_i are the coefficients of the characteristic polynomial of A, and $z = T^{-1}x$ for $T = \mathcal{O}_x^{-1}\mathcal{O}_z$, where $\mathcal{O}_x$ and $\mathcal{O}_z$ are, respectively, the observability matrices of $\dot{x} = Ax + Bu, y = Cx$ and the system (3.30). This T can also be determined using

$$T = \begin{bmatrix} \eta & A\eta & \cdots & A^{n-1}\eta \end{bmatrix} \quad \text{in which } \eta = \mathcal{O}_x^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Proof. The proof is analogous to that for the controllable canonical form: use that the system is observable if and only if the transposed system

$$\begin{cases} \dot{\tilde{x}} &= A^{\mathrm{T}}\tilde{x} + C^{\mathrm{T}}\tilde{u} \\ \tilde{y} &= B^{\mathrm{T}}\tilde{x} \end{cases}$$

is controllable. See also Appendix A.4. ■

Note that (3.30) is the state representation of the differential equation

$$y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_0 y = q_{n-1}u^{(n-1)} + \cdots + q_0 u \tag{3.31}$$

(Lemma 2.4.2). So the theorem above says that in an observable system $\dot{x} = Ax + Bu, y = Cx$ the relation between input and output can also be represented simply using an ordinary differential equation (3.31)! However, the main result of this theorem remains that observable systems can always be transformed into this special form. Due to its special structure, the observer canonical form is easier to analyze. We use this in the next chapter.

3.6 Exercises

3.1 Comprehension questions (on the whole chapter). Prove or give a counterexample.

(a) If in a system with $x(t) \in \mathbb{R}^2$, there exist an input that sends $x(0) = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ to $x(1) = \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$ and an input that sends $x(0) = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ to $x(1) = \left[\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right]$, then the system is controllable?

(b) If in a system with $x(t) \in \mathbb{R}^2$, there exists an input that sends $x(0) = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ to $x(1) = \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$ and an input that sends $x(0) = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ to $x(100) = \left[\begin{smallmatrix} 1 \\ -2 \end{smallmatrix}\right]$, then the system is controllable?

(c) If $C = 0 \in \mathbb{R}^{1\times n}$, then the system is not observable?

(d) Let $A \in \mathbb{R}^{n\times n}$. If the rank of $s_0 I - A$ is less than $n-1$, then (A, C) is unobservable?

(e) Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{1\times n}$. If the rank of $s_0 I - A$ is less than $n-1$, then (A, C) is unobservable?

(f) If (A, C_1) are (A, C_2) are observable, then $(A, C_1 + C_2)$ is also observable?

3.2 Let $t_1 > 0$. Show that $\mathbb{X}(t_1)$ as defined in (3.4) is a subspace of $\mathbb{R}^n$.

3.3 *Controllability.* Determine the controllability of the system $\dot{x} = Ax + Bu$ for the following pairs:

(a) $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

(b) $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

(c) $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

3.4 *Controllability.* Let M and N be matrices with the same number of rows, and assume that M is square. Show that the system

$$\dot{x} = \begin{bmatrix} 0 & M \\ M & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ N \end{bmatrix} u$$

is controllable if and only if M is invertible and

$$\dot{z} = M^2 z + Nu$$

is controllable.

3.5 *Controllability.* Determine the controllability of the systems $\dot{x} = Ax + Bu$ for the following pairs:

(a) $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

(b) $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

(c) $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

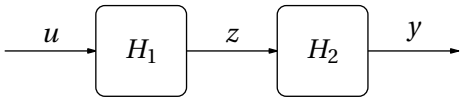

FIGURE 3.5: Serial interconnection.

3.6 Suppose $x_0 = 0$. Show that the u_* from (3.6) ensures that $x(t_1) = x_1$. Also show that $\|u_*\|^2 = x_1^{\mathrm{T}} P^{-1}(t_1) x_1$.

3.7 It is intuitively clear that the system $\dot{x} = Ax + Bu$ and transformed system $\dot{z} = T^{-1}ATz + T^{-1}Bu$ are either both controllable or both uncontrollable. In terms of the Hautus test this means that $\begin{bmatrix} sI - A & B \end{bmatrix}$ has full row rank for all $s \in \mathbb{C}$ iff $\begin{bmatrix} sI - T^{-1}AT & T^{-1}B \end{bmatrix}$ has full row rank for all $s \in \mathbb{C}$.

Specifically, show that for any given $s \in \mathbb{C}$, the matrix $\begin{bmatrix} sI - A & B \end{bmatrix}$ does not have full row rank iff $\begin{bmatrix} sI - T^{-1}AT & T^{-1}B \end{bmatrix}$ does not have full row rank. [Hint: use that a matrix W does not have full row rank iff $\eta^{\mathrm{T}} W = 0$ for some nonzero vector η.]

3.8 *Controllability.* Give a controllable system $\dot{x} = Ax + Bu$ such that for every column b of B, the system (A, b) is uncontrollable.

3.9 *Computation of the control signal u.* Compute the input signal $u : [0,1] \to \mathbb{R}$ for the system

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

that sends $x(0) = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ to $x(1) = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ with minimal $\|u\|^2$.

3.10 Determine the Kalman controllability decomposition of $\dot{x} = Ax + Bu$ for the following pairs:

(a) $A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

(b) $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

(c) $A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ -2 & 1 & 2 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

3.11 In analogy with the linear case, we call a nonlinear system

$$\dot{x}(t) = f(x(t), u(t))$$

controllable if for every pair of states x_0, x_1, there exists a $t_1 \geq 0$ and an input u such that if $x(0) = x_0$ then $x(t_1) = x_1$.

Is the nonlinear system $\dot{x} = -x + u^2$ controllable?

3.12 *Diagonalizable matrices.* Suppose that (A, B) is controllable and $n_u = 1$, and suppose that A has an eigenvalue of multiplicity more than 1. Can A be diagonalizable?

3.13 *Controllability of a serial interconnection.* Consider the serial interconnection of Figure 3.5, with subsystems H_1 and H_2 given by

$$\begin{aligned} H_1: &\quad \dot{x}_1 = A_1 x_1 + B_1 u, \quad z = C_1 x_1, \\ H_2: &\quad \dot{x}_2 = A_2 x_2 + B_2 z, \quad y = C_2 x_2. \end{aligned}$$

If the interconnected system (with input u, output y, and state $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$) is controllable is, are the subsystems H_1 and H_2 necessarily controllable?

3.14 Prove that (A, B) is uncontrollable if and only if there exists a $C \neq 0$ such that $Ce^{At}B = 0$ for all t.

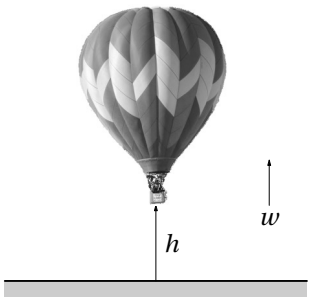

FIGURE 3.6: Hot air balloon. See Exercise 3.15.

3.15 *Hot air balloon.* The linearized equations for the motion of a hot air balloon are

$$\begin{aligned} \dot{\theta}(t) &= -\frac{1}{\tau_1}\theta(t) + u(t), \\ \dot{v}(t) &= -\frac{1}{\tau_2} v(t) + \sigma\theta(t) + \frac{1}{\tau_2} w(t), \\ \dot{h}(t) &= v(t). \end{aligned}$$

Here θ is the temperature in the balloon, u is the added heat, w is the speed of wind, h is the elevation of the balloon, and v is the vertical velocity of the balloon, and τ_1, τ_2, σ are postive paramters; see Figure 3.6.

(a) Is the system controllable at a given wind speed w?

(b) Is the system controllable if we view u and w as control variables?

3.16 Let E be an invertible $n \times n$ matrix. Furthermore, let, as usual, A be an $n \times n$ matrix and B be a matrix with as many rows as A. Show that the system described by the implicit state representation

$$E\dot{x}(t) = Ax(t) + Bu(t)$$

is controllable if and only if

$$\begin{bmatrix} sE - A & B \end{bmatrix}$$

has full row rank for all $s \in \mathbb{C}$.

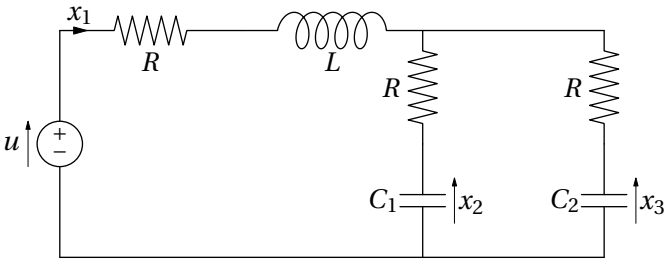

FIGURE 3.7: *RLC* circuit. See Exercise 3.17.

3.17 *Controllability of an RLC circuit.* Consider the *RLC* circuit of Figure 3.7. We take the voltage u over the voltage source as input. Straightforward modeling gives the model

$$\begin{bmatrix} L & RC_1 & 0 \\ 0 & RC_1 & -RC_2 \\ 0 & C_1 & C_2 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -R & -1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u$$

with u the voltage over the voltage source, $x_1 = i$ the current through the inductor, and x_2, x_3, respectively, the voltage over the capacitors C_1 and C_2. Assume that all constants R, C_1, C_2, L are greater than zero.

(a) Under which conditions on R, C_1, and C_2 is the system controllable? (Hint: Use the previous exercise.)

(b) Interpret your findings.

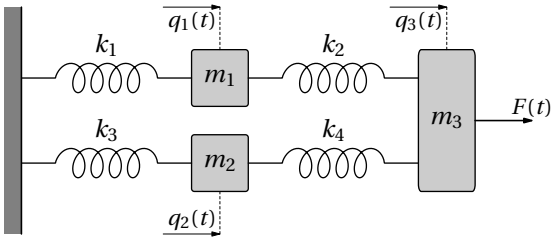

FIGURE 3.8: Spring-mass system. See Exercise 3.18.

3.18 *Controllability of a spring-mass system.* Consider the spring-mass system of Figure 3.8. Two masses m_1 and m_2 are attached to a wall (on the left) by springs with spring constants k_1 and k_3, and on the right to a mass m_3 by springs with spring constants k_2 and k_4. We can exert a force F on the mass m_3. The positions of the three masses with respect to their equilibrium points are denoted by q_i.

From Newton's second law and Hooke's law, we obtain the equations of motion

$$\begin{aligned} m_1\ddot{q}_1 &= -k_1 q_1 + k_2(q_3 - q_1), \\ m_2\ddot{q}_2 &= -k_3 q_2 + k_4(q_3 - q_2), \\ m_3\ddot{q}_3 &= -k_2(q_3 - q_1) - k_4(q_3 - q_2) \; + \; F. \end{aligned}$$

(a) Write this system in the form $\dot{x} = Ax + Bu$ with $u = F$ and state $x = \left[q_1, \dot{q}_1, q_2, \dot{q}_2, q_3, \dot{q}_3\right]^{\mathrm{T}}$.

(b) Suppose that $k_1 = k_2 = k_3 = k_4$. Under which conditions on the m_i is the system controllable? Interpret your findings.

(c) Suppose, once more, that $k_1 = k_2 = k_3 = k_4$. Choose q_3 as output. Under which conditions on the m_i is the system controllable? Interpret your findings.

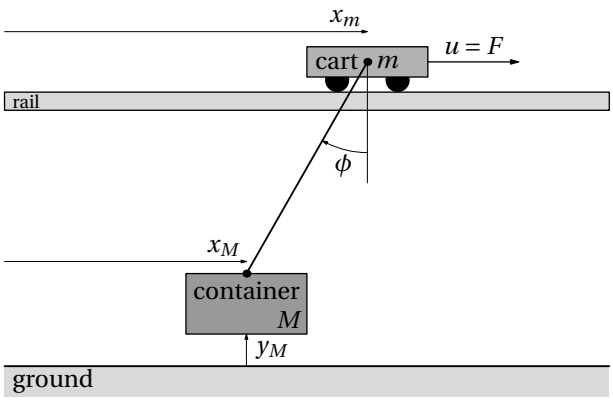

FIGURE 3.9: Container transfer. See Exercise 3.19.

3.19 *Container transfer.* Consider the problem of transferring containers (Figure 3.9). As input, we take the force we can exert on the cart, $u = F$. The linearized system is

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{M}{m}g & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{M+m}{Lm}g & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{m} \\ 0 \\ \frac{1}{Lm} \end{bmatrix} u \tag{3.32}$$

with $x = \left[x_m, \dot{x}_m, \phi, \dot{\phi}\right]^{\mathrm{T}}$.

(a) Is the linearized system controllable?

(b) Is the linearized system observable if we take only the position $y = x_m$ of the cart as output?

(c) In order to control the container using u, is it necessary to observe not only the position of the cart, but also that of the container (for example with cameras)?

3.20 *Observability.* Determine the observability of the systems $\dot{x} = Ax,\ y = Cx$ for the following pairs:

(a) $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$

(b) $A = \begin{bmatrix} 1 & 5 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$

(c) $A = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$

(d) $A = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n \times n}$, $C = \begin{bmatrix} 1 & 2 & \cdots & n \end{bmatrix}$

3.21 Determine the Kalman observability decomposition of $\dot{x} = Ax, y = Cx$ for the following pairs:

(a) $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 2 & 2 \end{bmatrix}$

(b) $A = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$, $C = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$

3.22 *Observability.* Prove the implications (2) $\Longrightarrow$ (3) and (3) $\Longrightarrow$ (4) of Lemma 3.4.2

3.23 *Hautus test for observability*

(a) Prove that the system (3.19) is observable if and only if $\operatorname{rank}\begin{bmatrix} sI-A \\ C \end{bmatrix} = n$ for all $s \in \mathbb{C}$.

(b) Under what conditions on $c_1, c_2, \ldots, c_n \in \mathbb{R}$ is the following system observable?

$$\dot{x} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n \end{bmatrix} x$$
$$y = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} x.$$

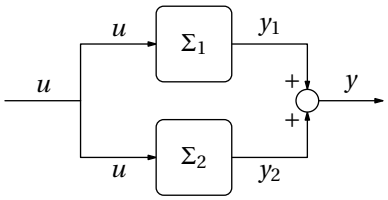

FIGURE 3.10: Parallel interconnection. See Exercise 3.24.

3.24 *Observability of a parallel interconnection.* Consider the configuration in Figure 3.10, with subsystems Σ_1 and Σ_2 that are both observable and controllable, with state representations

$$\Sigma_1: \quad \dot{x}_1 = A_1 x_1 + B_1 u, \quad y_1 = C_1 x_1,$$
$$\Sigma_2: \quad \dot{x}_2 = A_2 x_2 + B_2 u, \quad y_2 = C_2 x_2.$$

(a) Explain in words that the system with input u, output $y = y_1 + y_2$, and state $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is unobservable if $A_1 = A_2$ and $C_1 = C_2$.

(b) Prove this using Theorem 3.4.3.

3.25 Show that the controller canonical form (3.28) is controllable.

3.26 Prove the observability of (3.30) using the Hautus test (Theorem 3.4.8).

3.27 *Canonical forms.* Consider the system

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u, \\ y &= \begin{bmatrix} -1 & 2 \end{bmatrix} x. \end{aligned} \tag{3.33}$$

(a) Determine the controller canonical form of this system:

$$\begin{aligned} \dot{z} &= \begin{bmatrix} 0 & 1 \\ ? & ? \end{bmatrix} z + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \\ y &= \begin{bmatrix} ? & ? \end{bmatrix} z. \end{aligned} \tag{3.34}$$

(b) Determine the observer canonical form of the system.

Tougher Exercises

3.28 Suppose $n_u = n_y = 1$. If

$$\dot{x} = Ax + Bu,$$
$$y = Cx$$

is controllable and observable, is it isomorphic to

$$\dot{z} = A^{\mathrm{T}} z + C^{\mathrm{T}} u,$$
$$y = B^{\mathrm{T}} z?$$

3.29 *Reachability.* Consider Theorem 3.1.5.

(a) Use the fact that $\|u_*\|$ has minimal norm to prove that $x_1^{\mathrm{T}} P^{-1}(t) x_1$ is non-increasing as a function of t.

(b) Is $z^{\mathrm{T}} P(t) z$ non-increasing as a function of t?

(c) Suppose that A is stable. Does $\lim_{t\to\infty} P(t)$ exist?

(d) Show that $\dot{P}(t) = BB^{\mathrm{T}} + AP(t) + P(t)A^{\mathrm{T}}$.

(e) Show that if A is stable, then $P_\infty := P(\infty)$ satisfies $AP_\infty + P_\infty A^{\mathrm{T}} + BB^{\mathrm{T}} = 0$

3.30 *Dead beat control.* Dead beat control deals with constructing inputs u that bring the state to zero within a finite amount of time. We now consider discrete-time systems.

(a) Consider the *discrete*-time system

$$x[t+1] = Ax[t] + Bu[t], \qquad t \in \mathbb{Z}.$$

Show that "dead beat control" is possible for every initial state $x[0] \in \mathbb{R}^n$ if and only if

$$\begin{bmatrix} sI - A & B \end{bmatrix}$$

has full row rank for all $s \in \mathbb{C}, s \neq 0$.

(b) Does this same condition hold for continuous-time systems $\dot{x} = Ax + Bu$?

3.31 *Observing using outputs.* Consider the observable system with a single input and a single output

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

Prove that we can determine the state using *only* the knowledge of the output $y(t)$ for $t \geq 0$ if and only if $Ce^{At}B = 0$ for all t.

3.32 Theorem 3.1.5 on reachability introduced the *controllability Gramian* $P(t)$. The analogous theorem on observability (Theorem 3.4.3) lacks Gramians. We could have included them: prove that the following five statements are equivalent.

(a) The system is observable.

(b) $\ker(\mathcal{O}) = \{0\}$.

(c) The observability matrix $\mathcal{O}$ (3.20) has full column rank (rank n).

(d) The *observability Gramian* $Q(t)$ defined as

$$Q(t) = \int_0^t \mathrm{e}^{A^{\mathrm{T}}t} C^{\mathrm{T}} C \mathrm{e}^{At}\,\mathrm{d}t$$

is invertible for all $t > 0$

(e) The observability Gramian $Q(t)$ is invertible for some $t > 0$.

Now suppose that $u(t) = 0$ for all $t \in [0, t_1]$ and that the system is observable.

(f) Given a possible output $y : [0, t_1] \to \mathbb{R}^{n_y}$ show that the initial state $x(0)$ can be reconstructed from the output using

$$x(0) = Q^{-1}(t_1) \int_0^{t_1} \mathrm{e}^{A^{\mathrm{T}}t} C^{\mathrm{T}} y(t)\,\mathrm{d}t.$$

(g) In practice the measured output $\bar{y}(t)$ hardly ever equals the model $y(t) := C\mathrm{e}^{At}x(0)$. Show that

$$\bar{x}_0 := Q^{-1}(t_1) \int_0^{t_1} \mathrm{e}^{A^{\mathrm{T}}t} C^{\mathrm{T}} \bar{y}(t)\,\mathrm{d}t$$

solves the minimization problem

$$\min_{x_0} \|\bar{y} - C\mathrm{e}^{A\cdot}x_0\|.$$

Here the norm is the standard $\mathscr{L}_2$-norm of functions on the interval $[0, t_1]$:

$$\|z\| = \sqrt{\int_0^{t_1} z^{\mathrm{T}}(t) z(t)\,\mathrm{d}t}.$$

3.33 If a system is not reachable (i.e. not controllable) then, by definition, some states can not be reached. With this in mind it seems likely that "almost unreachable" systems require "large" inputs to control the state. This is confirmed by the following example. Let

$$\dot{x} = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} u.$$

Here both x and u have two components, and $\alpha, \beta \in \mathbb{R}$.

(a) Show that the system is uncontrollable iff $\beta = 0$.

(b) The smallest squared norm $\|u\|^2 = \int_0^{t_1} u_1^2(t) + u_2^2(t)\,\mathrm{d}t$ of all inputs that achieve $x(t_1) = x_1$ is given in Thm. 3.1.5. Compute this $\|u\|^2$.

