Systems and Control Theory

Lecture notes

Teacher: Prof. Franco Blanchini

Università degli Studi di Udine

the Italian version by
Luca Peretti

has been translated and revised by
Giulia Giordano, Flavio Pop, Franco Blanchini

November 6, 2023

Preface

These notes on *System Theory* are a revised version of the lecture notes first written by Dr. Luca Peretti after the course taught by Prof. Franco Blanchini in the academic year 2002/2003. Through the years, the notes have been deeply revised and integrated. Several students have pointed out typos or errors (most of which are due to the teacher).

The English translation was first released in 2016. Should the reader have any comments or find out errors, a message would be very appreciated. For this purpose, please contact

blanchini@uniud.it

The present notes will be updated/corrected in the future, tending to perfection in the very long run (with a very slow convergence speed).

Given their nature, these notes **SHOULD NOT** be considered as a substitute of a good textbook. They only provide a track of the arguments presented in the course.

The basic principles which inspire the work are:

- there are no error-proof manuscripts;
- no matter how big is your effort, someone will not approve your work;
- systems & control theory is as important as many other things in life;
- rule number one: have fun!

Franco Blanchini

Contents

List of Figures

Chapter 1

Introduction

In the subject name, *System Theory*, an important aspect is actually omitted. To be more precise, in fact, we should talk about the **Theory of Dynamical Systems**. The expression "dynamical system", or "dynamic system", generically refers to any entity that evolves over time according to the laws of causality, interacting with the environment in accord with the cause-and-effect principle.

From an external viewpoint, a dynamical system is characterised by a set of input functions, representing the cause, and a set of output functions that represent the effect (Figure 1.1). A key feature is that the link between input and output is not static, but is intrinsically linked to time and to the past history.

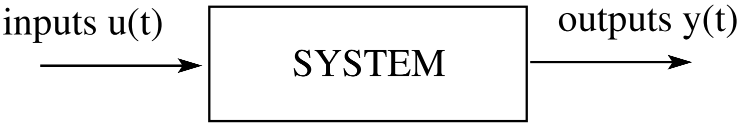

Figure 1.1: General scheme of a dynamic system.

For example, consider a system whose output $y(t)$ depends on the input $u(t)$ as per the differential equation[1]

$$\ddot{y}(t) + a\dot{y}(t) + by(t) = u(t)\,,$$

fairly representing a balance (Figure 1.2). When $u(t)$ is assigned, the solution of the equation describes the time evolution of $y(t)$. If we consider a constant input $u(t) = \bar{u}$, then this differential equation has a constant (static) solution, which can be easily computed by setting $\ddot{y}(t) = \dot{y}(t) = 0$. This solution is

$$\bar{y} = \frac{\bar{u}}{b}\,.$$

Typically, in order to determine the weight $\bar{u}$ of the object on the balance, we observe the output y, or a quantity proportional to y. Obviously, however, if we considered solutions of this type only, we would neglect a fundamental aspect, which is the transient response of the system to a given input. Indeed, a single given input may produce several (infinitely many) outputs, which is a peculiar feature of dynamical systems.

Dynamical systems were well studied already at the beginning of the XX century, but systems theory was deeply investigated and developed in the 1950s-1960s, with the main goal of explaining the dynamic behaviour of a broad class of entities and processes. Actually, systems theory is able

[1] We denote by $f'(x)$, $f''(x)$ the first and second derivative of the function $f(\cdot)$ with respect to the generic variable x, and by $\dot{g}(t)$, $\ddot{g}(t)$ the first and second derivative of $g(\cdot)$ as a *function of time*.

to provide really meaningful explanations in the case of specific classes of systems, such as regular dynamical systems (those that can be represented in terms of differential equations).

Linear Time-Invariant systems are an extremely interesting particular case, for which very powerful tools have been proposed. Some of the techniques developed for studying this type of systems can be applied to more general types of systems, thanks to a technique called *linearisation*. This is an extremely important aspect and justifies the fact that most of the course will be devoted to the study of Linear Time-Invariant (LTI) systems.

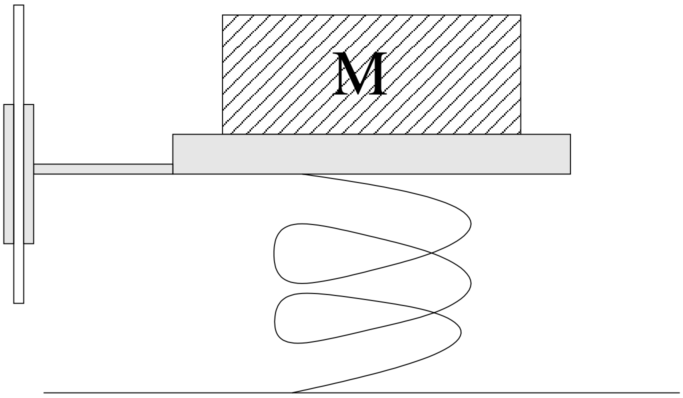

Figure 1.2: An example of linear time-invariant system: the balance.

Example 1.1. ***(The balance.)*** *The balance (scales) is a simple example of a linear time-invariant system. Consider Figure 1.2 and assume that (1) the mass of the balance plate is negligible compared to the mass of the object to be weighed and (2) the viscous friction is a linear function of the speed. Then, we can easily obtain the linear differential equation which governs the system:*

$$M\ddot{y}(t) + h\dot{y}(t) + ky(t) = Mg\,,$$

where $y(t)$ represents the vertical displacement of the plate, while the constants M, h, k and g respectively represent the mass of the object, the friction coefficient, the elastic constant of the spring and the gravitational acceleration. When measuring the weight of an object, we are interested in $y(t)$ after the system has settled: assuming $\ddot{y}(t) = \dot{y}(t) = 0$, we get

$$\bar{y} = \frac{Mg}{k}.$$

From this equation, expressing the steady-state *value of y, we derive the weight Mg. However, before reaching this static condition, the balance is subjected to a transient that depends on the value of the friction coefficient, as is qualitatively shown in Figure 1.3.*

Clearly, in the case of the scales, the most interesting aspect is the steady-state operation. However, in other cases the dynamic behaviour of the system has a fundamental role. An example is given by structural engineering. Buildings are normally designed assuming static loads. Limiting the study to this case is correct as long as the structure is not affected by forces (or accelerations) that vary over time. Yet, when the building is located in an earthquake region or when strong winds are present[2], the building must be designed by modeling it as a dynamic system.

[2] The Tacoma bridge is a famous example of this problem: `http://www.youtube.com/watch?v=j-zczJXSxnw`

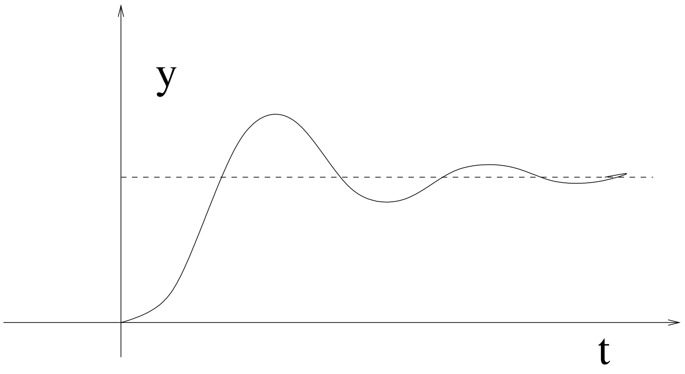

Figure 1.3: The balance system: a qualitative graph of the transient.

Outline

The course starts by introducing basic tools for modeling dynamical systems. These include differential and difference equations, vector spaces and transfer functions. The theory of linear systems and of their properties (superposition principle, free and forced responses, modes and stability) are explained in detail. The problem of sampling and sampled-data systems, along with the techniques for digital implementation, are then introduced.

A formal definition of a general dynamical system is given along with some examples of non-regular systems (*i.e.*, not expressed by differential or difference equations), such as automata.

Then, basic structural properties of linear systems are presented, in particular reachability and observability and their connection with the transfer function analysis.

The theory of realisation of transfer functions is subsequently introduced. The basic problem is how we can associate a (minimal) state representation with a transfer function (or a transfer function matrix).

These notions are preliminary and essential[3] to understand the fundamental concepts of regulation. The regulation problem via eigenvalue assignment with state feedback is initially presented for linear systems. The output feedback problem is then solved via state observer design.

The course then presents fundamental notions on the analysis and the control of nonlinear systems. The tools presented for the linear case can be successfully applied by means of the linearisation technique (a key point of the course). A strong support is given by Lyapunov theory,[4] which allows us to deal in an efficient and elegant way with systems whose solutions are impossible to compute analytically, or even numerically.

Finally, the course presents in a very simple way fundamental topics that cannot be discussed in depth. These include robustness and control optimality. Typically, the final lectures also present short seminars about specific topics, such as model-predictive control (MPC) and system identification.

About one third of the course is devoted to exercises. These include a survey of mathematical notions, in particular linear algebra. Many models of linear and nonlinear systems are described. Video of laboratory experiences are shown. Many exercises to better understand the theory are proposed, including three sessions of test exercises.[5]

[3] Although perhaps tedious...

[4] As you will notice, this is the part that is most appreciated by the teacher.

[5] The teacher strongly approves human cooperation in all occasions **but these**...

Chapter 2

Systems of differential equations

This chapter will briefly present the basics concept required as a technical support for what follows. The proposed theory should be the subject of previous courses. However, given its importance, some fundamental aspects are summarised here.

2.1 Differential equations in normal form

In this course, we will mainly consider is the class of *regular* systems, namely, systems represented by differential equations. A regular system has a number (m) of variables named inputs (associated with the causes) and a number (p) of variables named outputs (associated with the effects); note that the numbers m and p can be different in general. Input and output variables can be grouped in two vectors. We will denote by $u(t)$ and $y(t)$ the **input vector** and the **output vector**, respectively:

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_m(t) \end{bmatrix} \in \mathbb{R}^m \qquad y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{bmatrix} \in \mathbb{R}^p \tag{2.1}$$

These vectors provide an *external* representation of the system behaviour. A dynamical system is *internally* represented by a number n of variables named states, or **state variables**, which are also grouped in a vector $x(t)$:

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \in \mathbb{R}^n \tag{2.2}$$

Example 2.1. (Electric motor.) *Consider the system describing an electric motor, illustrated in Section 9.1. The input vector for this system is*

$$u(t) = \begin{bmatrix} v_f(t) \\ v_a(t) \\ C_m(t) \end{bmatrix},$$

whose components represent the excitation voltage, the armature voltage and the external torque. As an output vector, we can take

$$y(t) = \begin{bmatrix} \varphi(t) \\ \omega(t) \end{bmatrix},$$

representing the angle and the speed. The state vector is

$$x(t) = \begin{bmatrix} I_f(t) \\ I_a(t) \\ \omega(t) \\ \varphi(t) \end{bmatrix},$$

whose components are the excitation current, the armature current, the angular speed and the angle.

Note that some state variables can correspond to output variables (as in the example above). Conversely, a state variable cannot be an input.

In general, the relation between state variables and inputs is given by differential equations. In most cases, these equations are of the first order and the overall system can be expressed in the form

$$\begin{cases} \dot{x}_1(t) = f_1(t, x_1(t), \dots, x_n(t), u_1(t), \dots, u_m(t)) \\ \vdots \\ \dot{x}_n(t) = f_n(t, x_1(t), \dots, x_n(t), u_1(t), \dots, u_m(t)) \end{cases} \tag{2.3}$$

This is called the *normal form*, since the derivatives of the state are explicit functions of all of the other variables.

Remark 2.1. *A formal definition of* state *will be given later on. For regular systems in normal form, the states are the variables $x_k(t)$ whose derivatives appear on the left hand side of the differential equations.*

It is desirable to seek a compact form that allows us to easily manipulate the equations. To this aim, we consider the derivative of the state vector

$$\dot{x}(t) = \frac{d}{dt}x(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dots \\ \dot{x}_n(t) \end{bmatrix} \in \mathbb{R}^n, \tag{2.4}$$

so that the above equations reduce to the following compact vector representation:

$$\dot{x}(t) = f(t,\ x(t),\ u(t)). \tag{2.5}$$

This equation represents a generic regular system of n first order differential equations, with m inputs. It represents an instant relation between the state variables $x(t)$ and inputs $u(t)$ of the system.

Example 2.2. (Chemical reactions.) *Consider the following system of differential equations*

$$\begin{cases} \dot{x}_1(t) = -\alpha x_1(t)x_2(t) + u_1(t) \\ \dot{x}_2(t) = -\alpha x_1(t)x_2(t) + u_2(t) \\ \dot{x}_3(t) = \ \ \alpha x_1(t)x_2(t) - \beta x_3(t) \end{cases} \tag{2.6}$$

that involves three state variables ($x_1(t)$, $x_2(t)$ and $x_3(t)$) and two input variables ($u_1(t)$ and $u_2(t)$). If x_1, x_2 and x_3 are the concentration of the chemical species A, B and C, this system represents the time evolution of the concentrations when the chemical reactions

$$\emptyset \overset{u_1}{\rightharpoonup} A,\ \ \emptyset \overset{u_2}{\rightharpoonup} B,\ \ A + B \overset{\alpha}{\rightharpoonup} C,\ \ C \overset{\beta}{\rightharpoonup} \emptyset$$

are occurring: u_1 and u_2 are injections of A and B; C is produced from A and B with reaction speed α and C is consumed with speed β.

According to the theory of differential equations, in order to solve the equation (2.5) for a given value of the input vector $u(t)$, we need to know the initial conditions $x(t_0)$, namely, the value of the state at a given (initial) time instant t_0. Under regularity conditions, there will be a solution $x(t) = \varphi(x(t_0), u(t))$ that depends on the initial conditions $x(t_0)$ of the system and on the input $u(t)$. In the following, we will always assume that *the solution exists and is unique*. This is true if the function f has some regularity property (for instance, it is a Lipschitz function) and u is sufficiently regular (*e.g.*, piecewise-continuous).

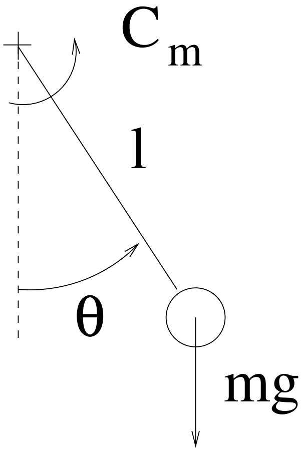

Figure 2.1: Model of a pendulum.

Example 2.3. ***(The pendulum.)*** *Consider Figure 2.1, where C_m is the torque (system input) and $y = \theta$ is the angle with respect to the vertical reference frame (system output). The equation governing this system in the absence of friction is*

$$\ddot{\theta}(t) = -\frac{g}{l}\sin(\theta(t)) + \frac{1}{ml^2}C_m(t)$$

Denoting by $\omega(t) = \dot{\theta}(t)$, this equation can be written in normal form as the system

$$\begin{aligned}\dot{\theta}(t) &= \omega(t)\\ \dot{\omega}(t) &= -\frac{g}{l}\sin(\theta(t)) + \frac{1}{ml^2}C_m(t)\end{aligned}$$

where θ and ω are the state variables.

In order to find the value of output at the generic time instant t, it is not enough to know the time evolution of the torque: we also need to know the initial position and the initial speed (namely, the initial state). Given these quantities, one can determine the time evolution of the position $\theta(t)$ and of the angular speed $\omega(t) = \dot{\theta}(t)$ by integrating the differential equation, thus obtaining a complete description of the trajectory.

The dynamical system described by equation (2.5) is said **invariant** o **autonomous** if the function f does not directly depend on time. Note, however, that it always indirectly depends on time, since t is an argument of both $x(t)$ and $u(t)$. A time invariant system can then be written as

$$\dot{x}(t) = f(x(t),\, u(t)) \tag{2.7}$$

Instead, if time enters directly in the equation, the system is said **non-autonomous**.

Example 2.4. *The following system*

$$\begin{cases} \dot{x}_1(t) = -t\, x_2(t) \\ \dot{x}_2(t) = u(t) \end{cases}$$

is non-autonomous.

The system is called **linear** if the function f is linear with respect to $x(t)$ and $u(t)$. In this case, it is expressed as:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{2.8}$$

A system is said to **linear time-invariant** (LTI) if it is both time invariant and linear:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.9}$$

In this case, then, the matrices A (dimension $n \times n$) and B (dimension $n \times m$) does not depend on time.

The output variables of a system are given by the components of a vector $y(t)$ whose components are statically[1] determined by the following *algebraic equations*

$$\begin{cases} y_1(t) = g_1(t, x_1(t), \dots, x_n(t), u_1(t), \dots, u_m(t)) \\ \vdots \\ y_p(t) = g_p(t, x_1(t), \dots, x_n(t), u_1(t), \dots, u_m(t)) \end{cases} \tag{2.10}$$

which, again, can be written in a compact form:

$$y(t) = g(t, x(t), u(t)). \tag{2.11}$$

The above equation is called the **output transformation**. If the system is time-invariant, the output transformation is of the form:

$$y(t) = g(x(t), u(t)). \tag{2.12}$$

Conversely, if the system is linear, we have:

$$y(t) = C(t)x(t) + D(t)u(t). \tag{2.13}$$

Finally, if the system is both linear and time-invariant, the output transformation is of the form:

$$y(t) = Cx(t) + Du(t), \tag{2.14}$$

where the matrices C (dimension $p \times n$) and D (dimension $p \times m$) are independent of time.

Putting together state and output equations, we achieve the following system:

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)) \\ y(t) = g(t, x(t), u(t)) \end{cases} \tag{2.15}$$

With few exceptions, this is the most general form of dynamical system we will consider in the course.

Remark 2.2. *There are systems involving partial derivatives, such as that given by the heat equation*

$$\frac{\partial}{\partial t}\phi(z, t) = a^2 \nabla^2 \phi(z, t),$$

where z is a space variable. In this case, the state at each time t is a function $\phi(z, t)$, *which can be seen as a vector in a proper (infinite dimensional) space. These systems are denoted as infinite dimensional systems, or distributed parameter systems, and will not be considered in this course.*

[1] Namely, no derivatives are involved.

2.2 General solution of a continuous-time linear time-invariant system

During the course, we will intensively discuss the case of linear time-invariant systems, whose the equations can be written in the form:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{2.16}$$

where matrix A has dimension $n \times n$, B has dimension $n \times m$, C has dimension $p \times n$ and D has dimension $p \times m$. We will sometime use the notations

$$\Sigma(A, B, C, D),$$

or

$$\Sigma\left[\begin{array}{c|c} A & B \\ \hline C & C \end{array}\right],$$

or even (A, B) or (A, C) when the properties we are discussing concern these matrices only.

For this class of systems, we can determine an analytic solution (while for most systems it is impossible to obtain a solution without resorting to numerical techniques). The solution we obtain will not be used numerically, but will be analysed qualitatively and will provide fundamental information about the system.

Suppose that $u(t)$ and $x(0)$ are known (since the system is time-invariant, we can choose $t_0 = 0$ without loss of generality). If we carefully look at the equations, we can easily see that, in order to understand the entire system behaviour, we only need to determine $x(t)$: then, $y(t)$ can be immediately determined from the algebraic linear output transformation. Therefore, our main aim is to solve the system:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ x(0) \text{ known} \end{cases} \tag{2.17}$$

A fundamental property of this class of systems[2] is the **superposition principle**.

For the vectors $u(t)$ and $x(0)$ representing input and initial conditions respectively, the symbol $\longrightarrow$ means "give rise to the solution":

$$x(0), \quad u(t) \quad \longrightarrow \quad x(t).$$

Assume that we have two different solutions with the different inputs and initial conditions

$$\begin{array}{rcccc} x^*(0), & u^*(t) & \longrightarrow & x^*(t) \\ x^{**}(0), & u^{**}(t) & \longrightarrow & x^{**}(t) \end{array} \tag{2.18}$$

and assume we combine inputs and initial vectors as follows

$$\begin{array}{rcl} u(t) & = & \alpha u^*(t) + \beta u^{**}(t) \\ x(0) & = & \alpha x^*(0) + \beta x^{**}(0) \end{array} \tag{2.19}$$

with coefficients α and β. Then,

$$\alpha x^*(0) + \beta x^{**}(0), \quad \alpha u^*(t) + \beta u^{**}(t) \quad \longrightarrow \quad \alpha x^*(t) + \beta x^{**}(t)$$

We can use this general principle for the natural-forced response decomposition: the problem of computing the solution can be decomposed into two simpler problems, and the overall solution

[2] Actually, of all *linear* systems, including heat equation, waves equation, transmission lines...

is the sum of the solutions of the two simpler problems. Any given initial condition and input $x(0)$ and $u(t)$ can be decomposed as follows:

$$\begin{array}{rcl} x(0) & : & x^*(0) = 0, \quad x^{**}(0) = x(0) \\ u(t) & : & u^*(t) = u(t), \quad u^{**}(t) = 0 \end{array}$$

We then have two-problems:

- **natural response** (assigned initial conditions, null input):

$$\begin{cases} \dot{x}_L(t) = Ax_L(t) \\ x_L(0) = x(0) \end{cases} \tag{2.20}$$

- **forced response** (null initial conditions, assigned input):

$$\begin{cases} \dot{x}_F(t) = Ax_F(t) + Bu(t) \\ x_F(0) = 0 \end{cases} \tag{2.21}$$

Applying the superposition principle with $\alpha = 1$ and $\beta = 1$, we get:

$$x(t) = x_L(t) + x_F(t). \tag{2.22}$$

2.2.1 Natural response

The natural response is the solution of the system

$$\begin{cases} \dot{x}_L(t) = Ax_L(t) \\ x_L(0) = x(0) \end{cases}$$

Consider the scalar case first. Then, we have

$$\begin{cases} \dot{x}_L(t) = ax_L(t) \\ x_L(0) = x(0) \end{cases} \Rightarrow x(t) = x(0)\, e^{at}$$

In fact, for $t = 0$ the exponential is equal to 1, and $\dot{x} = a\, x(0)\, e^{at} = ax(t)$. The solution in $\mathbb{R}^n$ is similar to that in the scalar case. Let us define the exponential of a matrix:

$$e^{At} \doteq I + At + \frac{1}{2}A^2t^2 + \frac{1}{3!}A^3t^3 + \ldots = \sum_{k=0}^{+\infty} \frac{A^k t^k}{k!} \tag{2.23}$$

This is a convergent series (the factorial dominates the exponential for $k \to +\infty$).[3] We can then verify that the natural response of the system is equal to:

$$x(t) = e^{At}x(0). \tag{2.24}$$

In fact, for $t = 0$,

$$e^{A0}x(0) = Ix(0) = x(0). \tag{2.25}$$

Moreover, if we write the derivative, we get

$$\frac{d}{dt}(e^{At}x(0)) = \frac{d}{dt}\left(\sum_{k=0}^{+\infty} \frac{A^k t^k}{k!}\right) = \left(\sum_{k=1}^{+\infty} A\frac{A^{k-1}t^{k-1}}{(k-1)!}\right) = A(e^{At}x(0)) \Rightarrow \dot{x}_L(t) = Ax_L(t) \tag{2.26}$$

The problem of computing e^{At} will be considered later.

[3] Note that the expression is different from the matrix obtained by taking the exponential of each element of the matrix.

2.2.2 Forced response

The forced response of the system is the solution of

$$\begin{cases} \dot{x}_F(t) = Ax_F(t) + Bu(t) \\ x_F(0) = 0 \end{cases} \tag{2.27}$$

This solution is equal to the convolution product

$$x_F(t) = \int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma \tag{2.28}$$

This can be verified by applying the Laplace transform to the system equations:

$$\begin{aligned} s\,x_F(s) - x_F(0) &= Ax_F(s) + Bu(s) \\ \Rightarrow (sI - A)\,x_F(s) &= Bu(s) \\ \Rightarrow x_F(s) &= (sI - A)^{-1}Bu(s) \end{aligned} \tag{2.29}$$

Since the Laplace transform of a convolution is equal to the product of the individual Laplace transforms, we have:

$$\begin{aligned} u(\sigma) &\xrightarrow{\mathcal{L}} u(s) \\ e^{At} &\xrightarrow{\mathcal{L}} (sI - A)^{-1} \\ \Rightarrow \mathcal{L}\Big[\int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma\Big] &= \mathcal{L}[e^{At}B]\mathcal{L}[u(t)] = \\ &= (sI - A)^{-1}B\,u(s) \end{aligned} \tag{2.30}$$

Note also that the Laplace transform of the matrix exponential is nothing but a generalisation of the Laplace transform of the scalar exponential:

$$e^{At} \xrightarrow{\mathcal{L}} (sI - A)^{-1}$$

2.2.3 Complete solution of the system

Since $x(t) = x_L(t) + x_F(t)$, by adding the obtained expressions we get:

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma \tag{2.31}$$

Then, the system output is

$$y(t) = Ce^{At}x(0) + C\int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma + Du(t) \tag{2.32}$$

Let us now introduce two assumptions: $x(0) = 0$ and $D = 0$ (the latter assumption is only for brevity). In this case we get:

$$y = \int_0^t Ce^{A(t-\sigma)}Bu(\sigma)d\sigma \tag{2.33}$$

This is the input-output relationship, which is very important and leads to the definition of **impulse response matrix** $W(t)$, having dimension $p \times m$:

$$W(t) = Ce^{At}B. \tag{2.34}$$

The reason of the name can be explained by considering the Dirac delta (impulse) as the system input. The Dirac delta $\delta(\cdot)$ is a distribution, having the property that, for any continuous function f defined on an interval $[a, b]$ and any t_0 in this interval, we have

$$\int_a^b f(t)\delta(t - t_0)dt = f(t_0). \tag{2.35}$$

Then, if $u(t) = \delta(t)$ (the impulse function at zero),

$$\int_0^t Ce^{A(t-\sigma)}B\delta(\sigma)d\sigma = Ce^{At}B. \tag{2.36}$$

The meaning of the elements of matrix $W(t)$ is then the following: *each element $W_{ij}(t-\tau)$ represents the ensuing response of the i-th output at time t, due to an impulse applied at time τ to the j-th input (Figure 2.2).*

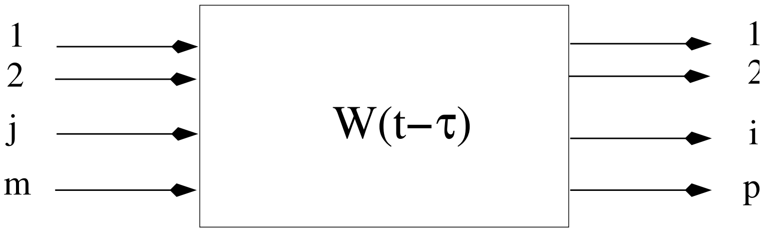

Figure 2.2: A system with m inputs and p outputs and its impulse response matrix.

In the general case with D non-zero, we have

$$W(t) = Ce^{At}B + D\delta(t),$$

where $\delta(t)$ is the impulse function at $\tau = 0$.

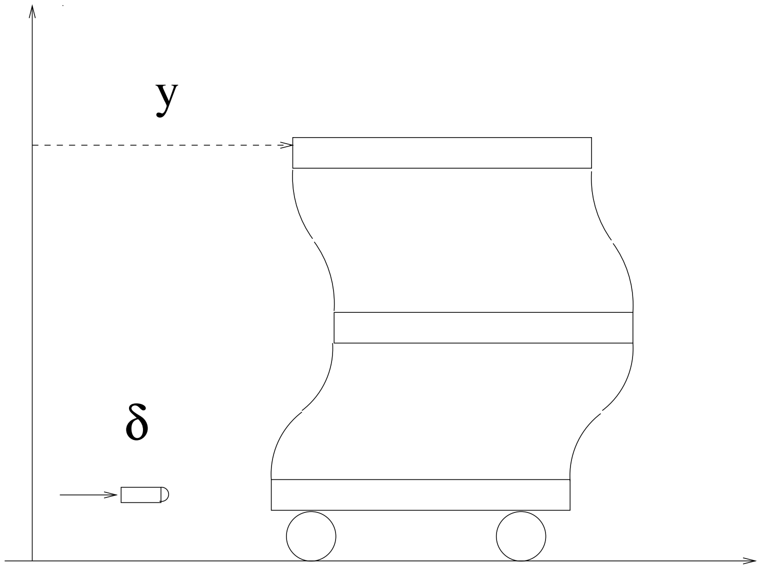

Figure 2.3: Scheme of a simple elastic structure.

Example 2.5. ***(Impulse on an elastic structure.)*** *Figure (2.3) represents a simple elastic structure, consisting of three storeys that can shift horizontally relatively to one another. The lower storey is placed on a cart. Take as inputs the forces u_1, u_2, u_3 acting on the storeys and as outputs the positions of the storeys, y_1, y_2, y_3. An impulsive input $u_1(t) = \delta(t - t_0)$ applied to the first storey (illustrated by an object hitting the first storey) causes the output W_{31} on the third storey.*

2.3 The exponential e^{At}

In order to obtain qualitative information on the general solution of the system, we need to analyse the exponential matrix e^{At}. This is a matrix of size $n \times n$.

Computing the eigenvalues and the right and left eigenvectors of A, we can derive an expression for e^{At} without using the power series that was previously introduced. Consider the case of A with n distinct eigenvalues, so that there are $t_1, \dots, t_n$ distinct right eigenvectors that are linearly independent and then form a basis. Let

$$T \doteq [t_1, \dots, t_n]$$

be the invertible matrix formed by the right eigenvectors. Let also

$$S = T^{-1} = \begin{bmatrix} s_1^\top \\ s_2^\top \\ \vdots \\ s_n^\top \end{bmatrix}$$

be its inverse matrix (note that $s_k^\top A = \lambda_k s_k^\top$, *i.e.*, the rows of matrix S are left eigenvectors of A). Denoting by Λ the diagonal matrix of the eigenvalues of A, we can verify that

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

By noticing that $A^k = T\Lambda T^{-1}T\Lambda T^{-1}T\Lambda T^{-1} \dots T\Lambda T^{-1} = T\Lambda^k T^{-1}$, we can derive

$$e^{At} = Te^{\Lambda t}S = \sum_{k=0}^{+\infty} \frac{T\Lambda^k S t^k}{k!} = T\left(\sum_{k=0}^{+\infty} \frac{\Lambda^k t^k}{k!}\right)S = [t_1, \dots, t_n]e^{\Lambda t}\begin{bmatrix} s_1^\top \\ \vdots \\ s_n^\top \end{bmatrix} = \sum_{i=1}^{n} t_i s_i^\top e^{\lambda_i t} = \sum_{i=1}^{n} Z_i e^{\lambda_i t}$$

(note that $e^{\Lambda t}$ is diagonal). Note that

$$Z_i = t_i s_i^\top,$$

where t_i is a column of matrix $T = [t_1\ t_2\ \dots\ t_n]$ and $s_i^\top$ is a row of matrix $S = T^{-1}$, are square matrices of size $n \times n$, named *components* of A.

Remark 2.3. *The product of a column vector and a row vector (unlike the product of a row vector and a column vector) forms a matrix (and not a scalar). For example, if we take the vectors*

$$t_i = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \qquad s_i^\top = \begin{bmatrix} 1 & 2 \end{bmatrix},$$

we obtain

$$Z_i = t_i s_i^\top = \begin{bmatrix} -1 & -2 \\ 2 & 4 \end{bmatrix},$$

while $s_i^\top t_i = [3]$.

Example 2.6. (Computing e^{At}.) *Consider the matrix*

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

Its characteristic polynomial is given by

$$\det(sI - A) = \det \begin{bmatrix} s+4 & 1 \\ -2 & s+1 \end{bmatrix} = s^2 + 5s + 6 = (s+2)(s+3)$$

Hence, the eigenvalues are $\lambda_1 = -2$, $\lambda_2 = -3$ and their respective eigenvectors can be computed as

$$(A - \lambda_1 I)\bar{t}_1 = 0 \Rightarrow \begin{bmatrix} -4+2 & -1 \\ 2 & -1+2 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = 0 \Rightarrow \bar{t}_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

and

$$(A - \lambda_2 I)\bar{t}_2 = 0 \Rightarrow \begin{bmatrix} -4+3 & -1 \\ 2 & -1+3 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = 0 \Rightarrow \bar{t}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

At this point we can easily find the matrices T and $S = T^{-1}$:

$$T = \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix}, \quad S = T^{-1} = \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix}$$

Finally, the exponential can be expressed as

$$\begin{aligned} e^{At} &= Te^{\Lambda t}S = \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} e^{-2t} & 0 \\ 0 & e^{-3t} \end{bmatrix} \begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix} = \\ &= \begin{bmatrix} 1 \\ -2 \end{bmatrix} \begin{bmatrix} -1 & -1 \end{bmatrix} e^{-2t} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} e^{-3t} = \\ &= \begin{bmatrix} -1 & -1 \\ 2 & 2 \end{bmatrix} e^{-2t} + \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} e^{-3t} = \\ &= \begin{bmatrix} -e^{-2t} + 2e^{-3t} & -e^{-2t} + e^{-3t} \\ 2e^{-2t} - 2e^{-3t} & 2e^{-2t} - e^{-3t} \end{bmatrix} \end{aligned}$$

The expression

$$e^{At} = \sum_{i=1}^{n} Z_i \, e^{\lambda_i t}, \tag{2.37}$$

valid in the case of distinct eigenvalues, reveals that the elements of matrix e^{At} are linear combinations of the functions $e^{\lambda_i t}$, where λ_i are the eigenvalues of A.

The scalar functions $e^{\lambda_i t}$ are called the **modes of the system**. They are of fundamental importance for the system analysis.

Note that, in general, the eigenvalue of a real matrix can be complex. If the i-th eigenvalue is complex, then

$$\lambda_i = \xi_i + j_i\omega_i \Rightarrow e^{\lambda_i t} = e^{(\xi_i + j\omega_i)t} = e^{\xi_i t}(\cos(\omega_i t) + j\sin(\omega_i t)) \tag{2.38}$$

If there is a complex eigenvalue λ, then also its complex conjugate λ^* is an eigenvalue. Moreover, if λ is associated with the matrix Z_i, then λ^* is associated with the conjugate matrix Z_i^*. Therefore, the exponential e^{At} can be decomposed into two summations that include, respectively, the contribution of real eigenvalues and the contribution of complex eigenvalues:

$$e^{At} = \sum_{i=1}^{r} Z_i e^{\lambda_i t} + \sum_{i=r+1,\ step\,2}^{n-1} (Z_i e^{\lambda_i t} + Z_i^* e^{\lambda_i^* t}), \tag{2.39}$$

where r is the number of real eigenvalues. Replacing $Z_i = M_i + jN_i$ and $\lambda = \xi + j\omega$ gives:

$$e^{At} = \sum_{i=1}^{r} Z_i e^{\lambda_i t} + 2 \sum_{i=r+1,\ step\,2}^{n-1} e^{\xi_i t}(M_i \cos(\omega_i t) - N_i \sin(\omega_i t)) \tag{2.40}$$

From this expression one can see that the exponential of a real matrix is always a real matrix, as expected.

In conclusion, from the analysis of the eigenvalues of a matrix we can obtain important information about the natural response of the system, which depends on the modes $e^{\lambda_i t}$.

2.3.1 The case of multiple eigenvalues

If the matrix A has distinct eigenvalues, then it can be diagonalised

$$A = T\Lambda T^{-1}, \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{bmatrix}$$

This is no longer true in general when there are coincident eigenvalues: in this case, we can only "block-diagonalise" A in the following way

$$A = TJT^{-1}, \qquad J = \begin{bmatrix} J_1 & 0 & 0 & 0 \\ 0 & J_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & J_n \end{bmatrix}, \tag{2.41}$$

where the blocks J_k are of the form

$$J_k = \begin{bmatrix} \lambda_k & 1 & 0 & 0 \\ 0 & \lambda_k & 1 & 0 \\ 0 & 0 & \ddots & 1 \\ 0 & 0 & 0 & \lambda_k \end{bmatrix}. \tag{2.42}$$

Then, we say that the matrix J is in **Jordan canonical form**, or in **Jordan normal form**.

Example 2.7. ***(Jordan form.)*** *The following matrix is in Jordan form*

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

In the presence of multiple eigenvalues, a fundamental parameter is the **ascent** $\deg(\lambda_k)$ of the eigenvalue λ_k, defined as the maximum size of the block associated with λ_k in the Jordan form. In Example 2.7, we have

$$\deg(2) = 3 \quad \text{and} \quad \deg(3) = 2.$$

In this case, the exponential e^{At} has the following expression:

$$e^{At} = \sum_{k=1}^{M} \sum_{j=0}^{\deg(\lambda_k)-1} Z_{kj}\, t^j\, e^{\lambda_k t} \tag{2.43}$$

where M is the number of distinct eigenvalues and $\deg(\lambda_k)$ the ascent of each of them.

The function

$$t^j\, e^{\lambda_k t}$$

is the most general form of a mode.

Example 2.8. *The modes associated with matrix A in Example 2.7 are:*

- $\lambda_1 = 2$*:* $e^{2t},\ t\,e^{2t},\ t^2 e^{2t}$*;*
- $\lambda_2 = 3$*:* $e^{3t},\ t\,e^{3t}$

Example 2.9. *Given the matrix in the Jordan form*

$$A = \begin{bmatrix} -4 & 1 & 0 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 1 & 0 & 0 \\ 0 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & -4 \end{bmatrix},$$

the associated modes are $e^{-4t},\ t\,e^{-4t},\ t^2 e^{-4t},\ t^3 e^{-4t}$.

We now have an expression for the impulse response matrix. In the case of distinct eigenvalues,

$$W(t) = Ce^{At}B = \sum_{i=1}^{n} Q_i\, e^{\lambda_i t}, \qquad Q_i = CZ_iB. \tag{2.44}$$

In the case of multiple eigenvalues,

$$W(t) = Ce^{At}B = \sum_{k=1}^{m} \sum_{j=0}^{\deg(\lambda_k)-1} Q_{kj}\, t^j\, e^{\lambda_k t} \qquad Q_{kj} = CZ_{kj}B. \tag{2.45}$$

The modes appear in the impulse response matrix as well, hence they affect not only the natural response, but also the forced response of the system.

2.4 Stability

Consider a linear time-invariant system with given initial condition and input:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ \bar{x}(0) \text{ known} \end{cases}$$

Denote by $\bar{x}(t)$ the corresponding solution. Assume now to perturb the initial condition: $x(t) \neq \bar{x}(t)$ will then be the new solution of the new system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ \bar{x}(0) + \Delta x(0) \text{ known} \end{cases}$$

The main question is how the two solutions are related: do they tend to stay close, to get closer, or to split apart? To investigate this problem, we can study the difference of the two solutions,

$$z(t) \doteq x(t) - \bar{x}(t),$$

and its behaviour as $t \to +\infty$. If we take the derivative of $z(t)$, we get the following relation:

$$\dot{z}(t) = \dot{x}(t) - \dot{\bar{x}}(t) = Ax(t) + Bu(t) - A\bar{x}(t) - Bu(t) = A(x(t) - \bar{x}(t)) = Az(t) \qquad (2.46)$$

Observation 2.1. *The difference between the nominal solution and the perturbed solution* **evolves as the natural response** *of the system, regardless of the input $u(t)$.*

We can now introduce the concept of stability of a system.

Definition 2.1.

- *The system is* ***stable*** *if $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ such that, if $\|x(0) - \bar{x}(0)\| = \|z(0)\| < \delta$, then $\|x(t) - \bar{x}(t)\| = \|z(t)\| \leq \varepsilon$.*
- *The system is* ***asymptotically stable*** *if the previous condition holds and moreover, for $\|z(0)\| < \delta$, we have $\lim_{t\to+\infty} \|z(t)\| = 0$;*
- *The system is* ***marginally stable*** *if it is stable, but not asymptotically stable.*
- *The system is* ***unstable*** *if it is not stable.*

Graphically, the situation is shown in Figure 2.4. For any sphere $\mathcal{S}_\varepsilon$ of radius $\varepsilon > 0$, there must exist a smaller sphere $\mathcal{S}_\delta$ of radius $\delta > 0$ such that, for any perturbation inside $\mathcal{S}_\delta$, the system solution is confined within $\mathcal{S}_\varepsilon$. If, on top of this, z converges to 0, then the system is asymptotically stable (see Figure 2.5).

Stability can be defined as a property of the system *for linear systems only*. In the nonlinear case, it is only possible to discuss *stability of a given equilibrium point*. A nonlinear system may have both stable and unstable equilibrium points, such as the pendulum in the upper and lower positions.

Example 2.10. *(The pendulum.) Figure 2.6 represents the pendulum in the lower equilibrium position (on the left) and the inverted pendulum in the upper equilibrium position (on the right). For small oscillations and small perturbations, the systems corresponding to the two cases can be considered linear. The "linearised" system associated with the lower position is stable. Conversely, the "linearised" system associated with the upper position is unstable, because for any disturbance, no matter how small, the difference between nominal solution and perturbed solution will leave a small sphere of radius ε.*

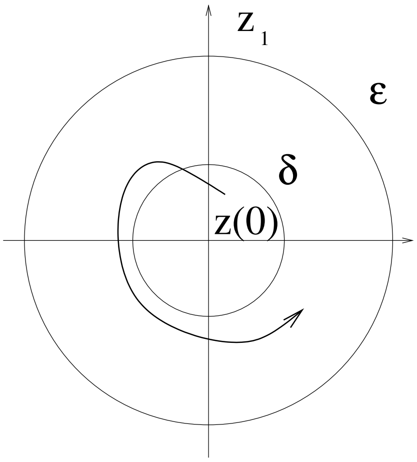

Figure 2.4: Stability of a system.

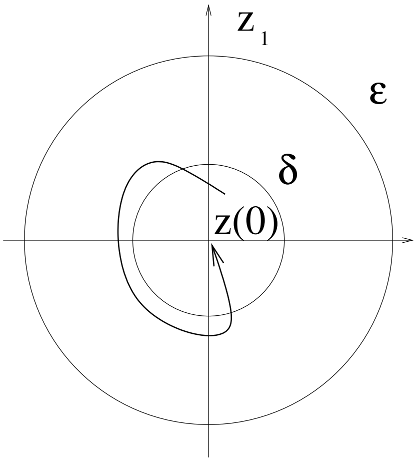

Figure 2.5: Asymptotic stability: convergence of $z(t)$ for $t \to +\infty$.

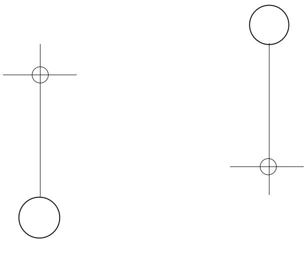

Figure 2.6: Pendulum and inverted pendulum.

Since the difference between nominal solution and perturbed solution evolves as the natural response, stability can be analysed by considering the modes of the system.

Definition 2.2. *The* ***spectrum of*** *A is the the set of eigenvalues of A:*

$$\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}.$$

We have to remind that any eigenvalue can be written as $\lambda = \xi + j\omega$, where $\omega = 0$ if the eigenvalue is real; we denote as $\Re(\lambda)$ the real part of a complex number. Hence, the most general expression of a mode is

$$t^k e^{\lambda t} = t^k e^{\xi t}[\cos(\omega t) + j\sin(\omega t)]$$

If we consider the magnitude, we have the following cases:

- if $\Re(\lambda) < 0$, then $|t^k e^{\lambda t}| \to 0$;
- if $\Re(\lambda) > 0$, then $|t^k e^{\lambda t}| \to \infty$;
- if $\Re(\lambda) = 0$, then $|t^k e^{\lambda t}| \to \infty$, unless $k = 0$, since in this case we have $|t^0 e^{\lambda t}| = |e^{\lambda t}| = 1$.

To have asymptotic stability, all of the modes must converge to 0; to have stability, they must be bounded. Then, the following fundamental properties hold.

Theorem 2.1.

- *The system is* ***asymptotically stable*** *if and only if, for any $\lambda \in \sigma(A)$, we have $\Re\{\lambda\} < 0$.*
- *The system is* ***stable*** *if and only if, for any $\lambda \in \sigma(A)$, we have $\Re\{\lambda\} \leq 0$ and, moreover, if some $\lambda \in \sigma(A)$ is such that $\Re\{\lambda\} = 0$, then it must be* $\deg(\lambda) = 1$.
- *The system is* ***unstable*** *if either there exists $\lambda \in \sigma(A)$ with $\Re\{\lambda\} > 0$ or there exists $\lambda \in \sigma(A)$ with $\Re\{\lambda\} = 0$ and* $\deg(\lambda) > 1$.

We stress that marginal stability should be carefully considered and it cannot be enough for engineering applications. Indeed, infinitely small perturbations of the system parameters can bring the eigenvalues with zero real part to the unstable (right) half-plane where $\Re(\lambda) > 0$. In addition, marginal stability does not guarantee the damping of the modes, *i.e.*, convergence, but only their boundedness.

Example 2.11. *Consider the following matrix in Jordan form:*

$$\begin{bmatrix} -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 0 & -j4 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & -j4 & 0 & 0 \\ 0 & 0 & 0 & 0 & j4 & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & j4 \end{bmatrix}$$

*The ascent of the eigenvalues having zero real part is equal to 2: hence, the system is unstable. If we replace the "**1**" entries with two zero entries, the system becomes stable (marginally).*

2.5 Transfer functions and frequency response

The Laplace transform is a fundamental tool to analyse linear time-invariant dynamical systems. Applying the Laplace transform to a LTI system equations gives

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \xrightarrow{\mathcal{L}} \begin{cases} s\,x(s) - x(0) = Ax(s) + Bu(s) \\ y(s) = Cx(s) + Du(s) \end{cases} \tag{2.47}$$

We stress that, for computing the transform, it is essential that the matrices A, B, C, D are constant over time, meaning that the system is linear and time-invariant. Then we get

$$\begin{aligned} (sI - A)x(s) &= x(0) + Bu(s) \\ \Rightarrow x(s) &= (sI - A)^{-1}x(0) + (sI - A)^{-1}Bu(s) \end{aligned} \tag{2.48}$$

which expresses the system state in terms of the Laplace transform. If we compare this with the result obtained in the time domain, we can observe the following matches:

$$\begin{aligned} e^{At} &\xrightarrow{\mathcal{L}} (sI - A)^{-1} \\ \int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma &\xrightarrow{\mathcal{L}} (sI - A)^{-1}Bu(s) \end{aligned} \tag{2.49}$$

The transformed output is given by

$$y(s) = C(sI - A)^{-1}x(0) + [C(sI - A)^{-1}B + D]u(s) \tag{2.50}$$

If we consider the forced response only, namely we set $x(0) = 0$, we have

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

where $W(s) = \dfrac{N(s)}{d(s)}$ is a matrix of rational transfer functions (where each entry expresses the ratio between the Laplace transform of an output and the Laplace transform of an input). The term $d(s)$ is the characteristic polynomial of matrix A.

The elements of matrix $W(s)$ are **proper** transfer functions, that is, the degree of numerator is less than or equal to the degree of the denominator:

$$\deg(n_{ij}) \leq \deg(d).$$

The strict inequality holds if and only if $D_{ij} = 0$:

$$\deg(n_{ij}) < \deg(d) \Leftrightarrow D_{ij} = 0.$$

In this case, the functions are called **strictly proper**. To summarise, the matrix of transfer functions $W(s)$ is proper, and it is strictly proper (it has all strictly proper components) if and only if $D = 0$.

The transfer function can be computed as follows. Consider the case with a single input and a single output $m = p = 1$

1. Compute $d(s) = \det(sI - A)$.

2. Compute the numerator as follows

$$n(s) = \det\left[\begin{array}{c|c} (sI - A) & -B \\ \hline C & D \end{array}\right]$$

To prove the last equation let $\psi(s) \doteq (sI - A)^{-1}B$ and note that

$$\frac{n(s)}{d(s)} = C(sI - A)^{-1}B + D = C\psi(s) + D$$

then

$$\left[\begin{array}{c|c} (sI-A) & -B \\ \hline C & D \end{array}\right] \left[\begin{array}{c} \psi(s) \\ \hline o \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \frac{n(s)}{d(s)} \end{array}\right] \quad \text{where } \; o = 1$$

Now let us pretend to ignore the value of $o = 1$ and let us apply Cramer's rule to determine o

$$o = \frac{\det \left[\begin{array}{c|c} (sI-A) & 0 \\ \hline C & \frac{n(s)}{d(s)} \end{array}\right]}{\det \left[\begin{array}{c|c} (sI-A) & -B \\ \hline C & D \end{array}\right]}$$

Now we suddlently remind that $o = 1$ and we notice that the numerator is the determinant of a block–triangular matrix, hence

$$\det \left[\begin{array}{c|c} (sI-A) & 0 \\ \hline C & \frac{n(s)}{d(s)} \end{array}\right] = \det(sI - A)\frac{n(s)}{d(s)} = n(s) = \det \left[\begin{array}{c|c} (sI-A) & -B \\ \hline C & D \end{array}\right]$$

It is not difficult to see that for generic m and n the numerators polynomials are

$$n_{ij}(s) = \det \left[\begin{array}{c|c} (sI-A) & -B_j \\ \hline C_i & D_{ij} \end{array}\right]$$

2.5.1 Frequency response

Consider a system with a single input and a single output (SISO) and assume $x(0) = 0$. For brevity, we also assume $D = 0$. The system output evolves as

$$y(t) = \int_0^t W(t - \sigma)u(\sigma)d\sigma \quad \xrightarrow{\mathcal{L}} \quad y(s) = W(s)u(s)$$

Now suppose that we apply to the system an input of the form

$$u(t) = \begin{cases} e^{\sigma t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases} \tag{2.51}$$

where σ is a complex number. Note that, if $\sigma = 0$, this input signal is the unit step. If $\sigma = j\omega$, we have

$$e^{j\omega t} = \cos(\omega t) + j\sin(\omega t).$$

Both are meaningful signals to test the system.