(c) For which $x_1 \in \mathbb{R}^2$ do we have $\lim_{\beta \to 0} \|u\| = \infty$, and explain why you are not surprised by the answer.

(d) Suppose $\alpha \neq 0$. Show that $\|u\|$ decreases as t_1 increases. (In words this means: the longer the time, the smaller the required control action.)

(e) The minimal $\|u\|$ is also distinctively different for positive and negative α if t_1 is large: show that

$$\lim_{t_1 \to \infty} \|u\|^2 = 0 \quad \text{if } \alpha > 0, \beta \neq 0 \tag{3.35}$$

and

$$\lim_{t_1 \to \infty} \|u\|^2 = 2|\alpha|(x_{11}^2 + x_{12}^2/\beta^2) \quad \text{if } \alpha < 0, \beta \neq 0.$$

(Here x_{11}, x_{12} are the entries of $x_1 \in \mathbb{R}^2$.)

(f) Explain in words why the limit (3.35) makes sense.

Chapter 4

State Feedback and Dynamic Observers

In the definition of controllability, we studied the existence of input signals $u : [0, t_1] \to \mathbb{R}^{n_u}$ that steer a given initial state $x(0) = x_0$ to a given desired state x_1 at some time t_1. This is a typical example of *open-loop control*: a time signal $u : [0, t_1] \to \mathbb{R}^{n_u}$ is programmed based on known system data and states x_0, x_1. This is depicted schematically in Figure 4.1.

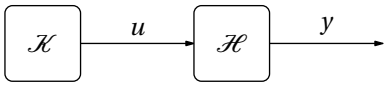

FIGURE 4.1: Open-loop control.

However, if such a computed input signal is used in an actual, real-life, system, then the end result will almost always be somewhat different from the computed and intended result. This is due to inevitable inaccuracies in the model, disturbances to the system, and an implementation of the computed input signal that is not 100% accurate. Although open-loop control is useful, in practice it will need to be supplemented with other methods.

What other control methods are there? Let us take another look at the example of the juggler (Example 3.2.2). We can imagine a juggler with a stick in his hand that he can move horizontally. Suppose that the juggler wants to balance the stick; in other words, he wants to keep the stick (approximately) upright. How should he proceed? In principle, he could, at any given time, determine the position and velocity of the stick and based on that (and on a precise knowledge of the equations of motion of the stick) compute how he needs to move his hand to put the stick upright and keep it there. Then he could—with his eyes closed—carry out this hand movement. This is an example of open-loop control.

Even if our juggler is capable of carrying out this computation, it is still clear that he cannot keep the stick balanced this way. After all, there will always be small errors or disturbances, and since the upright position of the stick is an unstable equilibrium the stick will fall anyway! The method he will use is completely different. He *observes* the position and velocity of the tip of the stick at every moment, and *based on this* adjusts his hand movement. This obvious method is the essence of *control theory*: the input $u(t)$ at every time t is determined as a function of a number of observed quantities $y(t)$ of the system up to this time t. This principle is called *feedback*, and the resulting control methodology is called *closed-loop control*. This is depicted schematically in Figure 4.2.

In fact, more is going on in our juggler example, as well as in many other examples. There is a *learning* or *adaptive* element in the determination of the necessary feedback: based on the result of the feedback, we will adjust the feedback if necessary. For example, the function that shows how the movement of the hand depends on the position and velocity of the tip of the stick can be adjusted, in particular, if the properties of the system that is being controlled

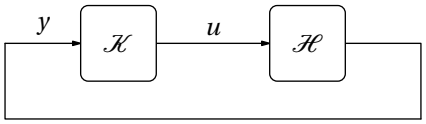

FIGURE 4.2: Feedback (closed-loop control).

change (for example because an object is added onto the tip of the stick).

This feedback mechanism (whether adaptive or not) can be found in many biological, physical, and technical systems. In control engineering, the controller is usually itself an automatic mechanism (in contrast to our juggler). One of the classic examples of a mechanical controller comes from James Watt, and concerns the control of the steam engine, the "motor" behind the industrial revolution of the nineteenth century. The steam engine can be seen as a state system with as input u the steam supply and as output y the angular velocity of the governor shaft; see Figure 4.3. A typical problem is keeping the output y as close as possible to a previously determined constant value y_0 (constant number of revolutions per minute). The flyball governor developed by James Watt[1] achieves this by having two metal weights rotating about a shaft operate a lever that controls the steam supply; see Figure 4.3. If the angular velocity increases, the weights move outwards, automatically decreasing the steam supply (and indirectly decreasing the rpm). Conversely, at a lower angular velocity, the weights move inwards, increasing the steam supply. By adding this flyball governor to the steam engine, there is feedback from the output y to the input u, and we can show mathematically that the input y converges to a constant value y_0, regardless of natural variations in the steam input u.

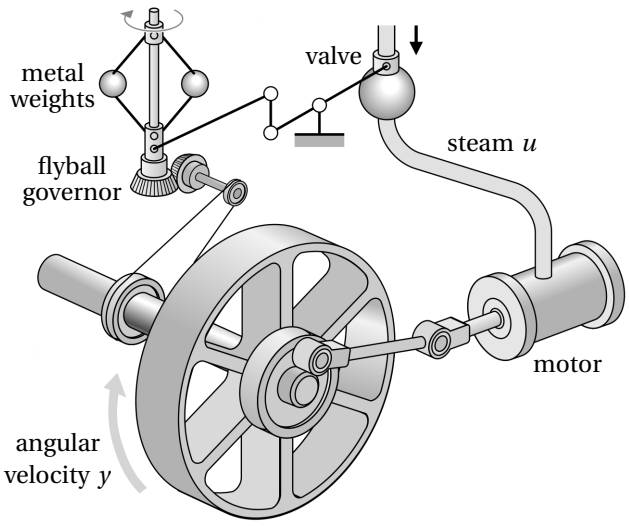

FIGURE 4.3: The steam engine with the flyball governor in the upper left corner. (This illustration is taken from the book *Modern Control Systems* by R.C. Dorf and R.H. Bishop.)

In the case of Watt's flyball governor, we only need to feed back the output y. In other examples, however, it may be necessary to feed back the entire state vector.

[1] James Watt adapted Christiaan Huygens' centrifugal governor.

In this chapter, we analyze feedback control of state systems of the form[2]

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx \end{aligned} \tag{4.1}$$

using a static control law of the form

$$u = -Fx \qquad \text{for some } F \in \mathbb{R}^{n_u \times n}$$

or a dynamic control law

$$\begin{aligned} \dot{z} &= A_{\mathscr{K}} z + B_{\mathscr{K}} y, \\ u &= C_{\mathscr{K}} z \end{aligned}$$

for certain matrices $A_{\mathscr{K}}, B_{\mathscr{K}}, C_{\mathscr{K}}$. For state systems, control breaks up into two more or less independent problems: 1) How can we send the state x to zero by choosing the input u appropriately as a function of x? 2) This assumes that we have x at our disposal for feedback, but what if we can only measure part of the state, say y; under what conditions is the knowledge of (u, y) sufficient to reconstruct x? Next, we combine these problems and arrive at the celebrated result that says that state systems that are controllable and observable can be stabilized fully automatically in the sense that $\lim_{t\to\infty} x(t) = 0$, $\lim_{t\to\infty} u(t) = 0$, and $\lim_{t\to\infty} y(t) = 0$ regardless of the initial conditions.

4.1 Stabilizability

In Chapter 2, we saw that $\lim_{t\to\infty} x(t) = 0$ for all solutions of $\dot{x} = Ax$ if and only if the eigenvalues of A have negative real part. An extension of stability that includes the freedom of choice of u is called stabilizability.

Definition 4.1.1 (Stabilizability). A system $\dot{x} = Ax + Bu$ is stabilizable if for every $x(0) = x_0 \in \mathbb{R}^n$, there exists a $u : [0,\infty) \to \mathbb{R}^{n_u}$ such that $\lim_{t\to\infty} x(t) = 0$. □

The input u is then called a stabilizing input. The definition of stabilizability does not explain how to choose u. If we can measure the entire state $x(t)$ at every time t, then we can try, as control law, a function of the form

$$u(t) = \mathscr{F}\big(t, x(\tau)|_{\tau \le t}\big),$$

that is, a control law that uses the state measured up to time t. The simplest control law of this form is the linear *static state feedback* law,

$$u(t) = -Fx(t) \qquad \text{for some } F \in \mathbb{R}^{n_u \times n}.$$

(The inclusion of a minus sign is a convention.) This is called static feedback because at every time t, the input $u(t)$ depends only on the "current" state $x(t)$ (and not on its past). The choice of a static state feedback is not unnatural, because by the definition of a state, $x(t)$ contains "all necessary" information from the past to determine the future.

Example 4.1.2 (Open loop versus closed loop). This example illustrates the fundamental difference between open loop control and closed loop control, and it demonstrates that closed-loop control is superior.

[2] Note that the direct feedthrough term D in $y = Cx + Du$ is assumed to be zero. This chapter's theory also works well for $D \neq 0$, but the formulas are then more complicated.

Consider the system

$$\dot{x} = x + u.$$

This system is *un*stable. The stabilize it we need to choose the input appropriately. There are many inputs u that stabilize the system. Two of them are

$$\begin{aligned} \textit{open loop:} \quad & u(t) = -3\mathrm{e}^{-2t}x(0), \\ \textit{closed loop:} \quad & u(t) = -3x(t). \end{aligned} \tag{4.2}$$

The two inputs u give the same result (verify this yourself):

$$x(t) = \mathrm{e}^{-2t}x(0),$$

and therefore both inputs stabilize the system. However, there is a big difference: in the open-loop control case, the input, $u(t) = -3\mathrm{e}^{-2t}x(0)$, is determined by the initial state $x(0)$ and this initial state fixes the input for the rest of time. This is like looking at the system once in your life, and then controlling it with your eyes closed for the rest of time. In the closed-loop control case, the input, $u(t) = -3x(t)$, at *every moment in time* is chosen depending on the state at that moment. This method requires the continuous observation of $x(t)$ and is, in that sense, more complicated. But it is also *much, much, much more robust*! Suppose, for example that the model $\dot{x} = x + u$ deviates a bit from the actual system, and that the actual system is

$$\dot{x} = 1.001x + u.$$

If we now apply the two control laws (4.2) to the actual system, we obtain (verify this yourself)

$$\begin{aligned} \textit{open loop:} \quad & x(t) = [\tfrac{3}{3.001}\mathrm{e}^{-2t} + \tfrac{0.001}{3.001}\mathrm{e}^{1.001t}]x(0), \\ \textit{closed loop:} \quad & x(t) = \mathrm{e}^{-1.999t}x(0). \end{aligned}$$

The open-loop control law destabilizes it (because $\mathrm{e}^{1.001t}$ diverges), but the closed-loop feedback control law still stabilizes the system, regardless of the small disturbance. So open-loop control can be extremely sensitive to modeling errors, while closed-loop control appears to be robust. This is typical, and in applications you always want to use closed loop control if the system itself is unstable. Open loop control is really not sufficient if the system is unstable. □

It is odd that the fundamental difference between open-loop and closed-loop control is not always well understood.

4.2 Static State Feedback

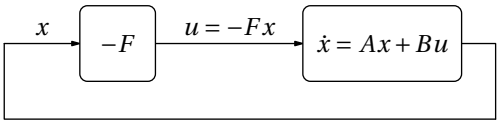

FIGURE 4.4: Static state feedback.

Probably the simplest closed loop control law is

$$u(t) = -Fx(t),$$

where F is some matrix. This is known as (linear, time-invariant) static state feedback. Application of this control to the system $\dot{x} = Ax + Bu$ modifies the dynamics into

$$\begin{aligned}\dot{x} &= Ax + Bu \\ &= Ax - BFx \\ &= (A - BF)x.\end{aligned}$$

If we can choose F in such a way that all eigenvalues of $A - BF$ have negative real part, then $u = -Fx$ is a stabilizing input.

Example 4.2.1 (Juggler). Consider, once more, the inverted pendulum of Example 3.2.2. For completeness, here are the system equations:

$$\begin{bmatrix}\dot{q}_1 \\ \dot{v}_1\end{bmatrix} = \begin{bmatrix}0 & 1 \\ \frac{g}{\ell_1} & 0\end{bmatrix}\begin{bmatrix}q_1 \\ v_1\end{bmatrix} + \begin{bmatrix}0 \\ -\frac{g}{\ell_1}\end{bmatrix} u. \tag{4.3}$$

It should be clear that this system is not asymptotically stable (otherwise juggling would be a piece of cake). Since this is a second-order system, the characteristic polynomial has degree 2. We want a state feedback that puts the two eigenvalues of $A - BF$ in $s = -1$. So, we want

$$\chi_{A-BF}(s) = (s+1)^2 = s^2 + 2s + 1.$$

We write the candidate state feedback as $u = -Fx = -\begin{bmatrix}f_1 & f_2\end{bmatrix}\begin{bmatrix}q_1 \\ v_1\end{bmatrix}$; then

$$A - BF = \begin{bmatrix}0 & 1 \\ \frac{g}{\ell_1}(1 + f_1) & \frac{g}{\ell_1} f_2\end{bmatrix}.$$

This has characteristic polynomial

$$\chi_{A-BF}(s) = s^2 - \frac{g}{\ell_1} f_2 s - \frac{g}{\ell_1}(1 + f_1).$$

It equals $s^2 + 2s + 1$ if we choose

$$f_1 = -1 - \frac{\ell_1}{g}, \qquad f_2 = -2\frac{\ell_1}{g}.$$

Done. □

In the example above, we have put the closed-loop poles at $s = -1$ (twice), but we could just as well have chosen another pair of poles[3]. We will soon see that this has everything to do with controllability.

Example 4.2.2 (Controller canonical form & pole placement). Suppose our system is in controller canonical form,

$$\dot{z} = \begin{bmatrix}0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1}\end{bmatrix} z + \begin{bmatrix}0 \\ \vdots \\ \vdots \\ 0 \\ 1\end{bmatrix} u.$$

Under the state feedback law

$$u = -\begin{bmatrix}r_0 - p_0 & r_1 - p_1 & \cdots & r_{n-1} - p_{n-1}\end{bmatrix} z \tag{4.4}$$

[3] In the case of complex poles, a pair of complex conjugates.

the controlled system is again in controller canonical form,

$$\dot{z} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -r_0 & -r_1 & \cdots & \cdots & -r_{n-1} \end{bmatrix} z.$$

This is interesting because the characteristic polynomial of the controlled system is $R(s) := s^n + r_{n-1}s^{n-1} + \cdots + r_0$. It shows we have complete "control" over the characteristic polynomial of the controlled system: choose your favorite $s^n + r_{n-1}s^{n-1} + \cdots + r_0$ and then (4.4) does the job! □

In Theorem 3.5.2 we saw that every controllable system (with $n_u = 1$) is isomorphic to a controller canonical form, so the next result is probably not a surprise (but realize that this next result allows any $n_u \geq 1$):

Theorem 4.2.3 (Pole placement). *Consider the system $\dot{x} = Ax + Bu$ with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n_u}$. For every polynomial*

$$R(s) := s^n + r_{n-1}s^{n-1} + \cdots + r_0, \qquad r_k \in \mathbb{R},$$

there exists an $F \in \mathbb{R}^{n_u \times n}$ such that

$$\det(sI - (A - BF)) = R(s)$$

if and only if the system is controllable.

Proof. First of all, note that controllability does not change under a state transformation $z = T^{-1}x$, and that a state feedback $u = -Fx$ remains a state feedback after transformation: $u = -Fx = -F_z z$ with $F_z := FT$. The characteristic polynomial of $\dot{x} = (A - BF)x$ also does not change. Indeed, the transformation

$$A_z := T^{-1}AT, \qquad B_z := T^{-1}B, \qquad F_z := FT \tag{4.5}$$

gives

$$\begin{aligned} \det(sI - (A_z - B_z F_z)) &= \det(sI - T^{-1}(A - BF)T) \\ &= \det(T^{-1}(sI - (A - BF))T) = \det(sI - (A - BF)). \end{aligned}$$

If $\dot{x} = Ax + Bu$ is uncontrollable, then it follows from the Kalman controllability decomposition (see Eqn. (3.3.1)) that $\chi_{A-BF}(s)$ always has a factor $\chi_{A_{22}}(s)$. So the eigenvalues of A_{22} are eigenvalues of $A - BF$ for every F. Placing the poles arbitrarily is therefore not possible for uncontrollable systems.

Next, suppose that the system is controllable. We first contruct an F for the case $n_u = 1$, so $B \in \mathbb{R}^{n\times 1}$. In the previous example we found that $u = -F_z z$ with

$$F_z = \begin{bmatrix} r_0 - p_0 & r_1 - p_1 & \cdots & r_{n-1} - p_{n-1} \end{bmatrix} \tag{4.6}$$

does the job. Then $F := F_z T^{-1}$ does the job for $\dot{x} = Ax + Bu$.

Now, suppose $n_u > 1$. Unfortunately, it does not follow from the controllability of (A, B) that (A, B_k) is controllable for at least one column B_k of B (see Exercise 3.8). It is a bit more complicated than that. By Heymann's lemma, see Appendix A.5, for every $u_0 \in \mathbb{R}^{n_u}$ for which $b := Bu_0$ is nonzero, there exists an $\tilde{F}$ such that $(A - B\tilde{F}, b)$ is a controllable pair. This reduces the problem to the case $n_u = 1$. Indeed, the controllability of $(A - B\tilde{F}, b)$ implies that for every nth degree monic $R(s)$, there exists an F_1 such that $\chi_{A-B\tilde{F}-bF_1} = R$. Take $F = \tilde{F} + u_0 F_1$. ■

An immediate consequence of this theorem is that every controllable system is stabilizable. Choose, for example, $R(s) = (s+1)^n$; all zeros of this polynomial are in the left half-plane. This theorem also tells us that controllable systems are always stabilizable through static state feedback $u = -Fx$. The question that comes up is: are there systems that are stabilizable, but not through static state feedback $u = -Fx$? We will see that the answer is no, although there are stabilizable systems that are not controllable. Suppose, for example, that in the Kalman controllability decomposition, the system is given by

$$\begin{bmatrix} \dot{z}_c \\ \dot{z}_{uc} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} z_c \\ z_{uc} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u. \tag{4.7}$$

This system is not controllable, but it can be stabilizable. A necessary condition for stabilizability is that $\dot{z}_{uc} = A_{22} z_{uc}$ is asymptotically stable, because that part cannot be influenced by u. This is also sufficient:

Theorem 4.2.4 (Stabilizability). *Consider the system $\dot{x} = Ax + Bu$. The following four statements are equivalent:*

1. *There exists an F such that $A - BF$ is asymptotically stable.*

   (So the system is stabilizable through static state feedback $u = -Fx$.)

2. *The system is stabilizable.*

3. *In the Kalman controllability decomposition (4.7) of $\dot{x} = Ax + Bu$ (with (A_{11}, B_1) controllable), the eigenvalues of A_{22} have negative real part.*

4. $\begin{bmatrix} sI - A & B \end{bmatrix}$ *has full row rank for all s with* $\operatorname{Re}(s) \geq 0$.

Proof. We prove the implications (1) $\Longrightarrow$ (2) $\Longrightarrow$ (3) $\Longrightarrow$ (1). The equivalence of (3) and (4) is an exercise (Exercise 4.11). (1) $\Longrightarrow$ (2): Obvious. (2) $\Longrightarrow$ (3): That $\dot{z}_{uc} = A_{22} z_{uc}$ must be asymptotically stable is obvious. (3) $\Longrightarrow$ (1): See Exercise 4.10. (3) $\Longleftrightarrow$ (4): See Exercise 4.11. ∎

We again note that this says that systems that are at all stabilizable — by open loop control or closed loop control, nonlinear or whatever method — are always stabilizable through static state feedback $u = -Fx$ as well. Nice.

4.2.1 Ackermann's Pole Placement Formula

The proof of the pole placement theorem (Theorem 4.2.3) is constructive. For $n_u > 1$, the construction of F is quite complicated, but for $n_u = 1$ it is simple and in essence does not use much more than a transformation to the controller canonical form. If we are only interested in F, then we do not even need to carry out the transformation, as we have the following result.