From the well-known properties of the Laplace transform, we have:

$$u_\sigma(s) = \frac{1}{s - \sigma}$$

Assuming $\sigma \notin \sigma(A)$ (σ is not in the spectrum of A, namely, it is not an eigenvalue), we obtain:

$$y(s) = W(s)\frac{1}{s-\sigma} = W(\sigma)\frac{1}{s-\sigma} + (W(s) - W(\sigma))\frac{1}{s-\sigma} \tag{2.52}$$

It can be shown that the second term does not have poles at $s = \sigma$, since the term $s - \sigma$ cancels out. Then

$$y(s) = W(\sigma)\frac{1}{s-\sigma} + \tilde{W}(s), \tag{2.53}$$

where $\tilde{W}(s) = \dfrac{\tilde{n}(s)}{p(s)}$ and $p(s)$ is the characteristic polynomial of matrix A. Hence, the poles of $\tilde{W}(s)$ are in the spectrum of matrix A. By applying the inverse Laplace transformation, we get:

$$y(t) = W(\sigma)e^{\sigma t} + \tilde{W}(t), \tag{2.54}$$

where $\tilde{W}(t)$ is a linear combination of the modes. In the case of distinct eigenvalues, we have

$$y(t) = W(\sigma)e^{\sigma t} + \sum_{i=1}^{n} \tilde{q}_i\, e^{\lambda_i t}. \tag{2.55}$$

The first term is called **steady-state response**, while the second is the **transient response** and depends on the modes only. If the system is asymptotically stable, the transient response tends to zero as $t \to +\infty$.
In particular, the steady-state response to the test signal for $\sigma = j\omega$ is

$$\begin{aligned} y_{j\omega}(t) &= W(j\omega)e^{j\omega t} = |W(j\omega)|e^{j\omega t + \varphi(j\omega)} = \\ &= |W(j\omega)|\cos(\omega t + \varphi(j\omega)) + j|W(j\omega)|\sin(\omega t + \varphi(j\omega)). \end{aligned} \tag{2.56}$$

This means that, if the input is a sine or cosine wave of pulsatance ω, the resulting output is a sine or cosine with the same pulsatance, amplified by a factor $|W(j\omega)|$ and shifted by an angle equal to $\varphi(j\omega)) = \arg\{W(j\omega)\}$. This can be represented in terms of phasors, as shown in Figure 2.7.

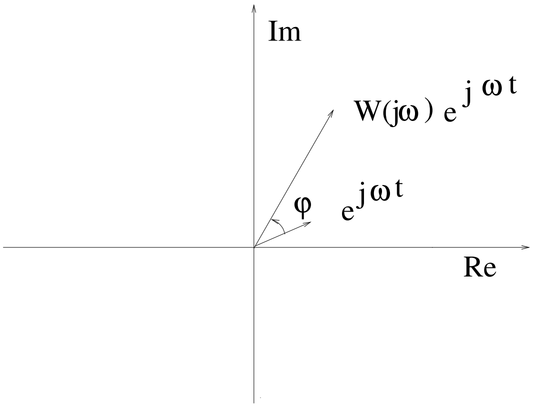

Figure 2.7: Phasors involved in the frequency response of a system.

2.6 Discrete-time systems

Discrete-time systems are similar to continuous-time systems. The substantial difference is that they are not described by differential equations, but by **difference equations**, in which input, output and state $u(k)$, $y(k)$, $x(k)$ are sequences defined for $k \in \mathbb{Z}$.

Example 2.12. (Fibonacci sequence.)

$$y(k+2) = y(k+1) + y(k)$$

is a difference equation that describes a discrete-time system. For $y(0) = y(1)$*, this difference equation gives the Fibonacci sequence.*

As in the case of continuous-time systems, the variables can be expressed in a vector form, $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^m, y(k) \in \mathbb{R}^p$. Then, the discrete-time equations are

$$\begin{cases} x(k+1) = f(k, x(k), u(k)) \\ y(k) = g(k, x(k), u(k)) \end{cases} \tag{2.57}$$

Again, this notation is equivalent to the following set of equations

$$\begin{array}{rcll} x_i(k+1) & = & f_i(k,\ x_1(k),\ \dots,\ x_n(k),\ u_1(k),\ \dots,\ u_m(k)) & \text{for} \quad i = 1,\ \dots,\ n \\ y_j(k) & = & g_j(k,\ x_1(k),\ \dots,\ x_n(k),\ u_1(k),\ \dots,\ u_m(k)) & \text{for} \quad j = 1,\ \dots,\ p \end{array}$$

Example 2.13. (Euler approximating system.) *The following equations*

$$\begin{cases} \dot{x}_1(k+1) = x_1(k) + \tau[-\alpha x_1(k)x_2(k) + u_1(k)] \\ \dot{x}_2(k+1) = x_2(k) + \tau[-\alpha x_1(k)x_2(k) + u_2(k)] \\ \dot{x}_3(k+1) = x_3(k) + \tau[\ \alpha x_1(k)x_2(k) - \gamma x_3(k)] \end{cases}$$

where τ is a small parameter, represent a nonlinear discrete-time system. If we assume that the time variables are taken from a continuous time axis that has been discretised with step τ,

$$t = k\tau, \qquad x(k) = x(k\tau), \quad u(k) = u(k\tau),$$

and τ is omitted to have a simpler notation, then this system is the Euler approximating system of the equations (2.6). For τ small enough, the sequence $x(k\tau)$ approximates the exact solution of the system (2.6).

In general, given any system $\dot{x}(t) = f(x(t), u(t))$, we can numerically approximate its solution by means of the Euler approximating system

$$x((k+1)\tau)) = x(k\tau)) + \tau f(x(k\tau), u(k\tau))$$

This is the simplest method to numerically compute the solution of a continuous-system (of course, there are more sophisticated methods that can be used).

A discrete-time system is **invariant** or **autonomous** if the functions f and g do not depend directly on time k:

$$\begin{cases} x(k+1) = f(x(k), u(k)) \\ y(k) = g(x(k), u(k)) \end{cases} \tag{2.58}$$

Example 2.14. *The system*

$$x(k+1) = \left(\frac{k}{k+1}\right)x(k) + \left(\frac{1}{k+1}\right)u(k)$$

provides at each time k the average of all the past input values:

$$x(k) = \frac{u(k-1) + u(k-2) + \cdots + u(0)}{k}.$$

This system is not autonomous. Conversely, the system providing the Fibonacci sequence is autonomous (time-invariant).

A discrete system is **linear** if f and g are linear with respect to x and u:

$$\begin{cases} x(k+1) = A(k)x(k) + B(k)u(k) \\ y(k) = C(k)x(k) + D(k)u(k) \end{cases} \tag{2.59}$$

A discrete system is **linear and time-invariant** if it satisfies both conditions:

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) \\ y(k) = Cx(k) + Du(k) \end{cases} \tag{2.60}$$

2.6.1 General solution of a discrete-time linear time-invariant system

The solution of a discrete-time system can be recursively computed. However, we are interested in qualitative information. As we have done for continuous-time systems, in the linear case we split the problem into two distinct parts:

- **natural response** (assigned initial conditions, null input):

$$\begin{cases} x_L(k+1) = Ax_L(k) \\ x_L(0) = x(0) \quad \text{known} \end{cases} \tag{2.61}$$

- **forced response** (null initial conditions, assigned inputs):

$$\begin{cases} x_F(k+1) = Ax_F(k) + Bu(k) \\ x_F(k) = 0 \end{cases} \tag{2.62}$$

We have

$$x(k) = x_L(k) + x_F(k)$$

It is not difficult to verify that the general solution is:

$$\begin{aligned} x(k) &= A^k x(0) + \sum_{h=0}^{k-1} A^{k-h-1} Bu(h) \\ y(k) &= CA^k x(0) + \sum_{h=0}^{k-1} CA^{k-h-1} Bu(h) + Du(k) \end{aligned} \tag{2.63}$$

Assume that matrix A has n distinct eigenvalues. Then there are n linearly independent eigenvectors that we can group in an invertible matrix T

$$T = [t_1, t_2, \dots, t_n]$$

Let us denote

$$S = T^{-1} = \begin{bmatrix} s_1^\top \\ s_2^\top \\ \vdots \\ s_n^\top \end{bmatrix}$$

as before. We remind that $A^k = T\Lambda^k S$. Hence, along the same lines as in the previous computation for the continuous-time case, we find

$$A^k = \sum_{i=1}^{n} Z_i \lambda_i^k, \tag{2.64}$$

where $Z_i = t_i s_i^\top$. The sequences λ_i^k are called **modes of the discrete-time system**, or discrete-time modes.

As expected, the modes affect both the natural response and the forced response. Indeed, for any eigenvalue $\lambda \in \mathbb{C}$ we have

$$\lambda = |\lambda| e^{j\theta},$$

where $\theta = 0$ if the eigenvalue is real. Then

$$\lambda^k = |\lambda|^k e^{j\theta k} = |\lambda|^k [\cos(\theta k) + j\sin(\theta k)]$$

Note that the magnitude of this sequence is increasing if $|\lambda| > 1$ or decreasing if $|\lambda| < 1$ and constant if $|\lambda| = 1$. Then we can already see the similarity of the theory for continuous and discrete systems: the only difference is that the real and imaginary part of eigenvalues are considered in the continuous-time case, while the magnitude and phase of the eigenvalues are considered in the discrete-time case.

In the case of distinct eigenvalues, we can write the following expression for A^k

$$A^k = \sum_{i=1}^{r} Z_i \lambda_i^k + 2 \sum_{i=r+1,\ step\,2}^{n-1} |\lambda_i|^k (M_i \cos(\theta_i k) - N_i \sin(\theta_i k)) \tag{2.65}$$

in the presence of r real eigenvalues.

If matrix A has multiple eigenvalues, we have a more general formula:

$$A^k = \sum_{i=1}^{m} \sum_{h=0}^{\deg(\lambda_i)-1} Z_{ih} \begin{pmatrix} k \\ h \end{pmatrix} \lambda_i^k \tag{2.66}$$

Note that

$$\begin{pmatrix} k \\ h \end{pmatrix} = \frac{k(k-1)\dots(k-h+1)}{h!} = p_h(k)$$

is a polynomial in k of degree h. The maximum degree is the ascent of the eigenvalue.

The expression for the modes can be obtained by taking into account that, for any matrix A, $A^k = TJ^kS$, where J is in Jordan form. Raising to a power k a matrix in Jordan form, we get

$$J^k = \text{blockdiag}\{J_1^k, J_2^k, \dots, J_s^k\}$$

a block diagonal matrix with the powers of the Jordan blocks. For each Jordan block

$$(\lambda I + J_0)^k = \sum_{h=0}^{k} \begin{pmatrix} k \\ h \end{pmatrix} \lambda^{k-h} J_0^h$$

where J_0 is

$$J_0 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

It is easy to check that, for h larger than the ascent of the eigenvalue, we have $J_0^h = 0$.

Example 2.15. *Consider of a matrix with a single eigenvalue in Jordan form*

$$J = \begin{bmatrix} 3 & 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

The three modes associated with the corresponding discrete-time system are

$$3^k, \quad \begin{pmatrix} k \\ 1 \end{pmatrix} 3^k = p_1(k)3^k, \quad \begin{pmatrix} k \\ 2 \end{pmatrix} 3^k = p_2(k)3^k.$$

We can also define the **impulse response matrix** as follows:

$$W(0) = D, \quad \text{and} \quad W(k) = CA^kB, \quad k = 1, 2, \ldots$$

Each element $W_{ij}(k) = C_iA^kB_j$ is the response of the i-th output when the discrete impulse $\{1,\ 0,\ 0,\ \ldots,\ 0\}$ is applied as the j-th input.

2.6.2 Discrete-time stability

To analyse the stability of discrete-time linear systems, note that the general expression of a mode is

$$p_h(k)\lambda^k = p_h(k)|\lambda|^k[\cos(\theta k) + j\sin(\theta k)]$$

Therefore, a mode converges to zero if and only if $|\lambda| < 1$ and is bounded if and only if $|\lambda| \leq 1$ and the polynomial has degree 0: $p_0 = \left(\begin{array}{c} k \\ 0 \end{array}\right) = 1$. We conclude the following.

Theorem 2.2. *A discrete system is*

- ***asymptotically stable*** *if and only if* $\forall\ \lambda \in \sigma(A)$ *it holds* $|\lambda| < 1$*;*
- ***stable*** *if and only if* $\forall\ \lambda \in \sigma(A)$ *we have* $|\lambda| \leq 1$ *and, if* $|\lambda| = 1$*, then* $\deg(\lambda) = 1$*;*
- ***unstable*** *if and only if there exists either* $\lambda \in \sigma(A)$ *with* $|\lambda| > 1$*, or* $\lambda \in \sigma(A)$ *with* $|\lambda| = 1$ *and* $\deg(\lambda) > 1$.

2.7 State transformation

For reasons that will become clear later on, it is often useful to introduce a change of variables in the natural representation of a system. Given an invertible matrix associated with a new basis for the system state space,

$$T = [\bar{t}_1,\ \bar{t}_2, \ldots,\ \bar{t}_n]\,,$$

we define the transformation

$$\begin{cases} \hat{x} = T^{-1}x \\ x = T\hat{x} \end{cases}$$

Applying this transformation to a linear autonomous system, we get:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \Rightarrow \begin{cases} T\dot{\hat{x}}(t) = AT\hat{x}(t) + Bu(t) \\ y(t) = CT\hat{x}(t) + Du(t) \end{cases}$$

$$\Rightarrow \begin{cases} \dot{\hat{x}}(t) = T^{-1}AT\hat{x}(t) + T^{-1}Bu(t) \\ y(t) = CT\hat{x}(t) + Du(t) \end{cases} \tag{2.67}$$

So the system has a new representation:

$$\begin{cases} \dot{x}(t) = \hat{A}x(t) + \hat{B}u(t) \\ y(t) = \hat{C}x(t) + \hat{D}u(t) \end{cases} \tag{2.68}$$

where the matrices have the following expressions:

$$\hat{A} = T^{-1}AT \tag{2.69}$$

$$\hat{B} = T^{-1}B \tag{2.70}$$

$$\hat{C} = CT \tag{2.71}$$

$$\hat{D} = D \tag{2.72}$$

Note that matrix D is unchanged. This was expected, because D represent the relationship between $y(t)$ and $u(t)$, which are not transformed.

Note also that, in the transition to new representation, matrix A undergoes a **similarity transformation**. Characteristic polynomial and (thus) eigenvalues are invariant under this transformation, hence the modes and the stability properties of the system are unchanged. Also the system transfer function matrix is unchanged, as expected since input and output are unaltered by the transformation.

2.8 Sampled-data systems

Nowadays, signal processing occurs almost always through the use of computers, which are known to evolve in discrete time. The theory for discrete- and continuous-time systems has strong analogies that can be exploited. Clearly, when a continuous-time system must interact with a discrete-time system, a fundamental aspect is to understand how it is possible to convert continuous-time signals (coming from the outside world into the digital system) in discrete-time signals (evolving within the computer) and vice versa.

2.8.1 Analog-to-Digital conversion

The conversion of an analog signal into a digital signal occurs according to the block diagram depicted in Figure 2.8.

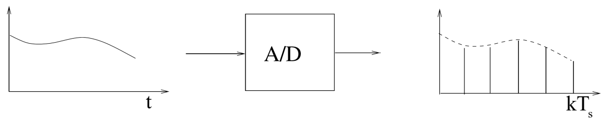

Figure 2.8: Analog-to-Digital conversion.

The block marked with A/D is an analog-to-digital converter and the implemented operation is called **sampling**. It is assumed here that the signal is sampled with a fixed step T_s. In general it is also possible that a vector signal (multiple input channels) is sampled with different rates for different components. These are called **multi-rate systems**, a case we will not consider. The sampling operator is linear, therefore it does not compromise the linearity of the system:

$$\begin{aligned} f(t) &\overset{A/D}{\longrightarrow} f(kT_s) \\ \alpha f(t) + \beta g(t) &\overset{A/D}{\longrightarrow} \alpha f(kT_s) + \beta g(kT_s) \end{aligned} \tag{2.73}$$

A basic scheme of an analog-to-digital converter is depicted in Figure 2.9. When the START signal switches to 1, it enables a ramp generator and a counter. The counter is stopped when the value r reaches the signal s. The counter provides then an integer number s^* proportional to s (with a roundoff error). The time required for conversion must be much smaller than the sampling time and it must be very small compared to the inverse of the signal bandwidth.

2.8.2 Digital-to-Analog conversion

This type of operation is certainly less straightforward than analog-to-digital conversion. In principle, given N data points, interpolation allows us to find a function (for example, a polynomial

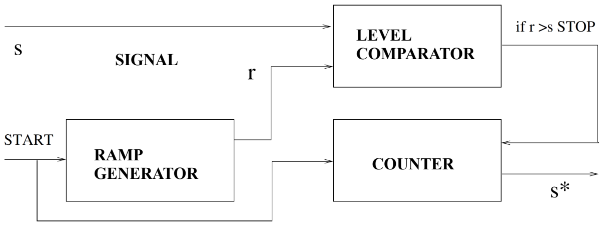

Figure 2.9: Qualitative scheme of an Analog-to-Digital Converter.

of degree $N - 1$) that passes through these N data points. In the case of **polynomial interpolation**, this reduces to the following linear equations

$$a_0 + a_1 x_i(kT_s) + a_2 x_i^2(kT_s) + \ldots + a_n x_i^{n-1}(kT_s) = f(kT_s) \quad \forall\, i = 1, \ldots, N \tag{2.74}$$

that allow us to determine the coefficients of the interpolating polynomial. However, in order to apply this method and find the polynomial, we must know all of the N samples. Then this method has to be adopted **off-line**. This is obviously not feasible in **real time** applications. It is necessary to find a way to convert **on-line** the samples, to generate the continuous-time signal in real time.

An easy way to convert on-line is to use the so-called **zero order hold** (Z.O.H.). This method simply creates a piecewise constant function, whose value is equal to the value of the last available sample (see Figure 2.10).

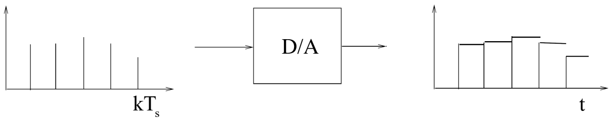

Figure 2.10: Zero Order Hold.

This operation creates from the samples $f(kT_s)$ a piecewise constant function whose value is

$$\hat{f}(t) = f(kT_s) \text{ for } kT_s \le t < (k+1)T_s.$$

There are also higher-order holders (see Figure 2.11 for the case of order 1), but nowadays the sampling time can be assumed so small that the information lost by adopting zero order holders is practically negligible (given that the error decreases when T_s decreases, as shown in Figure 2.12). In fact, methods based on higher order holders are seldom used.

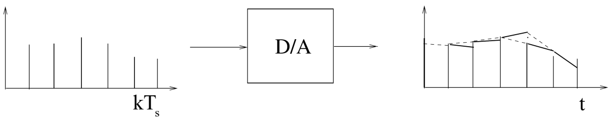

Figure 2.11: First Order Hold.

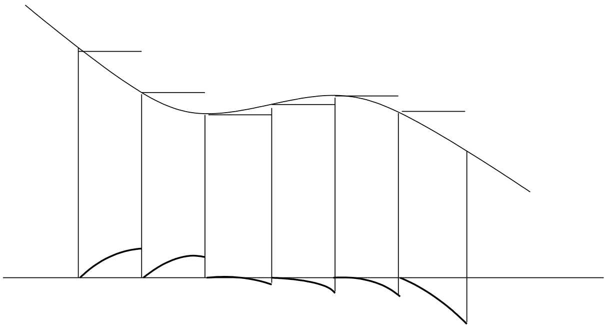

Figure 2.12: Error with Zero Order Hold.

2.8.3 Aliasing

We now briefly recall the problem of aliasing, which has to be carefully considered when we use a sampling device. We consider the case of a sinusoidal signal. This is not restrictive, since it is known that periodic functions can be described by a Fourier series expansion that consists of harmonic functions, and the most reasonable signals admit a Fourier transform that expresses their harmonic content.

Let us denote by

$$\Omega_c = \frac{2\pi}{T_s}$$

the **sampling pulsatance**. Consider the continuous-time signal $\cos(\omega_1 t)$: by sampling the signal with step T_s, we obtain a sequence $\cos(\omega_1 k T_s)$. It is easily seen that the *same sequence* can be obtained if we sample, with the same step T_s, a signal $\cos(\omega_2 t)$ with a different pulsatance ω_2, provided that the following equality holds:

$$\begin{aligned} \cos(\omega_2 k T_s) &= \cos(\omega_1 k T_s + mk2\pi) \\ \Rightarrow \omega_2 &= \omega_1 + m\frac{2\pi}{T_s} = \omega_1 + m\,\Omega_C, \quad \text{with}\, m \in \mathbb{Z} \end{aligned} \tag{2.75}$$

Therefore, if the expression

$$\omega_2 - \omega_1 = m\,\Omega_C, \quad m \in \mathbb{Z}$$

holds, the two signals are indistinguishable under sampling. In general, if a signal includes two components of this kind, it is impossible to correctly reconstruct the signal.

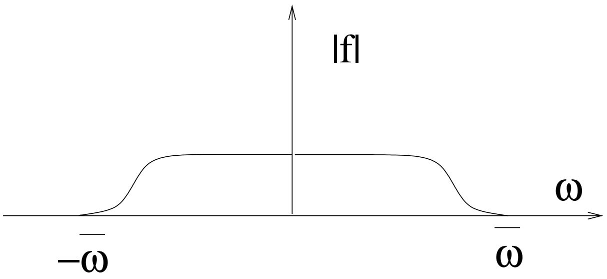

Figure 2.13: Spectrum of a band-limited signal.

Signals processed in real problems typically are band-limited: their Fourier transform, or their spectrum, has a bounded support, as is qualitatively shown in Figure 2.13. A band-limited signal can always be fully reconstructed from its samples, provided that a fundamental condition is satisfied (**Nyquist-Shannon sampling theorem**). For a band-limited signal whose band is between $-\bar{\omega}$ and $\bar{\omega}$, reconstruction is possible after sampling as long as

$$2\bar{\omega} < \Omega_C, \tag{2.76}$$

because in this case (2.75) can be satisfied only for $m = 0$, hence $\omega_2 = \omega_1$. Replacing $\bar{\omega} = 2\pi\bar{f}$ and $\Omega_c = 2\pi f_c$, where $\bar{f}$ is the maximum signal frequency and f_C is the sampling frequency, we get $\bar{f} < \frac{f_C}{2}$ or

$$f_C > 2\bar{f}, \tag{2.77}$$

namely, to be able to fully reconstruct the signal after sampling, the sampling frequency must exceed twice the maximum frequency in the band-limited signal. The minimum sampling frequency $f_{C_{min}} = 2\bar{f}$ is called the Nyquist rate. This fundamental result is a consequence of the **sampling theorem**, stating that the continuous-time signal is completely recoverable, at least in principle, starting from the individual samples provided that the condition above is satisfied.

The proof of sampling theorem, due to Shannon in 1949, was a fundamental step for signal theory and telecommunications, because it allowed the transition from analog to digital signals processing.

One final consideration concerns the condition $\bar{f} < \dfrac{f_C}{2}$ and its applicability. In reality, being the signals unknown and unpredictable, we cannot know the signal bandwidth in advance. In these cases, however, once the sampling frequency of the devices is known, the received signal is pre-filtered in order to eliminate all of the spurious spectral components whose frequency is higher than $\dfrac{f_C}{2}$. This filtering effect often occurs naturally in many devices, as microphones or sensors. Otherwise, the introduction of an analog filter is necessary.

Example 2.16. *The frequencies audible by human ears are up to about* 20 *kHz. This means that each spectral component whose frequency is above this value is not detectable by humans. Microphones have a cutoff frequency of about* 20 *kHz, so basically greater frequencies are not recorded. In view of the condition $f_C > 2\bar{f}$, the sampling frequency for digital music is usually chosen as $f_C = 44,1$ kHz.*

2.9 Discrete-time equivalent of a continuous-time system

Consider the scheme shown in Figure 2.14, representing a typical situation where a continuous-time process is controlled by means of a discrete-time device (*e.g.*, a computer). A/D and D/A converters in the figure are used to allow the communication between continuous-time and discrete-time subsystems.

To study the entire system, composed of both the process and the controller, it is convenient to have consistent equations (defined on the same time domain). Then, there are two possibilities:
a) compute a discrete-time system that is equivalent to the continuous-time process;
b) compute a continuous-time system that is equivalent to the discrete-time controller.

We now concentrate on case a) and compute the discrete-time equivalent of the continuous-time process. We assume that the digital-to-analog converter is a zero order holder, hence the signal it generates as an output is piecewise constant:

$$u(t) = u(kT_s) \quad \text{for} \quad kT_s \leq t < (k+1)T_s.$$

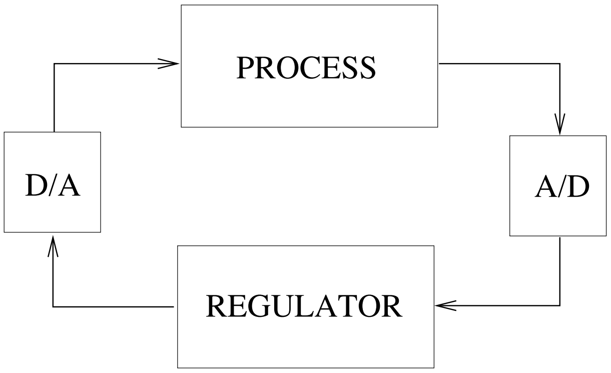

Figure 2.14: Discrete-time control of a continuous-time process.

The continuous-time linear process governed by equations

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

has then a constant input during the sampling interval. However, this does not mean that $x(t)$ and $y(t)$ are constant during the sampling interval as well. We can only observe that $x(t)$ is continuous and is differentiable everywhere, with the exception of the points kT_s (because u is not continuous at these points). The output $y(t)$ of the continuous-time process enters the analog-to-digital converter, and for this reason we can consider it to be relevant at instants kT_s that are multiples of the sampling time T_s. So, we look at the relationship between $u(kT_s)$, $x(kT_s)$, $x((k+1)T_s)$ and $y(kT_s)$.

If $u(kT_s)$ and $x(kT_s)$ are known, the continuous-time system solution at a time $t > kT_s$ is

$$x(t) = e^{A(t-kT_s)}x(kT_s) + \int_{kT_s}^{t} e^{A(t-\sigma)}Bu(\sigma)d\sigma \tag{2.78}$$

The time value kT_s has now been taken as the initial time, instead of $t = 0$ (this is allowed because the system is time invariant). If we evaluate this expression for $t = (k+1)T_s$, we get:

$$x((k+1)T_s) = e^{AT_s}x(kT_s) + \int_{kT_s}^{(k+1)T_s} e^{A((k+1)T_s-\sigma)}Bu(kT_s)d\sigma, \tag{2.79}$$

where we have set $u(\sigma) = u(kT_s)$ because the input is constant in the integration interval $[kT_s, (k+1)T_s)$. Hence, $u(kT_s)$ can be taken outside the integral

$$x((k+1)T_s) = e^{AT_s}x(kT_s) + \left[\int_{kT_s}^{(k+1)T_s} e^{A((k+1)T_s-\sigma)}Bd\sigma\right]u(kT_s) \tag{2.80}$$

and the term in brackets can be simplified by means of the change of variables $(k+1)T_s - \sigma = \xi$

$$\int_{kT_s}^{(k+1)T_s} e^{A((k+1)T_s-\sigma)}Bd\sigma = \int_{T_s}^{0} e^{A\xi}B(-d\xi) = \int_{0}^{T_s} e^{A\xi}Bd\xi \tag{2.81}$$

to see that the integral does not depend on k. Finally, we get:

$$x((k+1)T_s) = e^{AT_s}x(kT_s) + \left[\int_{0}^{T_s} e^{A\xi}Bd\xi\right]u(kT_s)$$

Therefore, we can write

$$\begin{cases} x((k+1)T_s) &= A_D x(kT_s) + B_D u(kT_s) \\ y(kT_s) &= C_D x(kT_s) + D_D u(kT_s) \end{cases} \tag{2.82}$$

where

$$A_D = e^{AT_s}, \qquad B_D = \int_0^{T_s} e^{A\xi} B d\xi, \qquad C_D = C, \qquad D_D = D \tag{2.83}$$

The graphic representation of the equivalent discrete-time system is shown in Figure 2.15.

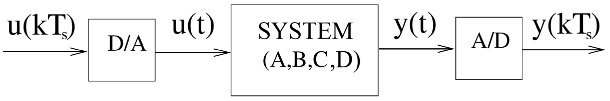

Figure 2.15: Discrete-time system equivalent to the continuous-time system.

Observe that the discrete-time system we have found is **exactly equivalent** to the continuous-time system to which the D/A and A/D converters have been applied (before the system and after the system, respectively), resulting in the digital input $u(kT_s)$ and the digital output $y(kT_s)$. Also, note that the matrices C and D are unchanged.

These results are very important in systems and control theory, since they allow us to design a regulator indifferently in the continuous- or in the discrete-time domain. Nowadays, a digital implementation is highly preferred for its flexibility, reliability and cheapness.

It can be shown that asymptotic stability of the continuous-time system represented by matrices A, B, C, D is equivalent (necessary and sufficient condition) to asymptotic stability of the discrete-time system with matrices A_D, B_D, C_D, D_D. This is true because, if A has eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, then $A_D = e^{AT_s}$ has eigenvalues $\{e^{\lambda_1 T_s}, e^{\lambda_2 T_s}, \ldots, e^{\lambda_n T_s}\}$ and, if λ_k has a negative real part, then the modulus of $e^{\lambda_1 T_s}$ is smaller than 1 (and vice-versa).

Example 2.17. ***(Calculation of the integral*** $\int_0^{T_s} e^{A\xi} B d\xi$***.)***
Using the expression $e^{A\xi} = T e^{\Lambda\xi} S$ *that we have previously derived, we get:*

$$\int_0^{T_s} e^{A\xi} B d\xi = \int_0^{T_s} T e^{\Lambda\xi} S\, B d\xi = T \left[\int_0^{T_s} e^{\Lambda\xi} d\xi \right] S\, B$$

and then we need only to calculate the individual integrals of the exponential $e^{\Lambda t}$*, which are placed on the diagonal.*

Example 2.18. ***(Formula for*** A ***invertible.)*** *If* A *is invertible, we have* $\displaystyle\int_0^{T_s} e^{At} dt = A^{-1}\left[e^{AT_s} - I\right]$. *This can be verified by considering that*

$$\begin{aligned} \frac{d}{dt} e^{At} = \frac{d}{dt} \sum_{k=0}^{+\infty} \frac{(At)^k}{k!} &= \sum_{k=0}^{+\infty} \frac{A^k}{k!} k\, t^{k-1} = \sum_{k=1}^{+\infty} \frac{A^k}{(k-1)!} t^{k-1} = \\ = A \sum_{k=1}^{+\infty} \frac{A^{k-1}}{(k-1)!} t^{k-1} &= A \sum_{h=0}^{+\infty} \frac{(At)^h}{h!} = A\, e^{At} \end{aligned}$$

In fact, we have that $\dfrac{d}{dt}\left[A^{-1} e^{At}\right] = e^{At}$*, hence the definite integral between 0 and* T_s *is the proposed expression.*

Example 2.19. ***(Discrete-time realisation of a controller.)***
Let us compute the discrete-time equivalent of the continuous-time PI (proportional-integral) controller:

$$K_P + K_I \frac{1}{s}.$$

Assuming that the input is $e(t)$ and output $u(t)$, in terms of Laplace transform we have $u(s) = \left(K_P + K_I \frac{1}{s}\right) e(s)$. A continuous-time state representation is

$$\begin{cases} \dot{x}(t) = e(t) \\ u(t) = K_P\, e(t) + K_I\, x(t) \end{cases}$$

associated with the matrices

$$A = [0]\,, \quad B = [1]\,, \quad C = [K_I]\,, \quad D = [K_P]$$

Therefore, the equivalent discrete-time representation requires the following matrices:

$$\begin{aligned} A_D &= \left[e^{AT_s}\right] = \left[e^{0T_s}\right] = [1] \\ B_D &= \left[\int_0^{T_s} e^{A\xi} B d\xi\right] = [T_s] \\ C_D &= [K_I] \\ D_D &= [K_P] \end{aligned}$$

and has the form

$$\begin{cases} x((k+1)T_s) = x(kT_s) + T_s\, e(kT_s) \\ u(kT_s) = K_I\, x(kT_s) + K_P\, e(kT_s) \end{cases}$$

To actually implement this controller on the computer, we can adopt the following algorithm.

1. *Initialise x;*
2. *read e;*
3. *compute $u = K_I\, x + K_P\, e$;*
4. *write u in the digital-analog converter;*
5. *update $x := x + T_s e$;*
6. *wait for the next sampling instant;*
7. *return to 2 if the job is still active, otherwise STOP.*

Chapter 3

Dynamical systems

3.1 General consideration

Although the systems considered in the course are mainly described by differential equations, or difference equations, it is appropriate to introduce a more general definition of dynamical system. Before giving such a definition, we discuss two simple examples.

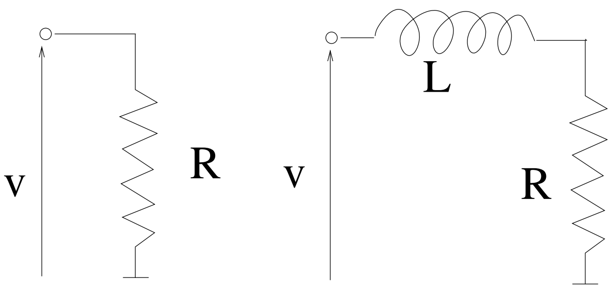

Figure 3.1: Two different types of systems.

Example 3.1. *Consider Figure 3.1, in which two different types of electrical circuit are represented, the first including just a resistor R and the second formed by a simple R-L circuit. The laws governing the two circuits are:*

- *for the first system,* $i(t) = \dfrac{v(t)}{R}$*;*
- *for the second system,* $L\dfrac{di(t)}{dt} = v(t) \;\Rightarrow\; i(t) = i(t_0) + \dfrac{1}{L}\displaystyle\int_{t_0}^{t} v(\sigma)d\sigma.$

In both cases, we can take as input (cause) the voltage and as output (effect) the current. To compute the output of the R circuit, we only need to know the function $v(t)$*. However, the R-L system is governed by a differential equation and the knowledge of* $v(t)$ *is not enough to uniquely determine the value of the current. In fact, we also need the value of* $i(t_0)$*, the initial condition of the system, for some initial time* t_0*. These type of systems, for which it is necessary to know the initial conditions at a given initial time instant, are called* ***dynamical****.*

For a dynamical system, the evolution after a time t_0 depends on both the input and the past history of the system, and not only on $u(t)$. One question we might ask is: besides $u(t)$, do we need to know the entire past history of the evolution of the system in order to calculate its future evolution? The answer is no: if we are interested in what happens *after* an initial time t_0, it is enough to know the **state** of the system at t_0: information concerning the previous instants is redundant. This concept is the notion of state of the system.

State of the system (meta-definition) The state of a system at time τ is given by all the information we need in order to uniquely determine the future evolution for $t > \tau$, given the input function $u(t)$.

Example 3.2. *In the case of the R-L circuit, a state variable is the current $i(t)$. Indeed we can write*

$$i(t) = \frac{1}{L}\int_{-\infty}^{t} v(\sigma)d\sigma = \frac{1}{L}\int_{-\infty}^{t_0} v(\sigma)d\sigma + \frac{1}{L}\int_{t_0}^{t} v(\sigma)d\sigma = i(t_0) + \frac{1}{L}\int_{t_0}^{t} v(\sigma)d\sigma \tag{3.1}$$

where we assumed $i(-\infty) = 0$. Note that we can arrive at the same value $i(t_0)$ via completely different trajectories, but this does not affect the future evolution.

3.2 Causal dynamical systems

In this section we provide a formal definition of a dynamical system. The main purpose is to make it clear that under the term "Dynamical Systems" we include objects which are much more general than the "simple" systems we analyse in the course.

Consider the following sets.

- A set of times T, which is totally ordered (that is, if $t_1 \neq t_2$, then either $t_1 < t_2$ or $t_2 < t_1$). In our case, we mainly consider $T = \mathbb{R}$ or $T = \mathbb{Z}$. Clearly, any sequence of ordered instants t_k, not necessarily equi-spaced, is also appropriate.
- A set of states X. It can be any set (finite or not). We will typically use $\mathbb{R}^n$.
- a set of inputs U. It can be any set (finite or not). We will mainly adopt $\mathbb{R}^m$.
- A set of outputs Y. We will mainly adopt $\mathbb{R}^p$.
- A set of input functions $\Omega = \{U(\ldots) : T \to U |\, \text{condition}\}$. As an example, in the case of an actuator with a minimum and a maximum value, this set will be the set of continuous functions defined on an interval.
- A set of output functions $\Gamma = \{Y(\ldots) : T \to Y |\, \text{condition}\}$. In the case of a sensor, output functions will take values in an interval between the minimum and the maximum values.

Then we introduce the **transition function** φ as

$$\begin{cases} \varphi : T \times T \times X \times \Omega \to X \\ x(t) = \varphi(t, \tau, x(\tau), u(\cdot)) \end{cases} \tag{3.2}$$

where t is the current time, τ the initial instant, $\bar{x}$ the initial state. The notation $u(\cdot)$ indicates the dependence on the entire function and not only on its value at time t. To define a dynamic system, the transition function must satisfy the following axioms.

- **Consistency**:

$$\varphi(\tau, \tau, \bar{x}, u(\cdot)) = \bar{x} \qquad \forall\, \tau \in T,\ \bar{x} \in X,\ u \in \Omega \tag{3.3}$$

  In essence, at the initial time $t = \tau$ the function must give the initial state $x(\tau) = \bar{x}$.

- **Composition**: Given three time instants $t_1 \leq t_2 \leq t_3$,

$$\begin{aligned} x(t_3) = \varphi(t_3,\, t_1,\, x_1,\, u(\cdot)) &= \varphi(t_3,\, t_2,\, x_2,\, u(\cdot)) && (3.4)\\ &= \varphi(t_3,\, t_2, \varphi(t_2,\, t_1,\, x_1,\, u(\ldots)), u(\ldots)) \\ \forall &\quad t_1 \leq t_2 \leq t_3 \in T,\, x \in X,\, u \in \Omega && (3.5) \end{aligned}$$

  This means that, if the state $x(t_2)$ is given, there is no need to know the history of the system before t_2 to determine the future $x(t)$ for $t \geq t_2$.

- **Forward definition**: the function φ must be defined for all $t \geq \tau \ \forall\, x \in X,\ u \in \Omega$. For $t < \tau$, the transition function may be not defined. However, it may be defined for all t as well; this can occur in many physical systems in which it is possible to trace the past evolution of a system: systems for which this happens are called **reversible**.

- **Causality**: given two inputs $u_1(t)$ and $u_2(t)$, overlapping ($u_1(t) = u_2(t)$) on the interval $t_1 \leq t \leq t_2$ (but they may differ outside of this range), we have:

$$\varphi(t, t_1, x, u_1(\ldots)) = \varphi(t, t_1, x, u_2(\ldots)), \quad \forall\, t \text{ in the interval} \tag{3.6}$$

  In practice, if we are interested in the range of time $[t_1, t_2]$, given the state in t_1, information on previous times is not required since it is included in $x(t_1)$, while information after t_2 (the future) is not relevant since it does not affect the past.

As far as the output of a system is concerned, we introduce the following function:

$$\begin{cases} \eta : T \times X \times U \to Y \\ y(t) = \eta(t,\, x(t),\, u(t)) \end{cases} \tag{3.7}$$

If $y(t) = \eta(t,\, x(t))$, the system is said **strictly proper** (in the case of linear systems, this means that matrix D is equal to zero).

Example 3.3. *The axioms above can be checked for the known expression of the solution of a linear time-invariant system*

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

This is a transition function. In fact, we have

- *$x(t_0) = Ix(t_0)$ because $e^{A(t_0-t_0)} = I$ and the integral vanishes, so we have consistency;*

- *given the instants $t_1 \leq t_2 \leq t_3$, we have*

$$\begin{aligned} x(t_3) &= e^{A(t_3-t_1)}x(t_1) + \int_{t_1}^{t_3} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma \\ &= e^{A(t_3-t_2)}e^{A(t_2-t_1)}x(t_1) + \int_{t_1}^{t_2} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma + \int_{t_2}^{t_3} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma \\ &= e^{A(t_3-t_2)}e^{A(t_2-t_1)}x(t_1) + \int_{t_1}^{t_2} e^{A(t_3-t_2)}e^{A(t_2-\sigma)}Bu(\sigma)d\sigma + \int_{t_2}^{t_3} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma = \\ &= e^{A(t_3-t_2)}\left[e^{A(t_2-t_1)}x(t_1) + \int_{t_1}^{t_2} e^{A(t_2-\sigma)}Bu(\sigma)d\sigma\right] + \int_{t_2}^{t_3} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma \\ &= e^{A(t_3-t_2)}\left[x(t_2)\right] + \int_{t_2}^{t_3} e^{A(t_3-\sigma)}Bu(\sigma)d\sigma \end{aligned}$$

- *$x(t)$ is surely defined for $t \geq t_0$ because the integral is defined for these values; in this case the system is reversible, so it is defined for $t < t_0$ as well;*

- *if $u_1(t) = u_2(t)$ in an interval $[t_1, t_2]$, then the integral does not change by taking different values $u_1(t) \neq u_2(t)$ outside the interval, so we have causality.*

The properties of the transition function are also verified for discrete-time systems. In this case, the exact calculation of φ is possible by recursion. However, for discrete-time systems the property of reversibility, typically verified in continuous-time, may fail. For example, the system described by the equation

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

is non-reversible, because the state matrix (A) is not invertible.

We may wonder when the system of equations

$$\begin{cases} \dot{x}(t) = f(t,\ x(t),\ u(t)) \\ x(\tau) \quad \text{known} \end{cases}$$

(which is not a transition function) can be associated with a dynamical system. Under the assumptions of existence and uniqueness of the solution, $x(t)$ is uniquely determined by the input and by $x(\tau)$. Therefore, in principle there exists a function $x(t) = \varphi(t,\ \tau,\ x(\tau),\ u(\ldots))$, which is precisely the transition function, even though most of the times it is impossible to know or to calculate analytically, without resorting to some numerical technique.

A mathematical problem that arises when we deal with differential equations is that, in pathological cases the uniqueness of the solution might be not insured. For example, the differential equation

$$\begin{cases} \dot{x}(t) = \sqrt{x(t)} \\ x(0) = 0 \end{cases}$$

has infinite solutions. This happens because the considered function f is not Lipschitz. This initial value problem is therefore not a candidate to describe a dynamic system. The differential equation

$$\begin{cases} \dot{x}(t) = x(t)^2 \\ x(0) = 1 \end{cases}$$

has a unique solution but it is not defined for all $t > 0$. Hence, this is not a candidate to describe a dynamic system if we take as T the real axis.

3.3 Automata as dynamical systems

A class of dynamical systems of interest are the finite automata, where the sets X, U, Y are finite:

$$\begin{aligned} U &= \{u_1,\ u_2,\ \ldots,\ u_n\} \\ Y &= \{y_1,\ y_2,\ \ldots,\ y_p\} \\ X &= \{x_1,\ x_2,\ \ldots,\ x_m\} \end{aligned} \tag{3.8}$$

An automaton is typically represented by a function of the form

$$x(k+1) = F(k,\ x(k),\ u(k)) \tag{3.9}$$

that provides the state at the next time instant given the current time, the current state and the current input. In the case of time-invariant automata, the expression has the form

$$x(k+1) = F(x(k),\ u(k)). \tag{3.10}$$

There are basically two ways to effectively represent an automaton: a table or a graph that represent the function F.

The table is a $m \times n$ matrix, whose n columns are associated with the system states, and whose m rows with the inputs. Every element F_{ij} of this matrix, called **transition table**, represents the future state in the case the current automaton state is $x = j$ and the current input is $u = i$.

	x_1	x_2	$\dots$	x_n
u_1	F_{11}	F_{12}	$\dots$	F_{1n}
u_2	F_{21}	F_{22}	$\dots$	F_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
u_m	F_{m1}	F_{m2}	$\dots$	F_{mn}

The complete description of the automaton can be also given, as said, in terms of flow graph. Consider for example Figure 3.2, where $X = \{1, 2, 3, 4\}$, $U = \{1, 2, 3\}$, $Y = \{0, 1\}$. This graph corresponds to the table

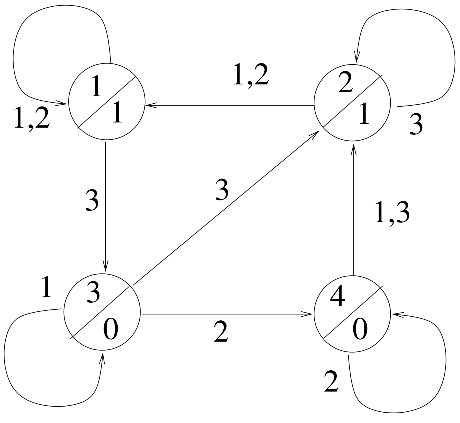

Figure 3.2: Flow graph for finite automata.

	x_1	x_2	x_3	x_n
u_1	1	1	3	2
u_2	1	1	4	4
u_m	3	2	2	2

The four possible states are represented as circles divided in two parts: the upper part shows the state index, the lower shows the output value associated with that state. Every input changes state of the system in accordance with the direction of the arcs. Each arch is associated with the value of input that causes the transition. It is important to note that, in this case, the output is linked to the current state only by an equation of the type $y = g(x)$, and then the system is proper. It is also possible to consider automata where the output depends on both the state and the input $y = g(x, u)$. The output of such a system can be represented by a matrix as well. The representation by means of flow graphs can be quite complex if the system has a high number of states.

Example 3.4. (Recognition of a string.) *Suppose we want to realise a finite state machine that recognises the string 358 within a longer string. The sets U, Y, X are:*

$$\begin{aligned} U &= \{0,\ 1,\ 2,\ \ldots,\ g\} \\ Y &= \{0,\ 1\} \\ X &= \{1,\ 2,\ 3,\ 4\} \end{aligned}$$

where the output is 0 *if the string is not recognised and* 1 *if the string is recognised. The flow graph is shown in Figure 3.3.*

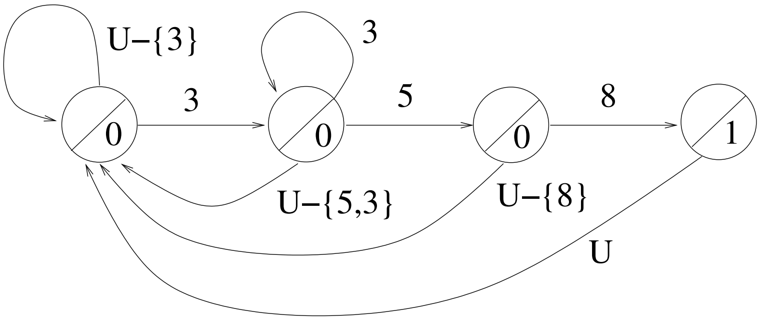

Figure 3.3: Example of automaton: recognition of a string.

Example 3.5. (Exercise: sum of two binary sequences.) *Suppose we want to calculate the sum of two sequences of bits that enter the machine, starting from the least significant digits. Sets U, Y, X are:*

$$\begin{aligned} U &= \left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\} \\ Y &= \{0,\ 1\} \\ X &= \{A,\ B,\ C,\ D\} \end{aligned}$$

As an exercise, derive the graph of the automaton.

Chapter 4

Reachability and Observability

The reachability and the observability problem are concerned with the relationships between input and state and between state and output, respectively. Hence, these two problems affect the input-output relationship. To understand the importance of these aspects, consider the following problem. Assume that matrices A (2×2), B (2×1) and C (1×2) are given and produce the input-output relation

$$y(s) = \frac{s-1}{(s+2)(s-1)}u(s) = \frac{1}{(s+2)}u(s)$$

expressed in terms of transfer function. The zero-pole cancellation reveals something pathological in the system. Although $\lambda = 1$ appears as a pole in the first expression, hence it is an eigenvalue of A, the cancellation implies that this pole does not affect the input-output behaviour. Indeed, the impulse response is

$$W(t) = e^{-2t}$$

At the end of the chapter we will understand that, in this situation, "something wrong" must occur in either the relation between input and state or the relation between state and output (or both).

4.1 Reachability of continuous-time systems

Consider a continuous-time linear autonomous system

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

In this section we will consider the following basic problem: given an initial state $\bar{x}_1$ and a target state $\bar{x}_2$, is there any choice of the input u that allows us to reach $\bar{x}_2$? If the answer is yes, we say that $\bar{x}_2$ is reachable from $\bar{x}_1$ and that $\bar{x}_1$ is controllable to $\bar{x}_2$. Formal definitions are given below for continuous-time systems. These definitions consider the reachability from zero and the controllability to zero. As we will see, they allow for a complete characterisation of the problem.

Definition 4.1. *The state $\bar{x}$ is **reachable** (from zero) in the time interval* $[0, \tau]$ *if* $\exists\, u(\cdot)$ *such that* $x(0) = 0 \;\Rightarrow\; x(\tau) = \bar{x}$.

Definition 4.2. *The state $\bar{x}$ is **controllable** (to zero) in the time interval* $[0, \tau]$ *se* $\exists\, u(\cdot)$ *such that* $x(0) = \bar{x} \;\Rightarrow\; x(\tau) = 0$.

Definition 4.3. *Given two states $\bar{x}_1$ and $\bar{x}_2$, we say that $\bar{x}_1$ is **controllable to** $\bar{x}_2$ and that $\bar{x}_2$ is **reachable from** $\bar{x}_1$ in the time interval* $[0, \tau]$ *if* $\exists\, u(\cdot)$ *such that* $x(0) = \bar{x}_1 \;\Rightarrow\; x(\tau) = \bar{x}_2$.

Note that, if $\bar{x}_1$ is controllable to zero and $\bar{x}_2$ is reachable from zero, then $\bar{x}_2$ is reachable from $\bar{x}_1$. Later, we will show that reachability and controllability properties are equivalent.

We now define the following set

$$X_r(\tau) = \{\bar{x} \in \mathbb{R}^n \ : \ \bar{x} \text{ is reachable (from 0) in } [0,\ \tau]\}$$

The set $X_r(\tau)$ is a subspace of $\mathbb{R}^n$. Indeed if $\bar{x}_1$ is reachable in $[0,\ \tau]$ with input $u_1(\cdot)$ and $\bar{x}_2$ is reachable in $[0,\ \tau]$ with input $u_2(\cdot)$, then $\alpha\bar{x}_1 + \beta\bar{x}_2$ is reachable in $[0,\ \tau]$ with input $\alpha u_1(\cdot) + \beta u_2(\cdot)$, given the linearity of the system.

This property can be verified by observing that the set of states that are reachable from zero can be expressed as follows:

$$\bar{x}(\tau) = 0 + \int_0^\tau e^{A(t-\sigma)} Bu(\sigma) d\sigma \tag{4.1}$$

Given the linearity of the integral, we notice that the whole set $X_r(\tau)$ is actually a subspace of $\mathbb{R}^n$.

The reachable states form a subspace

$$X_r(\tau) = \left\{ \bar{x} = \int_0^\tau e^{A(t-\sigma)} Bu(\sigma) d\sigma \right\}$$

In the same way we can define the set of controllable states

$$X_c(\tau) = \{\bar{x} \in \mathbb{R}^n \ : \bar{x} \text{ is controllable (to 0) in } \ [0,\ \tau]\}$$

The elements $\bar{x}$ of this set are such that:

$$0 = e^{A\tau}\bar{x} + \int_0^\tau e^{A(\tau-\sigma)} Bu(\sigma) d\sigma$$

Considering the invertibility of e^{At} (whose inverse is equal to e^{-At}) we have

$$\bar{x} = -e^{-A\tau} \int_0^\tau e^{A(\tau-\sigma)} Bu(\sigma) d\sigma = -\int_0^\tau e^{-A\sigma} Bu(\sigma) d\sigma \tag{4.2}$$

So we see that $X_c(\tau)$ is a subspace of $\mathbb{R}^n$ as well.

Let us now introduce the **reachability matrix** $\mathcal{R}$ as the following partitioned matrix

$$\mathcal{R} = \left[B | AB | A^2 B | \ldots | A^{n-1} B \right] \tag{4.3}$$

If B has dimension $n \times m$ and A has dimension $n \times n$, then $\mathcal{R}$ has dimension $n \times (n \cdot m)$. We remind that the **range** or **image of a matrix** M is the subspace $Ra(M) = \{y \ : \ y = Mx\}$. The following fundamental theorem holds.

Theorem 4.1. *$X_r(\tau) = X_c(\tau) = Ra(\mathcal{R})$ for each $\tau > 0$.*

Proof To prove the statement of the theorem, we equivalently demonstrate $X_r^\perp(\tau) = Ra(\mathcal{R})^\perp$.[1] Reachable states can be expressed as

$$\bar{x} = \int_0^\tau e^{A(\tau-\sigma)} Bu(\sigma) d\sigma$$

We have:

$$X_r^\perp(\tau) = \left\{ z : \ z^\top \int_0^\tau e^{A(\tau-\sigma)} Bu(\sigma) d\sigma = 0, \ \forall\, u(\cdot) \right\}$$

[1] This equivalence holds in finite-dimensional spaces in which the orthogonal of the orthogonal of X is $(X^\perp)^\perp = X$. This is not true in infinite-dimensional spaces.

We can characterise all of the elements of the orthogonal space using the following result:

$$\int_0^\tau z^\top e^{A(\tau-\sigma)}Bu(\sigma)d\sigma = 0\ \forall u(\cdot) \iff z^\top e^{A(\tau-\sigma)}B \equiv 0 \tag{4.4}$$

($\equiv$ means identically 0 for all $\sigma \in [0,\tau]$). The implication right to left is obvious. The implication left to right is not. Then denote with $W(\sigma) = z^\top e^{A(t-\sigma)}B$. Since the equality on the left is true for all inputs, can choose in particular $u(\sigma) = W^\top(\sigma)$, to get:

$$\int_0^\tau W(\sigma)u(\sigma)d\sigma = \int_0^\tau W(\sigma)W^\top(\sigma)d\sigma = \int_0^\tau \|W(\sigma)\|^2 d\sigma = 0 \tag{4.5}$$

Since the value $\|\cdot\|$ is always non negative and is zero only if the argument is the zero vector, we have that $W(\sigma) = z^\top e^{A(\tau-\sigma)}B = 0,\ \forall\, \sigma \in [0,\ \tau]$, the condition on the right.

Considering the expression of the matrix exponential, we can write:

$$z^\top \left[\sum_{k=0}^{+\infty} \frac{A^k(\tau-\sigma)^k}{k!}\right] B \equiv 0 \iff \sum_{k=0}^{+\infty} \frac{z^\top A^k B}{k!}(\tau-\sigma)^k \equiv 0 \tag{4.6}$$

In view of the principle of identity for power series this equality is true if and only if

$$z^\top A^k B = 0 \quad \forall k \geq 0 \tag{4.7}$$

From Cayley-Hamilton identity, the powers A^k with $k \geq n$ can be expressed as a linear combination of the first n powers of A. Therefore the previous condition holds true if and only if it is verified for $k = 0,\ 1,\ \ldots,\ n-1$. Putting these n zero terms together we have

$$\begin{aligned} \left[z^\top B | z^\top AB | z^\top A^2 B | \ldots | z^\top A^{n-1}B\right] &= 0 \\ \iff z^\top \left[B|AB|A^2B|\ldots|A^{n-1}B\right] &= 0 \\ \iff z^\top \mathcal{R} &= 0 \end{aligned} \tag{4.8}$$

This is equivalent to saying that any vector $z^\top$, belonging to the orthogonal subspace of $X_r(\tau)$, is orthogonal to the columns of $\mathcal{R}$, hence orthogonal to $Ra(\mathcal{R})$. On the other hand, all the previous implications are reversible and therefore any vector $z^\top$, belonging to the orthogonal subspace of $\mathcal{R}$, is orthogonal to $X_r(\tau)$. Hence $Ra(\mathcal{R})^\perp = X_r^\perp(\tau)$ and therefore $X_r(\tau) = Ra(\mathcal{R})$.

Along the same lines, using (4.2), we can arrive to the conclusion that $X_c(\tau) = Ra(\mathcal{R})$. □

Key points that emerge from the theorem are:

- X_r and X_c are equal;
- they are the image of $\mathcal{R}$;
- they are independent of τ.

The fact that reachability does not depends on the time horizon is suspect. This would mean, for example, that any state reachable by an aircraft can be reached in an arbitrarily small amount of time. This is obviously not true: it is in contradiction with the physical limits of the actuators in the plane. The apparent contradiction can be explained by the fact that the theorem does not take into account upper limits for the magnitude of the system input. If such limits are taken into account (for instance, $\|u\| \leq u_{max}$), then one can see that $X_r(\tau)$ is a bounded convex set that depends on τ.

Example 4.1. *Let*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and $|u| \leq 1$. These equations represent the motion of an object, where x_1 is the position, x_2 the speed and u a bounded force. The reachable set is the set of all states $[x_1(\tau)\ x_2(\tau)]$ with

$$x_1(\tau) = \int_0^\tau \sigma u(\sigma) d\sigma, \quad x_2(\tau) = \int_0^\tau u(\sigma) d\sigma,$$

It is not difficult to see that (as is physically obvious) this set is bounded if u is bounded and it gets larger as we increase τ.

We have just seen that the subspaces X_r and X_c do not depend on τ, and are the image of reachability matrix $\mathcal{R}$.

Definition 4.4. *A system is said* ***reachable (resp. controllable)*** *if all of the possible states are reachable (resp. controllable).*

In view of the previous theorem, since the reachable (controllable) space is the image of $\mathcal{R}$, we have the following.

Theorem 4.2. *The system is reachable (controllable) if and only if*

$$Rank(\mathcal{R}) = n \tag{4.9}$$

This means that we have reachability of the whole space: all possible state vector for the system can be generated by the columns of $\mathcal{R}$, namely, the columns of $\mathcal{R}$ are a set of generators for $\mathbb{R}^n$). This result is remarkable, because it transforms an analytical problem (finding the u to put in the integral) into an algebraic problem (easily solvable via standard software).

4.2 Kalman decomposition for reachability

If the system is not reachable, we can analyse the situation by means of the **Kalman Decomposition** for reachability.

Suppose that $\mathcal{R}$ has rank $r < n$; therefore, there are r linearly independent columns that form a basis of X_r. We place them in a matrix $T_r = [\bar{t}_1, \bar{t}_2, \dots, \bar{t}_r]$. To form a basis, consider now its complement $T_{nr} = [\bar{t}_{r+1}, \bar{t}_{r+2}, \dots, \bar{t}_n]$. Putting together the columns of T_r and T_{nr}, we get a matrix T of size $n \times n$,

$$T = [\bar{t}_1, \bar{t}_2, \dots, \bar{t}_r | \bar{t}_{r+1}, \bar{t}_{r+2}, \dots, \bar{t}_n] = [T_r | T_{nr}], \tag{4.10}$$

whose columns are a basis of $\mathbb{R}^n$. The space generated by T_r is the reachable subspace and the space generated by T_{nr} is the unreachable subspace. Notice that T_{nr} is any subspace that is complementary to T_r, so there is some freedom in this choice.

Each vector $x \in \mathbb{R}^n$ can then be expressed as a linear combination of the basis vectors, as follows

$$x = \left[\begin{array}{c|c} T_r & T_{nr} \end{array}\right] \left[\begin{array}{c} \hat{x}_r \\ \hline \hat{x}_{nr} \end{array}\right] \tag{4.11}$$

and, in particular, each vector belonging to the reachable subspace can be written as:

$$x = \left[\begin{array}{c|c} T_r & T_{nr} \end{array}\right] \left[\begin{array}{c} \hat{x}_r \\ \hline 0 \end{array}\right] \tag{4.12}$$

If we apply the state transformation T to the system we get the new representation

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

where

$$\begin{aligned}\hat{A} &= T^{-1}AT\\ \hat{B} &= T^{-1}B\\ \hat{C} &= CT\\ \hat{D} &= D\end{aligned}$$

We can write the system in the form

$$\frac{d}{dt}\begin{bmatrix}\hat{x}_r(t)\\ \hat{x}_{nr}(t)\end{bmatrix} = \left[\begin{array}{c|c}A_r & A_{r,nr}\\ \hline \phi_1 & A_{nr}\end{array}\right]\begin{bmatrix}\hat{x}_r(t)\\ \hat{x}_{nr}(t)\end{bmatrix} + \left[\begin{array}{c}B_r\\ \hline \phi_2\end{array}\right]u(t) \tag{4.13}$$

We demonstrate now that ϕ_1 and ϕ_2 are zero matrices. If we start from zero initial conditions, $\hat{x}_r(0) = 0$ and $\hat{x}_{nr}(0) = 0$, then after a time $t > 0$, by construction, only reachable states are reached. In the new representation, these are of the form

$$x = T\begin{bmatrix}\hat{x}_r(t)\\ 0\end{bmatrix}$$

This means that $\hat{x}_{nr}(t) = 0$ for all $t > 0$ and then $\dot{\hat{x}}_{nr}(t) = 0$ for all $t > 0$. Consider the second equation in (4.13)

$$\begin{aligned}\dot{\hat{x}}_{nr} &= \phi_1\hat{x}_r + A_{nr}\hat{x}_{nr} + \phi_2 u && (4.14)\\ \Rightarrow 0 &= \phi_1\hat{x}_r + \phi_2 u && (4.15)\end{aligned}$$

and write it in the form

$$\begin{bmatrix}\phi_1 & \phi_2\end{bmatrix}\begin{bmatrix}\hat{x}_r(t)\\ u(t)\end{bmatrix} = 0 \tag{4.16}$$

This relationship is valid $\forall\, u(t)$ and $\forall\, \hat{x}_r(t)$. Note that $\hat{x}_r(t)$ can be assumed arbitrary, because it is the reachable component and t is arbitrary. The only matrix which gives zero when multiplied by any vector is the zero matrix, hence ϕ_1 and ϕ_2 are zero. Then the system can be written in the form:

$$\begin{aligned}\frac{d}{dt}\begin{bmatrix}\hat{x}_r(t)\\ \hat{x}_{nr}(t)\end{bmatrix} &= \left[\begin{array}{c|c}A_r & A_{r,nr}\\ \hline 0 & A_{nr}\end{array}\right]\begin{bmatrix}\hat{x}_r(t)\\ \hat{x}_{nr}(t)\end{bmatrix} + \left[\begin{array}{c}B_r\\ \hline 0\end{array}\right]u(t)\\ Y(t) &= \left[\begin{array}{c|c}C_r & C_{nr}\end{array}\right]\begin{bmatrix}\hat{x}_r(t)\\ \hat{x}_{nr}(t)\end{bmatrix} && (4.17)\end{aligned}$$

The first subsystem is the reachable subsystem, the second the unreachable subsystem. Note that the unreachable subsystem influences the reachable one through the block $A_{r,nr}$.

The block diagram corresponding to the whole system is reported in Figure 4.1.

If a matrix is block-triangular, then its determinant is equal to the product of the determinants of the diagonal blocks:

$$\det(sI - \hat{A}) = \det(sI - A_r)\det(sI - A_{nr}) \tag{4.18}$$

The eigenvalues of $\hat{A}$, which are the same eigenvalues as A, are thus those of A_r and those of A_{nr}; more precisely, we can split the eigenvalues in two distinct sets:

- $\{\lambda_1, \lambda_2 \ldots, \lambda_r\} = \sigma(A_r)$ are the **reachable eigenvalues**, associated with **reachable modes**;
- $\{\lambda_{r+1}, \lambda_{r+2} \ldots, \lambda_n\} = \sigma(A_{nr})$ are the **unreachable eigenvalues**, associated with **unreachable modes**.

This result has a fundamental consequence when the system has to be controlled. The evolution of the second subsystem is completely independent of the input u and evolves in natural response with modes $\sigma(A_{nr})$. As we will see, a feedback controller may change the reachable eigenvalues, but not the unreachable ones (if any). Therefore, if the system has unstable unreachable modes, no controller will be able to stabilise it.

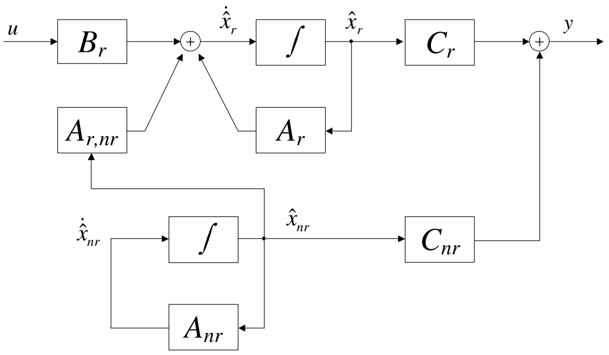

Figure 4.1: Kalman decomposition for reachability.

From the diagram in Figure 4.1, we can notice the absence of the interaction between the input and $\dot{\hat{x}}_{nr}$, which should be given by ϕ_2, and of the interaction between $\hat{x}_r$ and $\dot{\hat{x}}_{nr}$, which should be given by ϕ_1. The part consisting of (A_r, B_r, C_r) is called **reachable subsystem**, while that consisting of $(A_{nr}, 0, C_{nr})$ is called **unreachable subsystem**. A simplified block diagram is depicted in Figure 4.2.

Both in the block diagram reported in Figure 4.1 and in its simplified version in Figure 4.2, there is no arrow from u to the unreachable subsystem, in accordance with the definition of unreachable state. Also, there is no arrow from the reachable subsystem to the unreachable one: in fact, this would indirectly cause reachability of some states of the unreachable subsystem. Instead, no information can reach the unreachable subsystem transiting through the reachable one. However, note that the unreachable subsystem is able to influence the reachable subsystem (and the output of the system). This motivates us to study stability of the unreachable subsystem as well (as said, if the unreachable subsystem is not asymptotically stable, it is impossible to stabilise the system by means of a feedback controller).

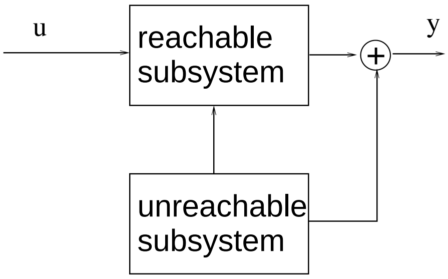

Figure 4.2: Simplified block diagram of the Kalman decomposition for reachability.

It can be shown that the transfer function depends on the reachable subsystem only

$$W(s) = C(sI - A)^{-1}B + D = C_r(sI - A_r)^{-1}B_r + D \tag{4.19}$$

because cancellations occur related to eigenvalues associated with the unreachable subsystem. This is obvious, because the transfer function represents the link between the system input and the system output when the initial conditions are zero and, by definition, the unreachable subsystem is not affected by the applied input.

Example 4.2. *Consider the system with the following matrices:*

$$A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix},\ B = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ C = \begin{bmatrix} 1 & 0 \end{bmatrix},\ D = [0]$$

The reachability matrix is equal to

$$\mathcal{R} = \left[\begin{array}{c|c} B & AB \end{array}\right] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

and its rank is 1 (hence, less than n $= 2$*), therefore the system is not completely reachable. The set of reachable vectors can be easily interpreted in a geometric fashion: it is bisector of the first and third quadrant of the Cartesian plane.*

We can complete the basis for $\mathbb{R}^2$ *by choosing, together with the only linearly independent column of* $\mathcal{R}$*, any vector that is not linearly dependent on that column. For example, we can choose:*

$$T = \left[\begin{array}{c|c} 1 & 1 \\ 1 & 0 \end{array}\right],$$

but also

$$T = \left[\begin{array}{c|c} 1 & 0 \\ 1 & 1 \end{array}\right], \quad or \quad T = \left[\begin{array}{c|c} 1 & -1 \\ 1 & 1 \end{array}\right]$$

are valid choices. Any vector of the state space, the plane $\mathbb{R}^2$*, can be expressed as a linear combination of the two vectors in* T*.*

Before applying T*, let us consider the system transfer function.*

$$\begin{aligned} d(s) &= \det(sI - A) = s^2 + 2s \\ n(s) &= \det\left[\begin{array}{c|c} sI - A & -B_j \\ \hline C_i & D_{ij} \end{array}\right] = \det\begin{bmatrix} s+1 & -1 & -1 \\ -1 & s+1 & -1 \\ 1 & 0 & 0 \end{bmatrix} = s + 2 \\ \Rightarrow W(s) &= \frac{s+2}{s(s+2)} = \frac{1}{s} \end{aligned}$$

We see that there is a cancellation of the factor $(s + 2)$*. We expect that this is due to the lack of reachability.*

We can easily verify that matrix $\hat{A} = T^{-1}AT$*, calculated with a* T *chosen as above, gives*

$$\hat{A} = \left[\begin{array}{c|c} 0 & \# \\ \hline \underline{0} & -2 \end{array}\right]$$

where the underlined $\underline{0}$ *is expected, and corresponds to the identically zero* ϕ_1 *matrix, while the* 0 *and* -2 *on the diagonal are the eigenvalues of matrix A. The element # in the upper right position depends on matrix T, and in particular on the choice of the second column. Similarly,* $\hat{B} = T^{-1}B$ *has the form*

$$\hat{B} = \left[\begin{array}{c} \$ \\ \hline \underline{0} \end{array}\right]$$

where, again, the zero $\underline{0}$ *is expected since it corresponds to the identically zero matrix* ϕ_2*, while* \$ *depends the chosen second column of T. The reader is invited to perform the computation with two different T matrices among those proposed above.*

4.3 Reachability of discrete-time systems

The previous definitions of reachability and controllability for continuous-time systems can be extended to the case of discrete systems

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

Definition 4.5. *A state $\bar{x}$ is said* ***reachable from zero*** *in an interval* $[0, T]$ *if* $\exists\, u(\cdot)$ *such that* $x(0) = 0 \;\Rightarrow\; x(T) = \bar{x}$.

Definition 4.6. *A state $\bar{x}$ is said* ***controllable to zero*** *in an interval* $[0, T]$ *if* $\exists\, u(\cdot)$ *such that* $x(0) = \bar{x} \;\Rightarrow\; x(T) = 0$.

There are strong analogies with the continuous-time case, but some important differences as well.

Consider the expression of the discrete time solution from 0 to k:

$$x(k) = A^k x(0) + \sum_{h=0}^{k-1} A^{k-h-1} B u(h)$$

If the state $\bar{x}$ is reached at time T from $x(0) = 0$ we have:

$$x(T) = \bar{x} = 0 + \sum_{h=0}^{T-1} A^{T-h-1} B u(h) \tag{4.20}$$

This expression can be written in matrix form:

$$\bar{x} = \left[\begin{array}{c|c|c|c} B & AB & \dots & A^{T-1}B \end{array}\right] \begin{bmatrix} u(T-1) \\ u(T-2) \\ \vdots \\ u(1) \\ u(0) \end{bmatrix} = \mathcal{R}_{(T)} U_{(T)} \tag{4.21}$$

Since vector $U_{(T)}$ can be arbitrary, the set of reachable states in T steps is given by:

$$X_r(T) = Ra(\mathcal{R}_{(T)}) \tag{4.22}$$

A first difference emerges between continuous-time systems and discrete-time systems: the reachability set, which is a subspace of the state space, depends on the time T, and gets bigger as T increases:

$$\begin{aligned} X_r(1) &= Ra\,([B]) \\ X_r(2) &= Ra\Big(\left[\begin{array}{c|c} B & AB \end{array}\right]\Big) \qquad (4.23) \\ &\;\vdots \qquad (4.24) \end{aligned}$$

Note that $X_r(T) \subseteq X_r(T+1)\ \forall\, T$. This is obvious, because if a state can be reached in T steps, then it is reachable in $T+1$ steps as well.[2]

The key point is that the reachability subspace of the system does not increase beyond $T = n$, where n is the system size. This happens because, by the identity of Cayley-Hamilton, the columns $A^k B$ with $k \geq n$ are a linear combination of the first n powers ($k = 0, 1, \dots, n-1$). In other words, matrix $\mathcal{R}_n$ has the maximum achievable rank, equal to n, and any reachable state can be reached in at most n steps. Hence, the range of matrix $\mathcal{R} = \mathcal{R}_n$ is the set of all reachable states.

[2] You can just do nothing for the first step and then use the remaining T steps to reach the state.

Theorem 4.3. *In a discrete-time system, the set of all reachable states is*

$$X_r = Ra(\mathcal{R}). \tag{4.25}$$

In this sense there is a strong analogy with the continuous-time case.

Example 4.3. ***(Inventory with ageing.)*** *Consider the system with the following matrices:*

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The quantity $u(k)$ is the supply of goods at time k, $x_3(k)$ is the quantity of goods arrived the previous day, $x_2(k)$ is the quantity of goods arrived two days before, $x_1(k)$ is the quantity of goods arrived three days before. Then, the goods are eliminated due to ageing (in $x_h(k)$, h represents the number of days for which the goods can survive). It is not difficult to see that the natural response $A^k x(0)$ goes to zero in finite time, as expected in a perishable goods inventory.

The set of states that are reachable in one step is

$$X_r(1) = Ra\left([B]\right) = Ra\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right)$$

and can be geometrically interpreted as the set of vectors $\mathbb{R}^3$ on the z axis. The set of states that can be reached in two steps is

$$X_r(2) = Ra\left(\begin{bmatrix} B \mid AB \end{bmatrix}\right) = Ra\left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\right)$$

and, geometrically, is the set of vectors in the yz plane. Finally, the set of states that are reachable in three steps is

$$X_r(3) = Ra\left(\begin{bmatrix} B \mid AB \mid A^2B \end{bmatrix}\right) = Ra\left(\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}\right)$$

namely the whole $\mathbb{R}^3$.

To summarise, in discrete time (A, B) is fully reachable if and only if $Ra(\mathcal{R}) = \mathbb{R}^n$, hence, if and only if $\text{rank}(\mathcal{R}) = n$. In addition, either a state can be reached in n or it cannot be reached at all.

The controllability problem in discrete time requires more caution. In fact, the set of controllable states, for a discrete-time system, is different from the set of reachable sets, as can be seen from the following example.

Example 4.4. *Given the system with matrices*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

the reachability matrix is equal to

$$\mathcal{R} = \begin{bmatrix} B \mid AB \end{bmatrix} = \left[\begin{array}{c|c} 0 & 0 \\ 0 & 0 \end{array}\right],$$

from which it is clear that the set of reachable states is the null set (only 0 *can be reached). However, the set of controllable states is equal to all* $\mathbb{R}_2$*. Indeed, for* $u(k) = 0$*,*

$$x(k) = A^k \bar{x} = 0 \quad \text{for } k \geq 2 \quad \text{because } A^k = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{for } k \geq 2.$$

Hence, any $x(0)$ *can be driven to zero in at most* 2 *steps. Matrix A in this example is* ***nilpotent****.*[3]

The above example highlights the discrepancy between the sets X_r and X_c in discrete time. In general, in fact, we have

$$X_r \subseteq X_c.$$

It can be proved that, if A is invertible (*i.e.*, it has no null eigenvalues), then $X_r = X_c$ in the discrete-time case as well.

There is no a simple formula to express the set of all controllable states in T steps. This set is the subspace of all vectors $\bar{x}$ for which

$$0 = A^T \bar{x} + \sum_{h=0}^{T-1} A^{T-h-1} B u(h)$$

for some choice of u. If we take $T = n$, this set is the kernel of

$$\begin{bmatrix} A^n & A^{n-1}B & A^{n-2}B \ldots AB & B \end{bmatrix}$$

projected on the subspace of the first n components.

4.4 Other reachability criteria

We return now to the case of continuous-time systems.

Theorem 4.4. *The following implications are equivalent*

- *System* (A, B) *is reachable (or, equivalently, controllable)*
- *Rank* $\begin{bmatrix} \lambda I - A \mid B \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}$
- *Rank* $\begin{bmatrix} \lambda I - A \mid B \end{bmatrix} = n \quad \forall \lambda \in \sigma(A)$

Proof. It is very easy prove that the third condition is equivalent to the second. Obviously the second implies the third. Now, if the third holds, then the rank condition holds even for $\lambda \notin \sigma(A)$, because in this case the matrix $[\lambda I - A \; B]$ has the full block $\lambda I - A$ having full rank.

We prove by contradiction that the first statement implies the third: if the third statement fails, the first must fail as well.[4] Assume that Rank$\begin{bmatrix} \lambda I - A \mid B \end{bmatrix} < n$, namely, the n rows of this matrix are not linearly independent. Therefore there exists a vector $z^\top \neq 0$ such that:

$$z^\top \begin{bmatrix} \lambda I - A \mid B \end{bmatrix} = 0$$

By the properties of the block matrix, we can therefore write

$$\begin{cases} z^\top (\lambda I - A) = 0 \\ z^\top B = 0 \end{cases}$$

[3] Matrix A is nilpotent if $A^k = 0$ for some k. This is true if and only if all of its eigenvalues are zero.

[4] $\{A\} \Rightarrow \{B\}$ if and only if not$\{B\} \Rightarrow$ not$\{A\}$.

and then it appears that $z^\top$ is the left eigenvector of matrix A associated with the eigenvalue λ. We can write

$$\begin{array}{rcl} z^\top A & = & \lambda z^\top \\ z^\top A^2 & = & z^\top A \cdot A = \lambda z^\top A = \lambda^2 z^\top \\ z^\top A^3 & = & z^\top A \cdot A^2 = \lambda z^\top A^2 = \lambda^3 z^\top \\ & \vdots & \\ z^\top A^k & = & \lambda^k z^\top \end{array}$$

and also

$$z^\top A^k B = \lambda^k z^\top B = 0$$

Therefore,

$$z^\top \left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array} \right] = \left[\begin{array}{c|c|c|c} z^\top B & z^\top AB & \dots & z^\top A^{n-1}B \end{array} \right] = 0,$$

hence the reachability matrix $\mathcal{R}$ does not have full rank n, and the system is not reachable.

To prove that the third condition implies the first, again we proceed by contradiction. Assume that the system is not reachable and write it in Kalman form. This can be done, since it can be easily seen that Rank$\left[\begin{array}{c|c} \lambda I - A & B \end{array} \right]$ = Rank$\left[\begin{array}{c|c} \lambda I - \hat{A} & \hat{B} \end{array} \right]$. We get the matrix

$$\left[\begin{array}{c|c} \lambda I - \hat{A} & \hat{B} \end{array} \right] = \left[\begin{array}{c|c|c} \lambda I - A_r & -A_{r,nr} & B_r \\ \hline 0 & \lambda I - A_{nr} & 0 \end{array} \right]$$

So, if the system is not reachable, then the last block-row only has the square matrix $[\lambda I - A_{nr}]$ of dimension $n - r$ that is not identically zero. Then, if we take $\lambda \in \sigma(A)$ among the unreachable eigenvalues, matrix $[\lambda I - A_{nr}]$ is singular and so the last $n \times r$ rows are linearly dependent. Therefore, the rank must be less than n. □

From the previous proof, we have the **Popov criterion**: Rank$\left[\begin{array}{c|c} \lambda_{nr} I - A & B \end{array} \right] < n$ for all the unreachable eigenvalues $\lambda_{nr} \in \sigma(A)$. This criterion has several applications. It is very useful to test if an eigenvalue λ (and the corresponding mode) is reachable or not.

For instance, assume that matrix A has an eigenvalue $\bar{\lambda}$ with positive or nonnegative real part. Consider $\left[\begin{array}{c|c} \bar{\lambda} I - A & B \end{array} \right]$ and observe that:

- if its rank is less than n, this eigenvalue $\bar{\lambda}$ is unreachable, hence there is no way to design a feedback controller that asymptotically stabilises the system;
- if its rank is equal to n, this eigenvalue is reachable, and therefore we can hope to realise the controller (we will deal with this problem in one of the following chapters).

The corresponding discrete-time theorem is the following.

Theorem 4.5. *The discrete system* $(A,\ B)$ *is*

- *reachable if and only Rank*$\left[\begin{array}{c|c} \lambda I - A & B \end{array} \right] = n \quad \forall\, \lambda \in \sigma(A)$*;*
- *controllable if and only if Rank*$\left[\begin{array}{c|c} \lambda I - A & B \end{array} \right] = n \quad \forall\, \lambda \in \sigma(A),\ \lambda \neq 0.$

Example 4.5. *We want to find the set of all reachable states for the system with matrices*

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

The reachability matrix is:

$$\mathcal{R} = \begin{bmatrix} -1 & -1 & -1 \\ 0 & -3 & 3 \\ 2 & -1 & 5 \end{bmatrix}$$

To compute the rank of this matrix, we can use the Gauss method for triangulation. By adopting elementary operations that do not change the matrix rank, we can reduce the matrix to a triangular form. In this case, multiplying the first row by 2 *and adding it to the last row, we have:*

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

and multiplying the second row by -1 *and adding it to the last we get*

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

The number of rows that are not identically zero represents the rank of matrix. In this case, it is equal to 2*. A basis of the reachable subspace is then obtained by considering two linearly independent columns of the original matrix* $\mathcal{R}$*. We can take, for instance,*

$$X_r = \left\{ \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} -1 \\ -3 \\ -1 \end{bmatrix} \right\}$$

This selection may be completed with an additional column that must be linearly independent to the other two, so as to form a basis of $\mathbb{R}^3$*:*

$$T = \left[\begin{array}{cc|c} 2 & -1 & 1 \\ 0 & -3 & 0 \\ -1 & -1 & 0 \end{array} \right]$$

This matrix is invertible and we have:

$$T^{-1} = \frac{1}{-3} \begin{bmatrix} 0 & -1 & 3 \\ 0 & 1 & 0 \\ -3 & 3 & -6 \end{bmatrix}$$

We can easily check that

$$\hat{A} = T^{-1}AT = \left[\begin{array}{cc|c} 0 & 2 & 1 \\ 1 & -1 & 1 \\ \hline \underline{0} & \underline{0} & -1 \end{array} \right], \quad \hat{B} = T^{-1}B = \left[\begin{array}{c} 1 \\ 0 \\ \hline \underline{0} \end{array} \right]$$

The underlined zeros are not a coincidence: they result from the Kalman decomposition. In fact, since the reachability matrix has rank 2*, there must be a matrix* A_{nr} *of size* 1×1*, and also a row of zeros in the first two positions of the last row of* $\hat{A}$*.*

As mentioned earlier, the eigenvalues of $\hat{A}$ *are same as those of* A *and are equal to the union of the eigenvalues of* A_r *and those of* A_{nr}*:*

$$\begin{aligned} \det(sI - A) &= \begin{bmatrix} s & -2 & -1 \\ -1 & s+1 & -1 \\ 0 & 0 & s+1 \end{bmatrix} = \\ &= (s+1)(s(s+1)-2) = (s+1)(s^2+s-2) \end{aligned}$$

In this case the eigenvalues are $\lambda_1 = 1$, $\lambda_2 = -2$, $\lambda_3 = -1$; the eigenvalue $\lambda_3 = -1$, which is asymptotically stable, is associated with the unreachable part of the system. If we had known the eigenvalues and wanted to check the feasibility of a feedback controller to be applied to system, we could just have checked the reachability of the eigenvalue λ_1, which is responsible of the system instability. To this aim, we can use the Popov criterion explained above. According to this criterion, to test reachability of the unstable eigenvalue, the matrix $\left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]$ must have rank n for $\lambda = 1$. We have:

$$\left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]\Big|_{\lambda=1} = \left[\begin{array}{ccc|c} 1 & 2 & -1 & 2 \\ 3 & 0 & 3 & 0 \\ 2 & -2 & 4 & -1 \end{array}\right]$$

We use again Gauss elimination to triangularise the system. Multiplying the first row by -2 and adding it to the last row, and multiplying the first row by -3 and adding it to the second, we obtain

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 2 \\ 0 & -6 & 6 & -6 \\ 0 & -6 & 6 & -5 \end{array}\right]$$

Then, multiplying the second row by -1 and adding it to the last one, we get

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 2 \\ 0 & -6 & 6 & -6 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

Finally, the matrix rank is equal to $3 = n$, hence the eigenvalue $\lambda_1 = 1$ is reachable. Therefore we expect that, in principle, there is a controller able to stabilise the system.

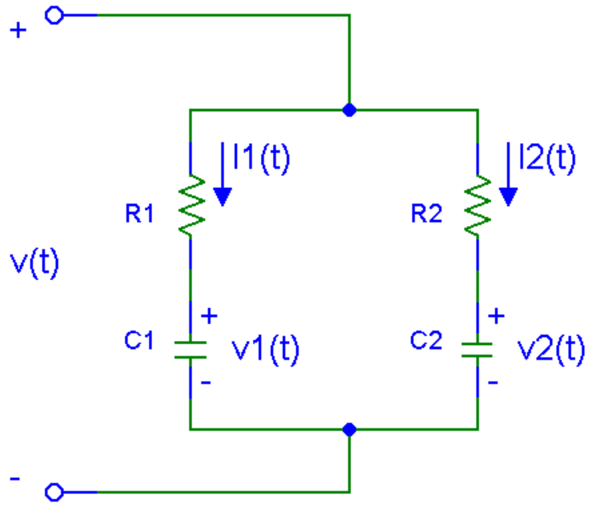

Figure 4.3: An electric circuit (example for discussing reachability).

Example 4.6. *The circuit in Figure 4.3 is governed by the following equations:*

$$\begin{aligned} v(t) &= R_1 I_1(t) + v_1(t) = R_1 C_1 \dot{v}_1(t) + v_1(t) \\ v(t) &= R_2 I_2(t) + v_2(t) = R_2 C_2 \dot{v}_2(t) + v_2(t) \end{aligned}$$

Set $v(t) = u(t)$ and consider the state vector:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix}$$

The system can be written in the following matrix form

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -\frac{1}{R_1C_1} & 0 \\ 0 & -\frac{1}{R_2C_2} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{R_1C_1} \\ \frac{1}{R_2C_2} \end{bmatrix} u(t)$$

The reachability matrix is

$$\mathcal{R} = \begin{bmatrix} \frac{1}{R_1C_1} & -\frac{1}{(R_1C_1)^2} \\ \frac{1}{R_2C_2} & -\frac{1}{(R_2C_2)^2} \end{bmatrix}$$

In order to have full rank (equal to 2*), matrix* $\mathcal{R}$ *must have a non-zero determinant*

$$\begin{aligned} \det(\mathcal{R}) &= -\frac{1}{R_1C_1}\frac{1}{(R_2C_2)^2} + \frac{1}{R_2C_2}\frac{1}{(R_1C_1)^2} = \\ &= \frac{1}{R_1C_1}\frac{1}{R_2C_2}\left(\frac{1}{R_1C_1} - \frac{1}{R_2C_2}\right) \neq 0, \end{aligned}$$

hence $R_1C_1 \neq R_2C_2$*. This condition means that, if the circuit presents a symmetry (*$R_1C_1 = R_2C_2$*), it is not possible to independently control both of the capacitor voltages. This situation is very similar to the case of two identical tanks with identical drainpipes, and filled by a single tap that supplies both of the tanks with the same amount of water. It is intuitive that, if the tanks start from the same level, by symmetry, we cannot unbalance the level of one tank relatively to the other. If, instead, the tanks are different, with an appropriate flow input function one can bring both tanks at the desired levels at some time* τ*.*

The situation of the previous example can be generalised. Consider two systems (A_1, B_1), (A_2, B_2), where B_1 and B_2 are columns (single input), and consider the parallel of the two. The parallel system equation is

$$\left[\begin{array}{c} \dot{x}_1(t) \\ \hline \dot{x}_2(t) \end{array}\right] = \left[\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array}\right] \left[\begin{array}{c} x_1(t) \\ \hline x_2(t) \end{array}\right] + \left[\begin{array}{c} B_1 \\ \hline B_2 \end{array}\right] u(t) \tag{4.26}$$

We can prove the following:

Theorem 4.6. *Consider two systems with a single input variable. The parallel system is reachable if and only if the two systems are reachable and they have no common eigenvalues.*

Proof. (⇒) Assume that A_1 has size $n_1 \times n_1$ and A_2 has size $n_2 \times n_2$. If the parallel system is reachable, then the matrix

$$\left[\begin{array}{cc|c} \lambda I - A_1 & 0 & B_1 \\ 0 & \lambda I - A_2 & B_2 \end{array}\right]$$

has full rank equal to $n_1 + n_2$, according to the Popov criterion. This means that the number of linearly independent rows of the matrix is $n_1 + n_2$. So, if we take the submatrix

$$\left[\begin{array}{cc|c} \lambda I - A_1 & 0 & B_1 \end{array}\right]$$

it must have n_1 linearly independent rows (and the same number of linearly independent columns). Removing the 0 column, we have that

$$\left[\begin{array}{c|c} \lambda I - A_1 & B_1 \end{array}\right]$$

has rank n_1, which implies the reachability of system (A_1, B_1). The same reasoning applies to the system (A_2, B_2).

It is also necessary that (A_1, B_1) and (A_2, B_2) have no common eigenvalues. Indeed, if there were a common eigenvalue λ, the matrices

$$\begin{bmatrix} \lambda I - A_1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ \lambda I - A_2 \end{bmatrix}$$

would have rank less than n_1 and than n_2, respectively. This means that, in the most favourable case, the sub matrix

$$\begin{bmatrix} \lambda I - A_1 & 0 \\ 0 & \lambda I - A_2 \end{bmatrix}$$

would have at most $n_1 + n_2 - 2 = n - 2$ linearly independent columns, so that the addition of the column

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

would bring the rank at most to to $n_1 + n_2 - 1$. As a consequence,the parallel system would be unreachable, in contradiction with the hypothesis.

($\Leftarrow$) Assume that the two systems are reachable and do not have common eigenvalues. Take λ eigenvalue of A_1. Then,

$$\left[\begin{array}{c|c} \lambda I - A_1 & B_1 \end{array}\right]$$

has rank n_1 because system 1 is reachable. On the other hand, matrix

$$\begin{bmatrix} \lambda I - A_2 \end{bmatrix}$$

has rank n_2, because λ is not an eigenvalue of A_2; hence, matrix

$$\left[\begin{array}{cc|c} \lambda I - A_1 & 0 & B_1 \\ 0 & \lambda I - A_2 & B_2 \end{array}\right]$$

has full rank $n_1 + n_2 = n$, because, in view of how the blocks are arranged, the first n_1 rows are independent and none of them can be generated as a combination of the last n_2, and vice versa. A similar reasoning applies to any eigenvalue of A_2; therefore, the Popov criterion is satisfied. □

4.5 Observability and detectability of continuous-time systems

The problem of observability amounts to studying the relationship between the system state and the system output. Consider a linear autonomous continuous-time system with matrices A, B, C, D. The assumption is that the signals $y(t)$, $u(t)$ are known in an interval $[0, \tau]$. Is it possible to determine the initial state of the system at time $t = 0$? Is it possible to determine the final state of the system at time $t = \tau$?

Definition 4.7. *The system is* ***observable*** *in the interval* $[0, \tau]$ *if, given known* $u(t)$ *and* $y(t)$ *for* $t \in [0, \tau]$*, it is possible to* ***uniquely*** *determine* $x(0)$*.*

Definition 4.8. *The system is* ***detectable*** *in the interval* $[0, \tau]$ *if, given known* $u(t)$ *and* $y(t)$ *for* $t \in [0, \tau]$*, it is possible to* ***uniquely*** *determine* $x(\tau)$*.*

Before proceeding further, we set matrix D to zero. In fact, $y(t) = Cx(t) + Du(t)$ reads as $y(t) - Du(t) = Cx(t)$, so we can consider the new "output" $\tilde{y}(t) = y(t) - Du(t)$. Hence, it is not restrictive to choose $D = 0$.

A first question is whether the problem of observability can be solved by adopting the punctual relationships at time $t = 0$:

$$y(0) = Cx(0).$$

The answer is clearly negative, because, most of the times, matrix C is not square and therefore not invertible, therefore it is impossible to uniquely determine $x(0)$ from the condition $y(0) = Cx(0)$. Then, we need more information and we have to observe the system evolution in an interval.

The system general solution is

$$\begin{aligned} x(t) &= e^{At}x(0) + \int_0^t e^{A(t-\sigma)}Bu(\sigma)d\sigma \\ y(t) &= Ce^{At}x(0) + \int_0^t Ce^{A(t-\sigma)}Bu(\sigma)d\sigma \end{aligned}$$

From this expression we can deduce the following fact.

Observation 4.1. *Since the continuous-time system is reversible, if we can find $x(0)$, then we can calculate $x(\tau)$, and vice versa. Therefore, in the case of continuous-time systems, the observability problem is equivalent to the detectability problem.*

Note also that the observability problem is conceptually different from the reachability problem discussed earlier. The reachability problem mainly concerns the *existence* of an input that allows us to reach a certain state. Here, the existence is a certain fact, because the system has indeed been in some initial state $x(0)$. The issue now is *uniqueness*. In fact, given the same input and observed output, there can be several initial states that are consistent with the observation.

To analyse this problem, we define

$$g(t) = \int_0^t Ce^{A(t-\sigma)}Bu(\sigma)d\sigma$$

and we discuss the uniqueness of the solution $x(0)$ of the equation

$$y(t) = Ce^{At}x(0) + g(t),$$

where $y(t)$ and $g(t)$ are known on the observation interval $[0, \tau]$.

Suppose that there is no uniqueness, namely there are two different states $\bar{x}_1$ and $\bar{x}_2$ such that:

$$y(t) = Ce^{At}\bar{x}_1 + g(t) \tag{4.27}$$

$$y(t) = Ce^{At}\bar{x}_2 + g(t) \tag{4.28}$$

that is, two different initial states that produce the same output when the same input is applied. Subtracting the first equation from the second and denoting by $\bar{x} = \bar{x}_2 - \bar{x}_1$ we obtain:

$$0 = Ce^{At}\bar{x} \tag{4.29}$$

Then uniqueness is ensured if there are no vectors $\bar{x} \neq 0$ which satisfy (4.29). Indeed if (4.27) and (4.28) are true, then $\bar{x} = \bar{x}_2 - \bar{x}_1 \neq 0$ is one of such vectors. The opposite implication is quickly verified: assume that $\bar{x} \neq 0$ satisfies (4.29). If one considers the equations:

$$\begin{aligned} 0 &= Ce^{At}\bar{x} \\ y(t) &= Ce^{At}x(0) + g(t) \end{aligned} \tag{4.30}$$

then their sum is

$$y(t) = Ce^{At}(x(0) + \bar{x}) + g(t) \tag{4.31}$$

and both the vectors $x(0)$ and $x(0) + \bar{x}$ produce the same output for any g: then, uniqueness is missing. We conclude the following.

Observation 4.2. *The problem of observability is solvable if and only if there are no vectors $\bar{x} \neq 0$ such that $Ce^{At}\bar{x} = 0$ for all $t \in [0, \tau]$.*

Definition 4.9. *A vector* $\bar{x} \in \mathbb{R}^n$, $\bar{x} \neq 0$ *is called* ***not observable*** *(indistinguishable from zero) in the interval* $[0, \tau]$ *if* $Ce^{At}\bar{x} = 0 \; \forall \, t \in [0, \tau]$.

We have established that the problem has no solution if and only if there are non-observable vectors. From equation (4.29), it is easy to realise that any of these vectors, if chosen as an initial condition with zero input $u(t) = 0$, produces a zero output. The set of non-observable vectors X_{no} is a subspace of the state space, because it is the kernel of Ce^{At}, and any kernel is a subspace. In the light of this definition, the observability problem has solution if and only if such a kernel is trivial, namely, $X_{no} = \{0\}$.

The expression $Ce^{At}\bar{x}$ can be expanded using the exponential series

$$Ce^{At}\bar{x} = C\left(\sum_{k=0}^{+\infty} \frac{A^k t^k}{k!}\right)\bar{x} = \sum_{k=0}^{+\infty} \frac{CA^k\bar{x}}{k!}t^k = 0 \quad \forall \, t \in [0, \tau]$$

Using once again the principle of power identity, this is equivalent to

$$CA^k\bar{x} = 0 \quad \forall \, k \geq 0 \quad \Leftrightarrow \quad CA^k\bar{x} = 0 \quad \text{for } k = 0, 1, \ldots, n-1$$

in view of the Cayley-Hamilton theorem. Similarly to what was done for the reachability problem, we can incorporate all of these terms in a matrix:

$$\begin{cases} C\bar{x} = 0 \\ CA\bar{x} = 0 \\ \vdots \\ CA^{n-1}\bar{x} = 0 \end{cases} \Leftrightarrow \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \bar{x} = 0 \quad \Leftrightarrow \quad O\bar{x} = 0$$

where the matrix

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = O \tag{4.32}$$

is the **observability matrix**. Hence, the set of all vectors such that $Ce^{At}\bar{x} = 0$ in $[0, \tau]$ is the kernel of O. We conclude the following.

Theorem 4.7. *The system is observable if and only if* $\ker(O) = \{0\}$.

Matrix O has dimension $(n \cdot p) \times n$, and it is square if $p = 1$ (single output). If $\ker(O) = 0$, then it must be $\text{rank}(O) = n$. In fact, if it happened that $O\bar{x} = 0$ with $\bar{x} \neq 0$, this would mean that the columns of O are linearly dependent, namely, that $\text{rank}(O) < n$. Therefore we have the following.

Theorem 4.8. *The system is observable if and only if rank*$(O) = n$.

Once again, an analytical problem has been turned into an algebraic problem. Note the following facts:

- matrix B has no role in determining the observability of the system;
- the horizon length τ does not influence observability.

The previous considerations are somewhat analogous to what was found for the reachability problem. In particular, the second fact means that, if the initial state of a system can be determined, this is possible in an arbitrarily small observation period. This has to do with the ideal hypothesis of the theorem, and not with real-world situations. Each measure $y(t)$, in fact, is certainly affected by disturbances that are not always exactly quantifiable. As a result, the smaller is the observation period, the higher is the influence of the error. In reality, therefore, to determine the state we need to filter the measured signal and this requires the observation over a period of time that is not infinitesimal.