Lemma 4.2.5 (Ackermann's pole placement formula). *Suppose that (A, B) is controllable with $B \in \mathbb{R}^{n \times 1}$ (so $n_u = 1$). Write the characteristic polynomial of A as $\chi_A(s) = s^n + p_{n-1}s^{n-1} + \cdots + p_0$. Given a monic polynomial*

$$R(s) = s^n + r_{n-1}s^{n-1} + \cdots + r_0,$$

there is a unique F such that $\chi_{A-BF} = R$. This F equals

$$F = \begin{bmatrix} r_0 - p_0 & r_1 - p_1 & \cdots & r_{n-1} - p_{n-1} \end{bmatrix} \mathscr{C}_z \mathscr{C}_x^{-1} \tag{4.8}$$

with $\mathscr{C}_x$ and $\mathscr{C}_z$ the controllability matrices of, respectively, the pair (A,B) and (3.28). Equivalently, F can be determined using Ackermann's formula:

$$F = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathscr{C}_x^{-1} R(A). \tag{4.9}$$

Proof. By (4.5), the vector F is equal to $F = F_z T^{-1}$, with F_z as in (4.6). By Theorem 3.5.2, the matrix T is equal to $T = \mathscr{C}_x \mathscr{C}_z^{-1}$, and so (4.8) follows. The proof of (4.9) is more technical. Let $\eta := \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathscr{C}_x^{-1}$, that is,

$$\eta \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Then

$$\eta (A - BF)^k = \eta A^k \qquad \forall\, k < n \tag{4.10}$$
$$\eta (A - BF)^n = \eta A^n - F. \tag{4.11}$$

By the Cayley–Hamilton theorem, a matrix satisfies its own characteristic polynomial, $R(A - BF) = 0$. Hence, in particular, we have $\eta R(A - BF) = 0$. Using (4.10),(4.11), we can write this expression as

$$\begin{aligned} 0 &= \eta R(A - BF) \\ &= \eta (r_0 I + r_1 (A - BF) + \cdots + (A - BF)^n) \\ &= r_0 \eta I + r_1 \eta A + \cdots + r_{n-1} \eta A^{n-1} + (\eta A^n - F) \\ &= \eta (r_0 I + r_1 A + \cdots + r_{n-1} A^{n-1} + A^n) - F \\ &= \eta R(A) - F. \end{aligned}$$

It follows that $F = \eta R(A)$. ∎

We should note that for large n, Ackermann's formula (4.9) is numerically ill-conditioned.

Example 4.2.6. Consider the system

$$\dot{x} = \begin{bmatrix} -1 & 0 \\ 4 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ -1 \end{bmatrix} u. \tag{4.12}$$

We want a state feedback $u = -Fx$ that places the eigenvalues of the closed-loop system in -1 and -4. Ackermann's formula gives

$$\begin{aligned} F &= \begin{bmatrix} 0 & 1 \end{bmatrix} \mathscr{C}_x^{-1} R(A) \qquad \text{use that } R(s) = (s+1)(s+4): \\ &= \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}^{-1} (A + I)(A + 4I) \\ &= \underbrace{\begin{bmatrix} 0 & 1 \end{bmatrix} \frac{1}{2} \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 4 & 2 \end{bmatrix}}_{\begin{bmatrix} 1/2 & 1/2 \end{bmatrix}} \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \\ &\qquad\qquad \begin{bmatrix} 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 10 & 5 \end{bmatrix}. \end{aligned}$$

□

4.3 Observers

Many procedures for control of systems are based on the assumption that the entire state vector $x(t)$ can be measured. There is a good reason to use these types of control laws. Indeed, intuitively the state at any particular time contains all information necessary for the future behavior of the system. A control method that wants to influence the future behavior must therefore be based on the state. We have already seen an important example of such a control method in the previous section.

Often, however, we cannot measure the entire state vector. For physical systems, measuring certain quantities requires expensive measuring equipment, while economic systems, for example, require very extensive (statistical) measuring procedures. It may also happen that some state components with internal variables are not directly accessible for measurement. In all these cases, control must be based on the knowledge of *part* of the state vector. From here on, we will assume that this part consists of the output $y = Cx$ of the system (for convenience, we take $y = Cx$ and not the more general form $y = Cx + Du$; see Exercise 4.9).

Example 4.3.1 (Static output feedback). Consider the mass-spring-damper system of Example 2.2.8 and assume that the damping is zero:

$$\begin{bmatrix} \dot{q} \\ \ddot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{bmatrix} \begin{bmatrix} q \\ \dot{q} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u.$$

As output, we take the position $y := q$. By closing the loop with static output feedback, $u = -Hy = -Hq$, for some $H \in \mathbb{R}$, we obtain

$$\begin{bmatrix} \dot{q} \\ \ddot{q} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{bmatrix} \begin{bmatrix} q \\ \dot{q} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} (-Hq) = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} - \frac{H}{m} & 0 \end{bmatrix} \begin{bmatrix} q \\ \dot{q} \end{bmatrix}.$$

This has eigenvalues

$$\lambda_{1,2} = \pm\sqrt{-\frac{H+k}{m}}.$$

We see that the sum of the eigenvalues is always zero — regardless of how we choose H — and therefore that the system cannot be stabilized through static output feedback. □

This is a negative result. If only we could measure the entire state. How do we solve this problem? We know that if the system is observable, then based on the knowledge of the input and output during the time interval $[0, t_1]$, we can, *in principle*, uniquely determine the state at any time $t \in [0, t_1]$. The idea is now to let this determination of $x(t)$ be carried out automatically by a system we will call the *observer*. Figure 4.5 shows a block diagram of a system with observer. Both u and y are available to the observer, which uses them to construct an estimate $\hat{x}$ of x.

We do not require the observer to determine the state of the system $x(t)$ *exactly* at every point in time, but do require that it provides an estimate $\hat{x}(t)$ that improves with time. The idea is that with time, more and more information over the system is available to the observer, which should therefore be able to make better and better estimates. Systems that have such an observer are called detectable:

Definition 4.3.2 (Detectability). A system (4.1) is detectable if there exists an observer (a map from signals (u, y) to a signal $\hat{x}$) such that

$$\lim_{t\to\infty} \|\hat{x}(t) - x(t)\| = 0$$

for all initial conditions $x(0)$ and all inputs u. □

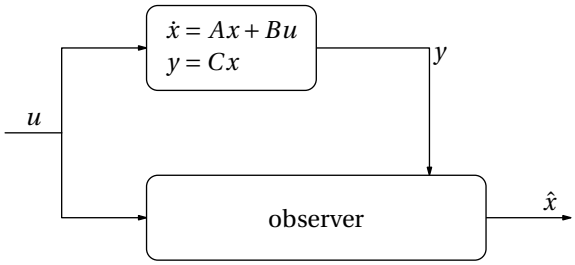

FIGURE 4.5: System with observer.

The definition of detectability does not say how to construct an observer. In principle we are allowed to use nonlinear and infinite-dimensional observers, but here we restrict ourselves to observers of the form

$$\textit{observer:} \quad \begin{cases} \dot{z} &= Pz + Qu + Ly, \\ \hat{x} &= Sz + Tu + Ry. \end{cases} \tag{4.13}$$

This is a dynamical system with input both u and y (because these are available), state z, and with output the estimate $\hat{x}$ of x. Actually, we restrict the search to an even simpler type of observer with $z = \hat{x}$, that is, $S = I$, $T = 0$, $R = 0$, so

$$\textit{observer:} \quad \dot{\hat{x}} = P\hat{x} + Qu + Ly, \tag{4.14}$$

and we also want it to satisfy the following condition:

Assumption 4.3.3. If $\hat{x}(t_0) = x(t_0)$ at some time t_0, then $\hat{x}(t) = x(t)$ for all $t \geq t_0$. □

In other words, "perfect once, perfect forever". This assumption says that if the estimation error $e(t)$, defined as

$$e(t) := x(t) - \hat{x}(t),$$

ever becomes zero, that is, $e(t_0) = 0$, then we want $e(t) = 0$ for all $t > t_0$ for all possible inputs. This estimation error satisfies

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} e &= \tfrac{\mathrm{d}}{\mathrm{d}t}(x - \hat{x}) \\ &= (Ax + Bu) - (P\hat{x} + Qu + LCx) \\ &= (A - LC)x - P\hat{x} + (B - Q)u \\ &= (A - LC)e + (A - LC - P)\hat{x} + (B - Q)u. \end{aligned}$$

Based on this we choose P and Q equal to

$$P = A - LC, \qquad Q = B,$$

because then Assumption 4.3.3 is satisfied and we have

$$\dot{e} = (A - LC)e.$$

Note that the dynamics of the error $e := x - \hat{x}$ is now disconnected from the input! With this choice of P and Q, the observer (4.14) takes the form

$$\dot{\hat{x}} = (A - LC)\hat{x} + Bu + Ly. \tag{4.15}$$

By the way, this form (4.15) is equivalent to

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - \hat{y}) \quad \text{with } \hat{y} := C\hat{x}. \tag{4.16}$$

Both representations (4.15) and (4.16) are depicted in Figure 4.6. Though equivalent, the two representations give rise to different interpretations. The first representation, (4.15), gives the observer in the standard state form. In particular, we see that the observer is asymptotically stable if $A - LC$ is asymptotically stable[4]. The second representation, (4.16), shows that the observer can also be seen as a duplicate of the original system with as additional input a term that depends only on the difference $y(t) - \hat{y}(t)$. As long as $\hat{y}(t)$ is equal to $y(t)$, there is apparently no reason to adjust the dynamics of $\hat{x}$. If, however, $\hat{y}(t)$ differs from $y(t)$, then we cannot have $\hat{x}(t) = x(t)$, and the correction term $L(y(t) - \hat{y}(t))$ in (4.16) then comes into play.

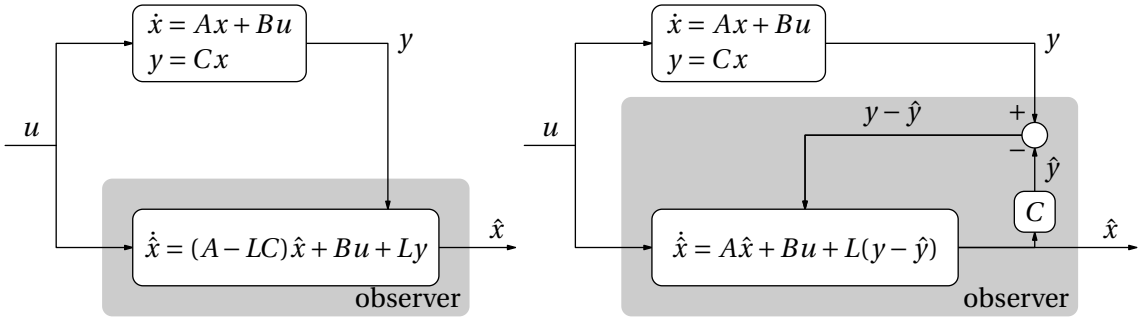

FIGURE 4.6: System with observer. The two observers are identical, but have been implemented differently.

The only freedom left in the observer is the matrix L. We still need to determine whether the estimation error $e(t) = x(t) - \hat{x}(t)$ goes to zero if $t \to \infty$. We have just seen that

$$\dot{e} = (A - LC)e.$$

Now, $\lim_{t\to\infty} e(t) = 0$ for every initial condition $e(0) = x(0) - \hat{x}(0)$ if and only if all eigenvalues of $A - LC$ have negative real part. The question is whether we can find a matrix L that achieves this.

Theorem 4.3.4 (Observer pole placement). *The pair (A, C) is observable if and only if for every polynomial $R(s) := s^n + r_{n-1}s^{n-1} + \cdots + r_0$, there exists an $L \in \mathbb{R}^{n\times n_y}$ such that* $\det(sI - (A - LC)) = R(s)$.

If $n_y = 1$, then the matrix L can be determined using Ackermann's formula,

$$L = R(A)\mathcal{O}_x^{-1}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \mathcal{O}_x^{-1}\mathcal{O}_z \begin{bmatrix} r_0 - p_0 \\ \vdots \\ r_{n-1} - p_{n-1} \end{bmatrix}$$

with $\mathcal{O}_x$ the observability matrix of (A, C) and $\mathcal{O}_z$ the observability matrix of the transformed system (3.30).

Proof. This result is the dual of Theorem 4.2.3 and Lemma 4.2.5. Recall that (A, C) is observable if and only if $(A^{\mathrm{T}}, C^{\mathrm{T}})$ is controllable. By Theorem 4.2.3, the controllability of $(A^{\mathrm{T}}, C^{\mathrm{T}})$ is equivalent to the existence of matrices $F \in \mathbb{R}^{n_y\times n}$ such that $\det(sI - (A^{\mathrm{T}} - C^{\mathrm{T}}F)) = R(s)$ for any choice of a monic nth degree polynomial $R(s)$. Take $L = F^{\mathrm{T}}$. ∎

[4] A matrix $A \in \mathbb{R}^{n\times n}$ is called asymptotically stable if all eigenvalues of A have negative real part.

Hence if (A, C) is observable, then we can certainly make $A - LC$ asymptotically stable. An immediate consequence of the observer pole placement theorem is therefore the following.

Theorem 4.3.5 (Observer). *If the system* (4.1) *is observable, then it is also detectable. In particular, any $\dot{\hat{x}} = (A - LC)\hat{x} + Bu + Ly$ is an observer provided we choose L such that $A - LC$ is asymptotically stable.* □

The zeros s of $\det(sI - (A - LC))$ are the eigenvalues of the observer and are also called the observer poles. As with state feedback, a logical question to ask is whether there are systems that are detectable—regardless of the form of observer—but for which there is no observer of the special form (4.15). The answer is no, which is good news:

Theorem 4.3.6 (Detectability). *Consider the system $\dot{x} = Ax + Bu, y = Cx$. The following four statements are equivalent:*

1. *There exists an L such that $A - LC$ is asymptotically stable.*

   (So the system is detectable with an observer of the form (4.15)*.)*

2. *The system is detectable.*

3. *In the Kalman observability decomposition, the eigenvalues of A_{22} have negative real part.*

4. *The matrix*

$$\begin{bmatrix} sI - A \\ C \end{bmatrix}$$

   has full column rank for all s with $\operatorname{Re}(s) \geq 0$.

Proof. We prove (1) $\implies$ (2) $\implies$ (3) $\implies$ (4) $\implies$ (1).

(1) $\implies$ (2) is trivial.

(2) $\implies$ (3): Suppose $u(t) = 0$ and $x(t) = 0$. Then $y(t) = 0$. Since the system is detectable, there is an observer signal $\hat{x}(t)$ such that $\lim_{t\to\infty} \hat{x}(t) = \lim_{t\to\infty} \hat{x}(t) - x(t) = 0$.

Now, suppose $u(t) = 0$ and $x(t) = \begin{bmatrix} 0 \\ e^{A_{22}t} z_{\text{uo}} \end{bmatrix}$. This state is in the unobservable space, so we now also have $y(t) = 0$. But then the observer produces the same signal $\hat{x}(t)$ as above, so $\lim_{t\to\infty} \hat{x}(t) = 0$. Because it is an observer, we must have $\lim_{t\to\infty} \hat{x}(t) - x(t) = 0$. Hence $x(t) = \begin{bmatrix} 0 \\ e^{A_{22}t} z_{\text{uo}} \end{bmatrix}$ also converges to zero when $t \to \infty$. Since z_{uo} can be chosen arbitrarily, A_{22} must be asymptotically stable.

(3) $\implies$ (4): See Exercise 4.12.

(4) $\implies$ (1): It suffices to show that condition (4) ensures the existence of an L for which $A - LC$ is asymptotically stable. This is the dual of (4) $\implies$ (1) of Thm. 4.2.4. ■

In analogy with the stabilizability, we have that systems that are at all detectable (using whatever form of observer such as linear, nonlinear, finite or infinite dimensional, etc.), can also be detected through a dynamical linear observer of the form (4.15). Consequently, choosing this type of dynamic linear observer is not restrictive. Good.

Example 4.3.7 (Juggler). Consider, once again, the inverted pendulum of Example 4.2.1 and suppose that we can measure only the position q of the tip of the pendulum, and not its velocity. So

$$y = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{C} \begin{bmatrix} q \\ v \end{bmatrix}.$$

The observer (4.16) for both the position and velocity is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\hat{q}\\ \hat{v}\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 1\\ \frac{g}{\ell_1} & 0\end{bmatrix}}_{A}\begin{bmatrix}\hat{q}\\ \hat{v}\end{bmatrix} + \underbrace{\begin{bmatrix}0\\ -\frac{g}{\ell_1}\end{bmatrix}}_{B}u + \underbrace{\begin{bmatrix}l_1\\ l_2\end{bmatrix}}_{L}(y-\hat{y}).$$

We determine (l_1, l_2) such that $A-LC$ has characteristic polynomial $R(s) = (s+2)^2 = s^2+4s+4$. (In particular, $A-LC$ is then asymptotically stable.) We have

$$\det(sI-(A-LC)) = \det\begin{bmatrix}s+l_1 & -1\\ -\frac{g}{\ell_1}+l_2 & s\end{bmatrix} = s^2 + l_1 s - \frac{g}{\ell_1} + l_2.$$

We must therefore choose $l_1 = 4$ and $l_2 = 4 + \frac{g}{\ell_1}$. Done!

Because this system is observable, by Theorem 4.3.4 we can also use Ackermann's formula to determine L:

$$L = (A^2+4A+4I)\mathscr{O}^{-1}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}4+g/\ell_1 & 4\\ 4g/\ell_1 & 4+g/\ell_1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}4\\ 4+g/\ell_1\end{bmatrix}.$$

The result is the same. □

Example 4.3.8 (Observer canonical form). Consider a system in observer canonical form

$$\begin{aligned}\dot{x} &= \underbrace{\begin{bmatrix}0 & \cdots & \cdots & 0 & -p_0\\ 1 & \ddots & & \vdots & -p_1\\ 0 & \ddots & \ddots & \vdots & \vdots\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & 1 & -p_{n-1}\end{bmatrix}}_{A} x + \begin{bmatrix}q_0\\ q_1\\ \vdots\\ \vdots\\ q_{n-1}\end{bmatrix} u, \qquad (4.17)\\ y &= \underbrace{\begin{bmatrix}0 & \cdots & \cdots & 0 & 1\end{bmatrix}}_{C} x.\end{aligned}$$

In this case, it is easy to find a column vector L such that $A-LC$ is asymptotically stable: write L as

$$L = \begin{bmatrix}l_0\\ l_1\\ \vdots\\ l_{n-1}\end{bmatrix}.$$

Then we have

$$A-LC = \begin{bmatrix}0 & \cdots & \cdots & 0 & -p_0-l_0\\ 1 & \ddots & & \vdots & -p_1-l_1\\ 0 & \ddots & \ddots & \vdots & \vdots\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & 1 & -p_{n-1}-l_{n-1}\end{bmatrix}.$$

This is again a companion matrix, and the characteristic polynomial is therefore

$$\chi_{A-LC}(s) = s^n + (p_{n-1}+l_{n-1})s^{n-1} + \cdots + (p_0+l_0).$$

It is clear that the coefficients of this polynomial can be chosen arbitrarily by choosing L accordingly. □

It looks like we can let observers act as quickly as we want. If we, for example, take $\det(sI-(A-LC)) = (s+100)^n$, then all eigenvalues of $A-LC$ are in $s=-100$ and consequently every solution $e(t)$ of $\dot{e} = (A-LC)e$ can be written as a linear combination of functions of the form $\mathrm{e}^{-100t}t^k$. This ensures that the estimation error $e(t)$ converges to zero very quickly, which implies that based on the output y, we will, in no time, have at our disposal an almost perfect estimate of x. This conclusion seems to contradict our discussion on page 91 that it is often impossible to measure entire state vectors. The problem is that fast (aggressive) observers are usually very sensitive to modeling errors and noise in the measurement of y, and although in theory they can be made arbitrarily fast, the presence of, for example, noise in the measurements limits the maximum speed. A proper consideration of this issue requires a stochastic setting—a subject in its own right. We will not discuss stochastics here, we just illustrate the problem:

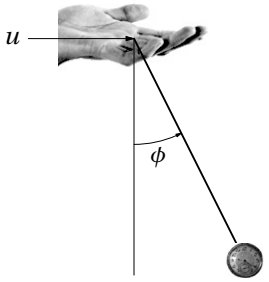

FIGURE 4.7: The hypnotist (Example 4.3.9).