4.6 Kalman decomposition for observability

Analogously to the reachability problem case, if $\ker(\mathcal{O}) \neq \{0\}$ (the system is not observable), then we can take the vectors forming a basis of $\ker(\mathcal{O})$ and group them in a matrix

$$T_{no} = [\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_r].$$

It is then possible to complete this basis with other $n - r$ linearly independent vectors

$$T_o = [\bar{t}_{r+1}, \bar{t}_{r+2}, \ldots, \bar{t}_n]$$

so as to get a basis of $\mathbb{R}^n$, given by the columns of

$$T = \left[\begin{array}{c|c} T_{no} & T_o \end{array}\right]. \tag{4.33}$$

Each vector in $\mathbb{R}^n$ can thus be expressed as:

$$x = \left[\begin{array}{c|c} T_{no} & T_o \end{array}\right] \left[\begin{array}{c} \hat{x}_{no} \\ \hline \hat{x}_o \end{array}\right] \tag{4.34}$$

In particular, each vector that is not observable can be expressed as

$$x = \left[\begin{array}{c|c} T_{no} & T_o \end{array}\right] \left[\begin{array}{c} \hat{x}_{no} \\ \hline 0 \end{array}\right], \tag{4.35}$$

from which it follows that

$$X_{no} = \left\{ x \,:\, x = \left[\begin{array}{c|c} T_{no} & T_o \end{array}\right] \left[\begin{array}{c} \hat{x}_{no} \\ \hline 0 \end{array}\right],\ \hat{x}_{no} \text{ arbitrary} \right\}.$$

We denote the subspace generated by T_o as the **observable** subspace. This nomenclature is purely conventional, because, if there exists unobservable vectors, the entire observability problem has no solution.

Note the duality with respect to the reachability problem: in that case, the basis of the reachability subspace is obtained by selecting the linearly independent columns of $\mathcal{R}$ and then the basis is completed (to get a basis of the whole state space) by adding new vectors that generate the "non-reachable" subspace (if the system is reachable, $\mathcal{R}$ is a basis of $\mathbb{R}^n$). Here, the "non-observable" subspace is determined as the kernel of $\mathcal{O}$ and the basis is completed (to get a basis of the whole state space) with vectors that generate the observable subspace.

If we transform the system by means of T, we get

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} \hat{x}_{no}(t) \\ \hat{x}_o(t) \end{bmatrix} &= \left[\begin{array}{c|c} A_{no} & A_{no,o} \\ \hline \phi_1 & A_o \end{array}\right] \begin{bmatrix} \hat{x}_{no}(t) \\ \hat{x}_o(t) \end{bmatrix} + \begin{bmatrix} B_{no} \\ B_o \end{bmatrix} u(t) \\ Y(t) &= \left[\begin{array}{c|c} \phi_2 & C_o \end{array}\right] \begin{bmatrix} \hat{x}_{no}(t) \\ \hat{x}_o(t) \end{bmatrix} \end{aligned} \tag{4.36}$$

where the matrices ϕ_1 and ϕ_2 are identically zero. To see this, consider $u(t) = 0$ and an unobservable initial vector equal to

$$x(0) = \begin{bmatrix} \hat{x}_{no}(t) \\ 0 \end{bmatrix}$$

By definition of unobservable vector, $y(t) = 0\ \forall\, t > 0$, and this is possible only if ϕ_2 is a null matrix. Moreover, if ϕ_1 were not zero in the same conditions we would have that, at some t, $\hat{x}_o(t) \neq 0$, and this would produce a nonzero output.

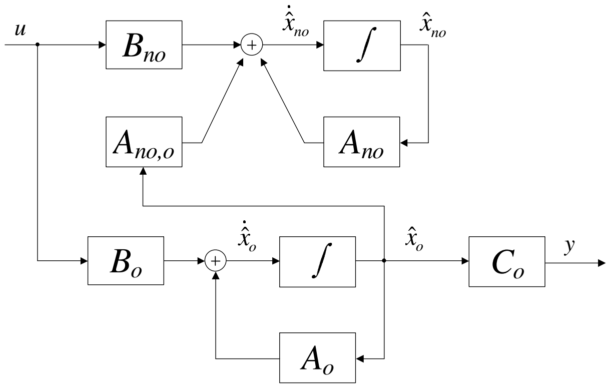

Figure 4.4: Kalman decomposition for observability.

The structural scheme of the system is shown in Figure 4.4.

The part consisting of (A_o, B_o, C_o) is called **observable subsystem**, while that composed of $(A_{no}, B_{no}, 0)$ is called **non-observable subsystem**. The observable subsystem is, as mentioned, the only one that is able to influence the output. Indeed, the unobservable subsystem does not affect the output y (this would be the contribution of submatrix ϕ_2, which is zero). Moreover, the unobservable subsystem states do not influence observable states (this would be the contribution of ϕ_1, which is zero), because otherwise we would have an indirect influence of the unobservable states on the output, through the observable subsystem.

The spectrum of matrix $\hat{A}$, obtained after the Kalman decomposition, includes the eigenvalues of A_{no} and those of A_o, because $\hat{A}$ is a block-triangular matrix. As in the reachability case, observability leads to cancellations. Indeed, the transfer function matrix depends on the observable subsystem only

$$W(s) = C(sI - A)^{-1}B = C_o(sI - A_o)^{-1}B_o \tag{4.37}$$

and the unobservable subsystem has no role in the input-output relationship.

4.7 Observability of discrete-time system

The observability problem for discrete-time systems is stated exactly as in the continuous-time case. The output of the system is now

$$y(k) = CA^k x(0) + \sum_{h=0}^{k-1} CA^{k-h-1}Bu(h) = CA^k x(0) + g(k)$$

and, assuming to perform T observations at different times. with the purpose of uniquely determining $x(0)$, we obtain the following equations:

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(T-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{T-1} \end{bmatrix} x(0) + \begin{bmatrix} g(0) \\ g(1) \\ \vdots \\ g(T-1) \end{bmatrix} \tag{4.38}$$

In a more compact form, we can write:

$$Y(T) = O_{(T)}x(0) + G(T), \tag{4.39}$$

which represents an observation experiment in T steps. It is intuitive that, if T is large, then we get more information on the system and more constraints on $x(0)$, hence we are more likely to determine it uniquely. It may happen that we do not obtain a unique solution in $T - k$ steps, while uniqueness holds in $T - k + 1$ steps. So, the question is how many steps are needed. For sure it is not necessary to continue for $T > n$, because the k-th power A^k of A is a linear combination of I, A, $\dots A^{n-1}$. Another interpretation is that, when we increase the number of steps, we try to increase the rank of matrix $\mathcal{O}$, in order to decrease its kernel dimension, with the hope of reducing it to the trivial subspace. This is hopeless after n steps. In conclusion, the problem of observability either has solution in n steps or has no solution. Therefore, we can consider n steps and the matrix $\mathcal{O} = \mathcal{O}_{(T)}$, and conclude that the problem has solution if and only if $\mathcal{O}$ has trivial kernel or, equivalently, has rank n. The following theorem holds.

Theorem 4.9. *The system is observable if and only if rank$(\mathcal{O}) = n$.*

The condition is thus the same obtained for continuous-time systems.

Analogously to reachability and controllability, also observability and detectability are not equivalent in general for discrete-time systems. They are equivalent if the system is reversible, *i.e.*, the matrix A is invertible, because in this case for any initial state we can determine the final one, and vice-versa.

Example 4.7. *Consider the following discrete-time system:*

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

$$y(k) = \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

Since we get no information from $y(k) = 0$, it is impossible to derive $x_1(0)$ and $x_2(0)$. However, A is a nilpotent matrix. Hence, for $k > 2$, the state vector is zero and the problem of detectability has a solution. From this simple case, we observe that the detectability problem is different from the observability problem for discrete-time systems.

4.8 Examples of reachability and observability analysis

Example 4.8. *Consider the following model of a direct-current electrical machine:*

$$\frac{d}{dt}\begin{bmatrix} i_a(t) \\ \omega(t) \\ \varphi(t) \end{bmatrix} = \begin{bmatrix} -\frac{R_a}{L_a} & -\frac{k\bar{i}_e}{L_a} & 0 \\ \frac{k\bar{i}_e}{J} & -\frac{f}{J} & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} i_a(t) \\ \omega(t) \\ \varphi(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{L_a} \\ 0 \\ 0 \end{bmatrix} v_a(t)$$

For simplicity, the external load torque is assumed to be zero. We first perform a reachability analysis. We have a single input u. The reachability matrix is

$$\mathcal{R} = \begin{bmatrix} \frac{1}{L_a} & -\frac{R_a}{L_a^2} & * \\ 0 & \frac{k\bar{i}_e}{JL_a} & * \\ 0 & 0 & \frac{k\bar{i}_e}{JL_a} \end{bmatrix}$$

where the determination of the entries $(*)$ is left as an exercise. This matrix is upper triangular, with non-zero diagonal elements, hence it has full rank and the system is reachable.

This systems can be used as an electric drive, but to this aim we need to measure the position, and possibly the speed and the current.

We first assume that an encoder is available as a sensor, namely, the angle is measured. The output matrix C takes the form

$$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

from which we get the observability matrix

$$O = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ \frac{k\bar{i}_e}{J} & -\frac{f}{J} & 0 \end{bmatrix}$$

This matrix has full rank, so we can find $\varphi(0)$ and the initial values of all of the other variables, based on measurement of the angular position $\varphi(t)$.

Conversely, assume that we choose the angular speed as an output. Then matrix C is

$$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

and the observability matrix is

$$O = \begin{bmatrix} 0 & 1 & 0 \\ \frac{k\bar{i}_e}{J} & -\frac{f}{J} & 0 \\ (*) & (*) & 0 \end{bmatrix}$$

where $(*)$ are non-zero numbers. Now, the matrix rank is equal to 2, hence the kernel has dimension 1 and its unique basis vector is

$$\bar{x} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

This means that $x(0)$ and $x(0) + \alpha\bar{x}$ produce the same output for any $\alpha \in \mathbb{R}$, namely, we can change the initial value of the angle and we will get the same output ω. Consequently, from the angular velocity we cannot determine the position. This is expected, because the angle can be determined from the speed by integration but we must know the initial value.[5] *Also if we measure the armature current, we cannot determine the angular position (this is left as an exercise). In conclusion, given the angle, we can find speed and current, but given any of these last two we cannot determine the position.*

Example 4.9. *One way to determine the angular velocity based on the angle (which is theoretically possible) is to use a pseudo-differentiator. Assuming $y(0) = 0$ (without restriction, since we measure its value), in time and frequency (Laplace) domain we have*

$$\begin{aligned} \omega(t) &= \frac{d}{dt}\varphi(t) \\ \omega(s) &= s\varphi(s) \end{aligned}$$

The transfer function above is not physically implementable, because it is not proper. However, we can alter the high-frequency Bode diagram by inserting two poles at higher frequencies, far away from the system bandwidth. For instance, we can take the strictly proper transfer function

$$W(s) = \frac{s}{(1+\tau s)^2} = \frac{1}{\tau^2}\frac{s}{(s+\frac{1}{\tau})^2} = \lambda^2 \frac{s}{(s+\lambda^2)}$$

What we get is a Bode diagram of the type shown in Figure 4.5.

[5] Recall that any primitive is defined up to a constant...

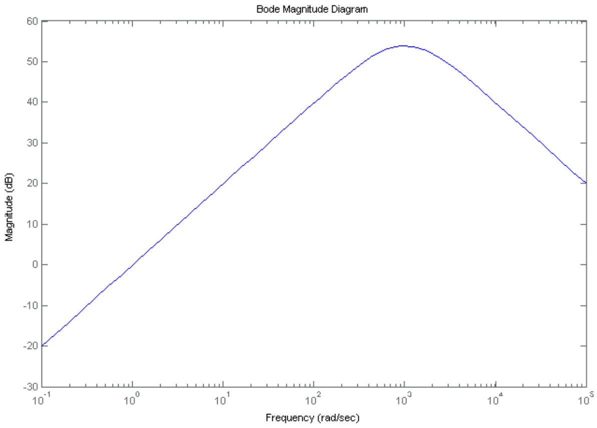

Figure 4.5: Bode plot of a real differentiator, with a pole at $\omega = 10^3 \,\frac{\text{rad}}{\text{s}}$.

A continuous-time state space representation corresponding to this transfer function is given (as we will show in the next chapter) by the matrices

$$A = \begin{bmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 0 & \lambda^2 \end{bmatrix}$$

We can then realise the corresponding discrete-time system by means of the equivalent discrete-time matrices

$$\begin{aligned} A_D &= e^{AT} \\ B_D &= \int_0^{\mathrm{T}} e^{A\xi} d\xi\, B \\ C_D &= C \end{aligned}$$

This type of pseudo-differentiator satisfactorily operates on a digital computer.

A different possibility is to choose

$$W(s) = \frac{s}{1 + 2\frac{\xi}{\omega_0} + \frac{s^2}{\omega_0^2}} = \frac{\omega_0^2 s}{s^2 + 2\xi\omega_0 s + \omega_0^2},$$

which has a sharper frequency cut. A suitable state space representation is

$$A = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\xi\omega_0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 0 & \omega_0^2 \end{bmatrix}$$

Example 4.10. *A method based on accelerometers allows us to track the position based on the acceleration, by means of a double integrator: acceleration → speed → position. This seems to be in contradiction with the observability notion, and impossible. Indeed, the position is not recoverable from the acceleration. As an exercise, one can write the state space representation of the force-acceleration equation*

$$\ddot{y}(t) = u(t)$$

and check non-observability. Where is the trick?

In these cases, as opposed to the observability problem, the position and the speed at the initial instant are assumed to be known. The differential problem to be solved is

$$\begin{aligned} v(t) &= \int_0^t a(\sigma)d\sigma + v(0) \\ x(t) &= \int_0^t v(t)dt + x(0) \end{aligned}$$

and requires an integrator.

A weakness of the methods is that a small measurement error affecting $a(t)$ has repercussions on the estimated position $\hat{y}(t)$. Assuming that $x(0) = x'(0) = 0$ (this is not restrictive), we actually obtain

$$\ddot{y}(t) = u(t) + \omega(t),$$

where $\omega(t)$ is the measurement noise. After a double integration, we have an estimated position equal to

$$\hat{y}(t) = \int_0^t \int_0^\xi u(\xi)d\xi\, d\sigma + \int_0^t \int_0^\xi \omega(\xi)d\xi\, d\sigma = y(t) + e(t),$$

where the error is $e(t)$. The double integration amplifies the noise effect, causing a drift. For this kind of systems, no longer in use after the introduction of satellite systems, the accelerometer must be reset at regular intervals, corresponding to positions where the location of the system is known exactly (in other ways, of course).

4.9 Duality and dual system

The reachability problem and the observability problem are quite similar. This analogy can be formalised by introducing the concept of dual system, which is a pure mathematical artifice, without physical meaning, but is very helpful. Given a system

$$\sum(A, B, C, D)$$

where the matrices are of dimensions $n \times n$, $n \times m$, $p \times n$, and $p \times m$ respectively, we define **dual system** represented by

$$\overset{*}{\sum} = \sum(A^\top, C^\top, B^\top, D^\top)$$

Namely, the primal and the dual system have the following equations

$$\sum = \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \quad \text{and} \quad \overset{*}{\sum} = \begin{cases} \dot{z}(t) = A^\top z(t) + C^\top v(t) \\ w(t) = B^\top z(t) + D^\top v(t) \end{cases} \tag{4.40}$$

An easy way to get the dual system matrices is take the transpose

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \xrightarrow{T} \left[\begin{array}{c|c} A^\top & C^\top \\ \hline B^\top & D^\top \end{array}\right] \tag{4.41}$$

Note that the number of inputs to the dual system is p, equal to the number of outputs of the primal system, while the number of outputs of the dual system is equal to m, the number of inputs to the primal system. The state of the dual system, instead, always belongs to $\mathbb{R}^n$, as in the primal system case. Also observe the following relationships:

- reachability concerns matrices (A, B) for the primal system, $(A^\top, C^\top)$ for the dual;

- observability concerns matrices (A, C) for the primal system and $(A^\top, B^\top)$ for the dual.

The reachability matrix $\mathcal{R}^*$ of the dual system,

$$\mathcal{R}^* = \left[\begin{array}{c|c|c|c|c} C^\top & A^\top C^\top & (A^\top)^2 C^\top & \dots & (A^\top)^{n-1} C^\top \end{array}\right] \tag{4.42}$$

is equal to the transpose of the observability matrix $\mathcal{O}^\top$ of the primal system: $\mathcal{R}^* = \mathcal{O}^\top$

$$\mathcal{R}^{*\top} = \begin{bmatrix} C \\ (A^\top C^\top)^\top \\ \vdots \\ \left((A^\top)^{n-1} C^\top\right)^\top \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = \mathcal{O} \tag{4.43}$$

The observability matrix of the dual system is the transpose of the reachability matrix of the primal: $\mathcal{O}^* = \mathcal{R}^\top$

$$\mathcal{O}^* = \begin{bmatrix} B^\top \\ B^\top A^\top \\ \vdots \\ B^\top (A^\top)^{n-1} \end{bmatrix} \Rightarrow \mathcal{O}^{*\top} = \left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = \mathcal{R} \tag{4.44}$$

Thus, fundamental properties hold:

- $\sum$ is reachable $\iff$ $\sum^*$ is observable;
- $\sum$ is observable $\iff$ $\sum^*$ is reachable

Observation 4.3. *Observability and reachability problems are* ***dual****.*

This duality property is useful from the mathematical standpoint, because every algorithm that we use for reachability is valid, with some modifications, for solving observability problems by means of the dual system.

A first application of the newly introduced duality concerns the Popov criterion. We have already seen that $\sum$ is reachable if and only if $\left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]$ has rank n for all eigenvalues λ of A. By duality, we can then say that $\sum$ is observable if and only if:

$$\operatorname{rank}\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n \tag{4.45}$$

for all eigenvalues λ of A. It is sufficient to apply the Popov criterion to the dual system and to note that the rank of a matrix is equal to the rank of its transpose, and that A and $A^\top$ have the same spectrum.

Another application concerns the parallel of systems. It has been shown that, given two systems (A_1, B_1) and (A_2, B_2), with B_1 and B_2 column matrices, the parallel system is reachable if and only if the two systems are reachable and have no common eigenvalues. Applying the same theorem to the dual system and exploiting duality properties, we can say that given two systems (A_1, C_1) and (A_2, C_2), with C_1 and C_2 row matrices, the parallel system is observable if and only if the two systems are observable and have no common eigenvalues.

Note also that, as far as the transfer function matrices are concerned,

$$W(s)^* = W^\top(s),$$

namely, the dual transfer function matrix is the transpose. Clearly, if $m = p = 1$, the primal system and the dual system have the same transfer function. The only difference in this case is that cancellations due to the unreachable modes in the primal system become cancellations due to unobservable modes in the dual system.

4.10 Joint Kalman form and transfer functions

Suppose that reachability and observability of a system have already been investigated, and the (uniquely determined) subspaces $X_r = \mathrm{Ra}(\mathcal{R})$ and $X_{no} = \ker(O)$ have been found. We can introduce the following subspaces :

- $X_1 = X_r \bigcap X_{no}$;
- X_2, such that $X_r = X_1 \bigoplus X_2$ (completion of X_1 in X_r);
- X_3, such that $X_{no} = X_1 \bigoplus X_3$ (completion of X_1 in X_{no});
- X_4, such that $X_1 \bigoplus X_2 \bigoplus X_3 \bigoplus X_4 = \mathbb{R}^n$.

We also have that $X_1 \bigoplus X_2 \bigoplus X_3 = X_r + X_{no}$. Furthermore, since the subspaces X_{nr} and X_o are complements of X_r and X_{no} in $\mathbb{R}^n$,

- $X_{nr} = X_3 \bigoplus X_4$
- $X_o = X_2 \bigoplus X_4$.

The relations between these subspaces are summarised in Table 4.1 (arrow means direct sum: *e.g.*, the arrow in the first row means $X_r = X_1 \bigoplus X_2$, and so on). It is now possible to choose a matrix T whose columns provide a basis for $\mathbb{R}^n$ and that is formed as follows:

reachable →	X_1	X_2
non reachable →	X_3	X_4
	↑ non observable	↑ observable

Table 4.1: Relations between subspaces.

$$T = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array}\right], \tag{4.46}$$

where the columns of the matrix T_k form a basis of the corresponding subspace X_k. Each vector $x \in \mathbb{R}^n$ can be represented as:

$$x = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array}\right] \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} \tag{4.47}$$

Each reachable vector can be expressed as:

$$x_r = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array}\right] \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ 0 \\ 0 \end{bmatrix} \tag{4.48}$$

Each non-observable vector can be expressed as:

$$x_{no} = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array}\right] \begin{bmatrix} \hat{x}_1 \\ 0 \\ \hat{x}_3 \\ 0 \end{bmatrix} \tag{4.49}$$

Similarly, the unreachable vectors are expressed as:

$$x_{nr} = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array} \right] \begin{bmatrix} 0 \\ 0 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} \tag{4.50}$$

While the observable vectors are expressed as:

$$x_o = \left[\begin{array}{c|c|c|c} T_1 & T_2 & T_3 & T_4 \end{array} \right] \begin{bmatrix} 0 \\ \hat{x}_2 \\ 0 \\ \hat{x}_4 \end{bmatrix} \tag{4.51}$$

Using the transformation T, we get $\hat{A} = T^{-1}AT$, $\hat{B} = T^{-1}B$, $\hat{C} = CT$, $\hat{D} = D$, hence

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} &= \begin{bmatrix} A_1 & A_{12} & A_{13} & A_{14} \\ 0 & A_2 & 0 & A_{24} \\ 0 & 0 & A_3 & A_{34} \\ 0 & 0 & 0 & A_4 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix} u \\ y &= \begin{bmatrix} 0 & C_2 & 0 & C_4 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} + Du \end{aligned} \tag{4.52}$$

All zeros in this matrix are structural. We can recognise the following subsystems:

- $\sum_1(A_1)$ is the reachable and unobservable subsystem;
- $\sum_2(A_2)$ is the reachable and observable subsystem;
- $\sum_3(A_3)$ is the unreachable and unobservable subsystem;
- $\sum_4(A_4)$ is the unreachable and observable subsystem;

Theorem 4.10. *The transfer function depends on the reachable and observable subsystem only*

$$W(s) = C(sI - A)^{-1}B = C_2(sI - A)^{-1}B_2 \tag{4.53}$$

The proof of the theorem requires tedious (not difficult) computations. This result is intuitive. Unreachable components, in fact, are not affected by the input, while unobservable components do not give any contribution to the output.

4.11 Cancellations

Consider the simplest case of a dynamic single input-single output (SISO) system, in which $m = p = 1$ and $D = 0$, hence $W(s) = C(sI - A)^{-1}B = \frac{n(s)}{d(s)}$. The denominator $d(s)$ is the characteristic polynomial of matrix A, hence it has degree n.

The polynomials $n(s)$ and $d(s)$ are **co-prime** if they do not have common roots, namely, $\nexists\, \lambda \in \mathbb{C}$ such that $n(\lambda) = d(\lambda) = 0$. Note that, if the two polynomial are not co-prime, then $n(s) = (s - \lambda)\tilde{n}(s)$ and $d(s) = (s - \lambda)\tilde{d}(s)$, hence we have a zero-pole cancellation and

$$W(s) = C(sI - A)^{-1}B = \frac{n(s)}{d(s)} = \frac{\tilde{n}(s)}{\tilde{d}(s)}$$

Theorem 4.11. *Given a linear system $\sum(A, B, C)$ with $m = p = 1$, the numerator $n(s)$ and the denominator $d(s)$ of the transfer function are co-prime if and only if (A, B, C) is reachable and observable.*

Proof. We prove the implication (⇒) by contradiction. Assume that $n(s)$ and $d(s)$ are co-prime and that the system is not reachable and observable. We have seen that the transfer function depends only on the reachable and observable part of the system:

$$W(s) = C(sI - A)^{-1}B = C_2(sI - A_2)^{-1}B_2$$

However, $\dim(A_2) < \dim(A)$. Then the denominator of the transfer function has degree at most $n_2 = \dim(A)$. This means that there is at least one cancellation. But then $n(s)$ and $d(s)$ cannot be co-prime: a contradiction.

We now prove the implication (⇐). We assume that the system is reachable and observable, with $n(s)$ and $d(s)$ not co-prime. We obtain a transfer function that is reducible through cancellations. This means that, for some λ,

$$\begin{aligned} d(s) &= \det(sI - A) = 0 \quad \text{if} \quad s = \lambda && (4.54)\\ n(s) &= \det\left[\begin{array}{c|c} sI - A & -B \\ \hline C & 0 \end{array}\right] = 0 \quad \text{if} \quad s = \lambda && (4.55) \end{aligned}$$

This implies that the matrix in (4.55) is singular and has a non-trivial kernel. We can write

$$\left[\begin{array}{c|c} \lambda I - A & -B \\ \hline C & 0 \end{array}\right]\left[\begin{array}{c} \bar{x} \\ \hline -\bar{u} \end{array}\right] = 0 \tag{4.56}$$

where $\bar{x}$ has n components and $\bar{u}$ is a scalar. The same applies to the matrix in (4.54): $[sI - A]$ is singular. We can write

$$\left[\begin{array}{c} \lambda I - A \end{array}\right]\tilde{x} = \left[\begin{array}{c} \lambda I - A \end{array}\right]\tilde{x} + B \cdot 0 = \left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]\left[\begin{array}{c} \tilde{x} \\ \hline 0 \end{array}\right] = 0 \tag{4.57}$$

where the last equality is always true due to the properties of partitioned matrices. We have two possible cases

- $\bar{u} \neq 0$: in this case the two vectors

  $$\begin{bmatrix} \bar{x} \\ \bar{u} \end{bmatrix}, \quad \begin{bmatrix} \tilde{x} \\ 0 \end{bmatrix}$$

  are linearly independent, therefore $\dim(\ker\left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]) \geq 2$. Since the number of columns of $\left[\begin{array}{c|c} \lambda I - A & B \end{array}\right]$ is equal to $n + 1$, it follows that the rank of this matrix is less than n, hence the system is not reachable (contradiction).

- $u = 0$: in this case

  $$\left[\begin{array}{c|c} \lambda I - A & -B \\ \hline C & 0 \end{array}\right]\left[\begin{array}{c} \bar{x} \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} \lambda I - A \\ \hline C \end{array}\right]\bar{x} + \left[\begin{array}{c} B \\ \hline 0 \end{array}\right]0 \tag{4.58}$$

  thus, if we consider only the first term, we get

  $$\left[\begin{array}{c} \lambda I - A \\ \hline C \end{array}\right]\bar{x} = 0 \tag{4.59}$$

  It follows that the matrix has rank less than n, and therefore, in view of the Popov observability criterion (see the section about the duality), the system is not observable: a contradiction.

The case of systems with $m > 1$ and $p > 1$ is much more involved. Consider the following example.

Example 4.11. *Given the system with matrices*

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

the reachability and the observability matrices are

$$\mathcal{R} = \begin{bmatrix} I & -I \end{bmatrix}, \quad \mathcal{O} = \begin{bmatrix} I & -I \end{bmatrix}$$

and they both have rank n, hence the system is reachable and observable. However, the transfer function matrix is

$$W(s) = \begin{bmatrix} \frac{s+1}{(s+1)^2} & \frac{0}{(s+1)^2} \\ \frac{0}{(s+1)^2} & \frac{s+1}{(s+1)^2} \end{bmatrix}$$

and we can see that there is a cancellation in each term. Therefore, it is no longer true that a reachable and observable system has no cancellations.

The case of multidimensional systems will not be analysed in detail here. However, we state some fundamental results:

- if either reachability or observability of the system fails, there must be cancellations in the transfer function matrix (but not vice versa, as already seen);
- if matrix A has distinct eigenvalues, then the system is reachable and observable if and only if there are no cancellations.

4.11.1 External stability

The input-output behaviour and the state representation of a dynamical system can be studied, as far as the asymptotic stability property is concerned, by introducing the following definition.

Definition 4.10. *Let $x(0) = 0$. We say that a system $\sum(A, B, C, D)$ is **externally stable** or **BIBO (Bounded Input, Bounded Output) stable** if, assuming that $\|u(t)\| \le \mu$, then there exists ν such that $\|y(t)\| \le \nu$.*

We have the following result.

Theorem 4.12. *A system is externally stable if and only if the reachable and observable subsystem is asymptotically stable. If the whole system is reachable and observable, then asymptotic stability is equivalent to BIBO stability.*

We will not prove the theorem, which is intuitive. Indeed, non-reachable and non-observable parts (if any) do not affect the input-output behaviour. Conversely, reachable unstable (or marginally stable) modes can become arbitrarily large due to a bounded input and, if they are observable, they will produce an arbitrarily large output.

We will prove the last part of the theorem, to get a very interesting formula, in the case $m = p = 1$. Assume that the system is reachable and observable, and that A is asymptotically stable. We show that the system is BIBO stable. Let $x(0) = 0$ and $\|u(t)\| \le \mu$. We have:

$$y(t) = \int_0^t W(t - \sigma)u(\sigma)d\sigma$$

hence

$$\begin{aligned}|y(t)| &= \left|\int_0^t W(t-\sigma)u(\sigma)d\sigma\right| \leq \int_0^t |W(t-\sigma)u(\sigma)|\, d\sigma = \int_0^t |W(\sigma)|\, |u(t-\sigma)|\, d\sigma \leq \\ &\leq \mu\int_0^t |W(\sigma)|\, d\sigma \leq \mu\int_0^{+\infty} |W(\sigma)|\, d\sigma = \mu\int_0^{+\infty} \left|Ce^{A\sigma}B\right| d\sigma = \mu\|(A,\, B,\, C)\|_1 = \nu\end{aligned}$$

where we have denoted by

$$\|(A,\, B,\, C)\|_1 \doteq \int_0^{+\infty} \left|Ce^{A\sigma}B\right| d\sigma$$

It can be shown that this quantity is a norm and can be quite useful to compute an output limitation in the presence of a limited input noise (typically due to non-quantifiable phenomena of which we know, at least, some bounds on the amplitude).

Observation 4.4. *The property of BIBO stability does not offer any kind of warranty on the proper functioning of the system, if the system is not reachable or is not observable. In fact, unstable unobservable modes do not affect the output, but may lead to exponentially increasing internal variables. Unstable unreachable modes theoretically remain at zero, because in the definition of BIBO stability we take* $x(0) = 0$*, but they may grow large when* $x(0) \neq 0$ *(no matter how small).* ***Non-observability and non-reachability is a pathology****.*

4.12 Controllable canonical form

Consider a system $\sum(A,\, B,\, C)$ with A of size $n \times n$, B of size $n \times 1$ (*i.e.*, scalar input), and C of size $1 \times p$ (*i.e.*, scalar output). We can check that, if the pair $(A,\, B)$ is reachable and the pair $(A,\, C)$ is observable, then there exists a transformation T such that $\hat{A}_F = T^{-1}AT$, $\hat{B}_F = T^{-1}B$ and $\hat{C}_F = CT$ have the form:

$$\hat{A}_F = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix}, \quad \hat{B}_F = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{4.60}$$

$$\hat{C}_F = \begin{bmatrix} c_0 & c_1 & c_2 & \dots & c_{n-1} \end{bmatrix} \tag{4.61}$$

This is called the **Frobenius form**. It is not difficult to prove the following

- the numbers a_k are the coefficients of the characteristic polynomial;
- the numbers c_k are the coefficients of the transfer function numerator.

The state transformation T can be computed as follows.

First notice that, given any transformation T, the reachability matrix in the new state form is

$$\hat{\mathcal{R}} = \left[\ \hat{B} \mid \hat{A}\hat{B} \mid \dots \mid \hat{A}^{n-1}B\ \right] = \left[\ T^{-1}B \mid T^{-1}AB \mid \dots \mid T^{-1}A^{n-1}B\ \right] = T^{-1}\mathcal{R}$$

hence

$$\hat{\mathcal{R}} = T^{-1}\mathcal{R} \tag{4.62}$$

By duality we also have

$$\hat{O} = OT \tag{4.63}$$

(in brief, $\mathcal{R}$ and O transform as B and C). If, as assumed, $m = 1$ and the system is reachable, the reachability matrix is square and invertible. Hence, $T\hat{\mathcal{R}} = \mathcal{R}$ and

$$T = \mathcal{R}\hat{\mathcal{R}}^{-1}. \tag{4.64}$$

Therefore, to determine the transformation T, we have to compute $\mathcal{R}$ and $\hat{\mathcal{R}}$. We know that the similarity transformation operation does not change the eigenvalues of the matrix A, so the characteristic polynomial of $\hat{A}$ is the same as that of A. To determine the unknown elements of $\hat{A}$, we need to compute the coefficients of the characteristic polynomial of A. Then, the transformation T can be computed through the following steps:

1. compute $\det(sI - A)$ and form matrix $\hat{A}$ with its coefficients;
2. compute $\mathcal{R}$;
3. compute $\hat{\mathcal{R}}$;
4. compute $T = \mathcal{R}\hat{\mathcal{R}}^{-1}$.

As mentioned, $\hat{A}$ is said to be in Frobenius form. This form is very important for three reasons, two of which will be analysed later. The remaining reason is that a matrix in Frobenius form is the simplest option to represent in a state form (n differential equations of the first order) a generic linear differential equations of order n. If, for example, we have

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \ldots + a_1 y^{(1)}(t) + a_0 y(t) = u(t), \tag{4.65}$$

then we can define the state variables

$$\begin{aligned} x_1(t) &= y(t) \\ x_2(t) &= y^{(1)}(t) \\ &\vdots \\ x_n(t) &= y^{(n-1)}(t) \end{aligned} \tag{4.66}$$

and write the equivalent system of n first order differential equations, thus obtaining

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= x_3(t) \\ &\ldots \\ \dot{x}_{n-1}(t) &= x_n(t) \\ \dot{x}_n(t) &= -a_0 x_1(t) - a_1 x_2(t) - \ldots - a_{n-1}x_n(t) + u(t), \end{aligned} \tag{4.67}$$

which is equivalent to the Frobenius form (check!).

Clearly, in view of duality, we can define the (Frobenius) observability canonical form, where A and C are the transposed matrices of A and B in the controllable canonical form.

$$\hat{A}_F^* = \begin{bmatrix} 0 & 0 & 0 & \ldots & -a_0 \\ 1 & 0 & 0 & \ldots & -a_1 \\ 0 & 1 & 0 & \ldots & -a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & -a_{n-1} \end{bmatrix}, \quad \hat{B}_F^* = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_2 \\ b_{n-1} \end{bmatrix} \tag{4.68}$$

$$\hat{C}_F^* = \begin{bmatrix} 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{4.69}$$

Again, a_k are the coefficients of the characteristic polynomial and b_k the coefficients of the numerator of the transfer function.

Chapter 5

Realisation

5.1 The problem in general

The realisation problem, in general, is the problem of associating a state-space representation with a given input-output description. We will consider here the case of linear time-invariant systems, whose input-output behaviour is described in terms of the transfer function matrix. For a linear autonomous system (in the time domain), the solution of the equations is given by

$$\begin{cases} x(t) = e^{At}x(0) + \int_0^{+\infty} e^{A(t-\sigma)}u(\sigma)d\sigma \\ y(t) = Cx(t) + Du(t) \end{cases}$$

By applying the Laplace transform, the expression becomes

$$\begin{cases} x(s) = (sI - A)^{-1}x(0) + (sI - A)^{-1}Bu(s) + D\delta(t) \\ y(s) = C(sI - A)^{-1}x(0) + C(sI - A)^{-1}Bu(s) \end{cases}$$

If $x(0) = 0$, the input-output relation expressed by the transfer function matrix $W(s)$:

$$y(s) = [C(sI - A)^{-1}B + D]u(s) = W(s)u(s).$$

Setting $x(0) = 0$ is not a strong restriction, because the essential information on the free evolution of the system, namely the modes, is preserved and can be found in the denominator of $W(s)$ (the characteristic polynomial of the matrix A, if there are no cancellations).

Based on a given state space representation, we can easily determine the corresponding transfer function matrix. Now the discussion concerns the opposite problem: can we derive a state representation from an assigned transfer function matrix? This is the realisation problem. Studying the realisation problem is useful from both a theoretical standpoint, to understand the link between the realisation A, B, C, D and $W(s)$, and from a practical standpoint, to physically implement filters or controllers designed in the frequency domain (as explained in the course of Automatic Control).

Given A, B, C, D, respectively with size $n \times n$, $n \times m$, $p \times n$ and $p \times m$, the transfer function matrix is

$$W(s) \quad = \quad \frac{N(s)}{d(s)} = C(sI - A)^{-1}B + D = \frac{N_0 + N_1 s + N_2 s^2 + \ldots + N_\nu s^\nu}{d_0 + d_1 s + d_2 s^2 + \ldots + d_\nu s^\nu},$$

where N_k are numeric matrices of dimension $p \times m$. Note that the elements of $W(s)$,

$$\left[W_{ij}(s)\right] = \left[\frac{n_{ij}(s)}{d(s)}\right],$$

are rational functions. Hence, the realisation problem can provide a state-space representation of the form $\Sigma(A, B, C, D)$ only if we start from a rational transfer function matrix. Moreover, the given rational transfer function matrix has to be proper.

Example 5.1. *It is legitimate to ask how we can realise the transfer function $W(s) = e^{-\tau s}$ (delay), but we will not find matrices $\Sigma(A, B, C, D)$ that provide a realisation.*

Example 5.2. *The PID controllers given by the transfer functions*

$$G(s) = G_D(s) + K_D + \frac{K_I}{s} = \frac{K_D s^2 + K_P s + K_I}{s}$$

are not directly implementable, because the above rational function is not proper. Then, it does not admit a realisation A, B, C, D. A physical implementation is possible if we introduce one or two additional poles, such as

$$\tilde{G}(s) = \frac{K_I + K_P s + K_D s^2}{s(1+\tau s)^2}.$$

These artificially introduced poles are located at high frequency, far from the band of frequencies of the system. They are not only necessary for the realisation, but also beneficial, since they have a low-pass effect.

Definition 5.1. *Given a proper rational matrix $W(s)$, we say that $\Sigma(A, B, C, D)$, is a realisation if*

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

The realisation problem corresponds to finding $\Sigma(A, B, C, D)$ for a given $W(s)$.

The first step for solving the realisation problem is the determination of matrix D. We have that

- if W is strictly proper, then $D = 0$;
- if W is weakly proper, then $D \neq 0$ and its value can be obtained by division.

Division means writing the transfer matrix as the sum of a constant matrix and a strictly proper transfer function:

$$\begin{aligned}
W(s) &= \frac{N_\nu s^\nu + N_{\nu-1}s^{\nu-1} + \ldots + N_0}{d_0 + d_1 s + \ldots + s^\nu} = \\
&= \frac{N_\nu(d_0 + d_1 s + \ldots + s^\nu) + (N_{\nu-1} - N_\nu d_{\nu-1})s^{\nu-1} + (N_{\nu-2} - N_\nu d_{\nu-2})s^{\nu-2} + \ldots + (N_0 - N_\nu d_0)}{d_0 + d_1 s + \ldots + s^\nu} \\
&= N_\nu + \frac{\tilde{N}_{\nu-1}s^{\nu-1} + \ldots + \tilde{N}_0}{d_0 + d_1 s + \ldots + s^\nu} = D + \tilde{W}(s) \qquad (5.1)
\end{aligned}$$

Then, the constant matrix D is obtained (which is obviously zero in the strictly proper case). The next step is finding the matrices A, B, C starting from the strictly proper part of the transfer function.

So let $W(s)$ be strictly proper and write it as

$$\frac{N_0 + N_1 s + \ldots + N_{\nu-1}s^{\nu-1}}{d_0 + d_1 s + \ldots + s^\nu} \qquad (5.2)$$

Determining a realisation is straightforward. We can just adopt the following matrices (generalised Frobenius form):

$$\begin{aligned}
A &= \begin{bmatrix} 0 & I_m & 0 & \ldots & 0 \\ 0 & 0 & I_m & \ldots & 0 \\ 0 & 0 & 0 & \ldots & I_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -d_0 I_m & -d_1 I_m & -d_2 I_m & \ldots & -d_{\nu-1}I_m \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ I_m \end{bmatrix}, \\
C &= \begin{bmatrix} N_0 & N_1 & N_2 & \ldots & N_{\nu-1} \end{bmatrix}, \qquad (5.3)
\end{aligned}$$

where I_m is the identity matrix of dimension $m \times m$. To demonstrate that this realisation returns $W(s)$ we just write:

$$\phi(s) = (sI - A)^{-1}B \Rightarrow (sI - A)\phi(s) = B. \tag{5.4}$$

By direct verification we can check that $\phi(s)$ is given by

$$\phi(s) = \frac{1}{d(s)} \begin{bmatrix} I \\ Is \\ s^2 I \\ \vdots \\ s^{\nu-1} I \end{bmatrix} \tag{5.5}$$

and then check the equation

$$W(s) = C\phi(s). \tag{5.6}$$

This realisation has a matrix A of dimension $m\nu \times m\nu$, where $m\nu$ is greater than ν (the degree of the denominator of the transfer functions). Therefore, we suspect that this is not a minimal realisation, in terms of state-space dimensions, and in general it is not.

Remark 5.1. *By duality, one can find a realisation as follows. Consider the transfer function matrix $W^\top(s)$ of the dual system and find a realisation (A_*, B_*, C_*, D_*), of dimension $p\nu \times p\nu$. Consider the dual realisation $(A_*^\top, C_*^\top, B_*^\top, D_*^\top)$: it is a realisation for W. This procedure is convenient if $p < m$ (and vice-versa).*

Example 5.3. *As a preliminary example, we show that the realisation found as above can be non minimal. Take*

$$W(s) = \begin{bmatrix} \frac{2}{(s+1)(s+2)} & \frac{s}{(s+1)(s+2)} \end{bmatrix} = \begin{bmatrix} \frac{[\,0\;2\,]}{(s^2+3s+2)} & \frac{[\,1\;0\,]s}{(s^2+3s+2)} \end{bmatrix}$$

The realisation has state dimension $2m = 4$.

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 0 & -3 & 0 \\ 0 & -2 & 0 & -3 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad C = \left[\begin{array}{cc|cc} 0 & 2 & 1 & 0 \end{array}\right]$$

Let us now consider the dual $W^\top$. The realisation is now $p \times 2 = 1$.

$$A_* = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \quad B_* = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad C_* = \left[\begin{array}{c|c} 0 & 1 \\ 2 & 0 \end{array}\right]$$

Then $(A_*^\top, C_*^\top, B_*^\top)$ is a smaller realisation of order 2. *We will see that this is minimal, in terms of number of state variables.*

5.2 Minimal realisation

In the problem of finding a state-space realisation for $W(s)$, the input dimension m and the output dimension p are the dimension of the transfer function matrix $W(s)$, hence they are fixed. However, the state space dimension n is not known. Since n is a nonnegative integer, there will be a minimal possible value of n that allows us to provide a realisation.

Definition 5.2. *Matrices $(A,\ B,\ C,\ D)$ are a **minimal realisation** of $W(s)$ if they are a realisation, namely $C(sI - A)^{-1}B + D = W(s)$, and there is no other realisation $(A',\ B',\ C',\ D)$ with* $\dim(A') < \dim(A)$.

Finding a minimal realisation amounts to minimising the size of matrix A, namely n. Once again, the number of columns B (equal to m) and the number of lines of C (equal to p) are fixed.

The following fundamental theorem characterises minimal realisations.

Theorem 5.1. *The realisation* $(A,\ B,\ C,\ D)$ *of* $W(s)$ *is minimal if and only if* $(A,\ B,\ C)$ *is reachable and observable.*

The implication ⇒ has already been discussed, because if the realisation is minimal then there are cannot be unreachable or unobservable parts in the system. Otherwise, adopting Kalman decomposition, we could just take the reachable (or observable) subsystem which has the same transfer function, but a smaller dimension. The implication ⇐ is quite difficult to prove: its proof can be found in specialised texts. The theorem tells us something interesting.

Observation 5.1. *If* $(A,\ B,\ C,\ D)$ *is a realisation, but not a minimal one, then the reachable and observable subsystem* $(A_2,\ B_2,\ C_2,\ D)$ *gives us a minimal realisation (because it has the same transfer function).*

The next question is the following. Given two minimal realisations, is there any difference between the two?

Assume to have obtained two minimal realisations, $(A_1,\ B_1,\ C_1,\ D_1)$ and $(A_2,\ B_2,\ C_2,\ D_2)$, of course with the matrices A_1 and A_2 having the same size. There is a fundamental link between the two solutions.

Theorem 5.2. *Let* $(A_1,\ B_1,\ C_1,\ D_1)$*, with state dimension n, be a minimal realisations of* $W(s)$*. Then* $(A_2,\ B_2,\ C_2,\ D_2)$ *is a minimal realisations of* $W(s)$ *if and only if: if has state dimension n,* $D_1 = D_2$ *and there exists T, invertible such that* $T^{-1}A_1T = A_2$, $T^{-1}B_1 = B_2\, C_1T = C_2$.

The previous theorem states that all minimal realisations are equivalent up to a state transformation.

The implication ⇐ is simple, because if $(A_1,\ B_1,\ C_1)$ is a minimal realisation and we apply a state transformation we obtain a system with the same transfer function and a state matrix having the same size as A_1. The implication ⇒ is much more difficult to prove and is not discussed.

We can say that if a given linear system $(A,\ B,\ C,\ D)$ is modelled by means of transfer functions, any information about the unreachable and unobservable parts is lost. If we assume that the system is reachable and observable, we still loose some information because we will not be able to determine the true realisation we started from, but we will be able to determine a realisation which is related to the original one by a state transformation.

We conclude with the general procedure to derive a minimal realisation.

1. Given $W(s)$ rational and proper, determine D by division: $W(s) = D + \tilde{W}(s)$, with $\tilde{W}(s)$ strictly proper.

2. Consider the strictly proper part $\tilde{W}(s)$ and take any realisation, not necessarily minimal (for instance the generalised Frobenius one).

3. Remove the unobservable and unreachable part to get a minimal realisation.

It is worth mentioning that the generalised Frobenius realisation shown above is always reachable. So, in this case, we need to remove the unobservable part only. By duality, the dual generalised Frobenius realisation is always observable and in this case we only need to remove the unreachable part.

5.3 Minimal realisation for SISO systems

The transfer function, in the case of single-input and single-output (SISO) systems ($m = p = 1$), can be written as

$$W(s) = \frac{n_0 + n_1 s + \ldots + n_{\nu-1} s^{\nu-1}}{d_0 + d_1 s + \ldots + s^{\nu}} + D \tag{5.7}$$

where D is zero if and only if W is strictly proper. As discussed earlier, one possible tool for the realisation and implementation is given by the Frobenius form

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -d_0 & -d_1 & -d_2 & \ldots & -d_{\nu-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \\ C &= \begin{bmatrix} n_0 & n_1 & n_2 & \ldots & n_{\nu-1} \end{bmatrix} \end{aligned} \tag{5.8}$$

From Theorem 4.11, since $m = p = 1$, this realisation is minimal if and only if there are no cancellations in $W(s)$.

A nice property of the Frobenius form is that it can be implemented not only by a digital device, but also by an analog device, by adopting the block diagram presented in Figure 5.1.

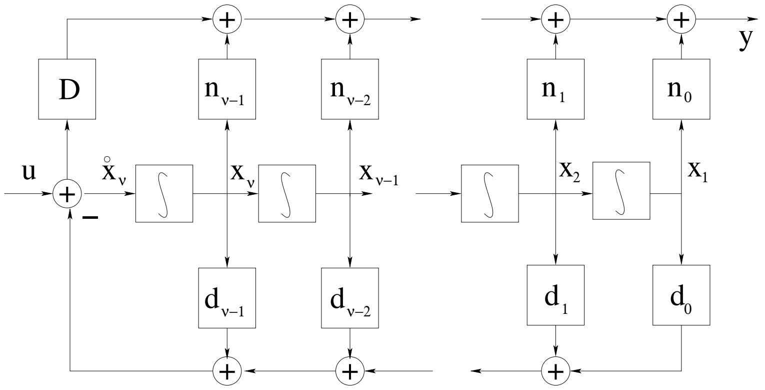

Figure 5.1: Block diagram of the minimal realisation with $m = p = 1$.

The diagram is easy to interpret if one considers that, given the form of the matrices, we obtain the following expressions

$$\begin{aligned} \dot{x}_i &= x_{i+1} \quad \forall\, i = 1, \ldots, n-1 \\ \dot{x}_n &= -\left(\sum_{i=0}^{\nu-1} d_i x_{i+1} \right) + u \\ y &= \sum_{j=0}^{\nu-1} n_j x_{j+1} + Du \end{aligned} \tag{5.9}$$

Each block in Figure 5.1 can be implemented by means of operational amplifiers.

5.3.1 Other techniques for SISO systems realisation

There are many different techniques for the realisation of systems, especially for SISO systems.

For instance, if we know the poles of the transfer function,

$$W(s) = \frac{n_0 + n_1 s + \ldots + n_{\nu-1} s^{\nu-1}}{(s-\lambda_1)(s-\lambda_2)\ldots(s-\lambda_n)}$$

if the poles are distinct we can compute the fraction expansion

$$W(s) = \sum_{i=1}^{n} \frac{r_1}{(s-\lambda_i)}$$

and notice that a realisation is

$$\begin{aligned} A &= \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \\ C &= \begin{bmatrix} r_1 & r_2 & r_3 & \ldots & r_n \end{bmatrix} \end{aligned} \tag{5.10}$$

If there are complex poles, the previous realisation is complex. To deal with this case, one can consider that pairs of complex poles give rise to real terms in the sum of degree two

$$\frac{r}{s-\lambda} + \frac{r^*}{s-\lambda^*} = \frac{\alpha s + \beta}{(s-\xi)^2 + \omega^2}$$

This partial term can be realised by means of a 2 × 2 Frobenius realisation or by

$$A = \begin{bmatrix} \xi & \omega \\ -\omega & \xi \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} \mu & \nu \end{bmatrix}$$

where μ and ν can be determined from α and β.

5.4 Realisation of discrete-time systems

The problem of the realisation of discrete-time systems is analogous to the realisation problem for continuous-time systems. The only difference is that the transfer function is obtained by using the Z-transform. Recall that the Z-transform is an operator that associates with a sequence $f(k)$ a function of complex variable $F(z)$ defined as follows

$$f(k) \xrightarrow{\mathcal{Z}} F(z) = \sum_{k=0}^{+\infty} f(k) \frac{1}{z^k}$$

The operator is linear and it has following property

$$f(k+1) \xrightarrow{\mathcal{Z}} zF(z) - zf(0)$$

If we consider the discrete-time system

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k)$$

and we set $x(0) = 0$, we have

$$y(z) = W(z)u(z)$$

where $W(z)$ is the discrete-time transfer function

$$W(z) = C(zI - A)^{-1}B + D,$$

which has the same expression as the transfer function for continuous-time systems. Then, the realisation problem is solved in exactly with the same approach adopted in the continuous-time case.

Chapter 6

Control Systems

6.1 General considerations

The main goal of a control system, in general terms, is to govern a system in order to achieve the desired behaviour, by suitably choosing the control input u. There can be many special goals in designing a control. The main are

- **regulation**: steer a system to a desired state and keep it there;
- **tracking**: steer a system along a desired trajectory;
- **disturbance rejection**: mitigate the effect of disturbances.

There are two main types of control.

- **Open loop control**: the input $u(t)$ applied to the system is synthesised by means of some *a priori* criterion based on the knowledge of the initial state and does not require the knowledge of $x(t)$ or $y(t)$ in real time.
- **Feedback control**: the input to the process is generated in real time (instant by instant) by a control algorithm that takes into account the current output (output-feedback) or the current state (state-feedback) of the system.

A simple example of open-loop control is given by a temporised oven, which heats a cake for a certain amount of time (decided at the beginning) and then stops.[1] The main feature of a feedback control is that it is able to modify the dynamics of a system to make it stable, or faster in terms of convergence. A feedback control makes the system less sensitive to disturbances and assures efficient tracking. In this chapter, we consider the case of feedback control for linear systems. This technique will be applied to nonlinear systems later on.

In general, the inputs and outputs of a process can be divided according to Figure 6.1.

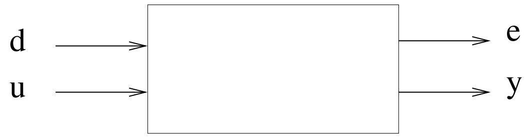

Figure 6.1: Subdivision of inputs and outputs in a system.

[1] Inconveniences of this approach are often experimented by inexperienced bakers.

- Vector $d(t)$ represents exogenous, **external inputs** not modifiable by the user. This category includes, for example, noise and external loads.
- Vector $u(t)$, instead, represents the **control inputs**, on which we can actually act in real time, in order to control the process.
- Vector $e(t)$ is the **performance output**, on which design specifications are assigned.
- Vector $y(t)$ is the **measured output**, namely, the signal measured in real time by sensors, which can be used to generate a feedback control law.

The input $u(t)$ corresponds therefore to the action of the **actuators**, while the output $y(t)$ originates from the **sensors** monitoring the system. The process needs to be equipped with these devices in order to be governed by a feedback controller.

Example 6.1. ***(Damper.)*** *Consider the damping control problem for a vibrating systems. The goal is to keep the coordinate $\theta(t)$ small in the presence of a disturbance $d(t)$. The measured output is the derivative $\dot{\theta}(t)$*

$$\begin{aligned}\ddot{\theta}(t) &= -\mu\theta(t) + d(t) + u(t)\\ e(t) &= \theta(t)\\ y(t) &= \dot{\theta}(t)\end{aligned}$$

A control can be any system with input y and output u, such as a simple passive damper: $u(t) = -ky(t)$. Needless to say, the eigenvalues play a fundamental role. For instance, in the case of the damper, the closed loop eigenvalues are the solutions of

$$s^2 + \kappa s + \mu = 0$$

The eigenvalues are constrained on the so called root locus $[\lambda_1(\kappa), \lambda_2(\kappa)]$. It can be shown that the real part is minimised by $\kappa^* = 2\sqrt{\mu}$. This control has one degree of freedom only. By adopting more sophisticated techniques, we will be able to arbitrarily place the eigenvalues of the system.

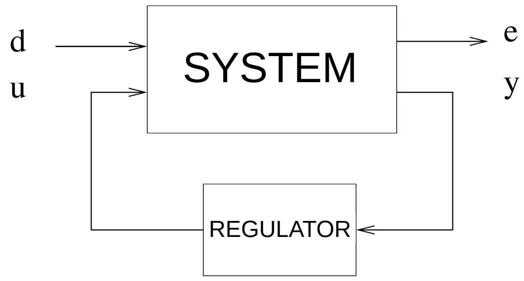

Figure 6.2: Linear system with feedback: block diagram.

In general, a feedback-regulated process can be described by the scheme in Figure 6.2. The first purpose of the feedback is the **stabilisation** of the process, if the system is not already stable. Even in the case of stable processes, feedback can have beneficial effects in improving the dynamics (yielding, for instance, faster convergence). In the case of linear systems, this essentially means **changing the position of the eigenvalues** in the complex plane, in order to modify the system transient. The allocation of the eigenvalues is possible only by means of feedback controllers (open-loop controllers do not modify the eigenvalues).

The generic linear process to be controlled has the form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t) \\ y(t) = Cx(t) \\ e(t) = Lx(t) \end{cases} \tag{6.1}$$

The controller is, in turn, a dynamical system

$$\begin{cases} \dot{z}(t) = Fz(t) + Gy(t) \\ u(t) = Hz(t) + Ky(t) \end{cases} \tag{6.2}$$

By connecting the two systems, the overall system becomes

$$\begin{aligned} \frac{d}{dt}\left[\begin{array}{c} x(t) \\ \hline z(t) \end{array}\right] &= \left[\begin{array}{c|c} A+BKC & BH \\ \hline GC & F \end{array}\right]\left[\begin{array}{c} x(t) \\ \hline z(t) \end{array}\right] + \left[\begin{array}{c} E \\ \hline 0 \end{array}\right] d(t) \\ e(t) &= \left[\begin{array}{c|c} L & 0 \end{array}\right]\left[\begin{array}{c} x(t) \\ \hline z(t) \end{array}\right] \end{aligned} \tag{6.3}$$

The closed-loop state matrix is

$$A_{CL} = \left[\begin{array}{c|c} A+BKC & BH \\ \hline GC & F \end{array}\right]$$

and the modes of the closed-loop system are given by the eigenvalues of A_{CL}. The first goal of a controller is to modify the original eigenvalues of matrix A and get new eigenvalues that are chosen appropriately. A more advanced task is to optimise the performance, under the constraints of closed-loop stability.

The basic problem, which we face for the moment, is the following.

Problem 6.1. Eigenvalues assignment problem. *Given a set Λ_{CL} of $N = dim(A) + dim(F)$ complex numbers, $\Lambda_{CL} = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$, determine the unknown matrices F, G, H, K so that the closed-loop eigenvalues are the element of Λ_{CL}:*

$$\sigma\left(\left[\begin{array}{c|c} A+BKC & BH \\ \hline GC & F \end{array}\right]\right) = \Lambda_{CL}$$

In the following, we will always assume that the chosen set Λ_{CL} satisfies the **conjugate eigenvalue constraint** (CEC):

$$\lambda \in \Lambda_{CL} \Rightarrow \lambda^* \in \Lambda_{CL}.$$

This means that either λ is real or, if it is complex, then also λ^* belongs to Λ_{CL}.[2] If we do not respect this condition, then the solution F, G, H, K will be given by complex matrices, hence useless from a practical standpoint.

A fundamental principle in dealing with the allocation problem is the following. The eigenvalues of the unreachable subsystem and of the unobservable subsystem cannot be changed.

Theorem 6.1. *The unreachable eigenvalues, given the pair (A, B), and the unobservable eigenvalues, given the pair (A, C), are **invariant under feedback**.*

There is no need to prove this theorem. Unreachable an unobservable subsystems can be removed without changing the input-output behaviour of the system, hence no regulator can change them. A fundamental consequence is the following.

[2] For instance, $-1, -2+j3, -2-j3, -3+2j, -3-2j, -4$ is a proper choice, while $-1, -2+j3, -2-j3, 6+2j, -3-2j, -4$ is not.

Theorem 6.2. *A process can be asymptotically stabilised if and only if the unreachable and the unobservable eigenvalues (if any) are asymptotically stable, namely, they have negative real part.*

Necessity follows from Theorem 6.1. The sufficiency part will be proved in a constructive way: we show how to arbitrarily modify reachable and observable modes (under CEC).

To study the allocation of the eigenvalues we will work, without restrictions, under the assumption that $\sum(A, B, C)$ is reachable and observable. Its unreachable and/or unobservable parts (whose eigenvalues are fixed) can be eliminated after having reduced the system in Kalman form.

We will solve the eigenvalue assignment problem in three steps.

- **State-feedback design**: we assume[3] that any state variable is measured and we solve the allocation problem.
- **Observer design**: we design a device that estimates the state variables based on output measurement.
- **Pole placement regulator**: we put together these two steps and we apply the state feedback to the estimated state variables.

In the first step, **state feedback** design, we consider the following equations:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ u(t) = Kx(t) \end{cases} \tag{6.4}$$

The corresponding diagram is shown in Figure 6.3.

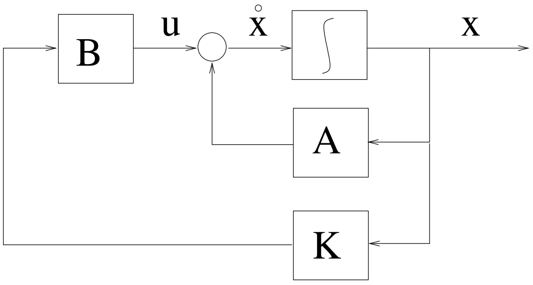

Figure 6.3: State feedback.

Note that a direct practical implementation of the state feedback $u(t) = Kx(t)$ requires a sensor for each state variable, and this may be unrealistic, expensive and even impossible in some cases. This leads to the second step, **state estimation**, in accordance to diagram shown in Figure 6.4. The estimation block returns an estimated value $\hat{x}(t)$ of the state $x(t)$ of the process. This value is subsequently used to operate the feedback $u(t) = K\hat{x}(t)$. This is convenient from the economic point of view, since it does not require sensors for all state variables. In the following sections we will study separately and in detail the two design phases above.

6.2 State feedback

We now assume that the measured output is the state x. This is actually true in some applications. To design a state feedback controller, we assume that a choice has been made for the closed loop system eigenvalues

$$\Lambda_c = \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \in C$$

[3] We pretend, because this is not true in general.

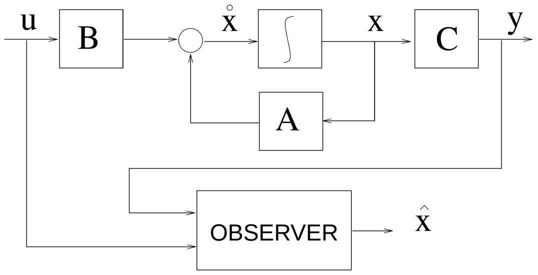

Figure 6.4: State observer.

where the subscript c stands for "compensator" and the CEC constraint are satisfied (so as to deal with real problems). We look for a real matrix K such that the system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ u(t) = Kx(t) \end{cases} \Rightarrow \dot{x}(t) = (A+BK)x(t) \tag{6.5}$$

has the desired set of eigenvalues

$$\sigma(A+BK) = \Lambda_c$$

The matrix K has dimension $m \times n$, hence there are $m \cdot n$ free parameters to be assigned in order to solve the problem.

Here we have much more freedom with respect to classical techniques, such as root locus and PID design, where the number of free parameters is limited.

Theorem 6.3. *The problem has a solution with Λ_c arbitrarily taken (provided that CEC constraints are satisfied), if and only if the system $\sum(A, B)$ is reachable. Instead, if the system is not reachable, then the problem has solution if and only if Λ_c includes all of the unreachable eigenvalues.*

Proof. If the system is not reachable, we can turn it into Kalman form. The equation $u(t) = Kx(t)$ turns into $u(t) = KT\hat{x}(t) = \hat{K}\hat{x}(t)$.[4] We get

$$\begin{aligned} \hat{A} + \hat{B}\hat{K} &= \left[\begin{array}{c|c} \hat{A}_r & \hat{A}_{r,nr} \\ \hline 0 & \hat{A}_{nr} \end{array}\right] + \left[\begin{array}{c} \hat{B}_r \\ \hline 0 \end{array}\right] \left[\begin{array}{c|c} \hat{K}_1 & \hat{K}_2 \end{array}\right] = \\ &= \left[\begin{array}{c|c} \hat{A}_r + \hat{B}_r\hat{K}_1 & \hat{A}_{r,nr} + \hat{B}_r\hat{K}_2 \\ \hline 0 & \hat{A}_{nr} \end{array}\right] \end{aligned} \tag{6.6}$$

This is a block-triangular matrix, whose spectrum is given by $\sigma(\hat{A}_r + \hat{B}_r\hat{K}_1) \cup \sigma(\hat{A}_{nr})$. The resulting spectrum includes the eigenvalues of $\hat{A}_{nr}$, the unreachable ones, which are invariant and cannot be changed with the feedback. So we cannot assign Λ_c if it does not include the eigenvalues of $\hat{A}_{nr}$.

We now assume that the system is reachable and we show how we can arbitrarily assign the eigenvalues to the system. We now consider the case $m = 1$, while we will discuss the general case later. The proof is constructive.

If the system is reachable, we can always find T such that the transformed system $A_F = T^{-1}AT$, $B_F = T^{-1}B$ is in the Frobenius form (note that the new feedback matrix is transformed as well,

[4] Note that matrix K is transformed as matrix C.

$\hat{K} = KT$). The closed-loop matrix is

$$A_F + B_F\hat{K} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_0 + \hat{k}_1 & -a_1 + \hat{k}_2 & -a_2 + \hat{k}_3 & \dots & -a_{n-1} + \hat{k}_n \end{bmatrix} \tag{6.7}$$

and it is therefore still a matrix in Frobenius form. Then the new coefficients of the closed-loop characteristic polynomial are the opposite of the entries of the last row of the matrix. These elements can be changed by selecting the individual elements $\hat{k}_i$. The problem of the arbitrary eigenvalues assignment can be then reduced to the **assignment of the characteristic polynomial** whose roots are the desired eigenvalues, which is $\hat{p}(s) = \prod_{i=1}^{n}(s - \bar{\lambda}_i)$.

The steps to solve the problem are the following.

1. Compute the coefficients a_0, a_1, ..., a_{n-1} of the characteristic polynomial of matrix A (with no feedback).

2. Choose $\Lambda_c = \{\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_n\}$ and compute the new characteristic polynomial

$$\hat{p}(s) = \prod_{i=1}^{n}(s - \bar{\lambda}_i) = \bar{a}_0 + \bar{a}_1 s + \dots + \bar{a}_{n-1}s^{n-1} + s^n$$

3. Compute $\hat{k} = [\,\hat{k}_1\,\hat{k}_2\,\dots\,\hat{k}_n]$, whose elements are derived as follows

$$\begin{aligned} \bar{a}_0 &= a_0 - \hat{k}_1 \Rightarrow \hat{k}_1 = a_0 - \bar{a}_0 \\ \bar{a}_1 &= a_1 - \hat{k}_2 \Rightarrow \hat{k}_2 = a_1 - \bar{a}_1 \\ &\dots \\ \bar{a}_{n-1} &= a_{n-1} - \hat{k}_n \Rightarrow \hat{k}_n = a_{n-1} - \bar{a}_{n-1} \end{aligned}$$

4. Return to the original form applying the inverse transformation: $K = \hat{K}T^{-1}$.[5]

If $m > 1$, the proof is more involved. The idea is to use one input at the time and proceed as follows.

- If the system is reachable from a single input, for instance u_1, then there is nothing to prove. The closed loop matrix will be $A + B_1K_1$ where B_1 is the first column of B.

- Alternatively, we assign all of the eigenvalues of the subsystem that are reachable from the first input, achieving the matrix $A^* + B_1K_1$. The system becomes

$$(A^* + B_1K_1, [B_2 \dots B_m]) = (A^{(1)}, B^{(1)})$$

  where A_1 has the eigenvalues already assigned. The first input is no longer considered, namely, we set $m := m - 1$, $A := A^* + B_1K_1$, $B := [B_2 \dots B_m]$.

- If there are still eigenvalues to assign, then go back to the previous step with $m - 1$ inputs, in order to assign all of the eigenvalue we can (by means of some K_2) and leave unchanged those already assigned. Otherwise STOP.

[5] This step is necessary because Frobenius form is only a tool to facilitate the solution of the problem, so we cannot apply $\hat{k}$ directly.

- The final closed-loop matrix will be

$$K = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_m \end{bmatrix}$$

Example 6.2. *Consider*

$$A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad B = \left[\begin{array}{c|c} 1 & -1 \\ 1 & -1 \end{array}\right]$$

The system is reachable (why?), but not by means of a single input. A has eigenvalues 0 *and* −2. *We wish to assign the eigenvalues* −4 *and* −4.

Consider the first input. 0 *is reachable with* B_1, *while* −2 *is not (use Popov tho check). So let us move* 0 *to* −4. *With the feedback matrix* $K_1 = [-2 \ \ -2]$ *we get*

$$A + B_1K_1 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} [-2 \ \ -2] = \begin{bmatrix} -3 & -1 \\ -1 & -3 \end{bmatrix}$$

having eigenvalues −4 *and* −2.

Consider the second input. −4 *is not reachable while* −2 *is reachable (use Popov to check). So let us move* −2 *to* −4 *in the* ***new*** *A (*−4 *remains unchanged!)*

$$A + B_1K_1 + B_2K_2 = \begin{bmatrix} -3 & -1 \\ -1 & -3 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} [\,1 \ \ -1] = \begin{bmatrix} -4 & 0 \\ 0 & -4 \end{bmatrix}$$

The eigenvalues are the desired ones. The overall feedback matrix is

$$K = \left[\begin{array}{cc} -2 & -2 \\ \hline 1 & -1 \end{array}\right]$$

Please check![6]

In general it can be proved that, if the system is reachable, the procedure is successful. These considerations are valid from a mathematical point of view. In practice, this procedure might not distribute the work among the different actuators in a proper way. For instance, if the system is reachable from the first input, the procedure would provide a feedback which uses the first input only. This multi-input control problem will be reconsidered in the optimal control theory.

6.2.1 A simple algorithm for eigenvalue assignment

Now we suggest a very simple algorithm for eigenvalue assignment for a reachable system $\sum(A,\, B)$, with $m = 1$ and matrix B having the form

$$B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{6.8}$$

[6] This is the English "please", which basically means that the reader is strongly invited to check...

We consider the unknown K and we write the characteristic polynomial of the feedback system

$$\begin{aligned}
&\det(sI-(A+BK)) = \\
&= \det\begin{bmatrix} s-a_{11}-b_1k_1 & -a_{12}-b_1k_2 & \dots & -a_{1n}-b_2k_n \\ -a_{21}-b_2k_1 & s-a_{22}-b_2k_2 & \dots & -a_{2n}-b_2k_n \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1}-b_nk_1 & -a_{n2}-b_nk_2 & \dots & s-a_{nn}-b_nk_n \end{bmatrix} = \\
&= a_0(k_1,k_2,\dots,k_n)+a_1(k_1,k_2,\dots,k_n)s+\dots+a_{n-1}(k_1,k_2,\dots,k_n)s^{n-1}+s^n
\end{aligned} \tag{6.9}$$

The characteristic polynomial coefficients are functions of the unknown elements k_i. It turns out that these functions are **affine**.[7] Now we simply match the coefficients of the desired polynomial, a_i, with those desired after feedback, $\bar{a}_k$: $\bar{p}(s) = \bar{a}_0 + \bar{a}_1 s + \bar{a}_2 s^2 \cdots + s^n$. Hence

$$\begin{cases} a_0(k_1,k_2,\dots,k_n) = \bar{a}_0 \\ a_1(k_1,k_2,\dots,k_n) = \bar{a}_1 \\ \dots \\ a_{n-1}(k_1,k_2,\dots,k_n) = \bar{a}_{n-1} \end{cases} \tag{6.10}$$

The solution of this linear system provides the elements k_i, hence solving the state feedback problem.

Observation 6.1. *The reachability test has to be performed before the algorithm. If the system is not reachable, then the matrix of the system is singular. In this case, we have a solution only if we choose a polynomial $\bar{p}(s)$ that has the non-reachable eigenvalues among its roots.*

Example 6.3. *Consider the unstable system described by the equations:*

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Then the closed-loop characteristic polynomial is

$$\det(sI-(A+BK)) = \det\begin{bmatrix} s-2-k_1 & -1-k_2 \\ -1 & s-2 \end{bmatrix} = s^2-(4+k_1)s+3+2k_1-k_2$$

Suppose that the desired new eigenvalues are $\bar{\lambda}_1 = -1$ and $\bar{\lambda}_2 = -2$. Then, the characteristic polynomial of the closed-loop system should match

$$\hat{p}(s) = (s+1)(s+2) = s^2+3s+2$$

We get the following linear system:

$$\begin{cases} -4-k_1 = 3 \\ 3+2k_1-k_2 = 2 \end{cases} \Rightarrow \begin{cases} k_1 = -7 \\ k_2 = -13 \end{cases}$$

6.3 State observer

A state observer is a fundamental device when the state variables are not all accessible, or are not measured for economic reasons (more sensors imply additional cost). Consider the scheme in Figure 6.5.

[7] An affine function is the sum of a constant and a linear function: *e.g.*, $2k_1 + 3k_2 + 1$ is affine.

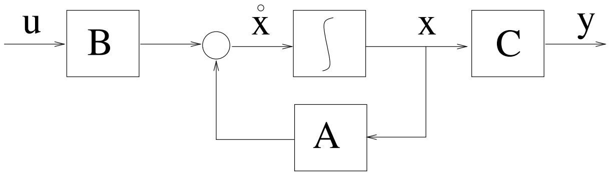

Figure 6.5: General scheme of a system.

The most favourable case that can occur is when C is the identical matrix (or has full row rank): this means that the outputs are precisely the state variables (or that the state variables can be immediately computed based on the outputs). Then, we can straightforwardly apply a state feedback controller. However matrix C, in general, has less rows than columns (often it has a single row). Hence, we have to proceed differently.

A first idea (which will not lead us to the solution) could be to replicate the system by implementing a device with the same equations, namely, with the same matrices of the process:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) \tag{6.11}$$

This is a **trivial observer (open-loop)**, in which $\hat{x}(t)$ is an estimate of the system state variables. To evaluate the effectiveness of this solution (or of any other estimator), we monitor the estimation error

$$e(t) = \hat{x}(t) - x(t)$$

Since the system is described by the equations

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

we can subtract the state equation from the estimated state equations. We obtain

$$\frac{d}{dt}(\hat{x}(t) - x(t)) = A(\hat{x}(t) - x(t)) + Bu(t) - Bu(t) = A(\hat{x}(t) - x(t)) \tag{6.12}$$

The estimation error is then described by the equation

$$\dot{e}(t) = Ae(t) \tag{6.13}$$

and therefore it evolves in natural response. We immediately understand that this open-loop observer is not applicable to unstable systems, because the error diverges. Even for asymptotically stable systems, trivial observers do not provide good results. Indeed, the state $\hat{x}(0)$ must be initialised in some way and cannot be initialised exactly to $x(0)$, which is unknown. In the absence of specific information, for the symmetry of the problem, the most obvious choice is to set $\hat{x}(0) = 0$. This means that $e(0) = \hat{x}(0) - x(0) = -x(0)$. Assuming, for example, $u(t) = 0$, we obtain

$$\begin{aligned} \dot{x} &= Ax(t), \\ \dot{e}(t) &= Ae(t), \end{aligned} \tag{6.14}$$

with $e(0) = -x(0)$, so the error has opposite initial conditions with respect to the state. By linearity, we get

$$e(t) = -x(t)$$

Then, the error (shown in Figure 6.6) has the same amplitude of the state $\|x(t)\| = \|e(t)\|$: the relative error is 100 %. We would like to design an observer able to provide an estimate of the state that

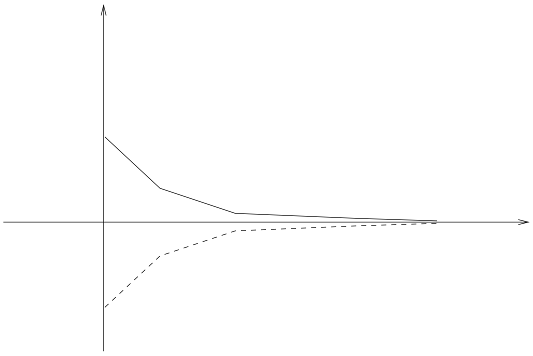

Figure 6.6: Evolution of the state (solid line) and of the error (dotted line) with the trivial observer.

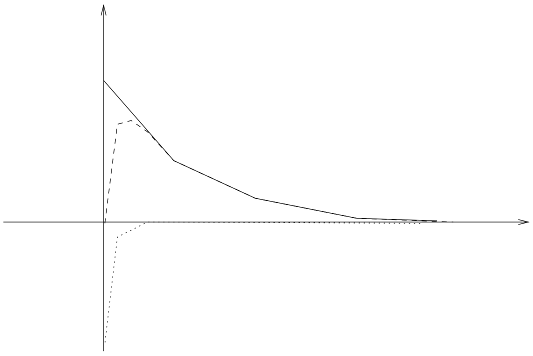

Figure 6.7: Evolution of the state (solid line), the estimated state (dashed line) and of the error (dotted line) with the desired observer.

converges to $x(t)$, also in the case of unstable systems. If the system is stable, we would like the observer to have a faster convergence than the natural convergence speed assured by the system modes, as qualitatively shown in Figure 6.7.

Then, the right solution is given by the **Luenberger observer**. The trivial observer uses information about $u(t)$, but not about $y(t)$. On the contrary, the Luenberger observer uses information about both the input and the output of the process. The differential equation that governs the state estimate is then

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) - LC(\hat{x}(t) - x(t)), \tag{6.15}$$

where L is a matrix to be determined. Note that the difference is the correction term

$$LC(\hat{x}(t) - x(t)) = L[\hat{y}(t) - y(t)]$$

The quantity $\hat{y}(t)$ is the estimated output, which is compared with the true output $y(t)$ and fed back. This device is often called **feedback observer**. Note that the state estimation error $\hat{x}(t) - x(t)$ is unknown, but the output estimation error $\hat{y}(t) - y(t) = C(\hat{x}(t) - x(t)) = C\hat{x}(t) - y(t)$ is known on-line, because $\hat{x}(t)$ is the known observer state and $y(t)$ is the measured output. For the observer implementation, we use the equivalent equation

$$\dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + Bu(t) + Ly(t) \tag{6.16}$$

Matrix L, having size $n \times p$, contains the free design parameters. For $L = 0$ we recover the trivial observer. The general scheme for the Luenberger observer is represented in Figure 6.8.

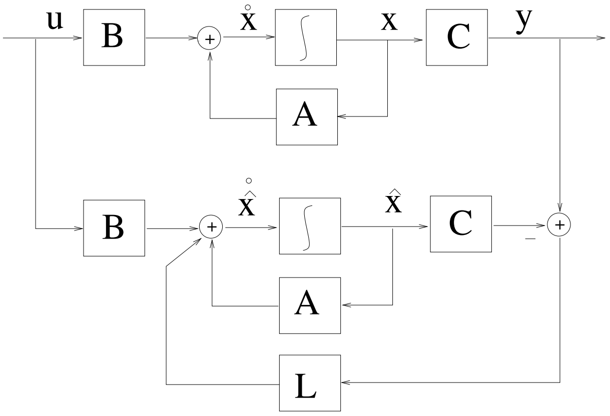

Figure 6.8: General scheme of a Luenberger observer.

To see how this new device works, we compute the error expression, subtracting the state derivative from the estimated state derivative:

$$\begin{aligned}\frac{d}{dt}(\hat{x}(t) - x(t)) &= (A - LC)\hat{x}(t) + Bu(t) + Ly(t) - Ax(t) - Bu(t) = \\ &= (A - LC)\hat{x}(t) - Ax(t) + LCx(t) = \\ &= (A - LC)\hat{x}(t) - (A - LC)x(t) \\ &= (A - LC)(\hat{x}(t) - x(t))\end{aligned}$$

and we get

$$\dot{e}(t) = (A - LC)e(t) \tag{6.17}$$

Therefore, the error does not depend on $u(t)$: it evolves in natural response according to the modes of matrix $(A - LC)$. Hence, the role of matrix L is similar to that of matrix K in a state feedback controller. In analogy to the problem of eigenvalue placement via state feedback, we would like to impose a set of eigenvalues Λ_o (O stands for observer) to matrix $(A - LC)$, so as to ensure fast converge of the error to zero. More formally, the problem is as follows.

Problem 6.2. Observer eigenvalues assignment *Given the set of complex numbers*

$$\Lambda_o = \{\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n\}$$

satisfying CEC (conjugate eigenvalues constraint), find matrix L so that the set of eigenvalues of $A - LC$ is

$$\sigma(A - LC) = \Lambda_o$$

Then we have the following result.

Theorem 6.4. *The set Λ_o (satisfying CEC) can be arbitrarily assigned if and only if (A, C) is observable. If the system is not observable, then Λ_o can be assigned if and only if it contains all of the eigenvalues of the unobservable subsystem.*

Proof. To prove the theorem, we exploit duality. Consider then the dual of the system that provides matrix $(A - LC)$:

$$\begin{aligned}A^* &= A^\top \\ B^* &= C^\top \\ K^* &= -L^\top\end{aligned} \tag{6.18}$$

The dual of matrix $(A - LC)$ is

$$(A - LC)^\top = (A^\top - C^\top L^\top) = (A^* + B^* K^*), \tag{6.19}$$

where $K^* \doteq -L$.

Then assigning the spectrum of $(A - LC)$ is equivalent to assigning the spectrum of $(A^* + B^* K^*)$, a problem we have already studied and solved. The proof follows by duality since observability of a system is equivalent to reachability of the dual system. Then, the problem is solvable with arbitrary Λ_o if and only if the dual system is reachable, hence, if and only the primal is observable. In the presence of unobservable modes, these correspond to unreachable modes in the dual, and the last part of the proof follows. □

The analogy between state feedback design and Luenberger observer design, due the duality, allows us to use the same algorithms to derive the matrices K and L. In fact, if we have any procedure that, given the input data A, B, Λ_c, returns matrix K of a state feedback controller, we can use the same procedure for observer design: by providing the input $A^\top$, $C^\top, \Lambda_o$, we obtain matrix K^* (the state feedback matrix for the dual system) and finally $L = (-K^*)^\top$.

- **State feedback gain**: $Procedure[A, B, \Lambda_c] \to K$
- **Observer gain**: $Procedure[A^\top, C^\top, \Lambda_o] \to K^* = -L^\top$

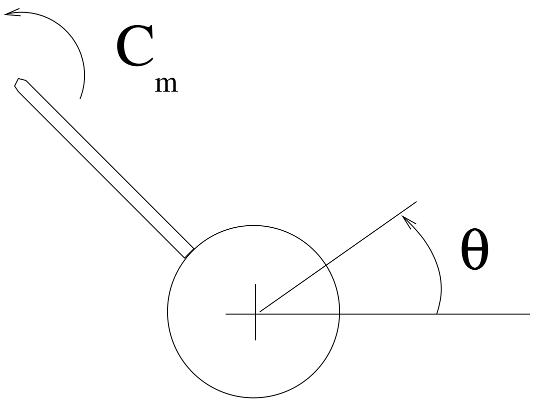

Figure 6.9: Positioning system.

Example 6.4. *Consider a servo system as in Figure 6.9. Assuming a reference angle $\theta = 0$, let us solve the problem of bringing the system to the zero state $\theta = \dot{\theta} = 0$. The equation describing this system is*

$$J\ddot{\theta}(t) = C_m(t) \Rightarrow \ddot{\theta}(t) = \frac{C_m(t)}{J} = u(t)$$

If we set $x_1(t) = \theta(t)$ and $x_2 = \dot{\theta}(t)$, we obtain a state representation with matrices

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Suppose we have sensors for both the position and the speed. We can implement a state feedback

$$u(t) = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

and we get the following closed-loop matrix

$$(A + BK) = \begin{bmatrix} 0 & 1 \\ k_1 & k_2 \end{bmatrix}$$

which is already in Frobenius form. Then, we just have to require that $-k_1$ *and* $-k_2$ *are equal to the coefficients of the desired characteristic polynomial. If we choose* $p(s) = (s+1)(s+2) = s^2 + 3s + 2$, *we immediately get* $k_1 = -2$ *and* $k_2 = -3$.

Employing both position and speed sensors is not convenient from an economic point of view. We can then expect to have a single sensor, for just one of the state variables, along with an observer estimating the variable that is not measured.

If we adopt a speed sensor, matrix C is

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

By means of a simple calculation, we see that the system is not observable (as is known, from the speed it is not possible to determine the position). In addition, the eigenvalue of the non-observable subsystem is zero, hence it is not asymptotically stable and the spectrum of $(A - LC)$ *must necessarily contain the zero eigenvalue. This is not good news, because the error system would not be asymptotically stable.*

If, instead, we consider the position sensor, matrix C takes the form

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

and in this case the observability matrix has full rank. We can then compute $(A - LC)$*:*

$$(A - LC) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} -l_1 & 1 \\ -l_2 & 0 \end{bmatrix}$$

The characteristic polynomial is:

$$\det(sI - A + LC) = \det \begin{bmatrix} s + l_1 & -1 \\ l_2 & s \end{bmatrix} = s^2 + l_1 s + l_2$$

Suppose that the desired eigenvalues are $\bar{\lambda}_3 = -3$ *and* $\bar{\lambda}_4 = -4$. *We have* $p(s) = s^2 + 7s + 12$, *hence* $l_1 = 7$ *and* $l_2 = 12$.

Remark 6.1. *The observer generates an estimate of vector* $\hat{x}(t)$ *whose components are the estimated position (which we already have as an output, thanks to the sensor) and the estimated speed. Hence, we could avoid estimating the position, since it is already measured, and estimate the speed only by means of a* reduced order observer, *which is not treated in the present course for two reasons. First, in view of the current technology, the complexity of the full order observer is not a problem in terms of implementation. Second, using the estimated position instead of the measured one has the advantage that the observer has a filtering (noise-suppressing) action.*

The further step for the control of the servo system is to feed the estimated state back:

$$u(t) = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} \hat{x}_1(t) \\ \hat{x}_2(t) \end{bmatrix}$$

Intuitively, this idea should work because, after all, the estimated state converges to the true one. We will prove that this scheme is successful in the following.

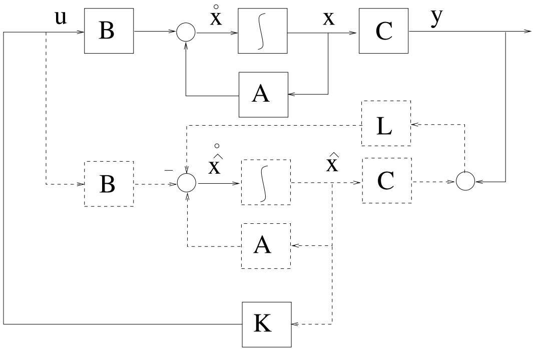

Figure 6.10: Block diagram of an observer-based controller.

6.4 Synthesis of regulators

After designing state feedback and state estimator (the latter is necessary if some state variables are not measured, that is, C is not the identity matrix), the next step is the overall (observer-based) controller synthesis. This is based on the feedback of the state estimate, as shown in Figure 6.10.

It is intuitive enough that everything is working properly. However, to verify the effectiveness of this strategy, we analyse the obtained overall system. The process equations are:

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

while those of the observer and the controller are

$$\begin{cases} \dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + Bu(t) + Ly(t) \\ u(t) = K\hat{x}(t) \end{cases}$$

The overall system has $n + n = 2n$ state variables (those of the process, together with those we introduce with the observer-based controller).

$$x_{overall}(t) = \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} \in \mathbb{R}^{2n} \tag{6.20}$$

The closed-loop matrix is

$$A_{CL} = \left[\begin{array}{c|c} A & BK \\ \hline LC & A - LC + BK \end{array}\right]$$

To carry out the computations, a state transformation is useful to obtain a state vector that includes both process state and estimated error. Mathematically, we have:

$$\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \left[\begin{array}{c|c} I & 0 \\ \hline -I & I \end{array}\right] \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} \tag{6.21}$$

The transformation matrix from the new to the old representation is then:

$$T = \left[\begin{array}{c|c} I & 0 \\ \hline I & I \end{array}\right] \qquad T^{-1} = \left[\begin{array}{c|c} I & 0 \\ \hline -I & I \end{array}\right] \tag{6.22}$$

(T is invertible). The overall system is governed by the following equations:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \\ \dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + Bu(t) + Ly(t) \\ u(t) = K\hat{x}(t) \end{cases} \tag{6.23}$$

Applying the state transformation (*i.e.*, manipulating the equations, with $\hat{x}(t) = x(t) + e(t)$) we get

$$\begin{cases} \dot{x}(t) = (A + BK)x(t) + BKe(t) \\ \dot{e}(t) = (A - LC)e(t) \end{cases} \tag{6.24}$$

In matrix form, we have

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \underbrace{\left[\begin{array}{c|c} A + BK & BK \\ \hline 0 & A - LC \end{array}\right]}_{=A_{cl}} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} \tag{6.25}$$

Matrix A_{CL} (CL means **Closed Loop**) is block triangular of size $2n \times 2n$. Therefore, its spectrum is equal to the union of the spectra of its diagonal blocks:

$$\sigma(A_{CL}) = \sigma(A + BK) \bigcup \sigma(A - LC) \tag{6.26}$$

Under reachability and observability assumptions, these spectra can be arbitrarily assigned. Hence, we can assign all of the eigenvalues of A_{CL}, as follows:

1. design a state feedback that assigns the eigenvalues $\sigma(A + BK) = \Lambda_c$;
2. design an observer for state estimation that assigns the eigenvalues $\sigma(A - LC) = \Lambda_o$;
3. design the overall controller as the feedback of the estimated state

$$\begin{aligned} \dot{\hat{x}}(t) &= (A - LC)\hat{x}(t) + Bu(t) + Ly(t) \\ u(t) &= K\hat{x}(t) \end{aligned}$$

Observation 6.2. *The assignment of the eigenvalues of the state feedback and of the observer is performed in two independent stages. This independence is called* ***separation principle****.*

The procedure leaves us complete freedom on how to choose the eigenvalues in Λ_c and Λ_r. As a guideline, the convergence of the observer must often be faster than the convergence of the system modes. Therefore, the negative real part of the eigenvalues of $(A - LC)$ is usually required to be larger (in magnitude) than the real part the eigenvalues of $(A + BK)$ (typically 4 times larger).

The observer-based controller is thus governed by the equations:

$$\begin{cases} \dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + Bu(t) + Ly(t) \\ u(t) = K\hat{x}(t) \end{cases}$$

Replacing the second equation in the first yields a useful mathematical simplification that is very important from a physical standpoint:

$$\begin{cases} \dot{\hat{x}}(t) = (A - LC + BK)\hat{x}(t) + Ly(t) \\ u(t) = K\hat{x}(t) \end{cases} \tag{6.27}$$

In fact, we notice that we do not need to have sensors capable of measuring the input $u(t)$, in order to send it to the observer, because the input itself is generated as an output of the regulator system. The regulator matrices are

$$(F, G, H, K) = (A - LC + BK, L, K, 0)$$

hence, the regulator is strictly proper.

The controller thus obtained can be implemented in practice by means of an analog circuit, or, much more realistically, can be implemented digitally on a computer, by exploiting the equivalent discrete-time system having matrices

$$\begin{aligned} F_D &= e^{FT} = e^{(A-LC+BK)T} \\ G_D &= \int_0^T e^{F\sigma} L d\sigma = \int_0^T e^{(A-LC+BK)\sigma} L d\sigma \\ H_D &= K \end{aligned}$$

For a computer implementation, we can adopt an algorithm of the following type.

1. Fix T and compute F_D, G_D, H_D.
2. Initialise $\hat{x}(0)$.
3. If $ONLINE = ON$, start.
4. Compute $u := H_D\hat{x}$.
5. Write the value of u in a memory location that will be read from the D/A converter.
6. Read the value of y written in a memory location by the A/D converter.
7. Compute the updated estimated state $\hat{x}^+ = F_D\hat{x} + G_D y$.
8. Set $\hat{x} = \hat{x}^+$.
9. Return to step 4 unless $ONLINE = OFF$.
10. Quit.

A final consideration concerns the effect of the observer. If we cannot completely assign the eigenvalues via state feedback, we need an observer. Hence, we need to assign additional modes: these affect the system with their transient (which is typically fast). To analyse the problem, consider an input of the form

$$u(t) = K\hat{x}(t) + v(t), \tag{6.28}$$

where $v(t)$ is a reference signal (often constant). It can be seen that the transfer function between $v(t)$ and $y(t)$ is the same in both the state feedback case and the estimated state feedback case, because the error e is not a reachable variable (this is immediate, since e evolves as $\dot{e} = (A - LC)e$).

Example 6.5. *Reconsider the previous example, where we have obtained the parameters of the matrices K and L for the servo-system. The controller implementation is immediate:*

$$\begin{aligned} F &= A + BK - LC = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} -2 & -3 \end{bmatrix} - \begin{bmatrix} 7 \\ 12 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} -7 & 1 \\ -14 & -3 \end{bmatrix} \\ G &= L = \begin{bmatrix} 7 \\ 12 \end{bmatrix} \\ H &= K = \begin{bmatrix} -2 & -3 \end{bmatrix} \end{aligned}$$

Then, the resulting modes of the overall feedback system correspond to

*a) the eigenvalues $\lambda_1 = -1$, $\lambda_2 = -2$: **modes of the regulator;***

*b) the eigenvalues $\lambda_3 = -3$, $\lambda_4 = -4$: **modes of the observer**.*

A final remark concerns the discrete-time systems case. The discrete-time observer-based control writes as

$$\begin{array}{rcl} u(k) & = & Kx(k) \\ \hat{x}(k+1) & = & (A-LC)\hat{x}(k)+Bu(k)+Ly(k) \end{array} \tag{6.29}$$

As in the continuous-time system case, the error is governed by the equation:

$$e(k+1)=(A-LC)e(k) \tag{6.30}$$

Once again, the theory for discrete-time systems is parallel to that for continuous-time systems. The only caution is that the eigenvalues must be assigned with the condition $|\bar{\lambda}_k|<1$ so as to ensure stability of the modes.

6.5 External disturbances

Any system is affected by disturbances. A noisy system can be described by equations of the form

$$\dot{x}(t) = Ax(t)+Bu(t)+Ed(t) \tag{6.31}$$

$$y(t) = Cx(t)+w(t) \tag{6.32}$$

where we have two external inputs (exogenous disturbances)

- $d(t)$, called process noise;
- $w(t)$, called measurement noise.

Process noise is typically an external signal affecting the system evolution, such as turbulence in aeronautics, voltage fluctuations in electrical circuits, external torque in mechanical systems. Measurement noise is due to the sensors (quantisation or electric noise). In this case, the error equation becomes

$$\dot{e}(t)=(A-LC)e(t)-Ed(t)+Lw(t) \tag{6.33}$$

If we have a bound for the noises

$$\|w(t)\|<\beta, \qquad \|d(t)\|<\delta$$

in general we still can achieve an error bound of the type

$$\|e(t)\|<\gamma$$

provided that $A-LC$ is asymptotically stable.

The overall controlled system is

$$\frac{d}{dt}\left[\begin{array}{c} x(t) \\ z(t) \end{array}\right]=\left[\begin{array}{c|c} A & BK \\ \hline LC & A-LC \end{array}\right]\left[\begin{array}{c} x(t) \\ z(t) \end{array}\right]+\left[\begin{array}{c|c} E & 0 \\ \hline -0 & L \end{array}\right]\left[\begin{array}{c} d(t) \\ w(t) \end{array}\right] \tag{6.34}$$

or, if we use e instead of z

$$\frac{d}{dt}\left[\begin{array}{c} x(t) \\ e(t) \end{array}\right]=\left[\begin{array}{c|c} A+BK & BK \\ \hline 0 & A-LC \end{array}\right]\left[\begin{array}{c} x(t) \\ z(t) \end{array}\right]+\left[\begin{array}{c|c} E & 0 \\ \hline -E & L \end{array}\right]\left[\begin{array}{c} d(t) \\ w(t) \end{array}\right] \tag{6.35}$$

It interesting to notice that there is a trade-off. If we attempt to assign the eigenvalues of $A+BK$ with very large negative parts to have a fast convergence, this requires a large K and results in a strong actuator exploitation, because $u=Kx$. If we attempt to assign the eigenvalues of $A-LC$ with very large negative part, this requires a large L and we amplify the effect of measurement noise.

6.6 Examples

Example 6.6. ***(Two tank system.)*** *As we will see in the examples chapter, the linearised equation of the two-tank system has matrices*

$$A = \begin{bmatrix} -p & p \\ p & -(p+\delta) \end{bmatrix} \quad B = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

The system poles are $-2p \pm \sqrt{4p^2 - 4p\delta}$. *Since* δ *is positive and small, one (negative) root is quite close to* 0.

To speed up the system, we can assign the eigenvalues: in particular, we can move the "slow" pole far from the imaginary axis (to the left).

$$\det[sI - A + BK] = \det\begin{bmatrix} s+p-k_1 & -p-k_2 \\ -p & s+(p+\delta) \end{bmatrix} = s^2 + (+2p - \delta k_1)s - (p+\delta)k_1 - pk_2 + \delta p$$

If the polynomial $\bar{p}(s) = s^2 + \bar{p}_1 s + \bar{p}_0$ *is to be assigned, we get the equations*

$$\begin{aligned} (+2p - \delta k_1) &= \bar{p}_1 \\ -(p+\delta)k_1 - pk_2 + \delta p &= \bar{p}_0 \end{aligned}$$

which the reader is invited to solve.

Example 6.7. ***(Electrical machine control.)*** *Consider the following model of a direct-current electrical machine*

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} -\alpha & -\beta & 0 \\ \gamma & -\delta & 0 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} \epsilon \\ 0 \\ 0 \end{bmatrix} u(t)$$

where $x_1(t) = i_a$, $x_2(t) = \omega$ *and* $x_3(t) = \varphi$ *are the armature current, the angular speed and the angle, respectively, and where the input* $u(t) = v_a(t)$ *is the armature voltage. We assume that* $y(t) = x_3(t) = \varphi$ *is the measured output, while* α, β, γ, δ *and* ϵ *are positive constants. Observability and reachability have been tested for this system. To assign the eigenvalues, we write the matrix*

$$[sI - (A + BK)] = \begin{bmatrix} s+\alpha-\epsilon k_1 & \beta-\epsilon k_2 & -\epsilon k_3 \\ -\gamma & s+\delta & 0 \\ 0 & -1 & s \end{bmatrix}$$

Its determinant is

$$s^3 + [\alpha - \epsilon k_1 + \delta]s^2 + [\alpha\delta - \delta\epsilon k_1 + \gamma\beta - \gamma\epsilon k_2]s + \epsilon k_3\gamma$$

Given the desired characteristic polynomial for the state feedback $s^3 + \bar{a}_2 s^2 + \bar{a}_1 s + \bar{a}_0$, *the solution is found by equating term by term*

$$\alpha - \epsilon k_1 + \delta = \bar{a}_2, \quad \alpha\delta - \delta\epsilon k_1 + \gamma\beta - \gamma\epsilon k_2 = \bar{a}_1, \quad \epsilon k_3\gamma = \bar{a}_0$$

therefore $k_1 = (\alpha + \delta - \bar{a}_2)\epsilon$, $k_3 = (\bar{a}_0/\gamma)\epsilon$ *and* $k_2 = (\alpha\delta - \delta(\alpha + \delta - \bar{a}_2) + \gamma\beta - \bar{a}_1)/(\epsilon\gamma)$.