Example 4.3.9 (Application: Hypnotist). A variation on the juggler is the hypnotist. Instead of stabilizing, the hypnotist wants to keep the pendulum in motion; see Figure 4.7. We denote the horizontal position of the hand by u and the angle the pendulum makes with the vertical axis by ϕ. The horizontal displacement of the pendulum (or watch) q is therefore $q = u + \ell\sin(\phi)$. Without derivation, we claim that the linearized equation of motion is

$$\ddot{q} + \frac{b}{m}\dot{q} + \frac{g}{\ell}q = \frac{g}{\ell}u.$$

Here ℓ is the length of the chord, b is a positive friction coefficient, g is the gravitational acceleration, and m is the mass of the watch. We take the values $m = 0.1$[kg], $\ell = 0.4$[m], $b = 0.05$[kg/s], and $g = 10$[m/s^2], so

$$\ddot{q} + 0.5\dot{q} + 2.5q = 2.5u.$$

With state variables q and $v := \dot{q}$, we get

$$\begin{bmatrix}\dot{q}\\ \dot{v}\end{bmatrix} = \begin{bmatrix}0 & 1\\ -2.5 & -0.5\end{bmatrix}\begin{bmatrix}q\\ v\end{bmatrix} + \begin{bmatrix}0\\ 2.5\end{bmatrix}u$$
$$y = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}q\\ v\end{bmatrix}.$$

We first take a nonaggressive observer, that is, one where the correction term $L(y-\hat{y})$ is not large. In this example, we could even take $L=0$, because the system itself is already asymptotically stable (with poles -0.25 ± 1.56i). We choose L "small", $L = \left[\begin{smallmatrix}0.5\\ 0.5\end{smallmatrix}\right]$. Then the observer is

$$\begin{bmatrix}\dot{\hat{q}}\\ \dot{\hat{v}}\end{bmatrix} = \begin{bmatrix}-0.5 & 1.0\\ -3.0 & -0.5\end{bmatrix}\begin{bmatrix}\hat{q}\\ \hat{v}\end{bmatrix} + \begin{bmatrix}0\\ 2.5\end{bmatrix}u + \begin{bmatrix}0.5\\ 0.5\end{bmatrix}y.$$

The observer poles are $-0.5 \pm 1.73\mathrm{i}$. Figure 4.8(top) shows the actual state (q, v) for $u(t) = \cos(\pi t)$ and $q(0) = 2, v(0) = 1$, and the state $(\hat{q}, \hat{v})$ estimated by the observer (red, dashed). It seems that the observer constructs a very good estimate of (q, v) in less than 10 seconds. Because the input is a sinusoid, the states do not converge to zero. It can be shown that the states converge to sinusoids. This convergence is slower than the convergence of $(\hat{q}, \hat{v})$ to (q, v) because the real part of the system poles

$$\mathrm{Re}\,(-0.25 \pm 1.56\mathrm{i}) = -0.25$$

is less negative than the real part of the observer poles

$$\mathrm{Re}\,(-0.5 \pm 1.73\mathrm{i}) = -0.5.$$

Next we take a more aggressive observer. We take $L = \left[\begin{smallmatrix} 5 \\ 5 \end{smallmatrix}\right]$. The observer now becomes

$$\begin{bmatrix} \dot{\hat{q}} \\ \dot{\hat{v}} \end{bmatrix} = \begin{bmatrix} -5 & 1 \\ -7.5 & -0.5 \end{bmatrix} \begin{bmatrix} \hat{q} \\ \hat{v} \end{bmatrix} + \begin{bmatrix} 0 \\ 2.5 \end{bmatrix} u + \begin{bmatrix} 5 \\ 5 \end{bmatrix} y.$$

and its poles are $-2.75 \pm 1.56\mathrm{i}$. With the same $u(t) = \cos(\pi t)$ and $q(0) = 2, v(0) = 1$ as before, the observer is much faster now; see Figure 4.8(bottom).

The more aggressive (faster) observer produces better results. But what happens if the measurement y of the position q is not perfect? We model the measurement error as an additional signal on q, and for simplicity assume that the measurement error is always $1/2$:

$$y = q + 1/2.$$

With this perturbed y, it is the less aggressive (slower) observer that outperforms the more aggressive (faster) observer; see Figure 4.9.

For our system, there is also an observer that is completely independent of measurement errors in y; see Exercise 4.1g. □

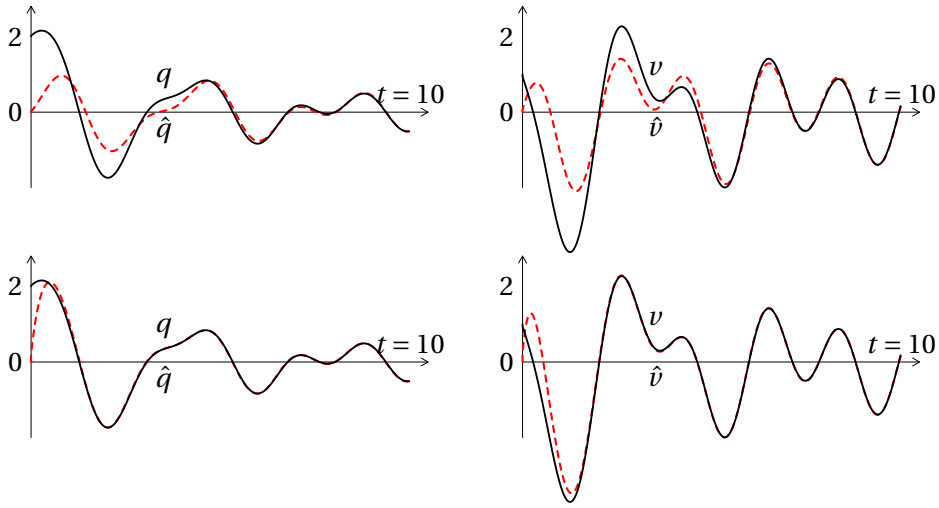

FIGURE 4.8: The actual (q, v) (black) and estimated $(\hat{q}, \hat{v})$ (red, dashed) for a slow observer (top) and for a fast observer (bottom). See Example 4.3.9.

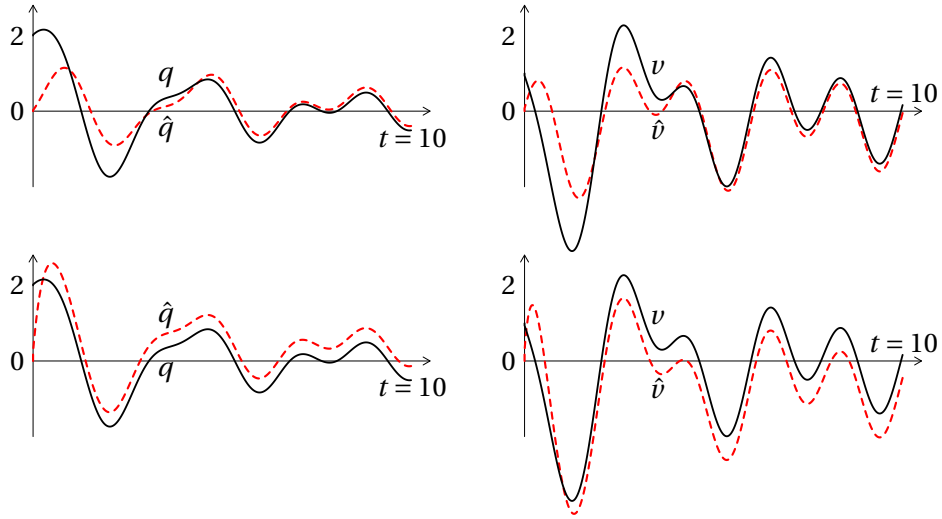

FIGURE 4.9: Actual (q, v) (black) and estimated $(\hat{q}, \hat{v})$ (red, dashed) for a slow observer (top) and for a faster observer (bottom). See Example 4.3.9.

4.4 Dynamical Output Feedback

We now have enough results to construct a control system $u = \mathscr{K}(y)$ that generates a signal u that stabilizes the given system, based only on the output y (and not on the entire state). See Figure 4.2. As before the given system is assumed of the form

$$\textit{given system:} \quad \begin{cases} \dot{x} &= Ax + Bu, \\ y &= Cx \end{cases} \tag{4.18}$$

and as control system (aka controller), we propose a dynamical system that determines an estimate $\hat{x}$ of the state x of the given system from (u, y) using an observer, and that provides $u = -F\hat{x}$ based on this estimate. So,

$$\textit{controller:} \quad \begin{cases} \dot{\hat{x}} &= (A - LC)\hat{x} + Bu + Ly \quad \text{(observer)}, \\ u &= -F\hat{x} \qquad\qquad\qquad\quad \text{(feedback)}. \end{cases}$$

We can eliminate the term Bu in the observer by substituting $u = -F\hat{x}$, and this gives the controller in the standard form (i.e. with input y and output u),

$$\textit{controller:} \quad \begin{cases} \dot{\hat{x}} &= (A - LC - BF)\hat{x} + Ly, \\ u &= -F\hat{x}. \end{cases} \tag{4.19}$$

The combination of the given system (4.18) and the controller (4.19) is called the closed-loop system, and it is described by

$$\textit{closed loop:} \quad \begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A & -BF \\ LC & A - BF - LC \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix},$$

(verify this yourself). In the previous section, we saw that the dynamics of the estimation error $e := x - \hat{x}$ satisfy $\dot{e} = (A - LC)e$ and that these dynamics do not depend on u. So the same must hold here. Indeed,

$$\begin{aligned} \dot{e} &= \dot{x} - \dot{\hat{x}} \\ &= (Ax - BF\hat{x}) - (LCx + (A - BF - LC)\hat{x}) \\ &= (A - LC)(x - \hat{x}) \\ &= (A - LC)e. \end{aligned}$$

The dynamics of x in terms of x and e simplify to

$$\begin{aligned}\dot{x} &= Ax - BF\hat{x} \\ &= Ax - BF(x-e) \\ &= (A-BF)x + BFe.\end{aligned}$$

The behavior of the closed loop can therefore equivalently be described by

$$\textit{closed loop:} \quad \begin{bmatrix}\dot{x}\\ \dot{e}\end{bmatrix} = \begin{bmatrix}A-BF & BF\\ 0 & A-LC\end{bmatrix}\begin{bmatrix}x\\ e\end{bmatrix}.$$

Because of zero block here we can infer that the eigenvalues of the closed-loop system are equal to the eigenvalues of $A-BF$ together with the eigenvalues of $A-LC$! The conclusion is that the eigenvalues of the closed-loop system are equal to the eigenvalues we would get through the state feedback $u=-Fx$ together with the eigenvalues of the observer. This leads us to the central result of this chapter:

Theorem 4.4.1 (Stabilizing dynamical controller). *If the system (4.18) is stabilizable and detectable, then there exist matrices F and L such that $A-BF$ and $A-LC$ are asymptotically stable. In that case, the controller (4.19) stabilizes the system (4.18), in the sense that $\lim_{t\to\infty} x(t) = 0$ and $\lim_{t\to\infty}\hat{x}(t) = 0$ for all initial conditions $x(0) = x_0$ and $\hat{x}(0) = \hat{x}_0$.*

This stabilization process is an example of a so-called separation principle: we can determine a state feedback law $u=-Fx$ and an estimate $\hat{x}$ for x independently of each other. Connecting the two using $u=-F\hat{x}$ gives the desired result.

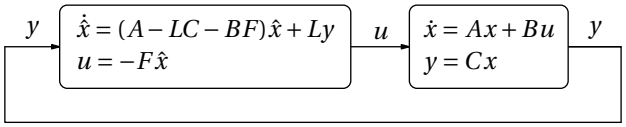

FIGURE 4.10: System with controller.

Figure 4.10 shows the closed-loop system schematically. The controller (the system at the top left) is also called a compensator. In contrast to observers, the controller has only y as input.

Example 4.4.2 (Juggler). In Example 4.2.1, we used $u=-Fx$ to place the eigenvalues of $A-BF$ in -1. For $\ell_1 = \frac{1}{2}g$, this gave $F = \begin{bmatrix}-3/2 & -1\end{bmatrix}$. In Example 4.3.7, we placed the eigenvalues of the observer in -2. For $\ell_1 = \frac{1}{2}g$, this gave $L = \begin{bmatrix}4\\6\end{bmatrix}$. Combining the observer and feedback $u=-F\hat{x}$ then gives (still for $\ell_1 = \frac{1}{2}g$)

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}q\\ v\\ \hat{q}\\ \hat{v}\end{bmatrix} = \underbrace{\left[\begin{array}{cc|cc}0 & 1 & 0 & 0\\ 2 & 0 & -3 & -2\\ \hline 4 & 0 & -4 & 1\\ 6 & 0 & -7 & -2\end{array}\right]}_{=\left[\begin{smallmatrix}A & -BF\\ LC & A-BF-LC\end{smallmatrix}\right]}\begin{bmatrix}q\\ v\\ \hat{q}\\ \hat{v}\end{bmatrix}$$

with eigenvalues $-1,-1,-2,-2$.

The controller is

$$\dot{\hat{x}} = \underbrace{\begin{bmatrix} -4 & +1 \\ -7 & -2 \end{bmatrix}}_{A-LC-BF} \hat{x} + \underbrace{\begin{bmatrix} 4 \\ 6 \end{bmatrix}}_{L} y,$$
$$u = \underbrace{\begin{bmatrix} 3/2 & 1 \end{bmatrix}}_{-F} \hat{x}.$$

One can check that here, the controller itself is also asymptotically stable. It is good to realize that this does not need to hold. In fact, it can be shown that certains systems can only be stabilized with an unstable controllers! □

In practice, controllers are also applied to *stable* systems. The aim is then to regulate the behavior in some other way. For instance, to speed up to convergence of the signals (see Exercise 4.15) or the steer the output to some *nonzero* value, possibly set by the user (think of the heating system where you set the desired room temperature).

4.5 Exercises

4.1 Comprehension questions (on the whole chapter). Prove or give a counterexample.

(a) If the system $\dot{x} = Ax + Bu$ is stabilizable, then for every $x(0)$ there exists a u such that $x(10) = 0$.

(b) If (A, B) is stabilizable, then so is $(A - BM, B)$.

(c) If $(A - MC, C)$ is detectable, then so is (A, C).

(d) If (A, B) is stabilizable, then so is $(A - LC, B)$.

(e) If $-A$ is asymptotically stable and (A, B) is stabilizable, then (A, B) is controllable. [This one is complicated!]

(f) If A is asymptotically stable, then (A, C) is detectable.

(g) If A is asymptotically stable, then $\dot{\hat{x}} = A\hat{x} + Bu$ is an observer.

4.2 Consider the system

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

(a) For which β is the system controllable?

(b) Show that $\frac{\mathrm{d}}{\mathrm{d}t}(-\beta x_1 + x_2) = (x_1 - \beta x_2)$, and explain why this implies that the system is not controllable if $\beta = \pm 1$.

(c) Is the system observable?

(d) Take $\beta = 2$. Give a state feedback $u = -Fx$ such that the closed loop has characteristic polynomial $s^2 + s + 1$.

(e) Take $\beta = 2$. Give an observer with double eigenvalue -1.

4.3 Consider the system

$$\dot{x} = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u,$$
$$y = \begin{bmatrix} 1 & 1 \end{bmatrix} x$$

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

(a) For which α is the system controllable?

(b) For which α is the system observable?

(c) Determine the characteristic polynomial of $\left[\begin{smallmatrix} 1 & \alpha \\ 0 & 1 \end{smallmatrix}\right]$.

(d) Determine the observer canonical form of the system (if it exists).

(e) Take $\alpha = 1$. Determine the controller canonical form of the system (if it exists).

(f) Take $\alpha = 1$. Determine a state feedback $u = -Fx$ that places the eigenvalues of $A - BF$ in $s = -2$ [twice].

(g) Show that if this system is not controllable, then the system is not stabilizable through a static state feedback $u = -Fx$.

4.4 *Third-order system.* We are given a system with

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}. \tag{4.20}$$

(a) Is the system observable? Is it controllable?

(b) Is $x = 0$ an asymptotically stable equilibrium point of $\dot{x} = Ax$?

(c) Is it possible to make the system asymptotically stable using a static output feedback $u(t) = -Hy(t)$? [Hint: you may want to know that for polynomials of the form $\lambda^n - c\lambda^{n-1} + \cdots$ the constant c equals the sum of all zeros of the polynomial[5].]

(d) Determine a state feedback $u = -Fx$ such that the eigenvalues of $A - BF$ are in $-1 \pm 2\mathrm{i}, -2$. [Hint: use properties of the companion matrix.]

4.5 *State feedback.* We are given the system $\dot{x} = Ax + Bu$ with

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

(a) Is $x = 0$ an asymptotically stable equilibrium point of $\dot{x} = Ax$?

(b) Is the system $\dot{x} = Ax + Bu$ controllable?

(c) Can we use the state feedback $u = -f_0x_1 - f_1x_2 - f_2x_3 - f_3x_4$ to place the eigenvalues of $\dot{x} = Ax + Bu$ in

i. $-2, -2, -1, -1$

ii. $-2, -2, -2, -1$

iii. $-2, -2, -2, -2$

4.6 *Deadbeat control.* In some aspects, discrete-time systems are fundamentally different from continuous-time systems. Consider the n-dimensional discrete-time system

$$x[t+1] = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix}}_{A} x[t] + \underbrace{\begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{B} u[t].$$

[5] Just expand $\prod_{i=1}^{n}(\lambda - \lambda_i)$.

Suppose that the input quantity is chosen to be $u[t] = -Fx[t]$. Determine $F \in \mathbb{R}^{1\times n}$ such that all eigenvalues of the resulting feedback system lie in the origin. Verify that $x[t] := (A-BF)^t x[0]$ is then zero for all $t \geq n$.

4.7 *Feedback-2.* Consider (4.1) with A, B, C as in (4.20).

(a) Give an observer with eigenvalues -4, -5, -1.

(b) Give a state feedback for the system such that after applying state feedback, the system has eigenvalues $-1 \pm \mathrm{i}, -2$.

(c) Give a state representation of the stabilizing controller obtained using parts (a) and (b).

4.8 *Feedback-3.* Consider (4.1) with A, B, C as in (4.20). Give an observer with eigenvalues -2, -2, -3.

4.9 In Section 4.3, observers are constructed for systems of the form $\dot{x} = Ax + Bu,\ y = Cx + Du$ with $D = 0$. What adjustments of those observers are needed for the case $D \neq 0$?

4.10 *Stabilizability.* Prove the implication (3) $\Longrightarrow$ (1) of Thm. 4.2.4.

4.11 *Stabilizability.* Prove the equivalence of parts (3) and (4) of Thm. 4.2.4.

4.12 *Detectability.* Prove the implication (3) $\Longrightarrow$ (4) of Thm. 4.3.6.

4.13 *Mass-spring-damper system.* Consider the mass-spring-damper system of Example 2.2.8 and take $m = 1$.

(a) For which values of $k \geq 0, r \geq 0$ (and $m = 1$) is the system of Example 2.2.8

i. asymptotically stable

ii. controllable

iii. observable

(b) Determine a state feedback $u = -Fx$ that places the poles of the closed-loop system in -1 and -2. (Don't forget that F also depends on k and r.)

(c) Determine an observer with observer poles in -4 and -5. (Here too, the answer depends on k and r.)

(d) Using parts (b) and (c), determine a controller that stabilizes the system.

(e) Suppose that $u = -Fx$ stabilizes the system. For which constant v does $u = -Fx + v$ bring the mass to a rest 1 meter to the right of the equilibrium point?

4.14 Construct an observer for the nonlinear system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + g(u(t)), \\ y(t) &= Cx(t) \end{aligned}$$

with g an arbitrary function. (Assume that (A, C) is observable.)