For the state observer we proceed in the dual way and we consider the determinant of matrix

$$[sI - (A - LC)] = \begin{bmatrix} s+\alpha & \beta & l_1 \\ -\gamma & s+\delta & l_2 \\ 0 & -1 & s+l_3 \end{bmatrix}$$

which we equate to the desired observer characteristic polynomial $s^3 + \hat{a}_2 s^2 + \hat{a}_1 s + \hat{a}_0$. *The reader is invited to complete the exercise. The resulting regulator is*

$$\dot{z}(t) = [A + Bk - LC]z + Ly, \qquad u(t) = Kz(t),$$

where $z(t)$ *is a state estimate.*

Example 6.8. ***(Unknown constant signal.)*** *As we have seen, the presence of an external signal or disturbance may prevent the estimation error from converging to zero. This is true in the absence of information on the disturbance. An interesting case is that of constant disturbances.*

Assume that we have the system

$$\dot{x}(t) = Ax(t) + Bu(t) + Ed, \qquad y(t) = Cx(t) \tag{6.36}$$

where d is unknown but constant. This allows us to write an additional equation, $\dot{d} = 0$. *We get*

$$\begin{bmatrix} x(t) \\ d(t) \end{bmatrix} = \begin{bmatrix} A & E \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ d(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u(t) \qquad y(t) = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ d(t) \end{bmatrix}$$

If the augmented system is observable, we can reconstruct the constant signal by means of an observer.[8] *Consider the servo system subject to a constant external torque (*e.g.*, a lift with an unknown load)*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad E = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Then the augmented system is

$$A_{aug} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_{aug} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad C_{aug} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

This system is observable. It is essential to notice that the augmented system is ***not*** *reachable with input B. In fact, by assumption, the external torque is constant and cannot be modified. Still, we can design an observer that estimates position, speed and torque. Then we can use state feedback to control the system position and speed.*

[8] If (A, E) is reachable and (A, C) is observable, then the augmented system is observable. The proof is not easy: it can be done via Popov criterion.

Chapter 7

Nonlinear systems

7.1 General considerations

Most of the systems encountered in nature are nonlinear. Nonlinear systems are much harder to study than linear systems. Indeed, in the case of nonlinear systems, it is in general impossible to find analytic solutions. This is a limitation of system theory. However, systems that cannot be studied analytically can be approached numerically, to obtain an approximated solution of the corresponding differential equations. A powerful tool, which is fundamental in the analysis of nonlinear systems and very important for control purposes, is linearisation. In this chapter, we will describe this approach in detail.

From a conceptual standpoint, a substantial difference between linear and nonlinear system is behavioural homogeneity. For linear systems, in fact, it is possible to consider modal functions and frequency analysis, and properties such as stability, oscillations and exponential decay are a property of a system. For nonlinear system, instead, every property is typically referred to a certain equilibrium point (or, more in general, to a certain working condition, such as a periodic orbit). In other words, a nonlinear system can be stable or not, sensitive to noise or not, oscillating or not (and so on), depending on the equilibrium we are considering. Therefore, the global theory that holds for linear systems cannot, in general, be extended and we most often need to resort to local theories.

Example 7.1. ***(Population evolution.)*** *Population dynamics are interesting to analyse, for instance, when studying colonies of bacteria in a laboratory or animal species in a natural environment. These phenomena can be modeled by more or less complicated equations. We consider a very simple model, which describes the evolution of a single population $x(t)$. In the simplest case, a population evolves according to the following linear equation: $\dot{x}(t) = ax(t)$, where $a > 0$ is a scalar coefficient. This represents the fact that the population grows proportionally to the population itself. The solution is then a classic exponential $x(t) = e^{at}x(0)$. However, this model does not take into account the limited resources of the surrounding environment. A well accepted model that includes the dependence on the available resources is represented by the following* ***nonlinear system****:*

$$\dot{x}(t) = a\left[1 - \frac{x(t)}{C}\right]x(t)$$

Note that, for $x(t) \ll C$, the term $\frac{x(t)}{C}$ is negligible and the model is almost linear. On the other hand, if $x(t)$ increases, the same term represents a limitation for the birth rate, due to the shortage of resources. Even more, if $x(t) > C$, we have a trend inversion, and the number of births is less than the number of deaths.

This is one of the simple and rare cases in which is possible to give an analytical solution, which is the following:

$$x(t) = \frac{e^{at}x(0)}{\left(1 - \frac{x(0)}{C}\right) + \frac{x(0)}{C}e^{at}}$$

For general nonlinear systems, there are few chances of finding analytic solutions like this one.

The evolution of the population with different initial conditions is reported in Figure 7.1.

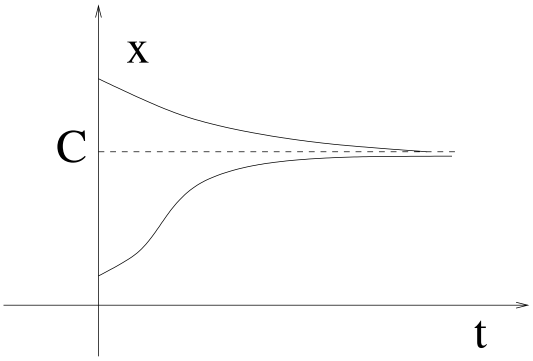

Figure 7.1: Evolution of a population for different initial conditions.

To present the spirit of the investigation of nonlinear systems, we show that the same qualitative considerations that we can derive from the closed-form solution can be drawn without solving the equation, but just analysing it. Indeed, from the equation we immediately see that

- *for* $x(t) > C$, $\dot{x}(t) < 0$*;*
- *for* $x(t) < C$, $\dot{x}(t) > 0$*;*
- *both* $x(t) = C$ *and* $x(t) = 0$ *yield* $\dot{x}(t) = 0$.

Then, there are two equilibria characterised by $\dot{x} = 0$*. Above the upper equilibrium the population decreases, between the two equilibria the population increases, in* 0 *the population remains at* 0*. The solution* $x(t) = C$ *is a stable equilibrium point, because all the solutions with initial condition grater than zero converge to* $x(t) = C$*. Vice versa,* $x(t) = 0$ *is an unstable equilibrium point, because any small perturbation brings the solution far from zero. Actually, what we can see is that all the conditions that we obtained from the analytical solution can be deduced based on a qualitative analysis as well. Such a qualitative analysis is the main subject of this chapter.*

We will consider autonomous continuous-time systems of the form

$$\begin{cases} \dot{x}(t) = f(x(t), u(t)) \\ y(t) = g(x(t), u(t)) \end{cases} \tag{7.1}$$

and autonomous discrete-time systems of the form

$$\begin{cases} x(k+1) = f(x(k), u(k)) \\ y(k+1) = g(x(k), u(k)) \end{cases} \tag{7.2}$$

With few exceptions, we neglect the case where there is a direct time dependence of $\dot{x}(t) = f(x(t), u(t), t)$ or $x(k+1) = f(x(k), u(k, k))$, on time (non-autonomous systems), since it is much more complicated and less important in practice.

7.2 Equilibrium points

Among all possible trajectories of a system, constant trajectories are of particular interest. These are called equilibria. Almost all dynamical systems (mechanical electrical, thermal, aeronautical) we encounter in practice operate at an equilibrium most of the time.

Definition 7.1. *Given a vector $\bar{x} \in \mathbb{R}^n$, we say that it is an* ***equilibrium point*** *for the system (either* (7.1) *or* (7.2)*) if there exists $\bar{u}$, called* ***equilibrium input****, such that, for $x(0) = \bar{x}$ and $u(t) = \bar{u}$, we have that $x(t) \equiv \bar{x}$. In this case, the pair $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathbb{R}^m$ is called* ***equilibrium pair****.*

The equilibrium conditions in a nonlinear system are easily obtained. Indeed, we have that, if $x(t) \equiv \bar{x}$ and $u(t) = \bar{u}$, then it is also true that $\dot{x} = 0$ for continuous-time systems, and $x(k+1) = x(k)$ for discrete-time systems. Then, the equilibrium conditions are

$$f(\bar{x}, \bar{u}) = 0 \tag{7.3}$$

the continuous-time systems and

$$\bar{x} = f(\bar{x}, \bar{u}) \tag{7.4}$$

for discrete-time systems. The equilibrium pairs are therefore the vector pairs that satisfy one of the above conditions (depending on the continuous- or discrete-time nature of the system).

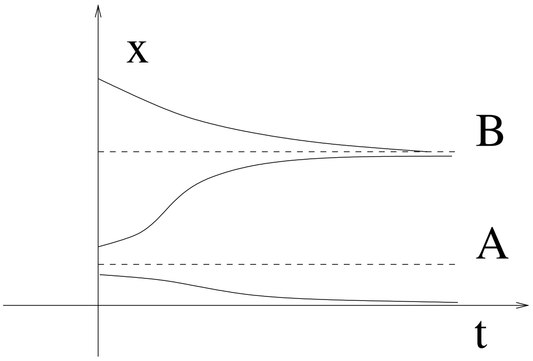

Figure 7.2: Evolution of a population with external input. Line A: extinction threshold; line B: survival threshold.

Example 7.2. ***(Population evolution, continued.)*** *Consider again the equation for population evolution in the previous example. We can easily find out that the equilibrium points are*

$$\left[1 - \frac{\bar{x}}{C}\right]\bar{x} = 0 \Rightarrow \bar{x} = 0,\ \bar{x} = C$$

Is is interesting to observe what happens if we add to the equation a term that represents an external input (for instance hunting or predation):

$$\left[1 - \frac{\bar{x}}{C}\right]\bar{x} + \bar{u} = 0$$

where $\bar{u} < 0$. In this case, the equilibrium points are the solutions of the equation above for a given $\bar{u}$:

$$\bar{x}_{1,2} = \frac{C}{a}\left[\frac{a}{2} \pm \sqrt{\frac{a^2}{2} + \frac{a}{C}\bar{u}}\right]$$

The new situation is depicted in Figure 7.2.

We can observe that, for initial conditions greater than $\bar{x}_2$, the death rate is larger than the birth rate, due to resource shortage. In the range between $\bar{x}_1$ and $\bar{x}_2$, repopulation occurs, while for initial conditions below $\bar{x}_1$ we have the species extinction. The extinction threshold $\bar{x}_1$ has been introduced by the presence of a negative external input. Moreover, $\bar{u}$ can have a different nature: if it is related to predation, then it is for sure influenced by $x(t)$. This typically requires the introduction of a new variable representing the predator population (prey-predator models). Conversely, if $\bar{u}$ is related to hunting, it will be an independent term.

Example 7.3. ***(Euler discrete-time system.)*** *The Euler approximating system of a continuous-time system is obtained by approximating the derivative $\dot{x}(t)$ as an incremental ratio:*

$$\dot{x}(t) \simeq \frac{x(t+\tau) - x(t)}{\tau} \tag{7.5}$$

By replacing this expression in the general formula of the continuous-time system and evaluating it only for $t = 0, \tau, 2\tau, \dots$, we obtain the discrete-time system:

$$x(t+\tau) = x(t) + \tau f(x(t), u(t)) \tag{7.6}$$

This method, known as ***Euler explicit method****, represents a good approximation if the step τ is small enough, while, if τ is too large, the discrete-time solution may not converge to the continuous-time solution.*

To assess the effectiveness of this method, we can check if the equilibrium conditions of the Euler system correspond to those of the original continuous-time system. Indeed, it is very easy to check that the equilibrium points of the approximating discrete-time system are the same of the continuous-time system:

$$\bar{x} = \bar{x} + \tau f(\bar{x}, \bar{u}) \Leftrightarrow f(\bar{x}, \bar{u}) = 0$$

Obviously, nothing can be said about the stability. In fact, stable equilibrium points of the continuous-time system could be unstable for the discrete-time approximation (in particular if $\tau > 0$ is not small enough).

Additional examples about equilibrium points will be presented in further sections.

Note that, for continuous-time linear systems, the equilibrium condition is

$$0 = A\bar{x} + B\bar{u},$$

namely, the equilibrium pairs are the kernel of matrix $[A \ B]$. The reader is invited to find out the equilibrium condition for discrete-time linear systems.

7.3 Lyapunov functions

A nonlinear system can have several equilibrium points, as said. We are interested in analysing their **local stability**, namely, in understanding what happens for small perturbations around the equilibrium point: does the system return to the equilibrium or does the trajectory escape? We stress (this concept will be repeated) that for nonlinear system stability is referred to an equilibrium of the system, and not to the system itself. Unless we consider very special cases, a nonlinear system can have both stable and unstable equilibria. Some examples of stable and unstable equilibrium points of the same "rolling sphere" system are shown in Figure 7.3.

Consider a nonlinear system admitting an equilibrium point characterised by the condition

$$0 = f(\bar{x}, \bar{u}) \tag{7.7}$$

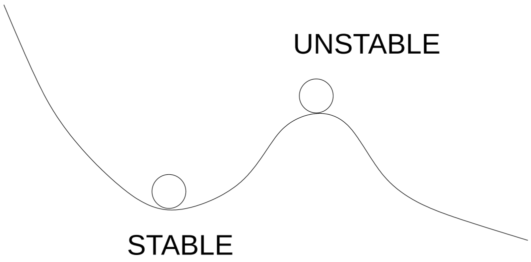

Figure 7.3: Stable and unstable equilibrium points for a rolling sphere.

We initially assume that $\bar{u}$ is fixed, so we can study the system:

$$\dot{x}(t) = f(x(t), \bar{u}) = F_{\bar{u}}(x(t)) \tag{7.8}$$

which is an autonomous system without inputs, with equilibrium point in $\bar{x}$. We can apply a coordinate shift: we introduce then the new variable $z(t) = x(t) - \bar{x}$ and we get

$$\dot{z}(t) = F_{\bar{u}}(x(t)) = F_{\bar{u}}(z(t) + \bar{x}) \tag{7.9}$$

With this new definition, the equilibrium point is $\bar{z} = 0$. We define $G_{\bar{x},\bar{u}}(z(t)) = F_{\bar{u}}(z(t) + \bar{x})$. Therefore can study the following system:

$$\dot{z}(t) = G_{\bar{x},\bar{u}}(z(t)) \tag{7.10}$$

with equilibrium point in $0 = G_{\bar{x},\bar{u}}(0)$. Thanks to a suitable change of variables (as done above), we can always bring us back to the study of autonomous systems without inputs, of the form

$$\dot{x}(t) = f(x(t)), \quad f(0) = 0.$$

In the sequel we will be also concerned with the stabilisation problem: we will use a control that stabilises the equilibrium (if necessary). In this case, a change of variable is applied to the input as well:

$$v(t) = u(t) - \bar{u}$$

Hence,

$$\dot{x}(t) = \dot{z}(t) = f(z(t) + \bar{x}, v + \bar{u}) = G_{\bar{x},\bar{u}}(z(t), v(t))$$

and

$$\dot{z}(t) = G_{\bar{x},\bar{u}}(z(t), v(t)) \tag{7.11}$$

is the equation we use for stabilisation. Again, the stabilisation problem can be approached by considering a system for which $(0, 0)$ is an equilibrium pair

$$\dot{x}(t) = f(x(t), u(t)), \quad f(0, 0) = 0$$

To recap we have

- stability analysis: fix $u = \bar{u}$ and check local stability;
- stabilisability analysis: use v (generated by an algorithm on-line) to stabilise the system.

7.3.1 Lyapunov stability theorem

Let us consider the stability analysis problem first. To this aim, we consider the system

$$\dot{x}(t) = f(x(t)) \tag{7.12}$$

with equilibrium $\bar{x} = 0$ and f continuous.

Definition 7.2. *System* (7.12) *is* ***stable in the equilibrium point (I.E.P.)*** $\bar{x} = 0$ *if, given* $\varepsilon > 0$, $\exists\,\delta > 0$ *such that, if* $\|x(0)\| \le \delta \Rightarrow \|x(t)\| \le \varepsilon$.
System (7.12) *is* ***asymptotically stable in the equilibrium point (I.E.P.)*** $\bar{x} = 0$ *if is stable I.E.P. and, furthermore, for* $\|x(0)\| \le \delta$*, we have that* $\|x(t)\| \to 0$ *for* $t \to +\infty$.

The study of the stability of the equilibrium points of a nonlinear system is mainly based on the theory introduced by Aleksandr Mikhailovich Lyapunov at the beginning of XX century. Lyapunov functions generalise the concept of "energy of a system". We know (without resorting to equations) that dissipative systems encountered in physics, in the body, in circuit theory, and so on, without an external energy supply, reach an equilibrium where the system energy has a minimum. If we generalise this concept, we can have an efficient general tool for assessing stability of nonlinear systems.

Consider a continuously differentiable function V ($V \in C^1$) defined on the state space, $V : \mathbb{R}^n \to \mathbb{R}$. Assume that V is a positive definite function in a neighborhood W of the origin, *i.e.*,

- $V(0) = 0$;
- $V(x) > 0$ for all $x \in W$, $x \neq 0$.

Observe that all the sub-level sets of this function, $\{x : V(x) \le k\}$, for $k > 0$, include the origin (see Figure 7.4).

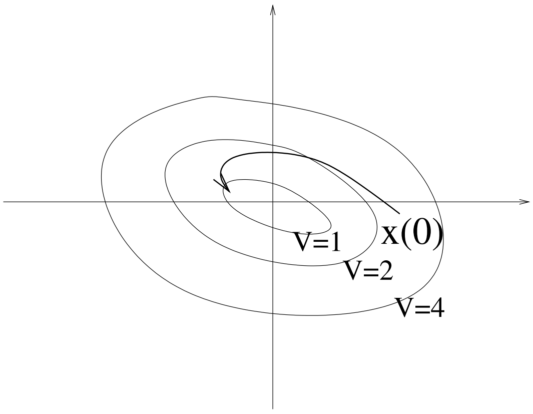

Figure 7.4: Level curves of a Lyapunov function.

The concept introduced by Lyapunov, which we will shortly formalise with a theorem, is based on an observation: if a trajectory $x(t)$ of the system, with $x(0)$ sufficiently small, is such that $V(x(t))$ is a non increasing function, then $x(t)$ is forced to stay inside a sub-level set. And, if $V(x(t))$ is decreasing, then the trajectory must converge to 0.

The fundamental problem is that, in most of the cases, the trajectory $x(t)$ of a nonlinear system is unknown, so in principle we are not able to tell whether $V(x(t))$ is a non increasing (or decreasing) function. However, to avoid this obstacle, we can consider the derivative of the function:

$$\frac{d}{dt}V(x(t)) = \dot{V}(x(t)) = \nabla V(x)\dot{x}(t) = \nabla V(x) f(x(t)), \tag{7.13}$$

where $\nabla V(x) = \left[\begin{array}{cccc} \frac{\partial V}{\partial x_1} & \frac{\partial V}{\partial x_2} & \dots & \frac{\partial V}{\partial x_n} \end{array} \right]$. We notice that, if at a certain instant t the system has state $x(t) = \tilde{x}$, then the derivative is easily computable

$$\frac{d}{dt}V(x(t))\bigg|_{x(t)=\tilde{x}} = \dot{V}(x(t))\bigg|_{x(t)=\tilde{x}} = \nabla V(\tilde{x})f(\tilde{x}) \tag{7.14}$$

Hence it is no longer necessary to know the full trajectory $x(t)$ in order to evaluate the derivative of V. This derivative, also known as **Lyapunov derivative**, is useful to establish the stability of a system: if $\dot{V}(x) \leq 0$, then $V(x)$ is a non increasing function, while if $\dot{V}(x) < 0$, then $V(x)$ is a decreasing function.

Remark 7.1. *The Lyapunov derivative $\dot{V}(x) = \nabla V(x)f(x)$ is the derivative of $V(x(t))$ if $x(t) = x$ along the system trajectory: it depends on both V and the system equation $\dot{x} = f(x)$.*

We can then formalise these ideas in the following theorem.

Theorem 7.1. *Given a neighborhood W of the origin, assume that a function $V \in C^1$, $V : \mathbb{R}^n \to \mathbb{R}$ is defined and such that $V(0) = 0$ and $V(x) > 0$ for $x \in W \backslash \{0\}$. Then*

- *if $\dot{V}(x) \leq 0$ for $x \in W$, then the system is stable I.E.P. $\bar{x} = 0$;*
- *if furthermore $\dot{V}(x) < 0$ for $x \in W \backslash \{0\}$, then the system is asymptotically stable I.E.P. $\bar{x} = 0$.*

Proof. We prove the first part of the theorem only: if $\dot{V}(x) \leq 0$, then the system is stable I.E.P. in zero. Let us take the sphere $S_\varepsilon = \{x : \|x\| \leq \varepsilon\}$ in the state space. It is not restrictive to choose $S_\varepsilon \subset W$. Let m be the minimum value of V on the external surface of the sphere, $\|x\| = \varepsilon$,

$$m \doteq \min_{x \in \partial S_\varepsilon} V(x).$$

Now we take a new sphere S_δ with radius $\delta < \varepsilon$. Define M as the maximum of V inside S_δ

$$M \doteq \max_{x \in S_\delta} V(x),$$

We can take δ small enough to assure that (by continuity of V, since $V(0) = 0$)

$$M < m.$$

By contradiction, assume that the system is not stable and that, for some initial condition inside S_δ, there is a trajectory that escapes from the sphere S_ε. Then there exists a time instant t' such that $x(t') \in \partial S_\varepsilon$. Let t_1 be the first positive instant in which $x(t)$ hits the boundary of S_ε

$$t_1 = \min\{t \geq 0 : x(t_1) \in \partial S_\varepsilon\} \tag{7.15}$$

Between the instants $t = 0$ and the instant $t = t_1$, the trajectory is contained in S_ε, and therefore also in W. Then we have that $\dot{V}(x(t)) \leq 0$ for all $t \in [0, t_1]$. On the other hand, by construction and by the definitions of m and M, we have

$$V(x(t_1)) \geq m > M \geq V(x(0)), \tag{7.16}$$

hence $V(x(t_1)) > V(x(0))$, in contradiction with the fact that $\dot{V}(x(t)) \leq 0$ in the interval.

The second part of the problem, which concerns asymptotic stability, will not be proved: the interested reader is referred to specialised books. However, the proof can be sketched as follows. First we notice that $\dot{V} = \nabla V(x)f(x)$ is continuous, since f is continuous. If $x(0) \in \{x : V(x) \leq \kappa\} \subset W$, namely, if we start inside a sphere that is contained in the κ sub-level set, then $V(x(t)) \leq \kappa$ for all $t > 0$, because $\dot{V} \leq 0$. For the same reason, if at some time t_1 the trajectory reaches a smaller sub-level set, $x(t_1) \in \{x : V(x) \leq \kappa_1\}$, $\kappa_1 < \kappa$, then it will be trapped inside this new sub-level set after t_1: $V(x(t)) \leq \kappa_1$ for all $t > t_1$. The proof can be carried out by showing that, no matter how small κ_1 is taken, the set $\{x : V(x) \leq \kappa_1\}$ will be ultimately reached. This can be proved by contradiction. If the set is not reached, then the solution has to stay in the intermediate set $x(t) \in \{x : \kappa_1 \leq V(x) \leq \kappa\}$, namely,

$$\kappa_1 \leq V(x(t)) \leq \kappa$$

for all $t \geq 0$. In this closed set, $\dot{V}$ has a maximum $(-\mu)$ that is strictly negative. Then

$$\dot{V}(x(t)) \leq -\mu.$$

By integration

$$V(x(t)) - V(x(0)) \leq \int_0^t \dot{V}(x(\sigma))d\sigma \leq \int_0^t -\mu d\sigma = -\mu t,$$

but this means that $V(x(t))$ becomes negative for $t > V(x(0))/\mu$, which is not possible. □

Example 7.4. *Consider the system*

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

and take $V(x_1, x_2) = x_1^2 + x_2^2$. The Lyapunov derivative is:

$$\begin{aligned} \dot{V}(x) &= \nabla V f(x) = \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix} \begin{bmatrix} -2x_1 + x_2^2 + x_1^4 \\ 3x_1^3 - 2x_2 + x_1^2 \end{bmatrix} = \\ &= -4x_1^2 - 4x_2^2 + 2x_1x_2^2 + 2x_1^5 + 6x_1^3x_2 + 2x_1^2x_2 \approx -(x_1^2 + 4x_2^2) \end{aligned}$$

In fact, in a sufficiently small neighborhood of the equilibrium point $(0, 0)$*, the quadratic terms dominate over the others. Therefore, the function $\dot{V}(x_1, x_2)$ is negative semi-definite and this proves the asymptotic stability of the system (I.E.P.).*

Example 7.5. ***(The pendulum.)*** *Given the pendulum shown in Figure 7.5, we want to study stability of its equilibrium points. The system is described by the equation*

$$l_m^2 \ddot{\vartheta}(t) = -l_m g \sin \vartheta(t)$$

Denoting by $x_1(t) = \vartheta(t)$ and $x_2(t) = \dot{\vartheta}(t)$, and assuming all constants equal to 1*, the system is:*

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -\sin x_1(t) \end{cases}$$

The natural candidate Lyapunov function, from physical considerations, is mechanical energy, equal to the sum of kinetic energy and potential energy:

$$V(x_1, x_2) = \frac{1}{2}x_2^2 + 1 - \cos x_1$$

*This function is positive definite in a neighborhood of the equilibrium, hence is a **candidate** Lyapunov function. We obtain:*

$$\dot{V}(x_1, x_2) = \begin{bmatrix} \sin x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_2 \\ -\sin x_1 \end{bmatrix} = x_2 \sin x_1 - x_2 \sin x_1 = 0$$

Hence, $\dot{V}(x_1, x_2)$ is negative semi-definite, and this guarantees the system stability I.E.P.

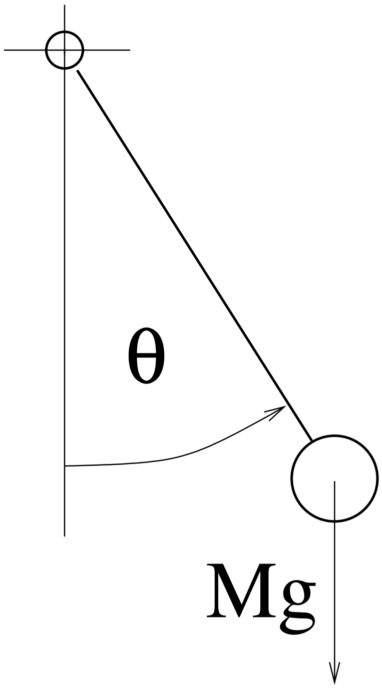

Figure 7.5: Pendulum (basic).

Example 7.6. ***Two-tank system*** *Let us consider the system with two water tanks, whose equations will be presented in Chapter 9.8.*

$$\begin{cases} \dot h_1 = -\frac{1}{\alpha S}\sqrt{h_1 - h_2} + u \\ \dot h_2 = \frac{1}{\alpha S}\sqrt{h_1 - h_2} - \frac{1}{\beta S}\sqrt{h_2} \end{cases}$$

The system admits a generic equilibrium point in $(\bar h_1, \bar h_2)$. *By translating the system with* $x_1 = h_1 - \bar h_1$ *and* $x_2 = h_2 - \bar h_2$,[1] *we have:*

$$\begin{cases} \dot x_1 = -\frac{1}{\alpha S}\sqrt{x_1 - x_2 + \bar h_1 - \bar h_2} + u \\ \dot x_2 = \frac{1}{\alpha S}\sqrt{x_1 - x_2 + \bar h_1 - \bar h_2} - \frac{1}{\beta S}\sqrt{x_2 + \bar h_2} \end{cases}$$

Consider the function $V(x_1, x_2) = x_1^2 + x_2^2$, $\dot V(x_1, x_2)$. *Then, with some manipulation we get*

$$\begin{aligned} \dot V(x_1, x_2) &= x_1\left(\frac{1}{\alpha S}\sqrt{x_1(t) + \bar h_1 - x_2(t) - \bar h_2} - \bar q\right) \\ &+ x_2\left(-\frac{1}{\alpha S}\sqrt{x_1(t) + \bar h_1 - x_2(t) - \bar h_2} + \frac{1}{\beta S}\sqrt{x_2(t) + \bar h_2}\right) \pm x_2\bar q \\ &= (x_1 - x_2)\left(\frac{1}{\alpha S}\sqrt{x_1(t) + \bar h_1 - x_2(t) - \bar h_2} - \bar q\right) - x_2\left(\frac{1}{\beta S}\sqrt{x_2(t) + \bar h_2} - \bar q\right) \end{aligned}$$

If we consider the equilibrium conditions, $\frac{1}{\alpha S}\sqrt{+\bar h_1 - \bar h_2} - \bar q = 0$ *and* $\frac{1}{\beta S}\sqrt{\bar h_2} - \bar q = 0$, *we see that the sign of the terms in round brackets is opposite to that of* $x_1 - x_2$ *and* x_2, *hence the products are negative. Then,* $\dot V(x_1, x_2) < 0$ *(for* $x_1, x_2) \neq (0, 0)$. *This proves asymptotic stability I.E.P.*

[1] The equilibrium conditions are derived in Chapter 9.8.

It is important to notice that a wrong choice of the function $V(x)$ does not lead us to any conclusion. If the Lyapunov derivative is possibly greater than zero in the neighborhood W, we cannot say anything about the system stability in the equilibrium point.

Example 7.7. *Consider system $\dot{x} = Ax$, where*

$$A = \begin{bmatrix} -1 & \alpha \\ -1 & -1 \end{bmatrix}$$

with $\alpha \geq 0$. This system is asymptotically stable (since $\det(sI - A) = s^2 + 2s + 1 + \alpha$*). If we take $V(x) = x_1^2 + x_2^2$, this is a Lyapunov function only for certain values of α. For example, it can be easily seen that, for $\alpha = 1$, the conditions of Lyapunov theorem are satisfied, while for large $\alpha > 0$ they might be not satisfied. Other examples are*

$$A_1 = \begin{bmatrix} 0 & 1 \\ -5 & -2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 \\ +5 & -2 \end{bmatrix}$$

The linear system associated with the first matrix is asymptotically stable, the one associated with the second is unstable. In both cases, if $V(x) = x_1^2 + x_2^2$, then $\dot{V}$ assumes both positive and negative values. Hence, if we take a positive definite V whose Lyapunov derivative $\dot{V}$ is sign indefinite, the only possible information we can get is: V is not a good candidate.

Then, the fundamental problem is the choice of a Lyapunov function that can prove stability of the system. Such a function must have particular properties: its level curves need to fit the trajectories of the system. There are inverse theorems that prove, under certain hypothesis, the existence of a Lyapunov function for asymptotically stable systems. Those results are theoretical and, in general, they are not of any help to actually determine $V(x)$.[2]

Example 7.8. *(Dissipative pendulum.) Let us consider again the pendulum equations, to which we add a term due to the friction and proportional to the speed:*

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -\sin x_1(t) - \alpha x_2(t) \end{cases}$$

The system is dissipative, so we might think that the energy, taken as a Lyapunov function, will prove the asymptotic stability of the system. Unfortunately, we obtain:

$$\dot{V}(x_1, x_2) = \begin{bmatrix} \sin x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_2 \\ -\sin x_1 - \alpha x_2 \end{bmatrix} = -\alpha x_2^2$$

This function is a negative semi-definite function: it is null in the origin and in the x_1-axis, $x_2 = 0$, and negative otherwise. This means that the energy variation when the velocity is null is zero, because there is no dissipation in the maximum elongation point. So, we can confirm stability of the system, but not asymptotic stability I.E.P., because $\dot{V}(x)$ is not negative definite.

There exists another criterion to ensure asymptotic stability of a system: the **Krasowskii criterion**.

Theorem 7.2. *We consider the same assumptions of the (weak) Lyapunov theorem: $\exists\, V \in C^1$, $V : \mathbb{R}^n \to \mathbb{R}$ with $V(x) > 0\ \forall\, x \in W \backslash \{0\}$ and $V(0) = 0$, with $\dot{V}(x) \leq 0\ \forall\, x \in W$. Consider now the set $N = \{x \neq 0 : \dot{V}(x) = 0\}$. If there are no trajectories entirely contained in $N \cap W$ (in other words, $x(t) \notin N \cap W\ \forall\, t \geq 0$) then the system is asymptotically stable.*

The criterion shows that, if the system does not stop in the region where there is no dissipation, sooner or later it will reach an equilibrium point. This criterion can be applied to the pendulum example.

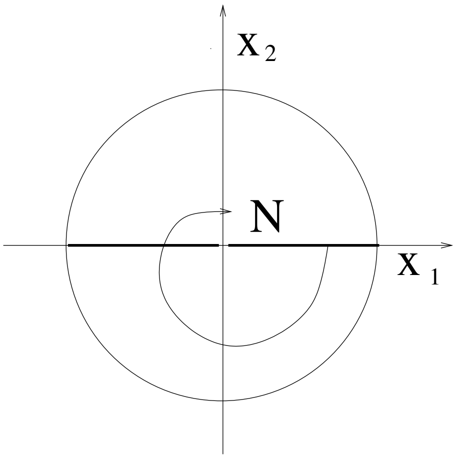

Figure 7.6: Dissipative pendulum: subspaces N and W and vector $\dot{x}(t)$.

Example 7.9. *Let us re-examine the dissipative pendulum and consider Figure 7.6.*

The figure shows the neighbourhood W of the origin (a circle), the set N (the set of points with $x_2 = 0$ *and* $x_1 \neq 0$*) and a trajectory of vector* $\dot{x}(t)$ *starting from an initial condition in N. For any point in N, such that* $x_2 = 0$ *and* $x_1 \neq 0$*, the derivative is*

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ -\sin \bar{x}_1 \end{bmatrix}$$

Hence, it must be $\dot{x}_2 \neq 0$ *and* $x_2(t)$ *cannot remain* 0*, hence the trajectory will leave the set N. In geometric terms, we see that the derivative vector has the direction of the* x_2*-axis, and the trajectory is forced to exit from N, leaving the set of states where there is no dissipation. Thanks to the Krasowskii criterion, we have then proved asymptotic stability. Figure 7.7 shows the trajectories of the dissipative pendulum.*

7.3.2 Instability theorems

There are also instability criteria, which complement stability criteria.

Theorem 7.3. ***(Cetaev criterion.)*** *Suppose that a function* $V(x) \in C^1$ *is defined in a neighbourhood W of the origin. Assume that there exists an open set A such that the origin is contained inside the closure of A (in other words,* 0 *is either in A or on the border of A,* ∂A*). Assume also that*

- $V(x) = 0 \ \forall \, x \in \partial A \cap W$*;*
- $V(0) = 0$*;*
- $V(x) > 0, \ \dot{V}(x) > 0 \ \forall \, x \in A \cap W$

then the system is unstable I.E.P.

We can give an intuitive explanation of this theorem. Suppose, to simplify, that the origin belongs to the border of A, which is the most interesting case, as shown in Figure 7.8. We see that every trajectory that starts from W near the border of A, $\partial A \cap W$, since V and $\dot{V}$ are positive and continuous, has to exit from W, so we cannot have local stability.

[2] Mathematicians often publish a scientific article when they find a good Lyapunov function...

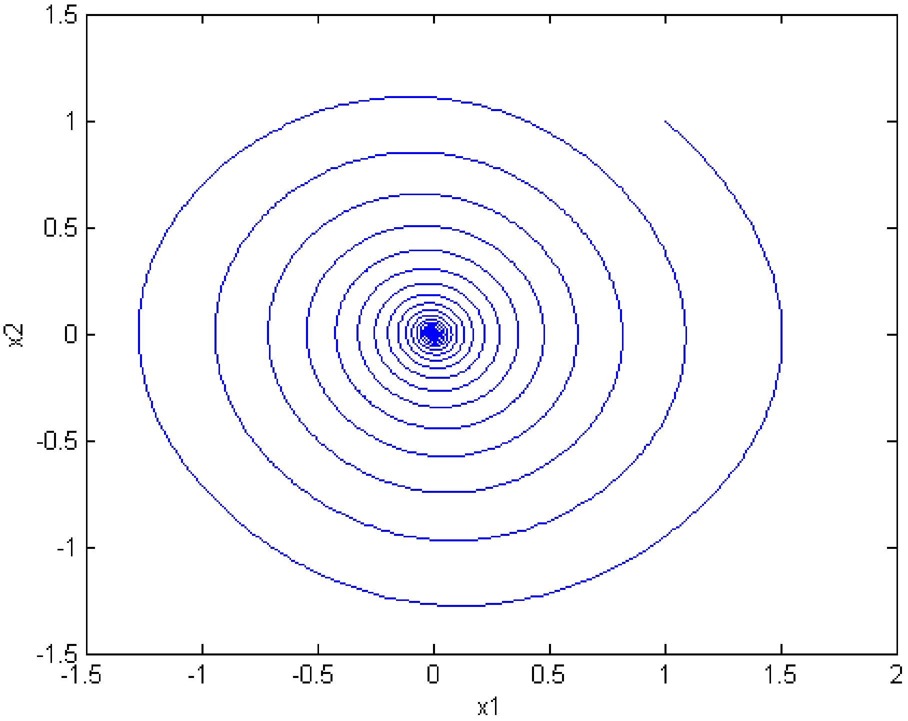

Figure 7.7: Trajectories of the dissipative pendulum.

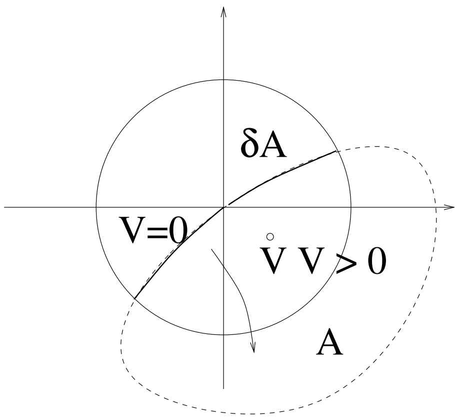

Figure 7.8: Intuitive representation of Cetaev criterion.

Example 7.10. *Consider the magnetic levitator shown in Figure 9.15, Chapter 9.9. If we neglect the electric equation (which means that we consider a permanent magnet), the system is described by following equations:*

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = g - \frac{k}{m}\frac{i^2}{x_1^2} \end{cases}$$

The equilibrium point of the system is:

$$\begin{cases} 0 = x_2 \\ 0 = g - \frac{k}{m}\frac{i^2}{x_1^2} \end{cases}$$

This means having null velocity and magnetic force equal and opposite to the gravity force. We want to prove the instability of the equilibrium point obtained with a constant value of the current $i = \bar{i}$. First of all, we shift the intersection of the axes to the equilibrium point:

$$\begin{aligned} z_1 &= x_1 - \bar{x}_1 \\ z_2 &= x_2 - \bar{x}_2 \\ &\Rightarrow \begin{cases} \dot{z}_1 = z_2 + \bar{x}_2 = z_2 \\ \dot{z}_2 = g - \frac{k}{m}\frac{\bar{i}^2}{(z_1+\bar{x}_1)^2} \end{cases} \end{aligned}$$

Now the equilibrium point is $(0, 0)$. *As a candidate Cetaev function, we take $V(x_1, x_2) = z_1 z_2$. We choose the set A as $A = \{z_1 > 0, z_2 > 0\}$. We have that:*

$$\dot{V}(z_1, z_2) = \begin{bmatrix} z_2 & z_1 \end{bmatrix} \begin{bmatrix} z_2 \\ g - \frac{k}{m}\frac{\bar{i}^2}{(z_1+\bar{x}_1)^2} \end{bmatrix} = z_2^2 + z_1\left(g - \frac{k}{m}\frac{\bar{i}^2}{(z_1 + \bar{x}_1)^2}\right)$$

V is null on the boundary (if either $z_1 = 0$ or $z_2 = 0$). Both V and $\dot{V}$ are positive in the interior of A. In view of Cetaev criterion, the system is unstable I.E.P.

7.4 Lyapunov criterion for discrete-time systems

The formulation of the Lyapunov criterion for discrete-time systems is similar to that given for continuous-time systems, with only one essential change: the Lyapunov derivative is replaced by the Lyapunov difference.

The discrete-time system has the following form:

$$x(k+1) = f(x(k), u(k)) \tag{7.17}$$

and the equilibrium points are found as the solutions of the equation

$$\bar{x} = f(\bar{x}, \bar{u}) \tag{7.18}$$

The shifting is done as for continuous-time systems, assuming $z = x - \bar{x}$, and assuming that $v(k) = u(k) - \bar{u}$ we can study a system of the following type:

$$z(k+1) = F_{\bar{x},\bar{u}}(z(k), v(k)) \tag{7.19}$$

with equilibrium pair $(0, 0)$.

Let us start with the analysis of system

$$x(k+1) = f(x(k)), \quad f(0) = 0,$$

in the 0 equilibrium state. For discrete-time systems, we take in consideration the **Lyapunov difference** defined as follows:

$$\Delta V(x(k)) \doteq V(x(k+1)) - V(x(k)) = V(f(x(k))) - V(x(k)) \tag{7.20}$$

Observe how in the last formula $x(k+1)$ has been replaced by the function $f(x(k))$, in view of the equation which describes the system: we do not need to know the trajectory. The main idea is that, if difference is not increasing, then the system is stable.

Theorem 7.4. *Assume that f is continuous and there exists a function $V(x)$ that is continuous in a neighborhood W of* 0 *and positive definite in W (*i.e., $V(0) = 0$ *and* $V(x) > 0\ \forall\, x \in W\backslash\{0\}$*). Then:*

- *if $\Delta V(x) \leq 0\ \forall\, x \in W$, then the system is stable I.E.P.;*
- *if $\Delta V(x) < 0\ \forall\, x \in W\backslash\{0\}$, then the system is asymptotically stable I.E.P..*

We do not prove the theorem.[3]

7.5 Lyapunov equations for linear systems

As we have previously seen, there is no systematic method to find a Lyapunov function for a nonlinear system. Normally, the determination of a good candidate is left to intuition and is typically based on the knowledge of the system (for example, physical considerations). Linear systems are an exception: for asymptotically stable linear systems, there is a systematic method to find a suitable Lyapunov function (to certifies asymptotic stability).

We consider the following quadratic function:

$$V(x) = x^\top P x \tag{7.21}$$

If matrix P is symmetric and positive definite, then $V(x)$ is a candidate Lyapunov function. It can be easily verified that the gradient is the row vector

$$\nabla V(x) = 2x^\top P. \tag{7.22}$$

Quadratic functions with symmetric and positive definite matrices are, without any doubt, the most used Lyapunov functions. Now, if we consider a linear system $\dot{x}(t) = Ax(t)$ (the input is not relevant to stability) and we compute the derivative of the Lyapunov function that we have just introduced, we obtain:

$$\dot{V}(x) = \nabla V(x) \cdot f(x) = 2x^\top PAx = x^\top PAx + x^\top PAx \tag{7.23}$$

The last two terms are scalars, therefore we can transpose the first one without changing the result. Then we have:

$$\begin{aligned} \dot{V}(x) &= x^\top A^\top P^\top x + x^\top PAx = x^\top A^\top Px + x^\top PAx = \\ &= x^\top (A^\top P + PA)x \doteq -x^\top Qx \end{aligned} \tag{7.24}$$

where we have defined $Q = -(A^\top P + PA)$. It can be easily seen that Q is a symmetric matrix, because it is the sum of $-A^\top P$ and its transpose $(-A^\top P)^\top = -PA$. We obtain the following fundamental equation

$$A^\top P + PA = -Q \tag{7.25}$$

[3] The proof is more involved than in the continuous-time case, because discrete time can "jump", so we cannot define the instant t_1 as we did the previous proof.

called **Lyapunov equation**.

Now, suppose that the system is asymptotically stable. Then, we could think about finding a matrix Q that is positive definite, so that $\dot{V}(x)$ is negative definite and verifies the conditions of Lyapunov theorem. This does not work in general because, again, it is not easy to find a matrix P such that Q is positive definite. For instance, if we consider matrix A_1 in Example 7.7 (which is asymptotically stable) and we take $P = I$, Q is not positive definite. Instead, it is appropriate to proceed in the other way: we fix matrix Q and then we compute the corresponding P, as a solution of the equation (7.25). This procedure is supported by the following theorem.

Theorem 7.5. *The following conditions are equivalent.*

1. *$\dot{x}(t) = Ax(t)$ is asymptotically stable;*

2. *for any symmetric and positive definite matrix Q, the corresponding matrix P that solves the Lyapunov equation $A^\top P + PA = -Q$ is symmetric and positive definite.*

Proof. We start by showing that 2 implies 1. If we arbitrarily fix a matrix Q and by solving the Lyapunov equation we find a matrix P that is symmetric and positive definite, then the quadratic function $V(x) = x^\top Px$ is a valid Lyapunov function and we have that $\dot{V}(x) < 0 \;\forall\, x \neq 0$, because Q is symmetric and positive definite. Therefore, the conditions of Lyapunov theorem are verified.

Now we prove that 1 implies 2. Assume that the system $\dot{x}(t) = Ax(t)$ is asymptotically stable, and fix a symmetric and positive definite matrix Q. We prove that the solution of the Lyapunov equation is the following:

$$P = \int_0^{+\infty} e^{A^\top t} Q e^{At}\, dt. \tag{7.26}$$

Indeed, by replacing this expression in the equation $A^\top P + PA = -Q$, we obtain:

$$\begin{aligned} A^\top P + PA &= A^\top \int_0^{+\infty} e^{A^\top t} Q e^{At}\, dt + \int_0^{+\infty} e^{A^\top t} Q e^{At}\, dt\, A = \\ &= \int_0^{+\infty} \left[A^\top e^{A^\top t} Q e^{At} + e^{A^\top t} Q e^{At} A\right] dt \end{aligned} \tag{7.27}$$

Noticing that $\frac{d}{dt} e^{At} = Ae^{At} = e^{At}A$,[4] the function to be integrated is a derivative, hence:

$$A^\top P + PA = \int_0^{+\infty} \frac{d}{dt}\left[e^{A^\top t} Q e^{At}\right] dt = \left[e^{A^\top t} Q e^{At}\right]_0^{+\infty} = 0 - Q, \tag{7.28}$$

where in the last equality we have used the fact that the system is asymptotically stable, hence $e^{At}\Big|_{t\to+\infty} = 0$.

Matrix P given by the previous expression is

- symmetric, since the function to be integrated is symmetric, because we can see (based on the exponential series) that $\left(e^{At}\right)^\top = e^{A^\top t}$;

- positive definite, because for any $\bar{x} \neq 0$ we have:

$$\bar{x}^\top \int_0^{+\infty} e^{A^\top t} Q e^{At}\, dt\, \bar{x} = \int_0^{+\infty} \bar{x}^\top e^{A^\top t} Q e^{At} \bar{x}\, dt = \int_0^{+\infty} x^\top(t) Q x(t)\, dt \tag{7.29}$$

  where we have denoted by $x(t) = e^{At}\bar{x} \neq 0$. Then the function to be integrated is positive, because by assumption Q is positive definite, therefore also the integral is positive, thus proving that P is positive definite.

[4] This is one of the few cases in which the product commutes and the position of matrix A is not important in the derivative.

□

The Lyapunov equation is linear in P, therefore we can avoid the computation of the integral and solve an algebraic linear system, where the unknowns are the elements of P. Actually, only the entries of the upper (or lower) triangle of P have to be considered, because P is symmetric. The total number of unknowns and equations is then $n + (n-1) + (n-2) + \ldots + 1 = \dfrac{(n+1)n}{2}$.

Example 7.11. *Consider the state matrix*

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

which yields an asymptotically stable system. Take $Q = I$, the Lyapunov equation is

$$\begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} + \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} = -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

namely

$$\begin{bmatrix} -2\beta & \alpha-\beta-\gamma \\ \alpha-\beta-\gamma & 2(\beta-\gamma) \end{bmatrix} = -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Equating term-by-term, we get $\beta = 1/2$, $\gamma = 1$ and $\alpha = 3/2$. The resulting matrix P is

$$P = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix},$$

which is positive definite, as expected.

7.5.1 Lyapunov equations for discrete-time systems

The case of discrete-time systems $x(k+1) = Ax(k)$ is similar. The Lyapunov function candidate is again $V(x) = x^\top Px$, but instead of the derivative we have to consider the Lyapunov difference. Therefore we obtain that:

$$\begin{aligned} \Delta V(x) &= V(x(k+1)) - V(x(k)) = V(Ax) - V(x) = \\ &= (Ax)^\top P(Ax) - x^\top Px = x^\top A^\top PAx - x^\top Px = \\ &= x^\top(A^\top PA - P)x = -x^\top Qx \end{aligned} \tag{7.30}$$

The fundamental equation is

$$A^\top PA - P = -Q \tag{7.31}$$

named the **discrete Lyapunov equation**. The following theorem holds.

Theorem 7.6. *The following conditions are equivalent:*

1. *$x(k+1) = Ax(k)$ is asymptotically stable;*
2. *for any positive definite matrix Q, the solution P of the discrete Lyapunov equation $A^\top PA - P = -Q$ is positive definite.*

Proof. It is similar to the proof in the continuous-time case, but we just have to consider, instead of the integral, the sum

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

□

The previous discussion allows us to perform a test to verify the asymptotic stability of a system. However, since we consider linear systems, such a test could not seem really useful, because stability can be simply studied by analysing the eigenvalues of matrix A. Moreover, almost all of the algorithms that compute matrix P by solving either continuous or discrete Lyapunov equations go through the computation of the eigenvalues of A, making the test pointless. Nevertheless, these results are of fundamental importance for their consequences, including the possibility to give a rigorous support to the linearisation theory.