4.15 *Two-tank system.* Figure 4.11 shows a serial interconnection of two water tanks. The variables u, q_1, q_2 denote the water flow, and h_1 and h_2 denote the water heights in

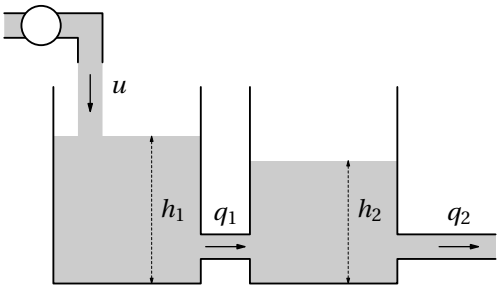

FIGURE 4.11: Two water tanks. See Exercise 4.15.

tanks 1 and 2, respectively. We linearize the system around a constant equilibrium solution (one of the many). That is, we write the variables as

$$\begin{aligned} u(t) &= u^* + \delta_u(t), \\ q_j(t) &= q_j^* + \delta_{q_j}(t), \\ h_j(t) &= h_j^* + \delta_{h_j}(t). \end{aligned}$$

Assuming that the tanks are identical and that q_1 depends only on the height *difference* $h_1 - h_2$, this gives the linearized model

$$\begin{bmatrix} \dot{\delta}_{h_1} \\ \dot{\delta}_{h_2} \end{bmatrix} = \frac{1}{S} \begin{bmatrix} -\frac{1}{R_1} & \frac{1}{R_1} \\ \frac{1}{R_1} & -\frac{1}{R_1} - \frac{1}{R_2} \end{bmatrix} \begin{bmatrix} \delta_{h_1} \\ \delta_{h_2} \end{bmatrix} + \frac{1}{S} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \delta_u, \tag{4.21}$$

where S is the area of the cross section of the tanks and the R_j are resistances.

For simplicity, take $S = R_1 = R_2 = 1$.

(a) Determine a state feedback that places the two poles in $-1 \pm \mathrm{i}$.

(b) Suppose that we can measure only the height of the first tank, $y := \delta_{h_1}$. Determine an observer with observer poles in -2 and -3.

(c) Using parts (a) and (b), determine a controller that stabilizes the system.

Remark. The given system (4.21) is itself already asymptotically stable, but the eigenvalues for $S = R_1 = R_2 = 1$ are -2.618 and -0.382, and as the latter is "close" to zero, fluctuation around the equilibrium point will die out only "slowly". For the controller you have constructed, the fluctuations die out more quickly because the eigenvalues of the closed-loop system are further away from the imaginary axis (are more negative).

4.16 Consider the system

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u, \\ y &= \begin{bmatrix} 1 & 1 \end{bmatrix} x \end{aligned}$$

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

(a) For which α_{nondetec} is the system not detectable?

(b) Determine, for arbitrary α, an L_α for which $A - L_\alpha C$ has double eigenvalue -1.

(c) What happens to L_α when $\alpha \to \alpha_{\text{nondetec}}$? Why is this not surprising?

Tougher Exercises

4.17 It is not unusual to only be able to measure a state (or part of it) with some delay. Consider the system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= x(t-\eta) \end{aligned}$$

with $\eta > 0$. From $t = 0$ on, we connect this system to the "observer"

$$\hat{x}(t) = \mathrm{e}^{A\eta} y(t) + \int_{\max(0,\,t-\eta)}^{t} \mathrm{e}^{A(t-\tau)} Bu(\tau)\,\mathrm{d}\tau, \qquad t > 0$$

Show that $\hat{x}(t) = x(t)$ for all $t > \eta$.

4.18 *Delays.* In practice, measurements are often accompanied by delay. Determine an observer for the system (4.1) that has input $u(t)$ and delayed input $y(t-1)$, and for which $\lim_{t\to\infty} \|\hat{x}(t) - x(t)\| = 0$.

4.19 *Container transfer.* Consider once more the container transfer system. In particular, consider the linearization (3.32) on page 78. A reasonable mathematical model for how someone would direct the cart is

$$u(t) = c\big(r(t) - x_m(t)\big) - k\dot{x}_m(t), \qquad k, c > 0. \tag{4.22}$$

The term $c\big(r(t) - x_m(t)\big)$ is proportional to the distance to the point to which we want to send the cart. It is positive if the cart is to the left of $r(t)$, and negative it it is to the right of $r(t)$. So $c(d(t) - x_m(t))$ is a force in the direction of the target $r(t)$. To prevent too large accelerations, we have added the term $-k\dot{x}_m(t)$.

(a) Does this $u(t)$ stabilize the system? [Difficult?]

(b) Do you have any idea whether this $u(t)$ stabilizes the nonlinear system from every x_0? [Very difficult.]

Chapter 5

Linear Quadratic Control

In the previous chapter we analysed and constructed stabilizing controllers, using static state feedback and dynamic output feedback. There is an enormous freedom in this design. For example, if the system is controllable then we can, in principle, achieve "any" set of closed loop poles. How to exploit this freedom to further the design? We should try to avoid "large" inputs, because systems usually cannot handle large inputs (think of the heating system in your house) and even if it can, large inputs are often costly (again think of your heating system). Also, *linear* models are usually inaccurate if signals are "large". We want, therefore, inputs that are "small" but that nevertheless stabilize and steer other signals "quickly" back to their equilibrium. *Linear Quadratic Control (LQ control)* has the potential to solve such problems. In short, the idea in LQ control is to minimize a "cost function" such as

$$\int_0^\infty y^2(t) + \rho u^2(t)\,\mathrm{d}t \tag{5.1}$$

over all stabilizing inputs u. Here ρ is a positive tuning parameter that we choose. If ρ is large then we put a large penalty on the input in the cost function, so the input that minimizes this cost is probably going to be "small". Conversely, if ρ is small (close to zero) then inputs are "cheap" and then the optimal input is probably "large" and possibly it is able to steer the output y to zero "quickly". By tuning ρ we can now hope to come up with a good compromise between small u and small y.

LQ control was initiated by Rudolf Kalman, and it attracted much attention during the period 1960–1980.

5.1 LQ problem with stability

Linear Quadratic (LQ) control is about minimization of a Quadratic cost over positive time for a given Linear system.

Definition 5.1.1 (LQ with stability). Consider the linear system with initial state

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= Cx. \end{aligned}$$

The LQ problem (with stability) is to minimize

$$\left\| \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right] \right\|_{\mathbb{L}_{2+}}^2 := \int_0^\infty y^{\mathrm{T}}(t)y(t) + u^{\mathrm{T}}(t)u(t)\,\mathrm{d}t$$

over all inputs u that stabilize the system, meaning inputs that achieve $\lim_{t\to\infty} x(t) = 0$. □

If y has m components then $y^{\mathrm{T}}y = y_1^2 + y_2^2 + \cdots + y_m^2$. In LQ we, therefore, want all components y_i, u_j of output and input to be small in some sense. In this chapter we develop a fairly complete theory of the LQ-problem, and we want to stress that this allows systems with multiple inputs and/or outputs, so with *matrices* B and C (not just column or row vectors).

5.2 Algebraic Riccati Equation

For certain special problems we can solve the LQ problem easily:

Example 5.2.1. Consider the system

$$\begin{aligned} &\dot{x} = u, \qquad x(0) = x_0, \\ &y = x. \end{aligned}$$

Then

$$y^2 + u^2 = x^2 + u^2 = (x+u)^2 - 2xu = (x+u)^2 - 2x\dot{x}.$$

The term $-2x\dot{x}$ has an explicit antiderivative, $-x^2$, so

$$y^2 + u^2 = \tfrac{\mathrm{d}}{\mathrm{d}t}(-x^2) + (x+u)^2. \tag{5.2}$$

Integrating this over $t \in [0,\infty)$ we see that the cost equals

$$\left\| \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right] \right\|_{\mathbb{L}_{2+}}^2 = x_0^2 + \|u + x\|_{\mathbb{L}_{2+}}^2. \tag{5.3}$$

Here we used that $\lim_{t\to\infty} x(t) = 0$, since u is assumed to stabilize the system. It is immediate from (5.3) that the cost is at least x_0^2, and it equals x_0^2 iff

$$u = -x. \tag{5.4}$$

Since the static state feedback $u := -x$ indeed stabilizes (because the closed loop system becomes $\dot{x} = -x$) we conclude that the state feedback (5.4) is the optimal control, and that the optimal (minimal) cost is

$$\left\| \left[\begin{smallmatrix} y \\ u \end{smallmatrix}\right] \right\|_{\mathbb{L}_{2+}}^2 = x_0^2.$$

Done. □

In this example we found that the minimal cost is quadratic in the initial state, and that the optimal input can be implemented as a static state feedback. Inspired by this example we conjecture that this is always the case. That is, we conjecture that the minimal cost is of the form

$$x_0^{\mathrm{T}} P x_0$$

for some matrix P, and that the optimal input equals $u(t) = -Fx(t)$ for some matrix F. Correct:

Theorem 5.2.2 (Solution of the LQ problem — Algebraic Riccati Equation). *Consider the system*

$$\begin{aligned} &\dot{x} = Ax + Bu, \qquad x(0) = x_0, \\ &y = Cx, \end{aligned}$$

and cost

$$\int_0^\infty y^{\mathrm{T}}(t)y(t) + u^{\mathrm{T}}(t)u(t)\,\mathrm{d}t. \tag{5.5}$$

Suppose P is a real symmetric matrix in $\mathbb{R}^{n\times n}$ that satisfies the Algebraic Riccati Equation

$$A^{\mathrm{T}}P + PA + C^{\mathrm{T}}C - PBB^{\mathrm{T}}P = 0 \tag{5.6}$$

with the property that

$$A - BB^{\mathrm{T}}P \quad \textit{is asymptotically stable.}$$

Then the static state feedback

$$u(t) = -B^{\mathrm{T}}Px(t)$$

solves the LQ problem, that is, it stabilizes the system and minimizes (5.5) *over all stabilizing inputs, and the optimal cost* (5.5) *equals $x_0^{\mathrm{T}}Px_0$.*

Moreover, if the system is stabilizable and detectable, then (5.6) *has unique solution $P \in \mathbb{R}^{n\times n}$ for which $A - BB^{\mathrm{T}}P$ is asymptotically stable, and this P is symmetric.*

Proof. We expect the optimal u to be a static state feedback $u = -Fx$ for some F, so with that in mind define $v := Fx + u$. (If our hunch is correct then optimal means $v = 0$.)

We write $y^{\mathrm{T}}y + u^{\mathrm{T}}u$ and $v^{\mathrm{T}}v$ and $\frac{\mathrm{d}}{\mathrm{d}t}(x^{\mathrm{T}}Px)$ as quadratic expressions in (x,u):

$$\begin{aligned}
y^{\mathrm{T}}y + u^{\mathrm{T}}u &= \begin{bmatrix} x^{\mathrm{T}} & u^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} C^{\mathrm{T}}C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}, \\
v^{\mathrm{T}}v &= (Fx+u)^{\mathrm{T}}(Fx+u) \\
&= \begin{bmatrix} x^{\mathrm{T}} & u^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} F^{\mathrm{T}}F & F^{\mathrm{T}} \\ F & I \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}, \\
\tfrac{\mathrm{d}}{\mathrm{d}t}(x^{\mathrm{T}}Px) &= \dot{x}^{\mathrm{T}}Px + x^{\mathrm{T}}P\dot{x} \\
&= (x^{\mathrm{T}}A^{\mathrm{T}} + u^{\mathrm{T}}B^{\mathrm{T}})Px + x^{\mathrm{T}}P(Ax+Bu) \\
&= \begin{bmatrix} x^{\mathrm{T}} & u^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A^{\mathrm{T}}P + PA & PB \\ B^{\mathrm{T}}P & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}.
\end{aligned}$$

Therefore

$$y^{\mathrm{T}}y + u^{\mathrm{T}}u - v^{\mathrm{T}}v + \tfrac{\mathrm{d}}{\mathrm{d}t}(x^{\mathrm{T}}Px) = \begin{bmatrix} x^{\mathrm{T}} & u^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A^{\mathrm{T}}P + PA + C^{\mathrm{T}}C - F^{\mathrm{T}}F & PB - F^{\mathrm{T}} \\ B^{\mathrm{T}}P - F & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}.$$

If P is symmetric and satisfies (5.6) with F equal to $F = B^{\mathrm{T}}P$ then the right-hand side of the above equation is zero. So then

$$y^{\mathrm{T}}y + u^{\mathrm{T}}u = -\tfrac{\mathrm{d}}{\mathrm{d}t}(x^{\mathrm{T}}Px) + v^{\mathrm{T}}v$$

and, hence, the cost (5.5) equals

$$\int_0^\infty y^{\mathrm{T}}(t)y(t) + u^{\mathrm{T}}(t)u(t)\,\mathrm{d}t = x_0^{\mathrm{T}}Px_0 + \int_0^\infty v^{\mathrm{T}}(t)v(t)\,\mathrm{d}t$$

whenever the input stabilizes the system. Given x_0 the above cost is minimal for $v = 0$, provided it stabilizes. It does: since $v = Fx + u$ we have $v = 0$ iff $u = -Fx = -B^{\mathrm{T}}Px$ and so the closed loop system is $\dot{x} = (A - BB^{\mathrm{T}}P)x$, which, by assumption on P, is asymptotically stable.

In the following section we show that stabilizability and detectability implies existence and uniqueness of P, and that it is symmetric: Thm. 5.3.1(4). ∎

The quadratic matrix equation (5.6) is famous in Systems Theory. It is known as an Algebraic Riccati Equation (or ARE for short) because of its connection with certain quadratic differential equations studied by Jacopo Riccati (1676–1754). We say that P is a stabilizing solution of the ARE if $A - BB^{\mathrm{T}}P$ is asymptotically stable.

This theorem is awesome! Notice, for example, that the LQ problem definition does not say we have to limit our inputs to linear static state feedbacks, $u(t) = -Fx(t)$. It just so happens to be the case. Nice. Also, Theorem 5.2.2 applies to multiple-input-multiple-output systems. That is, u and/or y may consist of several components. To say it differently, B may contain several columns, and C several rows.

Example 5.2.3 (Integrator system). Consider again the integrator system, $\dot{x} = u, y = x$. Then $A = 0, B = C = 1$, so $F := B^{\mathrm{T}}P = P$ and the ARE (5.6) becomes

$$0 = A^{\mathrm{T}}P + PA - C^{\mathrm{T}}C - PBB^{\mathrm{T}}P = 1 - P^2.$$

We find $F = P = \pm 1$, and since $A - BB^{\mathrm{T}}P = -P$ needs to be asymptotically stable we again find $F = P = +1$. Done. □

5.3 Hamiltonian Matrix & Stable Subspace

We have not yet proved existence and uniqueness of stabilizing solutions of AREs in case the system is stabilizable and detectable. This, and more, is resolved in this section.

The analysis starts with an alternative representation of the ARE: a matrix P satisfies the ARE iff $-C^{\mathrm{T}}C - A^{\mathrm{T}}P = P(A - BB^{\mathrm{T}}P)$, that is, iff

$$\underbrace{\begin{bmatrix} A & -BB^{\mathrm{T}} \\ -C^{\mathrm{T}}C & -A^{\mathrm{T}} \end{bmatrix}}_{H} \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} I \\ P \end{bmatrix} (A - BB^{\mathrm{T}}P). \tag{5.7}$$

We introduced here the $(2n)\times(2n)$-matrix H, and it is known as the Hamiltonian matrix of the LQ problem. This form is interesting, because if all matrices here are numbers, then it says that $\left[\begin{smallmatrix} I \\ P \end{smallmatrix}\right]$ is an eigenvector of H with eigenvalue $A - BB^{\mathrm{T}}P$. This idea generalizes perfectly to higher dimensions, and the result is the basis for numerical computation of P, and along the way it establishes that P exists and is unique and symmetric (if the system is stabilizable and detectable):

Theorem 5.3.1 (From stable subspace to P). *Define $H \in \mathbb{R}^{(2n)\times(2n)}$ as in* (5.7)*. If (A, B) is stabilizable and (C, A) detectable, then*

1. *H has no imaginary eigenvalues,*
2. *the stable subspace[1] of H has dimension n, and matrices $V \in \mathbb{R}^{(2n)\times n}$ exist whose columns span this stable subspace,*
3. *for any such $V \in \mathbb{R}^{(2n)\times n}$, if we partition V as $V = \left[\begin{smallmatrix} V_1 \\ V_2 \end{smallmatrix}\right]$ with $V_1, V_2 \in \mathbb{R}^{n\times n}$, then V_1 is invertible,*
4. *the ARE* (5.6) *has a unique stabilizing solution P. In fact*

$$P := V_2 V_1^{-1},$$

is the unique answer, and it is symmetric.

[1] The stable subspace $\mathscr{X}$ of H is the largest H-invariant subspace of $\mathbb{R}^{2n}$ restricted to which all eigenvalues of $H : \mathscr{X} \to \mathscr{X}$ are stable. If the columns of $X \in \mathbb{R}^{2n\times k}$ form a basis of this stable subspace $\mathscr{X}$ then $HX = X\Lambda$ for some $k \times k$ matrix Λ, and all eigenvalues of Λ are stable.

Proof. The proof is involved! It makes frequent use of properties of the corresponding Hamiltonian system

$$\begin{bmatrix}\dot{x}\\ \dot{z}\end{bmatrix} = \underbrace{\begin{bmatrix} A & -BB^{\mathrm{T}} \\ -C^{\mathrm{T}}C & -A^{\mathrm{T}}\end{bmatrix}}_{H}\begin{bmatrix}x\\ z\end{bmatrix}, \quad \begin{bmatrix}x(0)\\ z(0)\end{bmatrix} = \begin{bmatrix}x_0\\ z_0\end{bmatrix} \in \mathbb{C}^{2n}. \tag{5.8}$$

Its solution $\left[\begin{smallmatrix}x\\ z\end{smallmatrix}\right]$ has the property that

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}(z^*x) &= \dot{z}^*x + z^*\dot{x}\\ &= (-C^{\mathrm{T}}Cx - A^{\mathrm{T}}z)^*x + z^*(Ax - BB^{\mathrm{T}}z)\\ &= -x^*C^{\mathrm{T}}Cx - z^*BB^{\mathrm{T}}z\\ &= -\big(\|Cx\|^2 + \|B^{\mathrm{T}}z\|^2\big).\end{aligned} \tag{5.9}$$

Integrating (5.9) over $t \in [0,\infty)$ reveals that

$$\int_0^\infty \|Cx(t)\|^2 + \|B^{\mathrm{T}}z(t)\|^2 \,\mathrm{d}t = z_0^*x_0 - \lim_{t\to\infty} z^*(t)x(t), \tag{5.10}$$

provided the limit exists. Now we can prove the four statements.

1. Suppose $\left[\begin{smallmatrix}x_0\\ z_0\end{smallmatrix}\right]$ is an eigenvector of H with imaginary eigenvalue λ. Then $\left[\begin{smallmatrix}x(t)\\ z(t)\end{smallmatrix}\right] = \mathrm{e}^{\lambda t}\left[\begin{smallmatrix}x_0\\ z_0\end{smallmatrix}\right]$. Notice that $z^*(t)x(t)$ is constant. Equation (5.9) thus says that both Cx and $B^{\mathrm{T}}z$ are zero for all time. Inserting this into (5.8) shows that $\lambda x_0 = Ax_0$ and $\lambda z_0 = -A^{\mathrm{T}}z_0$. Hence $\left[\begin{smallmatrix}A-\lambda I\\ C\end{smallmatrix}\right]x_0 = 0$ and $z_0^*\left[\begin{smallmatrix}A+\overline{\lambda} I & B\end{smallmatrix}\right] = 0$. Stabilizability and detectability imply that $x_0 = 0, z_0 = 0$, but $\left[\begin{smallmatrix}x_0\\ z_0\end{smallmatrix}\right]$ is an eigenvector, so nonzero. Contradiction, therefore H has no imaginary eigenvalues.

2. Exercise 5.12 shows that $r(\lambda) := \det(\lambda I - H)$ equals $r(-\lambda)$. Thus, H has as many stable eigenvalues as unstable eigenvalues. Since H has no imaginary eigenvalues we conclude that the stable subspace has dimension n. This can be spanned by some (nonunique) $V \in \mathbb{R}^{(2n)\times n}$.

3. Suppose, to obtain a contradiction, that V_1 is singular. Then the stable subspace, $\mathrm{im}(\left[\begin{smallmatrix}V_1\\ V_2\end{smallmatrix}\right])$, of H contains a $\left[\begin{smallmatrix}x_0\\ z_0\end{smallmatrix}\right]$ with $x_0 = 0, z_0 \neq 0$. Being in the stable subspace, the solution $\left[\begin{smallmatrix}x(t)\\ z(t)\end{smallmatrix}\right]$ for this initial condition converges to zero as $t \to \infty$. Hence the integral in (5.10) equals $z_0^*x_0 = 0$. That can only be if Cx and $B^{\mathrm{T}}z$ are zero for all time. Equation (5.8) then implies that $\dot{z}(t) = -A^{\mathrm{T}}z(t), z(0) = z_0$. We claim that this contradicts stabilizability. Indeed, since $B^{\mathrm{T}}z(t) = 0$ for all time, we have

   $$\dot{z}(t) = -(A^{\mathrm{T}} - LB^{\mathrm{T}})z(t), \quad z(0) = z_0 \neq 0 \tag{5.11}$$

   for every L. By stabilizability there is an L such that $A - BL^{\mathrm{T}}$ is asymptotically stable. Then $-(A^{\mathrm{T}} - LB^{\mathrm{T}})$ is anti-stable, so the solution $z(t)$ of (5.11) diverges. But we also have $\lim_{t\to\infty} z(t) = 0$. Contradiction, so the assumption that V_1 is singular is wrong.

4. Let $P = V_2V_1^{-1}$. Since $V \in \mathbb{R}^{(2n)\times n}$ spans the stable subspace of H, also $VV_1^{-1} = \left[\begin{smallmatrix}I\\ P\end{smallmatrix}\right]$ does so. That is,

   $$\begin{bmatrix} A & -BB^{\mathrm{T}} \\ -C^{\mathrm{T}}C & -A^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}I\\ P\end{bmatrix} = \begin{bmatrix}I\\ P\end{bmatrix}\Lambda \tag{5.12}$$

for some asymptotically stable $\Lambda \in \mathbb{R}^{n\times n}$. Premultiplying the left and right-hand side of (5.12) with $\begin{bmatrix} -P & I \end{bmatrix}$ shows that

$$\begin{bmatrix} -P & I \end{bmatrix} \begin{bmatrix} A & -BB^{\mathrm{T}} \\ -C^{\mathrm{T}}C & -A^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} = 0.$$

This equation is the ARE! And P is a stabilizing solution because $A - BB^{\mathrm{T}}P = \Lambda$ which is asymptotically stable. So at least one stabilizing solution of the ARE exists (great news). Now suppose there would be another stabilizing solution of the ARE, say $\tilde{P}$. Then $H\begin{bmatrix} I \\ \tilde{P} \end{bmatrix} = \begin{bmatrix} I \\ \tilde{P} \end{bmatrix}(A - BB^{\mathrm{T}}\tilde{P})$. It would mean that H has two different stable subspaces: $\mathrm{im}(\begin{bmatrix} I \\ P \end{bmatrix}) \neq \mathrm{im}(\begin{bmatrix} I \\ \tilde{P} \end{bmatrix})$. Not the case, so there is precisely one stabilizing solution of the ARE. (Exercise 5.11 does another proof.)

About symmetry: $\begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} H$ equals $\begin{bmatrix} C^{\mathrm{T}}C & A^{\mathrm{T}} \\ A & -BB^{\mathrm{T}} \end{bmatrix}$ and so is symmetric. Then also the following is symmetric,

$$\begin{bmatrix} I & P^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} H \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} P^{\mathrm{T}} & -I \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} \Lambda = (P^{\mathrm{T}} - P)\Lambda.$$

Hence $(P^{\mathrm{T}} - P)\Lambda$ equals $\Lambda^{\mathrm{T}}(P - P^{\mathrm{T}})$, that is,

$$\Lambda^{\mathrm{T}}(P^{\mathrm{T}} - P) + (P^{\mathrm{T}} - P)\Lambda = 0.$$

In Exercise 5.10 we show that this, combined with asymptotic stability of Λ, implies that $P = P^{\mathrm{T}}$.

We really proved a central result in Systems Theory! ∎

Notice that *any* $V \in \mathbb{R}^{(2n)\times n}$ whose columns span the stable subspace of H does the job. That is, even though we can span the stable subspace of H in many different ways, whatever such V we take, we always have that V_1 is invertible and that P uniquely follows as $P = V_2 V_1^{-1}$. Usually we construct V from the eigenvectors of H with stable eigenvalues.

Example 5.3.2 ($n = 1$). Consider once again the integrator system $\dot{x}(t) = u(t)$ and cost $\int_0^\infty x^2(t) + u^2(t)\,\mathrm{d}t$. That is, $A = 0, B = C = 1$. The Hamiltonian matrix becomes

$$H = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

Its characteristic polynomials is $\lambda^2 - 1$, and the eigenvalues are $\lambda_{1,2} = \pm 1$. Its stable eigenvalue is $\lambda_{\mathrm{stab}} = -1$, and it is easy to verify that v is an eigenvector corresponding to this stable eigenvalue iff

$$v := \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} c, \qquad c \neq 0.$$

According to Theorem 5.3.1 the stabilizing solution P of the ARE is

$$P = v_2 v_1^{-1} = \frac{v_2}{v_1} = \frac{c}{c} = 1.$$

As predicted, P does not depend on the choice of eigenvector (the choice of c). As predicted, the (eigen)value of $A - BB^{\mathrm{T}}P = -1$ equals the stable eigenvalue of the Hamiltonian matrix, $\lambda_{\mathrm{stab}} = -1$. The optimal control is $u = -Fx := -B^{\mathrm{T}}Px = -x$. □

Example 5.3.3 ($n = 2$). Consider the controllable and observable system

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

and with standard cost

$$\int_0^\infty y_1^2(t) + y_2^2(t) + u^2(t)\,\mathrm{d}t.$$

(By the way, notice that now C is the 2×2 identity matrix.) The associated Hamiltonian matrix is,

$$H = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ \hline -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 \end{array}\right],$$

(verify this yourself). Its characteristic polynomial is $\lambda^4 - \lambda^2 + 1$, and the four eigenvalues turn out to be

$$\lambda_{1,2} = -\tfrac{1}{2}\sqrt{3} \pm \tfrac{1}{2}\mathrm{i}, \qquad \lambda_{3,4} = +\tfrac{1}{2}\sqrt{3} \pm \tfrac{1}{2}\mathrm{i}.$$

The first two, $\lambda_{1,2}$, are stable so we need eigenvectors corresponding to these two, $\lambda_{1,2}$. Not very enlightening manipulation shows we can take

$$v_{1,2} = \begin{bmatrix} -\lambda_{1,2} \\ -\lambda_{1,2}^2 \\ 1 \\ \lambda_{1,2}^3 \end{bmatrix}.$$

Now $V \in \mathbb{C}^{4\times 2}$ defined as

$$V = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} -\lambda_1 & -\lambda_2 \\ -\lambda_1^2 & -\lambda_2^2 \\ 1 & 1 \\ \lambda_1^3 & \lambda_2^3 \end{bmatrix}$$

spans the stable subspace of H. (The fact that these are complex is not a problem.) With V known, it is a piece of cake to compute the stabilizing solution of the ARE,

$$P = V_2 V_1^{-1} = \begin{bmatrix} 1 & 1 \\ \lambda_1^3 & \lambda_2^3 \end{bmatrix} \begin{bmatrix} -\lambda_1 & -\lambda_2 \\ -\lambda_1^2 & -\lambda_2^2 \end{bmatrix}^{-1} = \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix}.$$

The optimal input is $u = -B^{\mathrm{T}}Px = -x_1 - \sqrt{3}x_2$. The LQ-optimal closed loop system is described by

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & \sqrt{3} \end{bmatrix} x,$$

and its closed loop poles/eigenvalues are $\lambda_{1,2} = -\frac{1}{2}\sqrt{3} \pm \frac{1}{2}\mathrm{i}$ (the stable eigenvalues of H, as expected). □

In this example the characteristic polynomial, $\lambda^4 - \lambda^2 + 1$, is of degree $2n = 4$, but by letting $\mu = \lambda^2$ we see that it is effectively of degree $n = 2$: $\mu^2 - \mu + 1$. This always works, see Exercise 5.12.

5.4 Positive semi-definite matrices

A short section. So far we have not payed attention to a key property of P: if it is the solution of the LQ problem, then

$$x_0^{\mathrm{T}} P x_0 \geq 0 \qquad \forall\, x_0 \in \mathbb{R}^n. \tag{5.13}$$

This is immediate from the fact that all costs $\left\| \left[\begin{smallmatrix} y \\ u \end{smallmatrix} \right] \right\|_{\mathbb{L}_{2+}}^2$ are nonnegative, so also the minimal cost $x_0^{\mathrm{T}} P x_0$. Symmetric matrices P that satisfy (5.13) are called *positive semi-definite* matrices, notation:

$$P \geq 0.$$

This opens up other ways to determine the LQ-solution P.

Example 5.4.1. Consider once more the integrator system of Example 5.2.3. The ARE is $1 - P^2 = 0$. So P is either $+1$ or -1. Which one is it? Earlier we analyzed stability of $A - BB^{\mathrm{T}}P$ in order to determine the correct P, but because of (5.13) we can immediately discard $P = -1$. And since we know that in this example a stabilizing solution exists, it must be $P = +1$. Ready. □

The addition "since we know a stabilizing solution exists" is for a good reason. For example if $A = B = C = 0$ then the ARE has no stabilizing solution ((A, B) is not stabilizable) yet the ARE itself is trivial, $0P = 0$, and has infinitely many solutions, among which also positive solutions.

We usually assume both stabilizability and detectability, and then a solution does exist, and we have the following pleasant result:

Lemma 5.4.2 ("positive semi-definite = stabilizing"). *Suppose (A, B) is stabilizable and (C, A) detectable, and that P satisfies the ARE. Then P is a stabilizing solution of the ARE iff it is symmetric and positive semi-definite.*

Proof. ($\Longrightarrow$) This is the trivial part: if P is a stabilizing solution of ARE then Thm. 5.2.2 shows that $P = P^{\mathrm{T}}$ and that $x_0^{\mathrm{T}} P x_0$ is the optimal cost for every x_0. Clearly costs are nonnegative.

($\Longleftarrow$) Suppose $P = P^{\mathrm{T}} \geq 0$, and let x be an eigenvector of $A - BB^{\mathrm{T}}P$ with eigenvalue λ. We show that $\mathrm{Re}(\lambda) < 0$. The trick is to rewrite the ARE as

$$(A - BB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - BB^{\mathrm{T}}P) + C^{\mathrm{T}}C + PBB^{\mathrm{T}}P = 0.$$

Next, postmultiply this equation with the eigenvector x, and premultiply with x^*:

$$x^*\big((A - BB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - BB^{\mathrm{T}}P) + C^{\mathrm{T}}C + PBB^{\mathrm{T}}P\big)x = 0.$$

Since x is an eigenvector of $A - BB^{\mathrm{T}}P$ this simplifies to a sum of three terms,

$$(\lambda^* + \lambda)(x^*Px) + \|Cx\|^2 + \|B^{\mathrm{T}}Px\|^2 = 0.$$

If $\mathrm{Re}(\lambda) \geq 0$ then $(\lambda^* + \lambda)x^*Px \geq 0$, implying that all the above three terms must be zero: $(\lambda^* + \lambda)x^*Px = 0$, $Cx = 0$, and $B^{\mathrm{T}}Px = 0$ (and, consequently, $Ax = \lambda x$). This contradicts detectability. So it cannot be that $\mathrm{Re}(\lambda) \geq 0$. It must be that $A - BB^{\mathrm{T}}P$ is asymptotically stable. ■

Example 5.4.3. In Example 5.3.3 we found $P = \left[\begin{smallmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{smallmatrix}\right]$. The system is controllable and observable, hence the conditions of Lemma 5.4.2 are met. So the previous Lemma guarantees that $P = P^{\mathrm{T}} \geq 0$. We can verify this directly as well using completion of squares,

$$\begin{aligned} x_0^{\mathrm{T}} P x_0 &= \begin{bmatrix} x_{01} & x_{02} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} \\ &= \sqrt{3} x_{01}^2 + 2 x_{01} x_{02} + \sqrt{3} x_{02}^2 \\ &= \underbrace{\sqrt{3}}_{>0} (x_{01} + \tfrac{1}{\sqrt{3}} x_{02})^2 + \underbrace{(\sqrt{3} - \tfrac{1}{\sqrt{3}})}_{>0} x_{02}^2. \end{aligned}$$

The final expression clearly is nonnegative for all $\left[\begin{smallmatrix} x_{01} \\ x_{02} \end{smallmatrix}\right]$. In fact it is positive for all $\left[\begin{smallmatrix} x_{01} \\ x_{02} \end{smallmatrix}\right] \neq 0$. □

In Exercise 5.6 we analyze positive semi-definiteness for arbitrary 2×2 matrices. There are very interesting connections with Lyapunov functions, but we do not explore that here.

5.5 Applications

Our standard cost is of the form

$$\int_0^\infty y^{\mathrm{T}}(t) y(t) + u^{\mathrm{T}}(t) u(t) \, \mathrm{d}t,$$

but in applications we want to include a positive tuning parameter ρ,

$$\int_0^\infty y^{\mathrm{T}}(t) y(t) + \rho u^{\mathrm{T}}(t) u(t) \, \mathrm{d}t.$$

This does not complicate matters, because the new input $\tilde{u}$ defined as $\tilde{u} = \sqrt{\rho} u$ brings us back to the standard cost. As Exercise 5.7 explains, the ARE for the altered cost becomes

$$A^{\mathrm{T}} P + P A + C^{\mathrm{T}} C - \tfrac{1}{\rho} P B B^{\mathrm{T}} P = 0, \tag{5.14}$$

and we need P such that $A - \frac{1}{\rho} B B^{\mathrm{T}} P$ is asymptotically stable, and then the optimal input is

$$u = -\tfrac{1}{\rho} B^{\mathrm{T}} P x,$$

and the optimal cost "remains the same," $x_0^{\mathrm{T}} P x_0$.

Example 5.5.1 (Tuning the controller). Consider the observable and controllable system

$$\begin{aligned} &\dot{x} = u, \qquad x(0) = 1, \\ &y = 2x, \end{aligned}$$

with a cost that includes a tuning parameter $\rho > 0$,

$$\int_0^\infty y^2(t) + \rho u^2(t) \, \mathrm{d}t.$$

The ARE (5.14) and optimal input for this problem are

$$2^2 - \tfrac{1}{\rho} P^2 = 0, \qquad u = -\tfrac{1}{\rho} P x.$$

Clearly this means $P = \pm 2\sqrt{\rho}$. Since P needs to be nonnegative, we find that

$$P = +2\sqrt{\rho}, \qquad u = -\tfrac{2}{\sqrt{\rho}} x,$$

and that the closed loop pole is $A - \frac{1}{\rho}BB^{\mathrm{T}}P = -2/\sqrt{\rho}$.

If $\rho = 1$ then the control input u is "as cheap" as y. The closed loop pole is -2, and the optimal u and y have the same magnitude: $|u(t)| = |y(t)| = 2e^{-2t}$. (See the solid red graphs of Fig. 5.1.)

If $0 < \rho \ll 1$ then control input u is cheap. The closed loop system is fast now (the closed loop pole is $-2/\sqrt{\rho} \ll -2 < 0$), and both u, y converge to zero fast, but u initially is relatively large (in magnitude): $|u(0)| = 2/\sqrt{\rho} = |y(0)|/\sqrt{\rho}$. That is to be expected since control is cheap. (See the dotted black graphs of Fig. 5.1.)

Conversely, if $\rho \gg 1$ then the input u is expensive. The closed loop system is now slow (the closed loop pole is $-2/\sqrt{\rho} \approx 0$), and both u, y converge to zero slowly, although u initially is already small: $u(0) = -2/\sqrt{\rho} \approx 0$ but that is to be expected since control is expensive. (See the dashed blue graphs of Fig. 5.1.)

It is not hard to see that for the optimal solution we have $\|u\|^2_{\mathbb{L}_{2+}} = 1/\sqrt{\rho}$ and $\|y\|^2_{\mathbb{L}_{2+}} = \sqrt{\rho}$. Hence

$$\|y\|^2_{\mathbb{L}_{2+}} = \frac{1}{\|u\|^2_{\mathbb{L}_{2+}}}.$$

This relation establishes once more that small inputs (sluggish control) results in large outputs, and that large inputs (forceful control) results in small outputs. See Fig. 5.2. The parameter ρ has a nice geometric interpretation: the LQ-optimal solution is where the tangent has slope $-\rho$, see Fig. 5.2. □

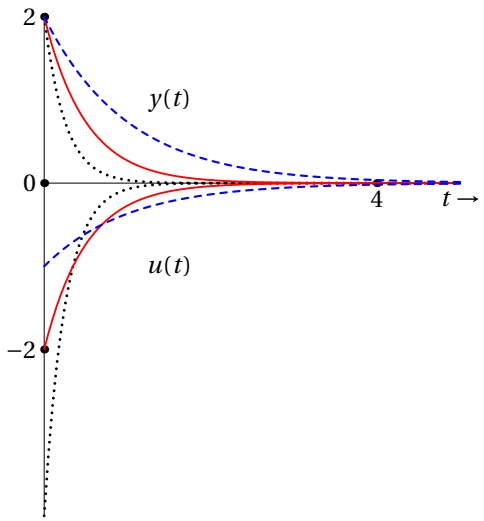

FIGURE 5.1: Graphs of optimal y and u for $\rho = 1$ (solid red), for $\rho = 1/4$ (dotted black), and $\rho = 4$ (dashed blue). The larger ρ is the slower the system is and the smaller $|u(0)|$ is. See Example 5.5.1.

A common strategy in practical LQ controller design is to select a limited number of signals $y_1, y_2, \ldots$ that we would like to be small. In this context, y is not necessarily the output that we can measure and have available for feedback. In the next example, for instance, we might be able to measure the entire state, but we only care to control the position of the car, so we take the position as our "y" in the LQ cost function.

Example 5.5.2 (Mass-spring system). Consider a car of mass m connected to a wall via a spring with spring constant k, see Fig. 5.3. The position of the car is denoted y and we can

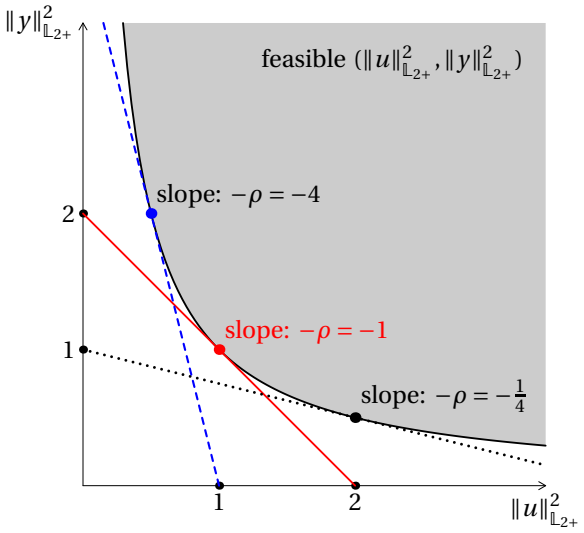

FIGURE 5.2: The shaded area are the feasible $(\|u\|^2_{\mathbb{L}_{2+}}, \|y\|^2_{\mathbb{L}_{2+}})$ by some choice of u. On the boundary we have $\|y\|^2_{\mathbb{L}_{2+}} = 1/\|u\|^2_{\mathbb{L}_{2+}}$. The LQ-optimal solution for a given $\rho > 0$ is the point on the boundary where the tangent has slope $-\rho$. See Example 5.5.1 and Fig. 5.1.