A final consideration concerns the choice of matrix Q in order to find P. Choosing a positive definite matrix is easy: a possibility is the identity matrix. As an alternative, we can consider a square and invertible matrix R and compute $Q = R^\top R$, which is

- symmetric, because $(R^\top R)^\top = R^\top R$;
- positive definite, because $\forall\, x \neq 0$ we have $x^\top (R^\top R)x = (Rx)^\top (Rx) = \|Rx\|^2 > 0$, because R is invertible (nonsingular).

7.6 Linearisation

The topics discussed in this section and in the next ones are of high importance, because they will justify our careful development of the theory for linear systems.

We start by reminding elementary notions of differential calculus. Consider a nonlinear function $f(x)$ in one dimension. It is well known that, if f is continuously differentiable, we can apply Taylor theorem and write the function as follows:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + R(x - x_0), \tag{7.32}$$

where $R(x - x_0)$ is the residual and is an infinitesimal of order greater that one:

$$\frac{|R(x - x_0)|}{|x - x_0|} \to 0 \text{ as } x \to x_0 \tag{7.33}$$

The first two terms form the **linear approximation**:

$$\bar{f}(x) = f(x_0) + f'(x_0)(x - x_0) \tag{7.34}$$

They represent a line passing through the point x_0 with slope $f'(x_0)$.

For nonlinear functions in two scalar variables, the situation is similar. With a Taylor expansion, we obtain:

$$\begin{aligned} f(x, u) &= f(x_0, u_0) + \nabla f \begin{bmatrix} x - x_0 \\ u - u_0 \end{bmatrix} + R(x - x_0, u - u_0) = \\ &= f(x_0, u_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0, u_0}(x - x_0) + \left.\frac{\partial f}{\partial u}\right|_{x_0, u_0}(u - u_0) + \\ &\quad + R(x - x_0, u - u_0) \end{aligned} \tag{7.35}$$

In this case, the linear approximation is the equation of the plane that passes through (x_0, u_0), with equation:

$$\bar{f}(x, u) = f(x_0, u_0) + \alpha(x - x_0) + \beta(u - u_0) \tag{7.36}$$

The partial derivatives computed in the point are the coefficients of the plane that locally approximates the curve $f = f(x, u)$.

This approximation can be applied to the case of nonlinear systems $\dot{x}(t) = f(x(t), u(t))$, where the linear approximation can be determined with reference to a particular equilibrium point identified by $0 = f(\bar{x}, \bar{u})$:

$$f(x, u) = \underbrace{f(\bar{x}, \bar{u})}_{=0} + \left[\frac{\partial f}{\partial x}\right]_{\bar{x}, \bar{u}} (x - \bar{x}) + \left[\frac{\partial f}{\partial u}\right]_{\bar{x}, \bar{u}} (u - \bar{u}) + R(x - \bar{x}, u - \bar{u}) \tag{7.37}$$

The terms

$$\left[\frac{\partial f}{\partial x}\right]_{\bar{x}, \bar{u}} \quad \left[\frac{\partial f}{\partial u}\right]_{\bar{x}, \bar{u}}$$

are Jacobian matrices, whose (i, j) terms are equal to

$$\left[\frac{\partial f_i}{\partial x_j}\right]_{\bar{x}, \bar{u}} \quad \left[\frac{\partial f_i}{\partial u_j}\right]_{\bar{x}, \bar{u}}$$

and become simple numbers once the derivatives are evaluated at the equilibrium point. Computing the derivatives is a simple operation.

If an output transformation is given,

$$y(t) = g(x(t), u(t))$$

is possible to linearise it at the equilibrium point. Let us assume

$$\bar{y} = g(\bar{x}, \bar{u})$$

as the equilibrium output. Then

$$y(t) = G(x(t), u(t)) = \bar{y} + \left[\frac{\partial g}{\partial x}\right]_{\bar{x}, \bar{u}} (x(t) - \bar{x}) + \left[\frac{\partial g}{\partial u}\right]_{\bar{x}, \bar{u}} (u(t) - \bar{u}) + S(x(t) - \bar{x}, u(t) - \bar{u}) \tag{7.38}$$

where S is the residual, which is an infinitesimal of order greater than one.

Now is possible to translate the variables as

$$\begin{aligned} z(t) &= x(t) - \bar{x} \\ v(t) &= u(t) - \bar{u} \\ w(t) &= y(t) - \bar{y} \end{aligned}$$

This transformation is nothing else than an axis shifting, in order to have the equilibrium point in $(z, v) = (0, 0)$ corresponding to the output $w = 0$. This means that, if in a thermal system $T = 20^oC$ is the equilibrium temperature, in the new reference the desired temperature will be 0^oC.[5]

Then the series expansion becomes:

$$\begin{aligned} \dot{z}(t) &= \dot{x}(t) = f(\bar{x}, \bar{u}) + \left[\frac{\partial f}{\partial x}\right]_{\bar{x}, \bar{u}} z(t) + \left[\frac{\partial f}{\partial u}\right]_{\bar{x}, \bar{u}} v(t) + R(z(t), v(t)) \\ &= 0 + A_{\bar{x}, \bar{u}} z(t) + B_{\bar{x}, \bar{u}} v(t) + R(z(t), v(t)) \end{aligned} \tag{7.39}$$

The residual has the property:

$$\frac{||R(z, v)||}{||z|| + ||v||} \to 0 \text{ as } (||v||, ||z||) \to 0$$

[5] With no complaints.

Therefore, for little shifts in the neighbourhood of the equilibrium point, we can neglect the residual term and obtain a linearised system of the form:

$$\dot{z}(t) = A_{\bar{x},\bar{u}} z(t) + B_{\bar{x},\bar{u}} v(t) \tag{7.40}$$

Now, by considering the variable $w(t) = y - \bar{y}$, we obtain the equation:

$$w(t) = \left[\frac{\partial g}{\partial x}\right]_{\bar{x},\,\bar{u}} z(t) + \left[\frac{\partial g}{\partial u}\right]_{\bar{x},\,\bar{u}} v(t) + S\,(z(t),\ v(t))$$

where

$$\frac{||S\,(z,\ v)||}{||z|| + ||v||} \rightarrow 0 \text{ as } (||v||,\ ||z||) \rightarrow 0$$

by neglecting the residual

$$w = C_{\bar{x},\bar{u}} z(t) + D_{\bar{x},\bar{u}} v(t)$$

To summarise, the nonlinear system written with reference to the equilibrium point is

$$\begin{cases} \dot{z}(t) = Az(t) + Bv(t) + R(z(t), v(t)) \\ w(t) = Cz(t) + Dv(t) + S\,(z(t), v(t)) \end{cases} \tag{7.41}$$

where the matrices **depend on the chosen equilibrium**, although we can remove the indices $\bar{x}, \bar{u}$ for a softer notation

$$A = A_{\bar{x},\bar{u}} \quad B = B_{\bar{x},\bar{u}}, \quad C = C_{\bar{x},\bar{u}}, \quad D = D_{\bar{x},\bar{u}} v(t)$$

and the linearised system (in the equilibrium point) is:

$$\begin{cases} \dot{z}(t) = Az(t) + Bv(t) \\ w(t) = Cz(t) + Dv(t) \end{cases} \tag{7.42}$$

The matrix entries are equal to:

$$\begin{aligned} [A]_{ij} &= \frac{\partial f_i}{\partial x_j}(\bar{x},\ \bar{u}) \\ [B]_{ij} &= \frac{\partial f_i}{\partial u_j}(\bar{x},\ \bar{u}) \\ [C]_{ij} &= \frac{\partial g_i}{\partial x_j}(\bar{x},\ \bar{u}) \\ [D]_{ij} &= \frac{\partial g_i}{\partial u_j}(\bar{x},\ \bar{u}) \end{aligned} \tag{7.43}$$

We will use the linearised system at the equilibrium point for two different purposes.

- Stability analysis: we analyse stability of the approximated system $\dot{z}(t) = A_{\bar{x},\bar{u}} z(t)$.
- Synthesis problem: by considering the linear approximation, we will build a regulator based on the techniques studied for linear systems. This regulator will be applied to the nonlinear system, as shown in Figure 7.9.

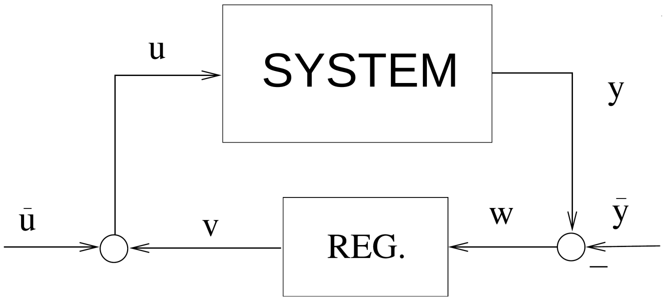

Figure 7.9: Controller for the linearised system at an equilibrium point.

Example 7.12. *Let us consider the magnetic levitator described in Chapter 9.9. The equations that describe the system are:*

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = g - \frac{k}{m}\frac{x_3^2(t)}{x_1^2(t)} \\ \dot{x}_3(t) = -\frac{R}{L}x_3(t) + \frac{1}{L}u(t) \end{cases}$$

The output $e(t)$ is the voltage of a photo-diode that detects the position of the metallic sphere, based on the light intensity that crosses a diaphragm and that is not covered by the sphere itself. The equilibrium points of the system are determined by:

$$\begin{cases} 0 = \bar{x}_2 \\ 0 = g - \frac{k}{m}\frac{\bar{x}_3^2}{\bar{x}_1^2} \\ 0 = -\frac{R}{L}\bar{x}_3 + \frac{1}{L}\bar{u} \end{cases}$$

In this case, it is natural to fix the position $\bar{x}_1$, in which we want to keep the sphere, and find $\bar{x}_3$ and $\bar{u}$ consequently. Setting $z_1 = x_1 - \bar{x}_1$, $z_2 = x_2 - \bar{x}_2$ e $z_3 = x_3 - \bar{x}_3$, we have the linearised system

$$\begin{bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \\ \dot{z}_3(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 2\frac{k}{m}\frac{\bar{x}_3^2}{\bar{x}_1^3} & 0 & -2\frac{k}{m}\frac{\bar{x}_3}{\bar{x}_1^2} \\ 0 & 0 & -\frac{R}{L} \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L} \end{bmatrix} v(t)$$

which can also be written in the following way:

$$\begin{bmatrix} \dot{z}_1(t) \\ \dot{z}_2(t) \\ \dot{z}_3(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ \alpha^2 & 0 & -\beta \\ 0 & 0 & -\gamma \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \delta \end{bmatrix} v(t)$$

where

$$\alpha^2 = 2\frac{k}{m}\frac{\bar{x}_3^2}{\bar{x}_1^3}, \quad \beta = 2\frac{k}{m}\frac{\bar{x}_3}{\bar{x}_1^2}, \quad \gamma = \frac{R}{L}, \quad \delta = \frac{1}{L}$$

The equilibrium point is unstable, because the eigenvalues are $-\gamma$, $-\alpha$, $+\alpha$. A simple reachability test proves that the system is reachable, hence it is possible to design a state feedback control that stabilises the linear system in a neighbourhood of the equilibrium point.

Concerning the output of the system, we have that $e(t) = \varphi(y(t))$, therefore:

$$e(t) - \bar{e} = \varphi(y(t)) - \varphi(\bar{y}) = \begin{bmatrix} \left.\frac{\partial\varphi}{\partial x_1}\right|_{\bar{x}_1,\bar{x}_2,\bar{x}_3} & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + R(y(t) - \bar{y})$$

The derivative inside matrix $C_{\bar{x}_1,\bar{x}_2,\bar{x}_3}$ can be found from the slope of the photo-diode curve given by the constructor. Matrix C is of the type

$$\left[\begin{array}{ccc} \mu & 0 & 0 \end{array}\right]$$

with $\mu \neq 0$, hence the system is observable, as it can be easily checked. Therefore the linearised system is stabilisable. The guess is that, if we can stabilise the linearised system, then we can also stabilise the nonlinear system around the equilibrium point. We will see that this is the case.

7.7 Stability analysis of an equilibrium point

The linear approximation is referred to a particular equilibrium point. Therefore, is intuitive that the linearised system can give informations exclusively about stability in such an equilibrium point, which is a local property.

Suppose that $v(t) = u(t) - \bar{u} = 0$, namely, $u(t) = \bar{u}$. Then

$$\dot{x}(t) = f(x(t), \bar{u}) \tag{7.44}$$

We have seen that, by means of the coordinate shift $z(t) = x(t) - \bar{x}$ and $v(t) = u(t) - \bar{u}$, we can bring the equilibrum point to 0 and obtain the equivalent system

$$\dot{z}(t) = f(z(t)) = Az(t) + R(z(t))$$

where A is the Jacobian evaluated at $x = \bar{x}$ (we should have written $A_{\bar{x}}$, but for brevity we will not write any index from now on).

Observation 7.1. *Remember: the linear approximation is valid* only *in a neighborhood of the considered equilibrium.*

Now we can associate with the nonlinear system

$$\dot{x}(t) = Ax(t) + R(x(t)), \tag{7.45}$$

such that $f(0) = 0$, the corresponding linearised model

$$\dot{x}(t) = Ax(t) \tag{7.46}$$

The fundamental question is now: by analysing the stability of (7.46), which information can we obtain on the stability of (7.45) **in the equilibrium point (I.E.P.)** $\bar{x} = 0$? The following is a fundamental result.

Theorem 7.7. *The two implications hold.*

- *If the linear system* (7.46) *is asymptotically stable, then the nonlinear system* (7.45) *is asymptotically stable I.E.P.*
- *If the linear system* (7.46) *is exponentially unstable (namely, $\exists\, \lambda \in \sigma(A)$ such that $\Re\{\lambda\} > 0$), then the nonlinear system is unstable I.E.P.*

Proof. We prove only the first claim, since the second one is more difficult. To this aim, we need two important properties. We remind that the norm of a matrix is defined as

$$\|M\| \doteq \sup_{\|x\|\neq 0} \frac{\|Mx\|}{\|x\|} = \left(\max \sigma(M^\top M)\right)^{\frac{1}{2}} \tag{7.47}$$

Hence $||Mx|| \leq ||M||\,||x||$ (the implication is quite easy to see) and $||MNx|| \leq ||M||\,||Nx|| \leq ||M||\,||N||\,||x||$. Moreover, we remind Schwartz inequality: if we denote by (x, y) the scalar product of two vectors x and y, then

$$|(x, y)| \leq ||x||\,||y||. \tag{7.48}$$

In view of Theorem 7.5, if the linear system is asymptotically stable, for any choice of Q positive definite, matrix P, solution of the Lyapunov equation, is positive definite. Then take $Q = I$, the identity matrix, and the solution P of the equation

$$A^\top P + PA = -I \tag{7.49}$$

provides the Lyapunov function $V(x) = x^\top Px$ for the *linear system*. We now show that $V(x)$ if a Lyapunov function*for the nonlinear system as well*, thus proving the stability of (7.45) I.E.P.

Compute $\dot{V}(x)$ for the nonlinear system:

$$\begin{aligned}\dot{V}(x) &= \nabla V(x) f(x) = \\ &= 2x^\top P(Ax + R(x)) = 2x^\top PAx + 2x^\top PR(x) = x^\top PAx + x^\top PAx + 2x^\top PR(x) = \\ &= x^\top A^\top Px + x^\top PAx + 2x^\top PR(x) = x^\top (A^\top P + PA)x + 2x^\top PR(x) = \\ &= -x^\top Ix + 2x^\top PR(x) \leq -||x||^2 + 2|x^\top PR(x)| \end{aligned} \tag{7.50}$$

We obtain, for $x \neq 0$,

$$\dot{V}(x) \leq -||x||^2 + 2|x^\top PR(x)| = -||x||^2 \left(1 - \frac{2|x^\top PR(x)|}{||x||^2}\right)$$

We show now that there exists a neighborhood W of zero in which

$$\frac{2|x^\top PR(x)|}{||x||^2} < 1 \tag{7.51}$$

In fact, by applying Schwartz inequality and the property of matrix norms,

$$\begin{aligned}\left|\frac{2x^\top PR(x)}{||x||^2}\right| &= 2\frac{\left|x^\top PR(x)\right|}{||x||^2} = \frac{|(x,\ PR(x))|}{||x||^2} \leq \\ &\leq \frac{2||x||\,||PR(x)||}{||x||^2} = \frac{2||PR(x)||}{||x||} \leq 2||P||\frac{||R(x)||}{||x||}\end{aligned}$$

We remind that $R(x)$ is an infinitesimal of order greater than one, hence $\frac{||R(x)||}{||x||} \to 0$ for $x \to 0$. Therefore, there exists a neighbourhood W of zero in which

$$\frac{||R(x)||}{||x||} < \frac{1}{2||P||}$$

(for $x \neq 0$). Hence, the inequality (7.51) holds in a set W for $x \neq 0$. In this set W, the derivative $\dot{V}(x)$ is less then zero (for $x \neq 0$), hence the conditions of Lyapunov theorem are satisfied, thus proving the asymptotic stability I.E.P. of the nonlinear system. □

The theorem does not take into consideration the case in which all of the eigenvalues of A have non-positive real part, but some have 0 real part. In this case, in fact, the linear system is not asymptotically stable: it can be either marginally stable or unstable, but not exponentially unstable. Then, there is no relation between the behaviour of the nonlinear system and of its linearisation: everything can happen.

Example 7.13. *System $\dot{x}(t) = \alpha x^3(t)$ can have different behaviours based on the values of α:*

- *if* $\alpha > 0$*, then for* $x(t) > 0$ *we have* $\dot{x}(t) > 0$*, and for* $x(t) < 0$ *we have* $\dot{x} < 0$*: the system is unstable I.E.P.;*
- *if* $\alpha < 0$*, then for* $x(t) > 0$ *we have* $\dot{x}(t) < 0$*, and for* $x(t) < 0$ *we have* $\dot{x} > 0$*: the system is asymptotically stable I.E.P., as we can see also by means of the Lyapunov function* $V(x) = x^2$*;*
- *if* $\alpha = 0$*, the system is marginally stable.*

If we linearise the system at the equilibrium point $\bar{x} = 0$*, we obtain that:*

$$\dot{x}(t) = \left[\frac{\partial f}{\partial x}\right]_0 x + R(x) = 0 + R(x)$$

Then the matrix of the linearised system is zero independently of α *(*$A = [0]$*, and so its eigenvalue is equal to zero for all* α*), in all of the three cases above. Therefore, it is obvious that the linearisation criterion does not give us any information about the stability of the nonlinear system.*

Remark 7.2. *In control theory, if we find some eigenvalues with zero real part (while the others have negative real part), we can come to a conclusion anyway. This is a situation that we should avoid. Indeed, small perturbations on the system parameters can lead the eigenvalues to have a real part greater than zero, thus leading to instability. Would you take a flight if the aircraft had some poles at* $\pm j\omega$*?*

7.8 Stabilisation

Suppose that we have already linearised a nonlinear system at the equilibrium point. To avoid complicated computations, we assume that $y = g(x)$ does not depend on u (strictly proper system). This immediately implies that $D = 0$. Hence,

$$\begin{cases} \dot{z}(t) = Az(t) + Bv(t) + R(z(t), v(t)) \\ y(t) = Cz(t) + S(z(t)) \end{cases} \tag{7.52}$$

By neglecting the residuals and working only with the approximated linear systems, we know how to design a linear regulator. A legitimate question is whether the regulator that has been designed based on the linear system is able to stabilise the nonlinear system in the equilibrium point as well.

The regulator, determined by means of *any* of the available methods for linear systems, will be a system of the form:

$$\begin{cases} \dot{z}_C(t) = Fz_C(t) + Gw(t) \\ v(t) = Hz_C(t) + Kw(t) \end{cases} \tag{7.53}$$

Assume that it stabilises the approximated linear system. The matrix of the closed-loop linear system is

$$\begin{bmatrix} \dot{z}(t) \\ \dot{z}_C(t) \end{bmatrix} = \begin{bmatrix} A + BKC & BH \\ GC & F \end{bmatrix} \begin{bmatrix} z(t) \\ z_C(t) \end{bmatrix} \tag{7.54}$$

and is stable if the regulator is properly designed.

This regulator, however, is applied to the true system, which is a nonlinear system. By joining (7.52) and (7.53), we get

$$\begin{aligned} \dot{z} &= Az + BHz_C + BKw + R(z, w) = \\ &= Az + BHz_C + BK(Cz + S(z)) + R(z, v) = \\ &= (A + BKC)z + BHz_C + R(z, v) + BKS(z) \\ \dot{z}_C &= Fz_C + G(Cz + S(z)) = Fz_C + GCz + GS(z) \end{aligned} \tag{7.55}$$

which can be written in the following matrix form:

$$\begin{bmatrix} \dot{z}(t) \\ \dot{z}_C(t) \end{bmatrix} = \begin{bmatrix} A + BKC & BH \\ GC & F \end{bmatrix} \begin{bmatrix} z(t) \\ z_C(t) \end{bmatrix} + \underbrace{\begin{bmatrix} R(z, v) + BKS(z) \\ GS(z) \end{bmatrix}}_{\text{infinitesimal of order greater than 1}} \tag{7.56}$$

where the last term is the residual, which is an infinitesimal of order greater than 1. So, if we apply the theorem about stability analysis at the equilibrium point, we can study the stability of the linearised system (neglecting the residual). But the matrix of such a linear system has eigenvalues with negative real part, because it is exactly the matrix obtained when designing the chosen regulator

$$A_{cl} = \begin{bmatrix} A + BKC & BH \\ GC & F \end{bmatrix}$$

Hence, the closed-loop system turns out to be stable at the equilibrium point. This confirms that the design of a regulator for the approximated linear system allows us to obtain a nonlinear system that is asymptotically stable I.E.P., as shown in Figure 7.10. In the figure, solid arcs represent linearisation, while dashed arcs represent the application of the regulator: by commuting the operations of linearisation and application of the regulator, we reach the same stable linear system.

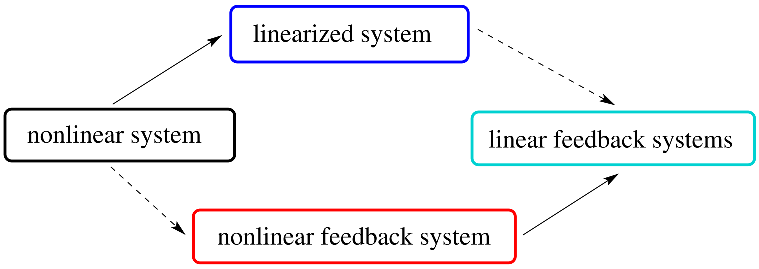

Figure 7.10: Linearisation and application of the regulator commute.

7.9 Robustness

All systems in practice are affected by uncertainties. Uncertainties are typically due to the following factors:

- unknown parameters, typically because they are not measurable;
- parameters that vary in time in an unpredictable way;
- model approximation.

In these cases, stability has to be ensured in a robust way, namely, under all possible cases and variations. This is the problem of **robust control design** and is fundamental in control applications.

We provide a general definition of robustness and then we give just a simple condition based on Lyapunov theory in a specific case.

First of all, to formulate a robustness problem we need

- a family $\mathcal{F}$ of systems, which can be all the possible actual models among which a single one will be the exact "true" system (we do not know which one);

- a property $\mathcal{P}$ of the system, which can be, for instance, stability, or maximum frequency response peak, maximum amplification of the input, non-overshoot, no oscillations,. . .

Definition 7.3. *Property $\mathcal{P}$ is* **robust** *if any system $f \in \mathcal{F}$ of the family has the property $\mathcal{P}$.*

Example 7.14. *Consider a linear system with constant unknown parameters and transfer function*

$$W(s) = \frac{s}{s^2 + as + b},$$

where we know that $a^- \leq a \leq a^+$, $b^- \leq b \leq b^+$, for some known bounds a^-, a^+, b^-, b^+. This is a family. The considered property can be asymptotic stability. Another property can be

$$\sup_{\omega \geq 0} |W(j\omega)| \leq \mu$$

for some μ, and so on.

Remark 7.3. *The definition seems a trivial tautology. A deeper insight shows that it points out two facts:*

- *we must have a family that is* **properly defined***;*

- *we must consider a* **specific** *property.*

For instance, in the example, the unknown parameters are constant. Completely different conclusions about stability can be drawn if we consider time-varying parameters $a^- \leq a(t) \leq a^+$, $b^- \leq b(t) \leq b^+$. We must also specify the property. For instance, a system may remain stable for any parameter variation, but may have a bad frequency response for some parameters.

Here we consider a simple case of robustness analysis. In the model

$$\dot{x}(t) = [A + E(t, x(t), w(t)]x(t) \tag{7.57}$$

A represents the nominal state matrix, while E is a perturbation matrix that can depend on time t, on the state and on some external signal w. We may assume that E is unknown, but has a norm that is bounded as follows:

$$\|E(t, x(t), w(t)\| \leq \rho \tag{7.58}$$

The following theorem ensures that, if the nominal system is stable, then there exists a margin in terms of the size ρ of the perturbation E for which robust stability is guaranteed.

Theorem 7.8. *If the nominal system*

$$\dot{x}(t) = Ax(t) \tag{7.59}$$

is asymptotically stable, then there exists $\rho > 0$ such that the perturbed model (7.57) is asymptotically stable for any E as in (7.58).

Proof. If (7.59) is asymptotically stable, then there exists P positive definite such that

$$A^\top P + PA = -I$$

We apply the Lyapunov function $V(x) = x^\top Px$ to the system (7.57). The Lyapunov derivative is then

$$\begin{aligned}
\dot{V}(x) &= 2x^\top P[A + E]x = x^\top(A^\top P + PA) + 2x^\top PEx = -x^\top Ix + 2x^\top PEx \\
&\leq -x^\top x + |2x^\top PEx| \leq -\|x\|^2| + 2|x^\top PEx| \leq 2\|x\|\,\|PEx\| \\
&\leq -\|x\|^2| + 2\|x\|\,\|P\|\,\|E\|\,\|x\| = -\|x\|^2\,(1 - 2\|P\|\,\|E\|)
\end{aligned}$$

In view of Lyapunov theorem, the derivative is negative definite if $2\|P\|\,\|E\| < 1$, namely

$$\rho < \frac{1}{2\|P\|},$$

which is a robustness bound. □

7.9.1 Lyapunov redesign

This technique is quite important in the control of mechanical systems, including robots, which present model uncertainties. For instance, in equation (9.18) we have the term Ω that is due to errors in the model, which cause a non exact cancellation.

Consider the system

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

(for brevity we assume $\Omega = \Delta x$), with $||\Delta|| \leq \delta$ an uncertain term. We assume that A is asymptotically stable, because it is the closed-loop matrix $A = A_0 + BK_0$ achieved by applying a stabilising control K_0 to the nominal system (with $\Delta = 0$). Then, we want to deal with the uncertain term Δ. The nominal closed-loop system satisfies the Lyapunov equation

$$A^\top P + PA = -I$$

We consider a "robustifying" control of this form

$$u = -\gamma^2 B^\top Px,$$

where γ is a parameter. Then

$$\begin{aligned}
\dot{V} &= x^\top(A^\top P + PA)x + 2x^\top PBu + 2x^\top PB\Delta x = -x^\top x - 2\gamma^2 x^\top PBB^\top Px + 2x^\top PB\Delta x \\
&= -||x||^2 - 2\left[\gamma^2 x^\top PBB^\top Px - x^\top PB\Delta x + \frac{x^\top\Delta^\top\Delta x}{4\gamma^2}\right] + 2\frac{x^\top\Delta^\top\Delta x}{4\gamma^2} \\
&= -||x||^2 - 2\left[\gamma B^\top Px - \frac{\Delta x}{2\gamma}\right]^\top\left[\gamma B^\top Px - \frac{\Delta x}{2\gamma}\right] + 2\frac{x^\top\Delta^\top\Delta x}{4\gamma^2} \\
&= -||x||^2 - 2\left\|\gamma B^\top Px - \frac{\Delta x}{2\gamma}\right\|^2 + \frac{||\Delta x||^2}{2\gamma^2} \leq -||x||^2 + 2\frac{||\Delta x||^2}{4\gamma^2} \leq -||x||^2 + \delta^2\frac{||x||^2}{2\gamma^2} \\
&= -||x||^2\left[1 - \frac{\delta^2}{2\gamma^2}\right] < 0
\end{aligned}$$

if we take $\gamma > \delta/\sqrt{2}$.

Chapter 8

Optimal control: basic notions

8.1 General considerations

When we apply a feedback control, typically we would like to assign eigenvalues with a real part that is "very negative" (negative and large in absolute value), in order to have fast transients. However, this implies that the elements of the state feedback matrix K must have large entries. For instance, if in a scalar system we assign $\lambda = -\mu$ with $\mu > 0$ large, by means of the feedback control $u = kx$, then in the equation

$$\dot{x} = ax + bu = (a + bk)x$$

we have $k = (-a - \mu)/b$, so the larger is $\mu > 0$, the larger is the magnitude of k. Too large values are not suitable, due to physical limitations introduced by the actuators:

- there is a trade-off between speed of convergence and control effort.

Moreover, the control action and/or the state variables are often constrained, such as

$$|u| \leq \bar{u}_{max}$$

This kind of problems can be faced with **optimal control** theory.

Essentially, there are two fundamental theories.

- Pontryagin's maximum principle;
- Bellman's dynamic programming;

The first theory provides an open-loop optimal control, while the second theory provides a closed-loop optimal control, which would be preferable in principle, but unfortunately is often intractable from a numerical point of view. For brevity, we will introduce the theory of dynamic programming only, and we will discuss the mentioned difficulties. Still, we will see that, for the fundamental linear-quadratic problem, Bellman's theory leads to an explicit, elegant and very efficient solution. We will also briefly sketch Pontryagin's approach in connection to the calculus of variation.

8.1.1 Lagrange Multipliers

We start the section with a simple consideration about the constrained optimality of functions. Consider the optimization problem

$$\min l(\theta)$$

where $l(\theta)$ is a continuously differentiable function of a vector $\theta \in \mathcal{R}^n$ defined on a domain $\theta \in \Theta$. A necessary condition for a minimum in θ^* internal to Θ is

$$\nabla l(\theta^*) = 0$$

This condition is known as the stationary condition. For instance the minimum of $l = \theta_1^2 + \theta_2^2 - 2\theta_1 - 4\theta_2$ is in the point where

$$\nabla l = [2\theta_1 - 2 \quad 2\theta_2 - 4] = 0$$

namely $\theta = (1, 2)^\top$.

Consider now the constrained optimization problem

$$\min l(\theta) \tag{8.1}$$

$$s.t.$$

$$h(\theta) = 0 \tag{8.2}$$

where l is a functional and $h : \mathcal{R}^n \to \mathcal{R}^m$, is a constraint function with $m < n$. Assume that both h and l are continuously differentiable. Then a necessary condition for a local minimum is that the Lagrangian

$$\mathcal{L}(\theta, \lambda) \doteq l(\theta) + \lambda^\top h(\theta)$$

has a stationary point. Vector $\lambda \in \mathcal{R}^m$ is called the Lagrange multiplier. Namely the constrained minima are computed by applying the stationary condition to $\mathcal{L}(\theta, \lambda)$. Differentiating with respect to λ, we derive $h(\theta) = 0$, namely (8.2), conversely differentiating with respect to θ we derive the equations

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \theta} = 0$$

Example 8.1. *Compute the minimum Euclidean norm solution of $A\theta = b$, $A = m \times n$ matrix, with $m < n$. Assume rank$[A] = m$. The problem is*

$$\min ||\theta||^2/2 \quad s.t. \quad A\theta - b = 0$$

The $1/2$ is added for clarity in the sequel. The Lagrangian is

$$||\theta||^2/2 + \lambda^\top[A\theta - b]$$

If we differentiate with respect to θ we get $\theta^\top = \lambda^\top A$. We transpose it

$$\theta = A^\top \lambda \tag{8.3}$$

and replace it in $A\theta = b$ to get

$$A(A^\top \lambda) = [AA^\top]\lambda = b$$

Since $AA^\top$ is invertible, we have $\lambda = [AA^\top]^{-1}b$, in view of the expression (8.3),

$$\theta^* = A^\top [AA^\top]^{-1} b$$

which is the expression of the minimum norm solution of the system of equations.

8.2 Variational calculus

As a sketch of Pontryagin's principle, we consider the standard problem of finding the scalar function x that minimises the cost

$$\min \int_0^T \, l(x(t), \dot{x}(t)) dt$$

under the boundary conditions $x(0) = a$ and $x(T) = b$ on the fixed interval $[0, T]$. The idea is that, if $\bar{x}$ is optimal, then any small variation

$$x \to x + \delta,$$

with $\delta(t)$ small, does not provide any advantage. Hence,

$$\int_0^T \left[l(\dot{x}(t) + \dot{\delta}(t), x(t) + \delta(t)) - l(\dot{x}(t), x(t))\right] dt \geq 0$$

The use of the first order approximation (assuming differentiability) provides

$$\int_0^T \left[\frac{\partial l(x, \dot{x})}{\partial \dot{x}}\dot{\delta} + \frac{\partial l(x, \dot{x})}{\partial x}\delta\right] \geq 0$$

Integrating by parts, to get rid of the term $\dot{\delta}$, we get

$$\int_0^T \left[\frac{d}{dt}\frac{l(x, \dot{x})}{\partial \dot{x}}\delta + \frac{\partial l(x, \dot{x})}{\partial x}\delta\right] dt - \left[\frac{\partial l(x, \dot{x})}{\partial \dot{x}}\delta\right]_0^T \geq 0$$

The inequality above has to be true for *any* δ, and in particular for those that are zero at the extrema $\delta(0) = \delta(T) = 0$. Hence, considering these perturbations, it must be

$$\int_0^T \left[\frac{d}{dt}\frac{l(x, \dot{x})}{\partial \dot{x}} - \frac{\partial l(x, \dot{x})}{\partial x}\right] \delta dt \geq 0$$

Now the condition must be true if we replace δ by $-\delta$, therefore the inequality is actually an equality

$$\int_0^T \left[\frac{d}{dt}\frac{l(x, \dot{x})}{\partial \dot{x}} - \frac{\partial l(x, \dot{x})}{\partial x}\right] \delta dt = 0$$

Since δ is arbitrary (with $\delta(0) = \delta(T) = 0$), the function between brackets must be 0:

$$\frac{d}{dt}\frac{\partial l(x, \dot{x})}{\partial \dot{x}} - \frac{\partial l(x, \dot{x})}{\partial x} = 0 \tag{8.4}$$

This is the fundamental Euler-Lagrange equation of the calculus of variations, which has to be solved along with the conditions $x(0) = a$ and $x(T) = b$.

The relation with the optimal control is the following. We can rewrite the problem as a special case of the optimal control problem

$$\min \quad \int_0^T l(x(t), u(t))dt \tag{8.5}$$

$$\dot{x}(t) = u(t) \tag{8.6}$$

$$x(0) = a, \quad x(T) = b \tag{8.7}$$

In general, in the optimal control problem, we have

$$J = \int_0^T l(x(t),\ u(t))\, dt \tag{8.8}$$

with the bound $u(t) \in \mathcal{U}$, a given set, and

$$\dot{x}(t) = f(x(t), u(t)),$$

with $x(0) = x_0$ and $x(T) = x_T$. We can conclude that the optimal control problem is more general and includes the calculus of variations as a special case, with the particular equation $\dot{x} = u$.

In this optimal control problem, the solution can be found by considering the Pontryagin equations as follows. Define the Hamiltonian function as

$$H(x, u, \lambda) \doteq l(x, u) + \lambda^\top f(x, u),$$

where $\lambda(t)$ is a vector function with values in $\mathbb{R}^n$, called the co-state. Then the Pontryagin equations are

$$\begin{aligned} \dot{x} &= \frac{\partial H(x, \bar{u}, \lambda)}{\partial \lambda} && (8.9)\\ \dot{\lambda} &= -\frac{\partial H(x, \bar{u}, \lambda)}{\partial x} && (8.10)\\ \bar{u} &= \arg\min_{u \in \mathcal{U}} H(x, u, \lambda) && (8.11)\\ & x(0) = x_0, \quad x(T) = x_T && (8.12) \end{aligned}$$

These equations are typically solved numerically. For $\dot{x} = u$ scalar (under suitable assumptions), we recover Euler-Lagrange equations.

We now derive the Pontryagin equations. As a first step we replace the cost function by

$$\bar{J} = \int_0^T l(x(t),\ u(t)) + \lambda(t)^\top \left[-\dot{x}(t) + f(x(t), u(t))\right] dt = \int_0^T \left[H(x(t), u(t)) - \lambda(t)^\top \dot{x}(t)\right] dt \quad (8.13)$$

and we notice that the problem is unchanged, because $-\dot{x}(t) + f(x(t), u(t)) = 0$. Then we apply the following principle.

Maximum principle. A pair (u, x) is optimal if for any infinitesimal perturbation δu

$$u(t) + \delta u(t) \to x(t) + \delta x(t), \quad \dot{x}(t) + \delta \dot{x}(t)$$

compatible with the constraints

$$\delta x(0) = \delta x(T) = 0$$

the cost does not decrease, namely the cost variation is nonnegative

$$\delta \bar{J} \ = \ \int_0^T \left[H(x + \delta x, u + \delta u, \lambda) - H(x, u, \lambda) + \lambda^\top \delta \dot{x}\right] dt \geq 0$$

Note that

$$\int_0^T \lambda^\top \delta \dot{x} dt = -\int_0^T \dot{\lambda}^\top \delta x dt + \left[\lambda^\top \delta x\right]_0^T = -\int_0^T \dot{\lambda}^\top \delta x dt$$

because $[\lambda^\top \delta x]_0^T = 0$ since no variations of x are possible at the extrema (constraints (8.12)). Then we have

$$\begin{aligned} \delta \bar{J} &= \int_0^T \left[H(x + \delta x, u + \delta u, \lambda) - H(x, u, \lambda) + \dot{\lambda}^\top \delta x\right] dt \\ &= \int_0^T \left[H(x + \delta x, u + \delta u, \lambda) - H(x, u + \delta u, \lambda) + \dot{\lambda}^\top \delta x\right] dt + \int_0^T \left[H(x, u + \delta u, \lambda) - H(x, u, \lambda)\right] dt \\ &\approx \int_0^T \left[\frac{\partial H(x, u, \lambda)}{\partial x} + \dot{\lambda}^\top\right] \delta x\, dt + \int_0^T \left[H(x, u + \delta u, \lambda) - H(x, u, \lambda)\right] dt \geq 0 \end{aligned}$$

Due to the optimality principle, if u is optimal the last inequality must be satisfied, as a necessary condition, for any admissible small variation and for any $\lambda(\cdot)$. If we take $\lambda(t)$ as the solution of the differential equation

$$\frac{\partial H(x, u, \lambda)}{\partial x} + \dot{\lambda}^\top = 0$$

which is (8.10), then the only condition to be ensured is

$$\int_0^T [H(x, u + \delta u, \lambda) - H(x, u, \lambda)]\, dt \geq 0$$

which means that u must minimize the Hamiltonian, namely (8.11).

Example 8.2. ***(Optimal height reaching.)*** *A craft has to reach the optimal-cruise height h after taking off. Assuming that the horizontal speed is constant, we can decide the slope of the trajectory. A height increase $\dot{x}(t)$, with $x(0) = 0$, implies a fuel cost. This cost has to be added to the cost of flying at height $x(t)$, which is smaller at the target height. Let us assume*

$$l(x, u) = \underbrace{\mu(x-h)^2 + \nu}_{\text{fuel consumption at height } x} + \underbrace{\sigma u^2}_{\text{rising consumption}}$$

with

$$\dot{x} = u$$

with no constraints ($u \in \mathbb{R}$). Euler-Lagrange equation (8.4) *becomes*

$$2\sigma\dot{u} + 2\mu(x - h) = 0$$

and then

$$\ddot{x} + \frac{\mu}{\sigma}(x - h) = 0$$

Let us change variable $y = (x - h)$ to get

$$\ddot{y} + \frac{\mu}{\sigma} y = 0$$

with conditions $y(0) = -h$ and $y(T) = 0$. The general solution is

$$y(t) = \alpha e^{\sqrt{\frac{\mu}{\sigma}}t} + \beta e^{-\sqrt{\frac{\mu}{\sigma}}t},$$

where α and β are constants to be determined based on the conditions

$$\alpha e^{\sqrt{\frac{\mu}{\sigma}}0} + \beta e^{-\sqrt{\frac{\mu}{\sigma}}0} = \alpha + \beta = -h \quad \alpha e^{\sqrt{\frac{\mu}{\sigma}}T} + \beta e^{-\sqrt{\frac{\mu}{\sigma}}T} = 0$$

whose the solution is left to the reader.

Note that this approach has some problems if it is formulated on an infinite horizon $T \to \infty$, because $e^{\sqrt{\frac{\mu}{\sigma}}T}$ tends to infinity. Conversely, there is nothing wrong in the fact that $e^{\sqrt{\frac{\mu}{\sigma}}t}$ diverges, since we are considering a finite horizon.

8.3 Dynamic programming

We now briefly present Bellman's approach. Consider the problem of minimising (or maximising) the following **index over an infinite horizon**:

$$J = \int_0^\infty l(x(t),\, u(t))\, dt \tag{8.14}$$

with the bound $u(t) \in \mathcal{U}$, which is a set of assigned constraints, and

$$\dot{x}(t) = f(x(t), u(t))$$

We assume that 0 is an equilibrium

$$0 = f(0, 0)$$

ant that the cost is positive definite: $l(0, 0) = 0$, and l positive elsewhere. One way to proceed is to define a cost-to-go function

$$\psi(x_0) \doteq \text{optimal value of the problem with initial conditions } x_0$$

This function must have the following properties. Let t be the initial instant and let $x(t + h)$ be an intermediate state of the optimal trajectory. Then it should be

$$\psi(x(t)) = \min_{u(t)\in\mathcal{U}} \left\{ \underbrace{\int_t^{t+h} l(x(t),\, u(t))\, dt}_{\text{partial cost}} + \underbrace{\psi(x(t+h))}_{\text{residual optimal cost}} \right\}$$

The minimum depends on the assigned constraints. Since $\psi(x(t))$ is independent of u, we can include it inside the brackets. Dividing by h, we have

$$\min_{u(t)\in\mathcal{U}} \left\{ \frac{\psi(x(t+h)) - \psi(x(t))}{h} + \frac{1}{h}\int_t^{t+h} l(x(t),\, u(t))\, dt \right\} = 0$$

Assume that the function ψ is differentiable. Then, for $h \to 0^+$, the first term gives the derivative and the second term gives $l(x(t),\, u(t))$ due to the mean value theorem

$$\min_{u(t)\in\mathcal{U}} \left\{ \dot{\psi}(x(t)) + l(x(t),\, u(t)) \right\} = 0$$

This has to be true for any t. Now we have $\dot{\psi}(x) = \nabla\psi(x)\dot{x} = \nabla\psi(x)f(x, u)$, which is the Lyapunov derivative of ψ, so we obtain the equation

$$\begin{aligned} &\min_{u\in\mathcal{U}} \{\nabla\psi(x) f(x, u) + l(x,\, u)\} = 0, \\ &\psi(0) = 0 \end{aligned} \tag{8.15}$$

named Bellman equation of dynamic programming. If the function $\psi(x)$ is known, then the control can be derived as the element of $u = u(x)$ that minimises the above expression. Unfortunately, for this equation it is not possible to determine an analytic solution in most cases. Numerical methods have to be adopted (which are computationally very heavy).

Example 8.3. *Consider the take-off problem in the y variable*

$$l(y, u) = \mu y^2 + \nu + \sigma u^2$$

with

$$\dot{y} = u$$

The Bellman equation is

$$\min_{u\in U} \left\{ \frac{d}{dy}\psi(y) u + \sigma u^2 + \mu y^2 \right\} = 0$$

and the minimiser $\bar{u}$ is achieved by setting the derivative to 0

$$\bar{u} = -\frac{1}{2\sigma}\psi'(y)$$

Replacing ū in the equation we get

$$-\frac{1}{4\sigma}\psi'(y)^2 + \mu y^2 = 0$$

Hence $\psi'(y) = 2\sqrt{\mu\sigma}y$ *and by integration (note that* $\psi(0) = 0$*)*

$$\psi(y) = \sqrt{\mu\sigma}y^2$$

The optimal control is

$$u = -\sqrt{\frac{\mu}{\sigma}}\; y \tag{8.16}$$

Note that the control (8.16) *is linear. The closed loop system is* $\dot{y} = -\sqrt{\frac{\mu}{\sigma}}y$*, hence it is stable. This is NOT the same achieved with Pontryagin's method. Why?*

8.3.1 Linear quadratic control

In the particular case of linear systems and quadratic cost, the solution of the problem can be easily determined. This is a remarkable property and this control is very popular and efficient.

Consider the problem

$$J = \frac{1}{2}\int_0^\infty \left(x^\top(t)Qx(t) + u^\top(t)Ru(t)\right)dt \tag{8.17}$$

with (8.18)

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ x(0) = x_0 \end{cases}$$

Matrix Q is assumed to be positive semi-definite (of course it may be also positive definite), while R is assumed to be symmetric positive definite. We assume that $U = \mathbb{R}^m$, namely there are no constraints. Equation (8.15) becomes

$$\min_{u(t)\in U}\left\{\nabla\psi(x)(Ax + Bu) + \frac{1}{2}(x^\top Qx + u^\top Ru)\right\} = 0 \tag{8.19}$$

We try a solution in the form $\psi(x) = x^\top Px/2$, with P positive definite, and we replace $\nabla\psi(x) = x^\top P$

$$\min_{u(t)\in U}\left\{x^\top P(Ax + Bu) + \frac{1}{2}(x^\top Qx + u^\top Ru)\right\} = 0 \tag{8.20}$$

For any fixed x, the minimum with respect to u is obtained by considering the gradient (with respect to u, namely assuming x constant) and setting it equal to zero

$$x^\top PB + u^\top R = 0$$

This means

$$u = -R^{-1}B^\top Px,$$

which is a linear control. By replacing this expression in the equation we have

$$x^\top PAx - x^\top PBR^{-1}B^\top Px + \frac{1}{2}(x^\top Qx + x^\top PBR^{-1}RR^{-1}B^\top Px) = 0$$

Since $x^\top PAx = (x^\top PAx + x^\top A^\top Px)/2$, we obtain

$$\frac{1}{2}x^\top[A^\top P + PA - PBR^{-1}B^\top P + Q]x = 0,$$

which has to be valid for each x. This brings us to the Riccati equation

$$A^\top P + PA - PBR^{-1}B^\top P + Q = 0 \tag{8.21}$$

We look for a positive definite solution P of this equation. Indeed, only positive definite solutions are meaningful, since the cost has to be positive. Note that in general there is more than one solution. To see this, it is sufficient to consider the scalar case

$$2pa - \frac{b^2}{r}p^2 + q = 0$$

(the products commute in this case). This second order equation has two solutions and we are interested in the positive one. In general, the Riccati equation has a single positive definite solution. We conclude with the following important result.

Theorem 8.1. *Assume that* (A, B) *is stabilisable and* Q *is positive definite. Then* (8.21) *admits a unique positive definite solution* P*. The optimal control is*

$$u = -R^{-1}B^\top Px = K_{opt}x \tag{8.22}$$

and the optimal cost with initial condition x_0 *is*

$$J_{opt} = x_0^\top P x_0 \tag{8.23}$$

We do not give a technical proof of the theorem. Just observe that, if Q is positive definite, then the integrand function in the cost is positive, unless $x_0 = 0$, hence $x_0^\top P x_0 > 0$, so P must be positive definite.

We can easily see that, if Q is positive definite, then the closed-loop system is asymptotically stable. Write (8.21) as

$$(A^\top P - PBR^{-1}B^\top P) + (PA - PBR^{-1}B^\top P) + Q + PBR^{-1}B^\top P = 0$$

The term $(A - BR^{-1}B^\top P) = (A + BK_{opt}) = A_{CL}$ is the closed-loop matrix. Then we obtain

$$A_{CL}^\top P + PA_{CL} + \underbrace{Q + PBR^{-1}B^\top P}_{=\hat{Q}} = 0 \tag{8.24}$$

Being Q positive definite and $PBR^{-1}B^\top P$ positive semi-definite, also $\hat{Q} = Q + PBR^{-1}B^\top P$ is positive definite. This means that (8.24) is a Lyapunov equation, hence the closed-loop system is asymptotically stable.

Remark 8.1. *It is not necessary to have* Q *positive definite. For instance, if we have a performance output* $z = Hx$*, the state term is*

$$z^\top z = x^\top H^\top H x = x^\top Q x$$

and in this case Q *is positive semi-definite. Still, matrix* $\hat{Q} = Q + PBR^{-1}B^\top P$ *and matrix* P *can be positive definite, and in this case the closed-loop is asymptotically stable. We need to solve the equation and check that* P *and* $\hat{Q}$ *are positive definite. A trick to avoid problems is to perturb the cost as* $H^\top H \to H^\top H + \epsilon I$*, which is positive definite. If* ϵ *is small, we are virtually not changing the problem.*

Example 8.4. *Consider*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

and $R = [1]$. The state variables are: x_1 the position, x_2 the speed and u the acceleration (controlled) of a cart. The cost is the integral of the square of the position plus the square of the acceleration. The positive definite solution of the Riccati equation is

$$P = \begin{bmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{bmatrix}$$

Q is not positive definite, however

$$\hat{Q} = Q + PBR^{-1}B^\top P = \begin{bmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}$$

is positive definite and the control is stabilising. The closed-loop poles are

$$\lambda_{12} = -\frac{1}{\sqrt{2}} \pm j\frac{1}{\sqrt{2}}$$

Conversely, if we take $H = \begin{bmatrix} 0 & 1 \end{bmatrix}$, we get $P = Q$ and the control is not stabilising. This is not surprising. We are penalising the speed and the control. So, if the initial condition is $x_1 \neq 0$ and $x_2 = 0$, the optimal control is $u = 0$ (do nothing and stay in position x_1), with cost 0, because both speed and acceleration are 0.