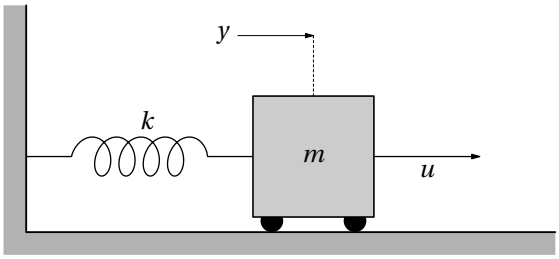

FIGURE 5.3: Car attached to a wall via a spring and with a force control u. See Example 5.5.2.

control the car with a force u. Newton's second law says that

$$m\ddot{y} + ky = u.$$

To keep matters simple we take $k = 1$ and $m = 1$, so

$$\ddot{y} + y = u.$$

For zero input the car would oscillate forever. The task of the controller is to bring the car quickly to a stand still at position $y = 0$ but without using excessive force u. We propose, therefore, to take as cost the classic one,

$$\int_0^\infty y^2(t) + \rho u^2(t)\,\mathrm{d}t.$$

As always, the tuning parameter $\rho > 0$ defines the trade-off between small y and small u. A state representation of our system, with output y, is

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

Here, the first state component is $x_1 = y$ and the second is $x_2 = \dot{y}$. For[2]

$$\rho = 1/3$$

the stabilizing solution of the ARE (5.14) happens to be

$$P = \frac{1}{3}\begin{bmatrix} 2\sqrt{2} & 1 \\ 1 & \sqrt{2} \end{bmatrix} = \begin{bmatrix} 2.8284 & 0.3333 \\ 0.3333 & 0.4714 \end{bmatrix},$$

and the closed-loop poles are $-\frac{1}{2}\sqrt{2} \pm \sqrt{3/2}\mathrm{i}$ and

$$F := \tfrac{1}{\rho} B^{\mathrm{T}} P = \begin{bmatrix} 1 & \sqrt{2} \end{bmatrix}.$$

Therefore the LQ-optimal state feedback is

$$u = -Fx = -F\begin{bmatrix} y \\ \dot{y} \end{bmatrix} = -y - \sqrt{2}\dot{y}.$$

One can implement this controller as a spring with spring constant $k_{\mathrm{LQ}} = 1$ parallel to a damper with damping coefficient $r_{\mathrm{LQ}} = \sqrt{2}$, see Fig. 5.4(left). For $x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ the LQ-optimal input and output converge to zero quickly, although there is some overshoot, see Fig. 5.4(right). □

Example 5.5.3 (Tuning position and speed). This example we took from lecture notes by Leonid Mirkin. Consider the system

$$\begin{aligned}\dot{x} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & -0.1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u \\ q &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} x.\end{aligned}$$

[2] We took this ρ because then explicit answers are fairly nice, but we skip the derivation because it is not fun.

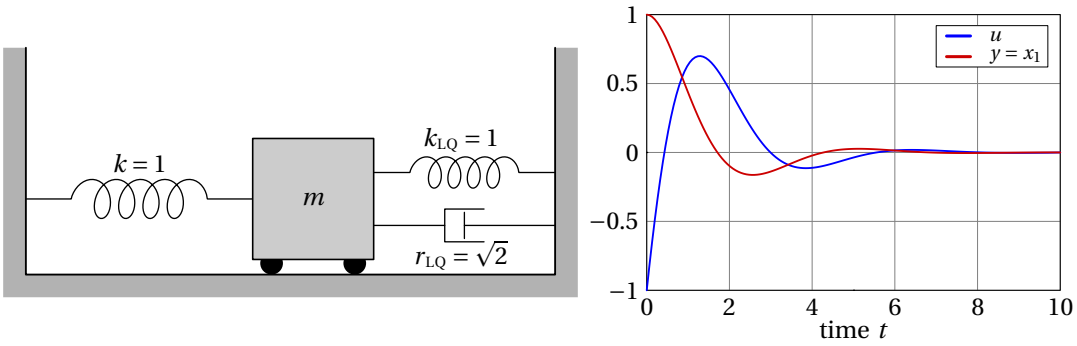

FIGURE 5.4: Left: car attached to a wall with LQ optimal force $u = -k_{\mathrm{LQ}}y - r_{\mathrm{LQ}}\dot{y}$ implemented as spring/damper system. Right: responses u and $y = x_1$ for initial state $x(0) = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, see Example 5.5.2.

We want to steer q to zero quickly but not too steeply, so $\dot{q}$ should be small as well and all that using small u. This calls for a cost function that combines q, $\dot{q}$ and u:

$$\int_0^\infty \lambda q^2(t) + (1-\lambda)\dot{q}^2(t) + \rho u^2(t)\,\mathrm{d}t.$$

Now we have two tuning parameters: besides the standard $\rho > 0$ we also have $\lambda \in [0,1]$, which defines a relative trade-off between small q and small $\dot{q}$. By the way, since $q = Cx$ we have $\dot{q} = C\dot{x} = CAx + CBu = \begin{bmatrix}0 & 1 & 1\end{bmatrix} x$. So we define

$$y := \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} := \begin{bmatrix} \sqrt{\lambda}q & 0 \\ 0 & \sqrt{1-\lambda}\dot{q} \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda} & 0 \\ 0 & \sqrt{1-\lambda} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} x.$$

Given $\lambda \in [0,1]$ and $\rho > 0$ the LQ solution can now be determined. In what follows we take as initial state

$$x_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Figure 5.5 shows the response of the optimal u and q for various combinations of λ and ρ. For $\lambda = 1$ the term $\dot{q}$ is not included in the cost so we can expect "steep" behavior in the output. The larger λ is the slower the output converges to zero. As for ρ, we see that smaller ρ means larger control action u and, consequently, faster convergence to zero of the output q.

Assuming we can live with control actions u of at most 2 then $\rho = 0.2$ might be a reasonable choice (the red graphs in Fig. 5.5(left)). Given that, a value of $\lambda = 0.75$ might be a good compromise between overshoot and settling time in the response q. For this $\rho = 0.2, \lambda = 0.75$ the optimal state feedback turns out to be

$$u = -Fx = -(1.9365x_1 + 3.0656x_2 + 2.6187x_3)$$

and the closed loop poles are -0.7468 and $-0.9859 \pm 1.2732\mathrm{i}$. □

5.6 Exercises

5.1 Determine the LQ optimal input and cost for the system $\dot{x} = 3x + 4u, y = x$.

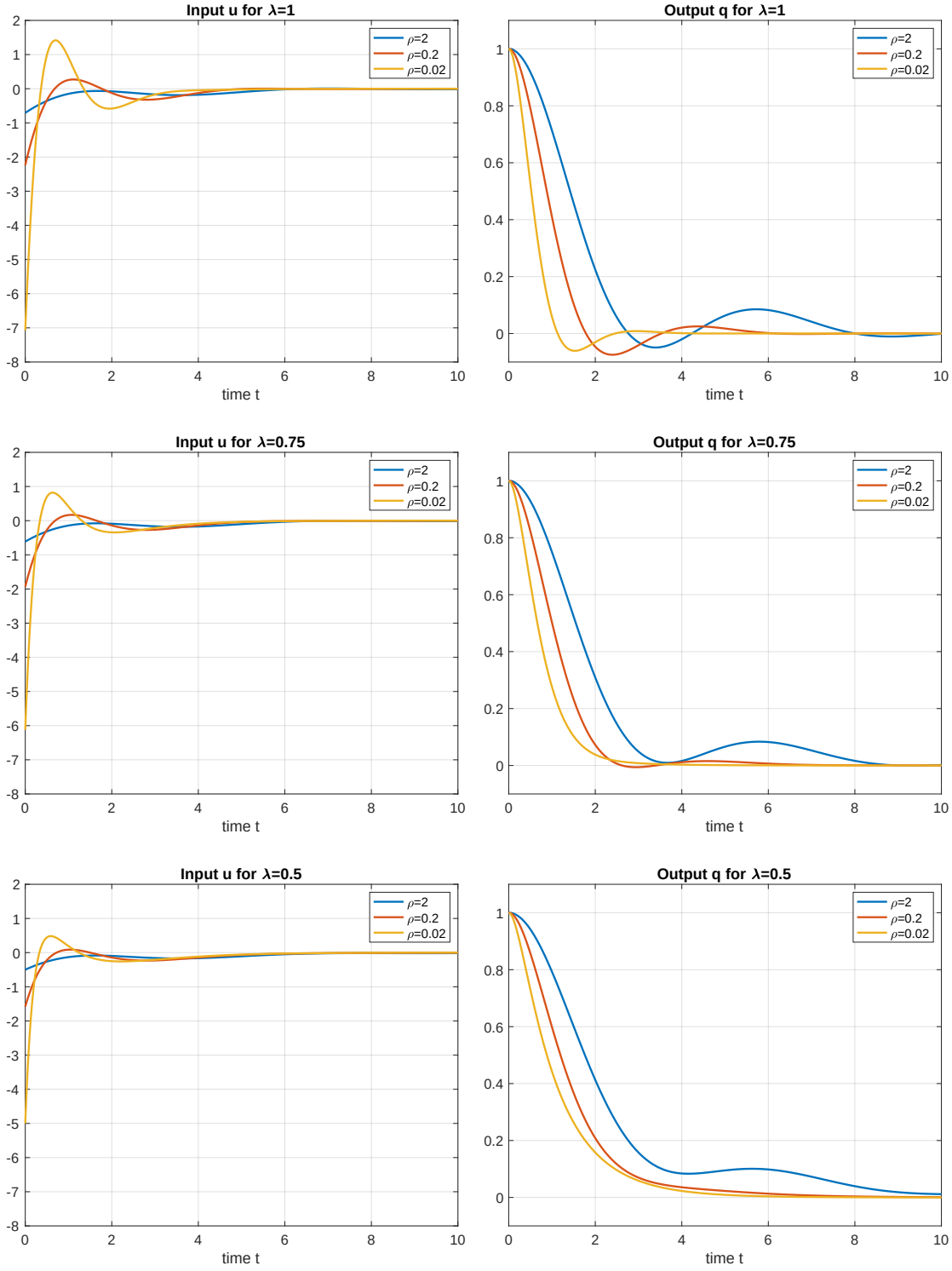

FIGURE 5.5: LQ optimal responses of u (left) and q (right) for various combinations of λ and ρ as explained in Example 5.5.3.

5.2 Minimize $\int_0^\infty x^2(t)+u^2(t)\,\mathrm{d}t$ over all stabilizing inputs for the system $\dot{x}=x+u, x(0)=x_0$.

5.3 Consider the system $\dot{x}=3x+2u, y=0$ and cost $\int_0^\infty u^2(t)\,\mathrm{d}t$.

(a) Is the system detectable? And what does this mean (considering Theorem 5.2.2)?

(b) Solve the LQ problem.

(c) The above system is unstable. Why is the LQ problem with cost $\int_0^\infty u^2(t)\,\mathrm{d}t$ for the stable systems, say, $\dot{x}=-3x+2u, y=0$ trivial?

5.4 Solve the LQ problem for system $\dot{x}=x/2+u, x(0)=x_0$ and modified cost $\int_0^\infty \mathrm{e}^{-t}(x^2(t)+u^2(t))\,\mathrm{d}t$. [Hint: introduce a new state $z(t):=\mathrm{e}^{-t/2}x(t)$.]

5.5 Consider the system

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

with standard cost $\|\left[\begin{smallmatrix}u\\ y\end{smallmatrix}\right]\|^2_{\mathbb{L}_{2+}}$.

(a) Is the system stabilizable and detectable?

(b) Show that $P:=\left[\begin{smallmatrix}2 & 1\\ 1 & 3\end{smallmatrix}\right]$ satisfies the ARE.

(c) Show that this P satisfies $P=P^{\mathrm{T}}\geq 0$.

(d) Is the above P a stabilizing solution?

(e) Determine the F such that $u:=-Fx$ is LQ-optimal.

(f) Determine the eigenvalues of the Hamiltonian matrix H. (This is an easy problem, given the above.)

(g) *Linear Algebra:* can we span the stable subspace of H with eigenvectors of H?

5.6 *Positive semi-definite* 2×2 *matrices.* In LQ problems where the state has $n=2$ components, the optimal cost is of the form

$$V(x):=\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}p & q\\ q & r\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$$

for some $p,q,r\in\mathbb{R}$.

(a) Verify that $V(x)=px_1^2+2qx_1x_2+rx_2^2$.

(b) Suppose $p<0$. Show that $V(x)$ is not positive semi-definite.

(c) Suppose $p=0$. Show that $V(x)$ is positive semi-definite iff $q=0, r\geq 0$

(d) Suppose $p>0$. Show that $V(x)$ is positive semi-definite iff $r\geq q^2/p$. [Hint: examine $V(x)-p(x_1+\frac{q}{p}x_2)^2$.]

(e) Suppose $p>0$. Show that $V(x)$ is positive semi-definite iff $\det(\left[\begin{smallmatrix}p & q\\ q & r\end{smallmatrix}\right])\geq 0$.

5.7 *Literature standard LQ problem.* In the literature (and in software packages) the cost function is assumed of the form

$$\int_0^\infty x^{\mathrm{T}}(t)Qx(t)+u^{\mathrm{T}}(t)Ru(t)\,\mathrm{d}t. \tag{5.15}$$

Here $Q\in\mathbb{R}^{n\times n}$ and $R\in\mathbb{R}^{n_u\times n_u}$, in fact both Q and R are assumed symmetric and positive semi-definite, and R invertible. It can be shown that this is equivalent to demanding

that $Q = C^{\mathrm{T}}C$ for some matrix C, and that $R = W^{\mathrm{T}}W$ for some square invertible matrix W.

Assume that (A, B) is stabilizable and (C, A) detectable, and consider the LQ problem with stability for the above cost (5.15).

(a) Suppose first that $R = \rho I_{n_u}$. Show that there is a unique $P \in \mathbb{R}^{n\times n}$ that satisfies

$$A^{\mathrm{T}}P + PA + Q - \tfrac{1}{\rho}PBB^{\mathrm{T}}P = 0$$

with the property that $A - \frac{1}{\rho}BB^{\mathrm{T}}P$ is asymptotically stable, and that $u := -\frac{1}{\rho}B^{\mathrm{T}}Px$ minimizes (5.15) over all stabilizing inputs.

(b) Consider the general case (so $Q = C^{\mathrm{T}}C, R = W^{\mathrm{T}}W$, and W invertible). Show that there is a unique $P \in \mathbb{R}^{n\times n}$ that satisfies

$$A^{\mathrm{T}}P + PA + Q - PBR^{-1}B^{\mathrm{T}}P = 0$$

with the property that $A - BR^{-1}B^{\mathrm{T}}P$ is asymptotically stable, and that $u := -R^{-1}B^{\mathrm{T}}Px$ minimizes (5.15) over all stabilizing inputs.

5.8 Suppose (A, C) is observable and that the stabilizing solution of P of ARE (5.6) exists. Show that P is invertible.

5.9 Theorem 5.3.1 assumes that (A, C) is detectable. Show that the theorem remains valid if we relax this assumption to that $\left[\begin{smallmatrix} A-sI \\ C \end{smallmatrix}\right]$ has full column rank for all *imaginary* eigenvalues of A. (Then we say "all imaginary eigenvalues of A are observable".)

5.10 *Lyapunov equation.* An import special case of the ARE is when $B = 0$. Then the ARE becomes

$$A^{\mathrm{T}}P + PA + C^{\mathrm{T}}C = 0. \tag{5.16}$$

While AREs are usually quadratic in P, now, with $B = 0$, it is linear. It is known as a *Lyapunov equation.*

(a) If A is asymptotically stable then Thm. 5.2.2 guarantees that the solution P exists and is unique. Why does that follow from Thm. 5.2.2?

(b) Let $B = 0$ and suppose A is asymptotically stable. Show that

$$\int_0^\infty \|Ce^{At}x_0\|^2 \,\mathrm{d}t = x_0^{\mathrm{T}}Px_0,$$

where P is the solution of (5.16).

In the rest of this exercise we analyse the Lyapunov equation (5.16) from scratch (so we do not need to worry about the very complicated Theorem 5.3.1). We assume that W is some $n \times n$ matrix.

(c) Argue that $\frac{\mathrm{d}}{\mathrm{d}t}(e^{A^{\mathrm{T}}t}We^{At}) = A^{\mathrm{T}}(e^{A^{\mathrm{T}}t}We^{At}) + (e^{A^{\mathrm{T}}t}We^{At})A$.

(d) Suppose A is asymptotically stable. Argue that P defined as $P = \int_0^\infty e^{A^{\mathrm{T}}t}C^{\mathrm{T}}Ce^{At}\,\mathrm{d}t$ is well defined and that it satisfies

$$A^{\mathrm{T}}P + PA + W = 0. \tag{5.17}$$

(e) Suppose A is asymptotically stable. The previous part shows that for every $W \in \mathbb{R}^{n\times n}$ there is a $P \in \mathbb{R}^{n\times n}$ that satisfies (5.17). Use the rank-nullity theorem from Linear Algebra to prove that for every $W \in \mathbb{R}^{n\times n}$ there is a *unique* $P \in \mathbb{R}^{n\times n}$ that satisfies (5.17).

(f) Suppose A is asymptotically stable. Argue that (5.16) for every C has a unique solution P and that it is symmetric.

5.11 The results from Exercise 5.10 indicates another proof of uniqueness of stabilizing solutions to AREs: suppose $P, \tilde{P}$ are two stabilizing solutions of ARE (5.6). Show that

$$(A - BB^{\mathrm{T}}P)^{\mathrm{T}}(P - \tilde{P}) + (P - \tilde{P})(A - BB^{\mathrm{T}}\tilde{P}) = 0$$

Use this identity and Exercise 5.10 to show that $P = \tilde{P}$.

5.12 In Example 5.3.3 the characteristic polynomial, $\lambda^4 - \lambda^2 + 1$, is of degree four, but by letting $\mu = \lambda^2$ we see that it is effectively of degree two: $\mu^2 - \mu + 1$. This is possible because odd powers of λ are absent in the characteristic polynomial. This is always the case as we will see in this exercise.

Define the polynomial $r(\lambda)$ as

$$r(\lambda) = \det(\lambda I - H) = \det \begin{bmatrix} \lambda I - A & BB^{\mathrm{T}} \\ C^{\mathrm{T}}C & \lambda I + A^{\mathrm{T}} \end{bmatrix}.$$

(a) Argue that

$$r(\lambda) = \det \begin{bmatrix} -C^{\mathrm{T}}C & -\lambda I - A^{\mathrm{T}} \\ \lambda I - A & BB^{\mathrm{T}} \end{bmatrix}.$$

(b) Show that $r(\lambda) = r(-\lambda)$.

(c) Argue that $r(\lambda)$ is a polynomial in λ^2.

Tougher Exercises

5.13 (This is a laborious exercise.) Consider the 2nd-order DE

$$\ddot{y} = u$$

and cost

$$\int_0^\infty \delta^4 y^2(t) + \sigma^2 \dot{y}^2(t) + \ddot{y}^2(t)\,\mathrm{d}t,$$

in which σ, δ are *positive* tuning parameters.

(a) Determine a state representation of the DE with two state components.

(b) Determine the Hamiltonian matrix H and show that its characteristic polynomial, $\det(sI - H)$ is $s^4 - \sigma^2 s^2 + \delta^4$.

(c) Show that the two stable eigenvalues of H are

$$s_{1,2} = -\sigma\sqrt{1/2 \pm \sqrt{1/4 - (\delta/\sigma)^4}}.$$

(d) In applications we sometimes prefer *real* closed loop poles (so with zero imaginary part), because then we do not have oscillations in the responses. Show that this is the case iff $\sigma \geq \sqrt{2}\delta$.

Among all those σ the "critically damped" case, $\sigma_* = \sqrt{2}\delta$, achieves the fasted $s_{1,2}$ (fasted as in $\max_{i=1,2} s_i$ being as small (negative) as possible). This is considered desirable.

(e) Let $\sigma = \sqrt{2}\delta$. Express the eigenvalues of H in terms of δ.

(f) Let $\sigma = \sqrt{2}\delta$. Determine the solution P of the LQ problem, and express P in terms of δ.

(g) Verify that P is positive semi-definite. [Hint: have a look at Exercise 5.6(e).]

(h) Let $\sigma = \sqrt{2}\delta$. What happens with the optimal cost as $\delta \downarrow 0$, and explain why this is to be expected.

5.14 Assume (A, B) is stabilizable and that (C, A) is observable.

(a) show the stabilizing solution of the ARE is positive definite, meaning $x_0^{\mathrm{T}} P x_0 > 0$ for every $x_0 \neq 0$.

(b) Show that $V(x) := x^{\mathrm{T}} P x$ is a Lyapunov function for the closed loop system $\dot{x} = (A - BB^{\mathrm{T}} P)x$ with equilibrium $\bar{x} = 0$. [Hint: other courses explain what it means to be Lyapunov function for a DE and given equilibrium.]

Appendix A

Some proofs and derivations

This appendix contains a number of proofs and derivations of results used in this course. They are included here because they fall beyond the scope of this course. We also give an overview of the MATLAB scripts we used.

A.1 Weak Solution (Thm. 2.2.4)

Do not read this section! It is included only for completeness. If u is piecewise continuous, then the x defined in (2.11) is well defined and continuous, but not differentiable in the classic sense. What does $\dot{x} = Ax + Bu$ then mean? The usual solution of this paradox is to say that x is a weak solution of $\dot{x} = Ax + Bu$ if x satisfies the integrated equation: $x(t) = x(t_0) + \int_{t_0}^{t} Ax(\tau) + Bu(\tau)\,\mathrm{d}\tau$. The choice of t_0 is irrelevant. One can verify that the following holds.

Lemma A.1.1 (Weak solution). *For every locally integrable*[1] *u, the x in* (2.11) *is well defined, is continuous, and is a weak solution of $\dot{x} = Ax + Bu$.*

Proof (sketch). Since $u(\tau)$ is locally integrable and all components of $\mathrm{e}^{A(t-\tau)}B$ are continuous as functions of τ, every component of $\mathrm{e}^{A(t-\tau)}Bu(\tau)$ is also locally integrable as a function of τ. Consequently, the $x(t)$ defined in (2.11) exists. This $x(t)$ is even continuous because

$$
\begin{aligned}
&\lim_{h\to 0} x(t+h) - x(t) \\
&= \lim_{h\to 0} (\mathrm{e}^{Ah} - I)x(t) + \int_0^h \mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau \\
&= \underbrace{\lim_{h\to 0} (\mathrm{e}^{Ah} - I)x(t)}_{=0} + \underbrace{\lim_{h\to 0} \int_0^h \mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau}_{=0} \\
&= 0.
\end{aligned}
$$

The last limit is zero by the dominant convergence theorem from the theory of Lebesgue integration. This theorem says that $\lim_{n\to\infty} \int f_n(\tau)\,\mathrm{d}\tau = \int f(\tau)\,\mathrm{d}\tau$ if the f_n are (locally) integrable and $f(t) = \lim_{n\to\infty} f_n(t)$ almost everywhere. If we take $f_h(\tau) := \mathrm{e}^{A(t-\tau)}Bu(\tau)$ for $\tau \in [0,h]$ and $f_h(\tau)$ zero elsewhere, and define $f(\tau) := \lim_{h\to 0} f_h(\tau)$, then $f(\tau) = 0$ for all $\tau \neq 0$. In this case, the dominant convergence theorem gives

$$
\lim_{h\to 0} \int_0^h \mathrm{e}^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau = \lim_{h\to 0} \int f_h(\tau)\,\mathrm{d}\tau = \int f(\tau)\,\mathrm{d}\tau = 0.
$$

■

[1] We say that u is locally integrable if $\int_a^b |u_i(t)|\,\mathrm{d}t < \infty$ for all $a, b \in \mathbb{R}$ and all components u_i of u.

In particular, the x in (2.11)) is a well-defined (weak) solution for every piecewise continuous u. The notion of weak solution may look like a trick, but it is not. During the modeling process—when the differential equations are determined—the functions are often assumed to be differentiable purely for convenience. However, they are not necessarily that smooth. In such cases, the weak solutions with their restricted smoothness condition are more natural and more general.

A.2 Routh–Hurwitz (Thm. 1.8.1)

The Routh–Hurwitz test (Thm. 1.8.1) is stated in terms of the Routh table because the test is then easier to carry out. To understand the test, it is better to state it in terms of polynomials.

Theorem A.2.1 (Routh–Hurwitz). *A non-constant polynomial $p(s) = p_0 s^n + p_1 s^{n-1} + \cdots + p_n$ ($p_i \in \mathbb{R}, p_0 \neq 0$) is asymptotically stable if and only if*

1. *p_0 and p_1 have the same sign (in particular, $p_1 \neq 0$);*
2. *the degree $n-1$ polynomial*

$$q(s) := p(s) - \frac{p_0}{p_1}(p_1 s^n + p_3 s^{n-2} + p_5 s^{n-4} + \cdots)$$

is asymptotically stable.

Proof. Let λ_i be the zeros of $p(s)$. Expanding the factorization

$$p(s) = p_0 \prod_{i=1}^{n} (s - \lambda_i) \tag{A.1}$$

shows that $p_1 = -p_0 \sum_i \lambda_i$ (this is one of Vieta's formulas). If p is asymptotically stable then all its zeros λ_i have negative real part so then by Vieta's formula p_1 has the same sign as p_0. This proves that part 1 holds for asymptotically stable polynomials.

Now assume that p_1 is nonzero and define the family of polynomials r_η by

$$r_\eta(s) := p(s) - \eta \tfrac{p_0}{p_1}(p_1 s^n + p_3 s^{n-2} + \cdots), \qquad \eta \in [0,1].$$

For $\eta = 0$ we have $r_\eta = p$, and for $\eta = 1$ we have $r_\eta = q$. A special property of this family of polynomials is that the *imaginary* zeros (including their multiplicity) do not depend on η. (The proof will be given later.) This means that when we vary η, none of the zeros of r_η can cross or land on the imaginary axis. The only way the number of stable zeros of r_η can change is if the degree drops. The polynomial r_η equals

$$r_\eta(s) = p_0(1-\eta)s^n + p_1 s^{n-1} + \cdots.$$

From this form it follows that the only value of η for which the degree drops is $\eta = 1$ (so when $r_\eta = q$) and then it drops precisely one degree because we assumed $p_1 \neq 0$. For all $0 \leq \eta < 1$, the polynomial r_η therefore has as many stable zeros as $r_0 = p$. By Vieta's formula, the zeros $\lambda_{i,\eta}$ (stable and unstable) of r_η add up to

$$\sum_i \lambda_{i,\eta} = \frac{-p_1}{p_0(1-\eta)}.$$

In the limit $\eta \uparrow 1$, precisely $n-1$ of these zeros go to the zeros of $r_1 = q$ and the remaining zero therefore goes to $-p_1/(p_0(1-\eta))$ minus those $n-1$ zeros. This remaining zero hence goes to

$\pm\infty$. This remaining zero is stable if and only if $-p_1/(p_0(1-\eta))$ is negative for $\eta \uparrow 1$, in other words, if and only if p_0 and p_1 have the same sign. Done.

Remains to prove that the imaginary zeros of r_η are independent of η: We prove it for even n (the proof for odd n is analogous). For even n, we can write r_η as

$$r_\eta(s) = [p_{\text{even}}(s) - \eta \tfrac{p_0}{p_1} s\, p_{\text{odd}}(s)] + p_{\text{odd}}(s).$$

(The even part of a polynomial is the sum of the even powers s^{2k} and the odd part is the sum of the odd powers s^{2k+1}.) An imaginary $s = \mathrm{i}\omega$ is a zero of r_η of multiplicity k if and only if it is a zero of multiplicity k of both the even part $[p_{\text{even}} - \eta \frac{p_0}{p_1} s p_{\text{odd}}]$ and the odd part p_{odd}. This is because for imaginary $s = \mathrm{i}\omega$, the even part is real and the odd part is imaginary. So $\mathrm{i}\omega$ is a zero of r_η of multiplicity k if and only if it is a zero of multiplicity k of both p_{odd} and $p_{\text{even}} - \eta \frac{p_0}{p_1} s p_{\text{odd}}$, but that is the case iff it is a zero of multiplicity k of both p_{odd} and of p_{even}. This is independent of η. ∎

Now, we need to connect this theorem to the Routh table. It is not difficult (but also no fun) to verify that the Routh table of $q(s)$ is exactly that of $p(s)$ minus the first row. Since the Routh–Hurwitz test is correct for first-degree polynomials (verify), it follows by induction that it is correct for every degree n.

A.3 Model of the Inverted Pendulum (Example 3.2.2)

The following is a straight-forward derivation. Let $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathbb{R}^2$ be the position of the tip of the pendulum. Then we have

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u + \ell \sin(\phi) \\ \ell \cos(\phi) \end{bmatrix}.$$

Differentiating with respect to time gives

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \dot{u} + \ell \cos(\phi)\dot{\phi} \\ -\ell \sin(\phi)\dot{\phi} \end{bmatrix}.$$

Differentiating again, we obtain

$$\begin{bmatrix} \ddot{x} \\ \ddot{y} \end{bmatrix} = \begin{bmatrix} \ddot{u} - \ell \sin(\phi)\dot{\phi}^2 + \ell \cos(\phi)\ddot{\phi} \\ -\ell \cos(\phi)\dot{\phi}^2 - \ell \sin(\phi)\ddot{\phi} \end{bmatrix}. \tag{A.2}$$

By Newton's second law, this is equal to the force $F \in \mathbb{R}^2$ divided by the mass. The force consists of the gravity $-mg$ in the y-direction and a compression force λ in the direction of the stick. So the total force is

$$\begin{bmatrix} 0 \\ -mg \end{bmatrix} + \lambda \begin{bmatrix} \sin(\phi) \\ \cos(\phi) \end{bmatrix},$$

for some compression force λ. This, divided by m, equals (A.2), so

$$\begin{bmatrix} 0 \\ -g \end{bmatrix} + \tfrac{1}{m}\lambda \begin{bmatrix} \sin(\phi) \\ \cos(\phi) \end{bmatrix} = \begin{bmatrix} \ddot{u} - \ell \sin(\phi)\dot{\phi}^2 + \ell \cos(\phi)\ddot{\phi} \\ -\ell \cos(\phi)\dot{\phi}^2 - \ell \sin(\phi)\ddot{\phi} \end{bmatrix}.$$

This is a linear equation in $(\lambda, \ddot{\phi})$. The solution is

$$\begin{aligned} \tfrac{1}{m}\lambda &= \cos(\phi)\, g + \sin(\phi)\, \ddot{u} - \ell \dot{\phi}^2 \\ \sin(\phi)\, g &= \cos(\phi)\, \ddot{u} + \ell \ddot{\phi}. \end{aligned}$$

The first of these two is of no importance to us (it determines λ, so the compression force of the stick). The second is the model we are were looking for.

A.4 Canonical Form (Formula (3.29))

We know that the desired T transforms the matrix A to

$$T^{-1}AT = A_z := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix}.$$

For $S = T^{-1}$, this becomes

$$SA = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & -p_{n-1} \end{bmatrix} S.$$

We write the rows of S as $s_1, \ldots, s_n$. The first row of the equation above says that $s_1 A = s_2$; the second says that $s_2 A = s_3$; etc. In other words,

$$S = \begin{bmatrix} s_1 \\ s_1 A \\ s_1 A^2 \\ \vdots \\ s_1 A^{n-1} \end{bmatrix}.$$

Of the controllability matrix $\mathscr{C}_z$, we only use that it has the following structure (verify this yourself):

$$\mathscr{C}_z = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & * \\ 0 & \iddots & \iddots & \vdots \\ 1 & * & \cdots & * \end{bmatrix}.$$

Since $S = \mathscr{C}_z \mathscr{C}_x^{-1}$, we see that the first row s_1 is equal to $\begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathscr{C}_x^{-1}$. Hence $s_1 = \eta$, as defined in (3.29).

The T from Thm. 3.5.4 can be determined in a similar way. We have $AT = TA_z$. We denote the columns of T by $T_1, T_2, \ldots, T_n$, and for A_z we use the companion matrix. Then $AT = TA_z$ becomes

$$A \begin{bmatrix} T_1 & T_2 & \cdots & T_n \end{bmatrix} = \begin{bmatrix} T_1 & T_2 & \cdots & T_n \end{bmatrix} \begin{bmatrix} 0 & \cdots & \cdots & 0 & -p_0 \\ 1 & \ddots & & \vdots & -p_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -p_{n-1} \end{bmatrix}.$$

The first column of this equation says that $AT_1 = T_2$; the second says that $AT_2 = T_3$; etc. Hence

$$T = \begin{bmatrix} T_1 & AT_1 & A^2 T_1 & \cdots & A^{n-1} T_1 \end{bmatrix}.$$

We take T_1 from the observability matrix $\mathscr{O}_z$. It has the following structure (verify):

$$\mathscr{O}_z = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & * \\ 0 & \iddots & \iddots & \vdots \\ 1 & * & \cdots & * \end{bmatrix}$$

Since $T = \mathcal{O}_x^{-1}\mathcal{O}_z$, the first column T_1 of T must be

$$T_1 = \mathcal{O}_x^{-1}\begin{bmatrix}0\\ \vdots \\ 0\\ 1\end{bmatrix}.$$

This is the η from Lemma 3.5.4.

A.5 Heymann's Lemma (Thm. 4.2.3)

The proof of Thm. 4.2.3 uses Heymann's lemma.

Lemma A.5.1 (Heymann's lemma). *If (A,B) is controllable, then for every $u_0 \in \mathbb{R}^{n_u}$ with $b := Bu_0 \neq 0$, there exists an $\tilde{F}$ such that $(A - B\tilde{F}, b)$ is controllable.*

Proof. This is a technical proof. We first show that for $x_0 = 0$ and some suitable input $u_0, u_1, \ldots, u_{k-1}$, the *discrete*-time system $x_{k+1} = Ax_k + Bu_k$ produces a series of states $x_1, x_2, \ldots, x_k$ that are linearly independent for all $k = 1, \ldots, n$. We prove this using induction on k: for $k = 1$, the result obviously holds because $x_1 = Bu_0 \neq 0$. Now, suppose that $x_1, x_2, \ldots, x_k$, for $k < n$, are linearly independent. Then there exists a u_k such that $x_{k+1} = Ax_k + Bu_k \notin \operatorname{span}\{x_1, x_2, \ldots, x_k\}$. Indeed, suppose that this is not the case, so that

$$x_{k+1} := Ax_k + Bu_k \in \underbrace{\operatorname{span}\{x_1, x_2, \ldots, x_k\}}_{\mathscr{L}_k :=} \quad \forall u_k.$$

This implies that $Ax_k \in \mathscr{L}_k$ (take $u_k = 0$) and therefore also that $\operatorname{Im}(B) \subseteq \mathscr{L}_k$. But this then implies that $Ax_j = x_{j+1} - Bu_j \in \mathscr{L}_k$ for all $j \leq k$. This means that

$$A\mathscr{L}_k \subseteq \mathscr{L}_k.$$

Together with the inclusion $\operatorname{Im}(B) \subseteq \mathscr{L}_k$, this gives

$$\operatorname{Im}\begin{bmatrix}B & AB & \cdots & A^{n-1}B\end{bmatrix} \subseteq \mathscr{L}_k.$$

Because of the controllability, this says $\mathbb{R}^n \subseteq \mathscr{L}_k$, or $\mathscr{L}_k = \mathbb{R}^n$. This implies that $k \geq n$, which is a contradiction. So for all $k \leq n$ there *does* exist a u_k such that $x_1, x_2, \ldots, x_k$ are linearly independent.

Given such u_i, x_i and an arbitrary u_n define F_0 as

$$F_0 = -\begin{bmatrix}u_1 & u_2 & \cdots & u_n\end{bmatrix}\begin{bmatrix}x_1 & x_2 & \cdots & x_n\end{bmatrix}^{-1}.$$

By construction it satisfies $-F_0 x_i = u_i$. We then have

$$x_{k+1} := Ax_k + Bu_k = (A - BF_0)x_k,$$

or $x_{k+1} = (A - BF_0)^k x_1 = (A - BF_0)^k b$. The controllability matrix of $(A - BF_0, b)$ is equal to the *invertible* matrix $\begin{bmatrix}x_1 & x_2 & \cdots & x_n\end{bmatrix}$, and therefore the pair $(A - BF_0, b)$ is controllable. ∎

A.6 All homogeneous solutions

Proof of Thm. 1.5.3. We rely on some properties proved in Exercise 1.17. Decompose the characteristic polynomial as

$$(\lambda-\lambda_1)^{m_1}(\lambda-\lambda_2)^{m_2}\cdots(\lambda-\lambda_k)^{m_k} \tag{A.3}$$

with all $\lambda_1,\ldots,\lambda_k$ distinct. We prove the theorem by induction in k.

For $k=1$ the DE is $(\frac{\mathrm{d}}{\mathrm{d}t}-\lambda_1)^{m_1}y=0$. Write $y(t)$ as $y(t)=\mathrm{e}^{\lambda_1 t}z(t)$, and note that this is without loss of generality because $\mathrm{e}^{\lambda_1 t}$ is invertible. Then $(\frac{\mathrm{d}}{\mathrm{d}t}-\lambda_1)^{m_1}y(t)=0$ becomes $(\frac{\mathrm{d}}{\mathrm{d}t})^{m_1}z(t)=0$. In other words $z^{(m_1)}(t)$ is the zero function. That is the case iff $z(t)$ is a polynomial of degree at most m_1-1. The set of such polynomials is a subspace and we denote it by $\mathbb{P}_{m_1-1}$. Now $y(t)=\mathrm{e}^{\lambda_1 t}z(t)$ is as claimed in the theorem.

The induction hypothesis is that the claim is valid for $k-1$ distinct zeros λ. We need to proof the claim for k distinct zeros. The DE for k distinct zeros is

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_1)^{m_1}\underbrace{(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_2)^{m_2}\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_k)^{m_k}y(t)}_{q(t)}=0. \tag{A.4}$$

The so defined $q(t)$ satisfies

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_1)^{m_1}q(t)=0.$$

We already proved that this is the case iff $q(t)\in\mathrm{e}^{\lambda_1 t}\mathbb{P}_{m_1-1}$. The solutions of DE (A.4) hence are precisely those for which

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_2)^{m_2}\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_k)^{m_k}y(t)\in\mathrm{e}^{\lambda_1 t}\mathbb{P}_{m_1-1}. \tag{A.5}$$

The general solution $y(t)$ of this final equation is the general homogeneous solutions plus a particular solution. By the induction hypothesis the general homogeneous solution is

$$\sum_{i=2}^{k}\mathrm{e}^{\lambda_i t}\mathbb{P}_{m_i-1}$$

so the proof is complete if we can show that the set of particular solutions is $\mathrm{e}^{\lambda_1 t}\mathbb{P}_{m_1-1}$ (nothing more, nothing less). That is we need to check that

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_2)^{m_2}\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_k)^{m_k}(\mathrm{e}^{\lambda_1 t}\mathbb{P}_{m_1-1})=\mathrm{e}^{\lambda_1 t}\mathbb{P}_{m_1-1}.$$

This is the same as

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_{2,1})^{m_2}\cdots(\tfrac{\mathrm{d}}{\mathrm{d}t}-\lambda_{k,1})^{m_k}\mathbb{P}_{m_1-1}=\mathbb{P}_{m_1-1}.$$

where $\lambda_{i,1}:=\lambda_i-\lambda_1$. The above can be expanded as

$$(\tfrac{\mathrm{d}}{\mathrm{d}t}^{(?)}+\cdots+d)\mathbb{P}_{m_1-1}=\mathbb{P}_{m_1-1}, \tag{A.6}$$

where $d=\prod_i-\lambda_{i,1}$. The so defined d is nonzero because λ_1 differs from the other λ_i's. Convince yourself of the fact that (A.6) holds iff $d\neq 0$..... ∎

Index