The optimal control has infinite gain margin. If we modify the control as

$$u = -\kappa R^{-1}B^\top Px$$

with $\kappa \geq 1$ arbitrarily large, the system remains stable. This can be easily checked, because we can derive a Lyapunov equation similar to (8.24)

$$(A - \kappa BR^{-1}B^\top P)^\top P + P(A - \kappa BR^{-1}B^\top P) + Q + (2\kappa - 1)PBR^{-1}B^\top P(\kappa - 1)PBR^{-1}B^\top P = 0$$

In some cases, increasing the gain κ improves robustness (see Chapter 7.9).

An interesting observation is the following. Consider a performance output $z(t) = Hx(t)$ and the cost

$$J = \frac{1}{2}\int_0^\infty \left(x^\top(t)H^\top Hx(t) + u^\top(t)u(t)\right)dt = \frac{1}{2}\int_0^\infty \left(\|z(t)\|^2 + \|u(t)\|^2\right)dt$$

This is a particular case with $R = I$ and $Q = H^\top H$. If we consider the “augmented output“

$$\begin{bmatrix} u(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} K \\ H \end{bmatrix} x(t)$$

we get the corresponding system

$$\begin{aligned} \dot{x}(t) &= [A + BK]x(t) + I\Delta(t) && (8.25) \\ u(t) &= Kx(t) && (8.26) \\ z(t) &= Hx(t) && (8.27) \end{aligned}$$

with input Δ. Denoting by $W(t)$ the matrix of impulse responses, the LQ control is the one that minimises

$$J_{tr} = \frac{1}{2}\int_0^\infty tr[W(t)^\top W(t)]dt = \frac{1}{2}\int_0^\infty \sum_{ij} W_{ij}^2(t)$$

where $tr[S]$ is the trace of the matrix S.[1] Note that

$$tr[M^\top M] = tr[MM^\top] = \sum_{ij} M_{ij}^2$$

The Riccati equation to be solved in this particular case is

$$A^\top P + PA - PBB^\top P + H^\top H = 0 \tag{8.28}$$

[1] The trace is the sum of the diagonal elements, $tr[S] \doteq \sum S_{ii}$.

8.3.2 Optimal observer: the Kalman-Bucy filter

Consider a system in the presence of noise

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Ev(t) \\ y(t) &= Cx(t) - w(t) \end{aligned}$$

where $v(t)$ and $w(t)$ are unknown noises or disturbances. If we consider the standard observer

$$\dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + Bu(t) + Ly(t),$$

defining the error as $e = \hat{x} - x$, we obtain the error equation

$$\begin{aligned} \dot{e}(t) &= [A - LC]e(t) - Lw(t) - Ev(t) && (8.29) \\ \eta(t) &= Ie(t) && (8.30) \end{aligned}$$

If we consider the dual system of (8.29)-(8.30)

$$\left[\begin{array}{c|c} A - LC & [-L\ E] \\ \hline I & 0 \end{array}\right]^* = \left[\begin{array}{c|c} A^\top + C^\top(-L)^\top & I \\ \hline \begin{bmatrix} (-L)^\top \\ H^\top \end{bmatrix} & 0 \end{array}\right] = \left[\begin{array}{c|c} A^* + B^*K^* & I \\ \hline \begin{bmatrix} K^* \\ H^* \end{bmatrix} & 0 \end{array}\right]$$

we get a system of the form (8.25)-(8.27), so its impulse response matrix is the transpose of the impulse response matrix of (8.29)-(8.30). It can be easily seen that $tr[M^\top M] = tr[MM^\top]$. So, to minimise

$$J_{tr} = \frac{1}{2}\int_0^\infty tr[W(t)W(t)^\top]dt,$$

we can exploit duality and replace K with $-L^\top$, H with $-E^\top$ and B with $C^\top$. By transposing (8.28), we obtain the dual Riccati equation

$$PA^\top + AP - PC^\top CP + EE^\top = 0 \qquad (8.31)$$

and the optimal filter gain

$$L = PC^\top.$$

Duality between optimal control and optimal filtering is a fundamental achievement in control theory. It can be shown that, by properly combining an optimal observer and an optimal state feedback control, we get an optimal output feedback control.

8.4 Connection filter-optimal control

In this section we examine the results we achieve by combining the optimal control with the Kalman filter. We start with a technical computation concerning the performance loss we achieve by replacing the optimal control by a generic stabilizing control u. For convenience we take $R = I$,

$Q = H^\top H$ and $x(0) = x_0$. Then

$$\begin{aligned}
J &= \int_0^\infty \left[x^\top H^\top Hx + u^\top u\right] dt \\
&= \int_0^\infty \left[x^\top H^\top Hx + \left[u + B^\top Px - B^\top Px\right]^\top \left[u + B^\top Px - B^\top Px\right] + \frac{d}{dt}x^\top Px\right] dt + x_0^\top Px_0 \\
&= \int_0^\infty \left[u + B^\top Px\right]^\top \left[u + B^\top Px\right] dt + x_0^\top Px_0 \\
&+ \int_0^\infty \left[x^\top H^\top Hx - 2\left[u + B^\top Px\right]^\top B^\top Px + x^\top PBB^\top Px + 2x^\top PAx + 2x^\top PBu\right] dt \\
&= \int_0^\infty \left[u + B^\top Px\right]^\top \left[u + B^\top Px\right] dt + x_0^\top Px_0 + \int_0^\infty x^\top \underbrace{\left[H^\top H - PBB^\top P + PA + A^\top P\right]}_{=0} xdt \\
&= \int_0^\infty \left[u + B^\top Px\right]^\top \left[u + B^\top Px\right] dt + x_0^\top Px_0
\end{aligned}$$

As expected any control is worse than $u_{opt} = -B^\top Px$, for which the optimal performance is $x_0^\top Px_0$ is ensured.

To evaluate the performance of the combination Kalman filter and optimal control we consider $u = -B^\top P\hat{x}$, the feedback of the estimated state to get

$$\begin{aligned}
J &= \int_0^\infty \left[u + B^\top Px\right]^\top \left[u + B^\top Px\right] dt + x_0^\top Px_0 = \int_0^\infty \left[B^\top P(x - \hat{x})\right]^\top \left[B^\top P(x - \hat{x})\right] dt \\
&= \int_0^\infty \|B^\top Pe\|^2 dt + x_0^\top Px_0
\end{aligned}$$

where $e = x - \hat{x}$ is the estimation error. On the other hand the Kalman filters minimizes

$$\int_0^\infty \|e\|^2 dt$$

hence the combination is optimal.

If we consider, instead of a generic initial condition, the impulse response, we can replace, for $m = 1$, $E = x_0$

$$J = \int_0^\infty \|B^\top Pe\|^2 dt + E^\top PE$$

For multiple input systems, we just need to take the trace

$$J = \int_0^\infty \|B^\top Pe\|^2 dt + \mathrm{tr}(E^\top PE)$$

8.5 Model predictive control

Model predictive control is a popular technique in which the control is computed by solving an optimization problem on-line. Consider the discrete–time problem with initial time k_0

$$\begin{aligned}
\min \quad & \sum_{k=k_0}^{\infty} g(x(k), u(k)) \\
s.t. \quad & \\
x(k+1) \;=\; & f(x(k), u(k)) \\
& x(k) \in \mathcal{X}, \quad u(k) \in \mathcal{U}, \\
x(k_0) \quad & \text{assigned}
\end{aligned}$$

where we assume that $(0,0)$ is the target equilibrium point. Sets $\mathcal{X}$ and $\mathcal{U}$ are constraint sets. In most cases they are of the form

$$\mathcal{X} = \left\{x : \ x_i^- \leq x \leq x_i^+\right\} \quad \mathcal{U} = \left\{u : \ u_i^- \leq u \leq u_i^+\right\}$$

The idea is the following. First consider problem with a finite horizon T

$$\begin{array}{rcl} \min & & \displaystyle\sum_{k=k_0}^{k_0+T} g(x(k),u(k)) + h(x(k_0+T)) \\ s.t. & & \\ x(k+1) & = & f(x(k),u(k)) \\ & & x(k) \in \mathcal{X}, \quad u(k) \in \mathcal{U}, \\ x(0) & & \text{assigned} \end{array}$$

with T integer large enough. The term $h(x(k_0+T))$ is a weight on the final state. It can be replaced by a constraint such as $x(k_0+T) = 0$. Assume that this problem can be efficiently solved on–line (real–time). Limiting ourself in solving it at the beginning $k = 0$, we would get an open–loop solution. To derive a feedback solution we can apply the following procedure.

1. At $k_0 = 0$, given $x(k_0)$ compute the optimal sequence $u^*(k_0)$, $u^*(k_0+1), \dots u^*(k_0+T-1)$;
2. Apply only the first input $u^*(k_0)$;
3. Measure the new state $x(k_0+1) = f(x(k_0),u^*(k_0))$ and consider it as the new initial state;
4. Set $k_0 := k_0 + 1$ and GOTO 1 (recompute the sequence);

This technique, known as model predictive control (or receding horizon control) is very popular. Its implementation requires the solution of an optimization problem at each step. This is reasonable in most applications in view of the current computer performances. The solution time must be smaller than the sampling time. Several commercial solvers are available to implement the technique.

One important case is that of linear systems with linear constraints and quadratic cost

$$\begin{array}{rcl} \min & & J = \displaystyle\sum_{k=k_0}^{k_0+T-1} x(k)^\top Qx(k) + u(k)^\top Ru(k) + x(k_0+T)^\top S\, x(k_0+T) \\ s.t. & & \\ x(k+1) & = & Ax(k) + Bu(k)) \\ y(k) & = & Hx(k) + Mu(k)) \\ & & y_i^- \leq y \leq y_i^+ \\ x(0) & & \text{assigned} \end{array}$$

where Q, R and S are assumed positive definite matrices. Matrix S weights the final state. An alternative possibility to this weight is the additional constraint $x(T) = 0$

The linear constraints ensure that by defining

$$\xi = \begin{bmatrix} x(k_0+1) \\ x(k_0+2) \\ \vdots \\ x(k_0+T) \\ u(k_0) \\ u(k_0+1) \\ \vdots \\ u(k_0+T-1) \end{bmatrix}, \quad \Psi = \left[\begin{array}{cccc|cccc} Q & 0 & \dots & 0 & 0 & \dots & \dots & 0 \\ 0 & Q & 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & S & 0 & \dots & \dots & 0 \\ \hline 0 & \dots & \dots & 0 & R & 0 & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 & R & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 0 & \dots & 0 & R \end{array}\right]$$

we can write the optimization problem as

$$\begin{aligned} \min \quad & J = \xi^\top \Psi \xi \\ s.t. \quad & \\ \Phi\xi = \phi(x(k_0)) & \\ \xi^- \leq \xi \leq \xi^+ & \\ x(k_0) \quad & \text{assigned} \end{aligned}$$

where Φ is a (large) matrix with nT rows and $(n+m)T$ columns, $\phi(x(k_0))$ is a nT–vector function of the initial condition. This is a linear–quadratic optimization problem. This is a convex optimization problem for which efficient solvers are available.

The next issue is how can we enforce stability of the scheme. Essentially, the following property holds. Assume that (A, B) is reachable and that ξ^- and ξ^+ are positive and negative vectors respectively (so $x = 0$ and $u = 0$ are feasible vectors). Assume that the optimization problem has a feasible solution for the initial condition $x(0)$. Note that under constraints this is not always ensured. For instance

$$x(k+1) = 2x(k) + u(k), \quad |u(k)| \leq 1$$

if we take $x(0) = 2$, there is no constrained control driving the state to 0.

The following general principle holds: The closed–loop stability is ensured if the horizon T is large enough for all initial states which belong to a convex set $\mathcal{D}$ including the origin in its interior.

There are several results which provide the length of the horizon T to have stability. We prove closed-loop stability with the additional assumption: the following constraint is added to the optimization

$$x(k_0 + T) = 0 \tag{8.32}$$

this means that the final state is imposed to be 0.

Theorem 8.2. *The model–predictive scheme with the additional constraint* (8.32) *is stabilizing.*

Proof Consider the optimal cost with initial condition x_0

$$V(x_0) = \text{optimal constrained cost } J \text{ with initial condition } x_0$$

This is a Lyapunov function for the closed–loop system. Consider the optimal sequences

$$u(k_0),\ u(k_0+1),\ \dots u(k_0+T-1), \quad x(k_0+1),\ x(k_0+2),\ \dots x(k_0+T) = 0$$

(the last equality is the constraint). If we apply the first input $u(k_0)$ we get the state new $x(k_0 + 1)$. This is the new initial state for the following optimization. Consider the new sequences from the new initial state $x(k_0 + 1)$

$$u(k_0+1),\ u(k_0+1),\ \dots u(k_0+T-2), 0, \quad x(k_0+2),\ x(k_0+3),\ \dots x(k_0+T+1) = 0,\ x(k_0+T+1) = 0$$

in which the last two states are at 0, and the last control is at zero. This is an admissible *new* sequence (it satisfies the constraints) because all elements are also in the previous sequence or 0. The cost $\tilde{J}$ of such a new sequence is smaller than the previous sequence because the former state $x(k_0 + 1)$ and input $u(k_0)$, which have positive costs, do not appear since they are replaced by the null state $x(k_0 + T + 1) = 0$ and by the null input $u(k_0 + T - 1) = 0$ at the end. All the other states and inputs of the new sequence are the shifted value of the previous sequence.

The new sequence is not optimized. If we optimize we find the optimal cost which is (of course) not greater than $\tilde{J}$, hence

$$V(x(k_0 + 1)) \leq \tilde{J} < V(x_0)$$

This means that the cost–to–go $V(x_0)$, namely the optimal cost is a Lyapunov function and this proves asymptotic stability. □

Chapter 9

Examples of dynamical systems

In this chapter, we present models and examples of dynamical systems. These examples are used as application benchmarks in the course.

9.1 DC electrical machine

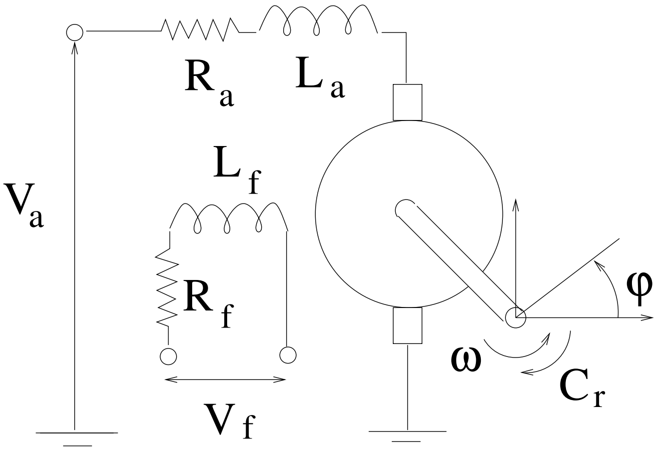

Figure 9.1: General model of a DC motor.

Consider Figure 9.1, representing the diagram of a DC (direct current) electrical machine. This device can be used in two different modes.

- **Generator mode**: mechanical power is provided and electrical power is produced.
- **Motor mode**: electrical power is provided and mechanical power is produced.

The current flows in the armature circuit of the rotor, connected to an external supplier (or an electrical load in the case of generator mode) through a collector. In view of the rotation, the rotor coils move with respect to the magnetic field. This creates a force, hence a torque, and an electromotive force. These simultaneous effects are the basic principle of the electromechanical energy conversion.

The equations of the machine are the following. The field circuit, which has the goal of generating the flow in the machine, has equation

$$L_f \frac{di_f(t)}{dt} = -R_f i_f(t) + v_f(t), \tag{9.1}$$

where L_f is the inductance, i_f is the field circuit current, R_f is the resistance, i_f the current and v_f the field generator voltage. In small machines, the field circuit is replaced by a permanent magnet that generates a constant flow, hence this equation is not present. The excitation circuit equation is

$$L_a \frac{di_a(t)}{dt} = -R_a i_a(t) - \hat{k}\, i_e(t)\, i_a(t) + v_a(t), \tag{9.2}$$

where L_a is the armature inductance, R_a is the resistance, i_a is the armature current and v_a the applied voltage (or the potential difference on the load). The term $\hat{k}\, i_e(t)\, i_a(t)$ is the electromotive force. There are two mechanical equations. The torque equation is

$$\begin{aligned} J\dot{\omega}(t) &= C_m(t) - f\omega(t) - C_r(t) \\ \dot{\varphi}(t) &= \omega(t) \end{aligned} \tag{9.3}$$

In the first equation J is the inertia, ω the rotational speed, f is a viscous coefficient, $C_r(t)$ is a torque due to the load. The term C_m is the torque produced by the machine. The second equation is the definition of rotational speed, which is the derivative of the machine angle.

These equations are nonlinear. We can assume, as input and state variables, the vectors

$$u(t) = \begin{bmatrix} v_e(t) \\ v_a(t) \\ C_r(t) \end{bmatrix} \in \mathbb{R}^3 \qquad x(t) = \begin{bmatrix} i_e(t) \\ i_a(t) \\ \omega(t) \\ \varphi(t) \end{bmatrix} \in \mathbb{R}^4 \tag{9.4}$$

As far as the output of the system is concerned, this depends the problem (and also on the applied sensors). Typically, if the goal is to design a control system for the machine, we can take as an output vector

$$y(t) = \begin{bmatrix} \varphi(t) \\ \omega(t) \end{bmatrix} \tag{9.5}$$

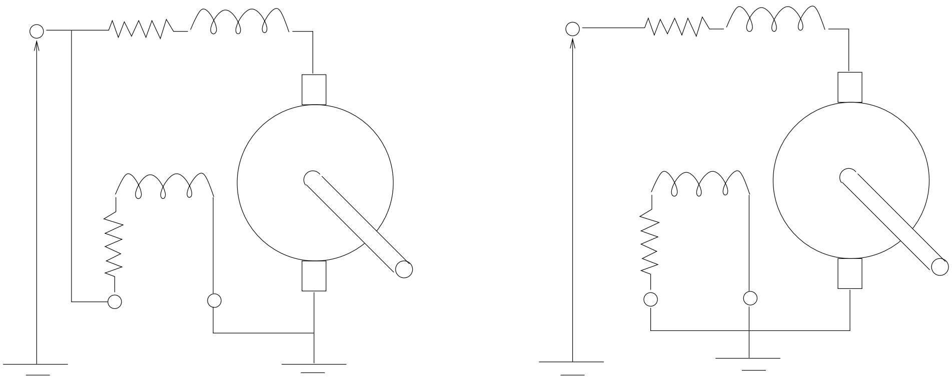

Figure 9.2: Series and parallel connections.

The two electrical circuits of the machine can be connected in different ways. There are three main possibilities.

- Parallel connection (Figure 9.2, left): $v_e = v_a$, hence the two inputs are reduced to one.
- Series connection (Figure 9.2, right): $i_e(t) = i_a(t)$; hence, two state variables are replaced by one.

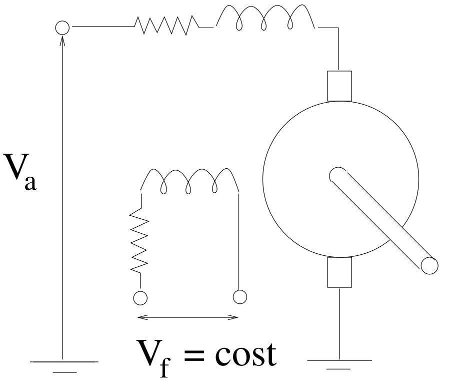

Figure 9.3: Independent connection.

- Independent connection (Figure 9.3): typically we set $v_e(t) = \bar{v}_e$ constant. In this case the field equation is removed and we assume i_e =constant.

In the independent configuration, the equation of the excitation circuit is decoupled from the other, as follows:

$$\frac{di_e}{dt} = \frac{R_e}{L_e} i_e(t) + \frac{\bar{v}_e}{L_e} \tag{9.6}$$

and yields the steady-state value $\bar{i}_e = \frac{\bar{v}_e}{R_e}$. Therefore, assuming that the transient for i_e is over, and $\bar{i}_e$ is at its steady-state value, this current can be considered as a constant. The equation for i_e is not necessary anymore. Let us then write $\hat{k}i_e = k$. We get

$$\begin{aligned} L_a \frac{di_a(t)}{dt} &= -R_a i_a(t) - k\omega(t) + v_a \\ J\dot{\omega}(t) &= C_m(t) - f\omega(t) - C_r \\ \dot{\varphi}(t) &= \omega(t) \end{aligned}$$

Let us multiply the first equation by i_a and the second by $\omega(t)$ to derive the power balance

$$\underbrace{\frac{d}{dt}\left(\frac{1}{2} L_a \frac{di_a}{dt}^2\right)}_{\text{magnetic power}} + \underbrace{R_a i_a^2}_{\text{resistor dissipated power}} + \underbrace{k\omega i_a}_{\text{converted power}} = \underbrace{v_a i_a}_{\text{supplied power}}$$

$$\underbrace{\frac{d}{dt}\left(\frac{1}{2} J\omega^2\right)}_{\text{kinetic power}} + \underbrace{f\omega^2}_{\text{mechanical dissipated power}} + \underbrace{C_r\omega}_{\text{used power}} = \underbrace{C_m\omega}_{\text{converted power}}$$

Considering the converted power as it appears in the two equations, we have $C_m\omega = k\omega i_a$, hence the following expression for the torque:

$$C_M = k i_a.$$

The equations are

$$\begin{aligned} \frac{di_a(t)}{dt} &= \frac{R_a}{L_a} i_a(t) - \frac{k}{L_a}\omega(t) + \frac{v_a(t)}{L_a} \\ \dot{\omega}(t) &= \frac{k}{J} i_a(t) - \frac{f}{J}\omega(t) - \frac{C_r(t)}{J} \\ \dot{\varphi}(t) &= \omega(t) \end{aligned} \tag{9.7}$$

with the same outputs previously chosen. It is possible to write the system in matrix form as follows. State, input and output vectors are

$$x(t) = \begin{bmatrix} i_a(t) \\ \omega(t) \\ \varphi(t) \end{bmatrix}, \qquad u(t) = \begin{bmatrix} v_a(t) \\ C_r(t) \end{bmatrix}, \qquad y(t) = \begin{bmatrix} \varphi(t) \\ \omega(t) \end{bmatrix} \tag{9.8}$$

Then, denoting by $\alpha = \frac{R_a}{L_a}, \beta = \frac{k}{L_a}, \gamma = \frac{k}{J}, \delta = \frac{f}{J}, \epsilon = \frac{1}{L_a}, \mu = \frac{1}{J}$,

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} -\alpha & -\beta & 0 \\ \gamma & -\delta & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} \varepsilon & 0 \\ 0 & -\mu \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$$

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} \tag{9.9}$$

To find the modes of the system, we look for the eigenvalues of matrix A, hence we write the characteristic polynomial

$$\det(sI - A) = \det \begin{bmatrix} s+\alpha & \beta & 0 \\ -\gamma & s+\delta & 0 \\ 0 & -1 & s \end{bmatrix} = s[(s+\alpha)(s+\delta) + \beta\gamma]$$

A root of the characteristic polynomial is $s = 0$. The other two eigenvalues, for machines of this type, are typically real and negative. The modes of the system are therefore the type $e^{\lambda_1 t}$, $e^{\lambda_2 t}$, 1 where the latest mode (due to $s = 0$) is always present for any parameter values. This depends on the fact that a machine that starts from an initial non-zero angle and 0 speed and current remains in the same position. The transfer function is:

$$W(s) = \frac{\begin{bmatrix} n_{11}(s) & n_{12}(s) \\ n_{21}(s) & n_{22}(s) \end{bmatrix}}{s[(s+\alpha)(s+\delta) + \beta\gamma]} \tag{9.10}$$

where the numerator entries can be computed according the formula

$$n_{ij}(s) = \det \begin{bmatrix} sI - A & -B_j \\ C_i & D_{ij} \end{bmatrix} \tag{9.11}$$

where C_i is the i-th row of matrix C, and B_j is the j-th column of matrix B. For example, considering the first input ($u_2 \equiv 0$) and the first output $y_1(t)$, we can compute $y_1(s) = W_{11}(s)\, u_1(s)$, where:

$$n_{11}(s) = \det \begin{bmatrix} s+\alpha & +\beta & 0 & -\varepsilon \\ -\gamma & s+\delta & 0 & 0 \\ 0 & -1 & s & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \varepsilon\gamma \tag{9.12}$$

The computation of the other elements n_{ij} is analogous.

9.2 Robot arm

Consider Figure 9.4, representing a robot arm. The equations of such a system, with three degrees of freedom, are not easy to write. In general, however, we can say that equations of the following form can be derived

$$M(q(t))\ddot{q}(t) + H(q(t), \dot{q}(t))\dot{q}(t) + K(q(t)) = \tau(t), \tag{9.13}$$

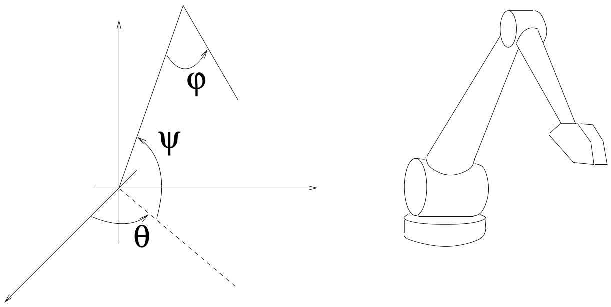

Figure 9.4: Scheme of a robot arm.

where the components of the vectors $q(t),\ \dot{q}(t),\ \ddot{q}(t)$ are, respectively, angles, angular velocities and angular accelerations. In general robotic systems, q is a vector of free coordinates.

Assuming that $M(q(t))$ is a non-singular matrix (therefore, it is invertible), we can choose the following control law:

$$\tau(t) = H(q(t), \dot{q}(t))\dot{q}(t) + K(q(t)) + M(q(t))u(t), \tag{9.14}$$

where u is a new control signal to be specified later. Note that, in order to implement such a control, we need sensors that measure angles and angular velocities. Replacing τ in the initial equations, we get

$$M(q(t))\ddot{q}(t) = M(q(t))u(t) \Rightarrow M(q(t))[\ddot{q}(t) - u(t)] = 0. \tag{9.15}$$

Since $M(q(t))$ is non-singular by assumption, it must be

$$\ddot{q}(t) = u(t) \tag{9.16}$$

The resulting system (included in the dotted rectangle in Figure 9.5) is linear and can be controlled by means of one of the (several) available techniques. This process is known as **cancellation of the nonlinearities**. Based on this, linear **feedback** allows us to effectively control the robot arm. Adopting the linearising feedback provides m decoupled equations (m is the number of degrees of freedom). Then we may decide to control all of these coordinates q_i independently. In general, the state-space representation is:

$$q(t) \in \mathbb{R}^m\ , \qquad u(t) \ \in \mathbb{R}^m, \quad x(t) = \begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} \in \mathbb{R}^{2m}$$

$$\frac{d}{dt}\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} u(t)$$

where $I \in \mathbb{R}^m$ is the identity matrix. Typically, the sensors for this system are encoders (measuring angles) hence

$$y(t) = \begin{bmatrix} I & 0 \end{bmatrix}\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix}$$

If also speed sensors are available, the output is the state $y(t) = x(t)$, that is

$$y(t) = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix}.$$

We can actually avoid the speed sensors, because the speed can be determined from the angles by means of an observer.

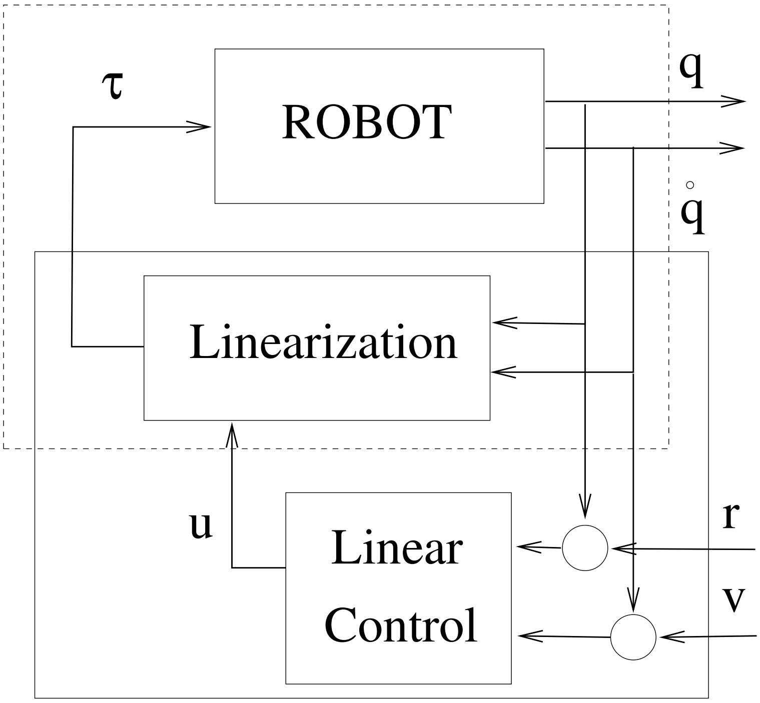

Figure 9.5: Control of a robot arm: r and v are position and speed references.

There is a major issue here. The cancellation (9.14) is not exact, because it is based on a model:

$$\tau(t) = \tilde{H}(q(t), \dot{q}(t))\dot{q}(t) + \tilde{K}(q(t)) + \tilde{M}(q(t))u(t) \tag{9.17}$$

Therefore, there is an error term in the resulting equation

$$\ddot{q}(t) = u(t) + \Omega \tag{9.18}$$

caused by a non exact cancellation, where

$$\Omega = M^{-1}[(\tilde{M} - M)u + (\tilde{H} - H)\dot{q} + (\tilde{K} - K)]$$

is an uncertain term. Hence, the control must be robust against this uncertainty. In Section 7.9, we consider a simple case in which $\Omega = \Delta x$.

9.3 Mortgage payment

Even economic phenomena can be studied by the theory of dynamical systems. For example, a bank account can be described by the following equation:

$$x(k+1) = x(k) + r(k) + i\,x(k), \tag{9.19}$$

where $x(k)$ represents the account balance, $r(k)$ the money transfer, and i the interest. This is a discrete-time system, *i.e.*, k time units represent k days (or months). The underlying hypothesis is that payments or withdrawals $r(k)$ are accounted daily (or monthly). We have

$$x(k+1) = (1+i)x(k) + r(k) \quad \Rightarrow \quad x(k+1) = \alpha x(k) + r(k), \quad \text{with} \quad \alpha = 1 + i, \tag{9.20}$$

which is a simple linear discrete-time system. According to the formula derived in the initial chapter on discrete-time systems, if we consider $r(k)$ as the system input, the evolution of the bank account balance is:

$$x(k) = \alpha^k x(0) + \sum_{h=0}^{k-1} \alpha^{k-h-1} r(h). \tag{9.21}$$

9.3.1 Mortgage installment computation

Now suppose that, in the account, we must consider the negative contribution of a mortgage of value C, which has to be extinguished in N time units (starting with with $-C$), by paying a constant installment $\bar{r}$. Which is the amount of the constant installment, to be paid at each time k, such that the loan is extinguished at time N?

We can simply put $x(0) = -C$ as the initial condition and consider that, at the instant $k = N$, we must have $x(N) = 0$ by contract:

$$0 = \alpha^N(-C) + \sum_{h=0}^{N-1} \alpha^{N-h-1}\bar{r} \tag{9.22}$$

Since the sum is the classic geometric series $\sum_{i=0}^{n} \alpha^i = \frac{\alpha^{n+1} - 1}{\alpha - 1}$, we get the installment expression

$$\bar{r} = \frac{\alpha^N(\alpha - 1)}{\alpha^N - 1} C = \frac{(1+i)^N}{(1+i)^N - 1} i C \tag{9.23}$$

An interesting observation comes out if we write the installment according to the evolution at a generic time

$$\bar{r} = -i\,x(k) + [x(k+1) - x(k)]. \tag{9.24}$$

When k is small, $x(k)$ is large (negative) and the term $-i\,x(k) > 0$ prevails. The situation reverses near the end of the loan (k close to N). One can thus conclude that, in the case of constant installment, we return more interest than capital at the beginning, and vice-versa at the end.

9.4 Positive linear systems and Markov chains

A linear discrete-time system is positive if it is represented by the equation

$$x(k+1) = Ax(k), \quad A_{ij} \geq 0$$

Nonnegative matrices have remarkable properties, such as positivity of the associated system:

$$x(0) \geq 0 \Rightarrow x(k) \geq 0, \quad \forall k > 0.$$

(the notation $x \geq 0$ for a vector is to be intended component-wise $x_i \geq 0$). A further property is the Perron-Frobenius theorem.

Theorem 9.1. *The eigenvalue with the largest modulus (dominating eigenvalue, also denoted as Frobenius eigenvalue) of a nonnegative matrix is nonnegative and real. In other words, if we order the eigenvalues by magnitude,*

$$\lambda_1 = |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$$

Moreover, if λ_1 is strictly dominating, i.e., $\lambda_1 = |\lambda_1| > |\lambda_2| \dots$ then the corresponding eigenvector (Frobenius eigenvector) has positive components.

Example 9.1. *A simplified population evolution is the following. There are three classes of age. Youth x_1, middle age x_2, and old age x_3.*[1] *At each time step, each population unit passes to the older stage. In the middle age the population is assumed to reproduce and give birth, in the average, to α new young individuals. In the old age the population is assumed to reproduce and give birth, in the average, to β new young individuals. Then*

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

It can be seen that for $\alpha + \beta > 1$ the population diverges ($\lambda_1 > 1$) while for $\alpha + \beta < 1$ the population extinguishes ($\lambda_1 < 1$).

Assume, for instance, that $\alpha + \beta > 1$. Then the free response is

$$x(k) = A^k x(0) = t_1 s_1^\top x(0)\lambda_1^k + t_2 s_2^\top x(0)\lambda_2^k + t_3 s_3^\top x(0)\lambda_3^k \approx t_1 s_1^\top x(0)\lambda_1^k$$

for large k, because the dominant real mode diverges faster than the others. Then, since $s_1^\top x(0) = \gamma_1$ is a scalar,

$$x(k) \approx t_1(s_1^\top x(0))\lambda_1^k = t_1\gamma_1\lambda_1^k$$

and we see that the asymptotic distribution of the population is given by the Frobenius eigenvector, which is positive. For $\alpha = 1$ and $\beta = 1$, the eigenvalues are 1.32472 and $-0.66236 \pm 0.56228j$. The complex eigenvalues have modulus 0.86884, hence their modes converge. The Frobenius eigenvector is $t_1 = [0.43016 \quad 0.32472 \quad 0.24512]^\top$ so the asymptotic distribution is proportional to these components: the components have been normalised so that their sum is 1, therefore we will have 43016% of the population in the first class and so on.

The reader can imagine a more sophisticated model with division in n age classes, each with a different contribution to birth.

9.4.1 Markov chains

Markov chains are a special class of positive linear system, in which the state vector components represent probability values. Given a family of independent events

$$\{E_1, E_2, \dots, E_n\}$$

that are a partition of the certain event, the vector x has components

$$x_i(k) = \Pr[E_k] \text{ (probability of event } E_k \text{ occurring)}$$

with

$$\sum_{i=1}^{n} x_i(k) = 1, \quad x_i(k) \geq 0. \tag{9.25}$$

We write the previous conditions as

$$\bar{1}^\top x_i(k) = 1, \quad x(k) \geq 0 \tag{9.26}$$

where $\bar{1}^\top = [1 \quad 1 \dots \quad 1]$, and the vector inequality is intended componentwise.

Assume that the event probabilities are evolving in time, according to a linear equation. The assumption is that, at each time, the probability distribution depends only on the previous distribution

[1] After old age, the population is assumed to reach a retired happy-leaving status and is no longer considered in the problem.

and that the transition probabilities are time-invariant. Assuming a discrete-time evolution, we get a model of the form

$$x(k+1) = Px(k). \tag{9.27}$$

Since constraints (9.25) have to be satisfied all the times, we need that

$$P_{ij} \geq 0, \quad \sum_{i=1}^{n} P_{ij} = 1, \forall j$$

namely the matrix has to be nonnegative and for each column the elements must sum up to 1. This is a necessary and sufficient condition for us to say that $x(k)$ represents a probability:

$$\bar{1}^\top x(0) = 1 \quad \text{and} \quad x(0) \geq 0, \quad \Rightarrow \quad \bar{1}^\top x(k) = 1 \quad \text{and} \quad x(k) \geq 0. \tag{9.28}$$

The entries of a discrete-time Markov matrix P represent transition probabilities

$$P_{ij} = \text{ transition probability from state } j \text{ to state } i.$$

The dominant eigenvalue of a Markov chain is 1. Indeed from (9.28), the system is marginally stable, so its eigenvalues are in the unit disk (border included). Also, the columns sum up to 1

$$\bar{1}P = \bar{1} = (1)\bar{1},$$

so 1 is an eigenvalue, and since it is real and dominant it is the Frobenius eigenvalue. Assume now that 1 is strictly dominant: $1 > |\lambda_2| \geq |\lambda_3| \ldots$. The Frobenius eigenvector t_1 is very important because it expresses the asymptotic probability distribution. Indeed, we can normalise the left eigenvector to $s_1 = \bar{1}$ and we have, by definition of probability, that the initial condition satisfies

$$s_1^\top x(0) = 1$$

hence

$$x(k) = t_1 s_1^\top x(0)\lambda_1^k + \underbrace{t_2 s_2^\top x(0)\lambda_2^k + \cdots + t_n s_n^\top x(0)\lambda_n^k}_{\to 0} \to t_1 s_1^\top x(0) = t_1 \tag{9.29}$$

Since $s_1^\top x(k) = 1$, necessarily, at the limit, $s_1^\top t_1 = 1$. Hence we have the following.

Proposition 9.1. *The* **asymptotic probability distribution** *is given by the Frobenius eigenvector normalised to* 1.

In general, if the eigenvalue 1 is not strictly dominant, then there can be infinite asymptotic distributions, for instance P_{Id} below, or no asymptotic distribution, for instance P_{Bd} (Buridan's donkey phenomenon) below:

$$P_{Id} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad P_{Bd} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Example 9.2. (Buffer of a processing unit.) *Consider the following processing unit (machine) in which objects to be processed are waiting in a buffer. The buffer states are supposed to be*

status	*description*
1	*0 parts, buffer empty*
2	*1 part, first slot occupied*
3	*2 parts, two slots occupied*
4	*3 parts, three slots occupied*
5	*4 part, buffer full*

Assume that each part is processed in one time unit $T = 1$. We assume that probabilities are assigned to the numbers of part arrivals:

probability	*description*
P_0	*0 parts arrival*
P_1	*1 part arrival*
P_2	*2 parts arrival*
$P_k = 0$	*$k \geq 3$ parts arrival*

This corresponds to the graph in Figure 9.6. The Markov matrix P can be derived as follows

- p_{11} *is the probability that the buffer is empty and remains empty. This happens if zero or one parts arrive:* $p_{11} = P_0 + P_1$.
- $p_{12} = P_2$ *is the probability that, if the buffer is empty, the next time there will be one part.*
- *column j is associated with the state in which there are $j - 1$ parts. If* 0 *parts arrive the buffer decreases by* 1*, if* 1 *part arrives the buffer remains in this state and if* 2 *parts arrive the buffer increases by* 1.
- $p_{55} = P_1 + P_2$ *is the probability that the full buffer remains full, namely if it receives* 1 *or* 2 *parts, in the latter case we assume that one of the two parts is rejected.*

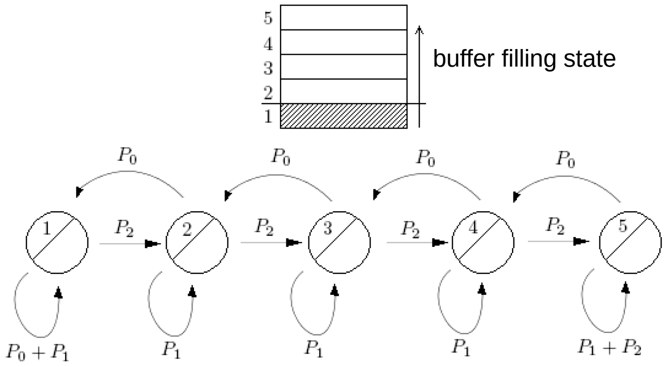

Figure 9.6: Buffer: graph of states and transitions.

The transition probabilities are reported in the following matrix governing the system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \\ x_5(k+1) \end{bmatrix} = \begin{bmatrix} P_0+P_1 & P_0 & 0 & 0 & 0 \\ P_2 & P_1 & P_0 & 0 & 0 \\ 0 & P_2 & P_1 & P_0 & 0 \\ 0 & 0 & P_2 & P_1 & P_0 \\ 0 & 0 & 0 & P_2 & P_1+P_2 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \\ x_5(k) \end{bmatrix}$$

Given the probabilities P_0, P_1 and P_2, the steady-state buffer distribution is determined by computing $\bar{t}$, the nonnegative eigenvector associated with $\lambda = 1$. We can examine the special cases

- $P_0 = 1$, $P_1 = 0$, $P_2 = 0$*: the buffer becomes empty;*
- $P_0 = 0$, $P_1 = 0$, $P_2 = 1$*: the buffer becomes full congested;*
- $P_0 = 0$, $P_1 = 1$, $P_2 = 0$*: the asymptotic vector does not exist, because $\lambda = 1$ is not simple. Figure 9.7 presents some other cases.*

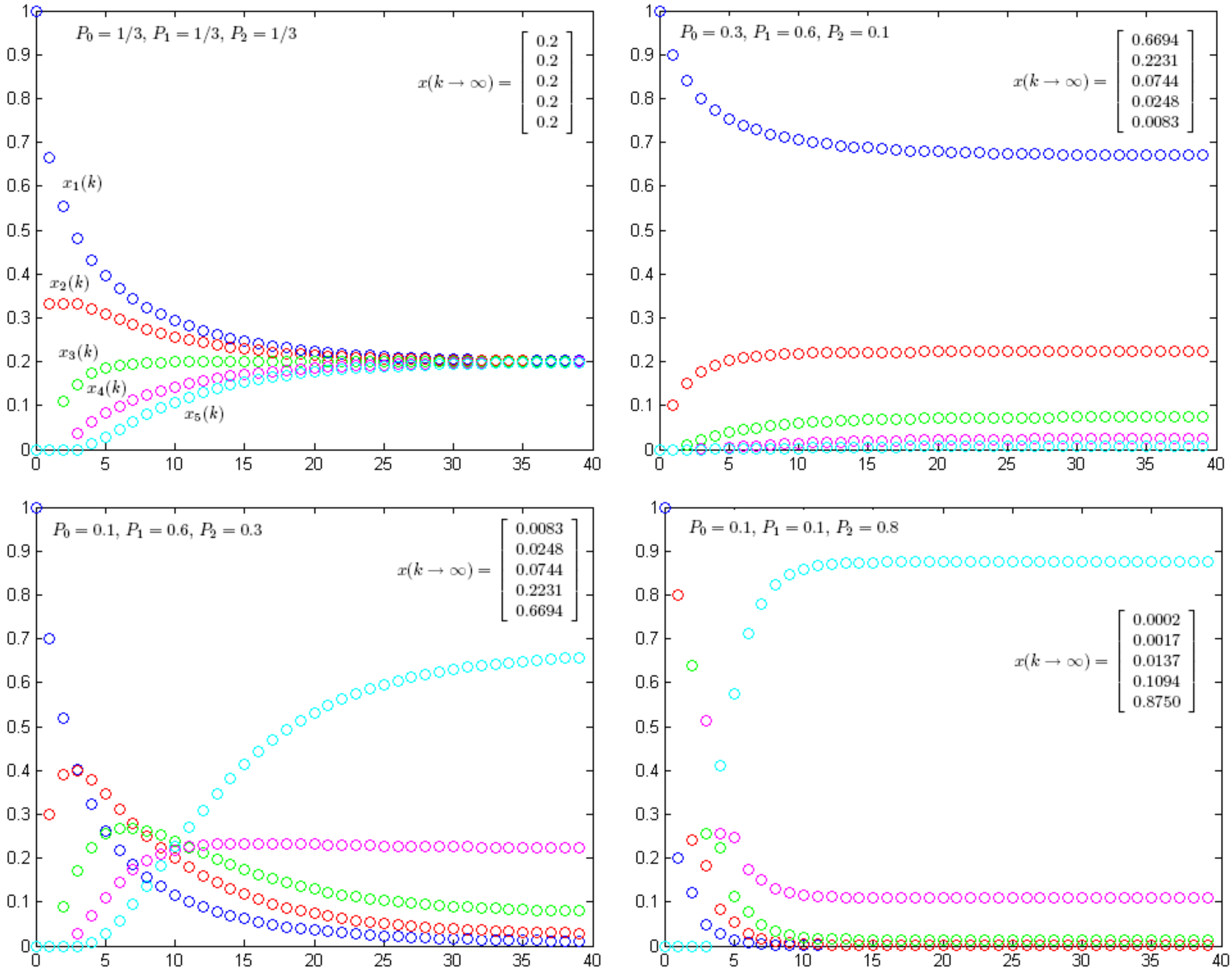

Figure 9.7: Buffer simulation for different values of P_0, P_1, P_2.

9.5 Vibrating systems

Consider Figure 9.8, which shows a structure (building) under seismic action. This system is governed by the following equations

$$\begin{cases} m_1\ddot{\vartheta}_1(t) &=& -k_1\vartheta_1(t) - k_{12}\left(\vartheta_1(t) - \vartheta_2(t)\right) - m_1 a(t) \\ m_2\ddot{\vartheta}_2(t) &=& -k_{12}\left(\vartheta_2(t) - \vartheta_1(t)\right) - k_{23}\left(\vartheta_2(t) - \vartheta_3(t)\right) - m_2 a(t) \\ m_3\ddot{\vartheta}_3(t) &=& -k_{23}\left(\vartheta_3(t) - \vartheta_2(t)\right) - m_3 a(t) \end{cases} \tag{9.30}$$

A ground reference has been chosen. This reference is non-inertial. The variables ϑ_1, ϑ_2, ϑ_3 represent the horizontal relative displacements of the floors. The coefficients m_1, m_2, m_3 are the masses of the floors while the coefficients k_1, k_2, k_3 are elastic coefficients. The term $k_1\vartheta_1$ represents the elastic force applied on the first floor from the ground, while the terms $\pm k_{ij}\left(\vartheta_i - \vartheta_j(t)\right)$ are elastic forces between floors. The term $m_i a(t)$ is a virtual force due to ground acceleration (the reference frame acceleration).

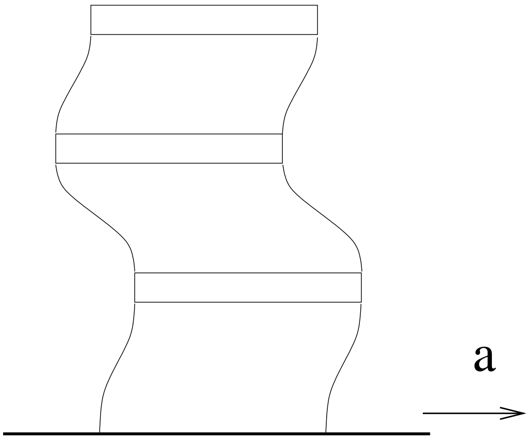

Figure 9.8: Building under seismic action.

This system can be written in the following matrix form:

$$\begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \ddot{\vartheta}(t) \;=\; \begin{bmatrix} -(k_1+k_{12}) & k_{12} & 0 \\ k_{12} & -(k_{12}+k_{23}) & k_{23} \\ 0 & k_{23} & -k_{23} \end{bmatrix} \vartheta(t) + \begin{bmatrix} -m_1 \\ -m_2 \\ -m_3 \end{bmatrix} a(t) \quad (9.31)$$

Note that no friction has been considered and we will be back on this later.

In general, a vibrating system can be written in the following form

$$M\ddot{\vartheta}(t) = -K\vartheta(t) + Ru(t) \quad (9.32)$$

To explain these equations let us write them as follows: after pre-multiplication by $\dot{\vartheta}^\top(t)$ (the row vector of the speeds), we have

$$\dot{\vartheta}^\top(t)M\ddot{\vartheta}(t) + \dot{\vartheta}^\top(t)K\vartheta(t) = \underbrace{\frac{d}{dt}\left(\frac{1}{2}\dot{\vartheta}^\top(t)M\dot{\vartheta}(t) + \frac{1}{2}\vartheta^\top(t)K\vartheta(t)\right)}_{\text{energy=kinetic+elastic}} = \underbrace{\dot{\vartheta}^\top(t)Ru(t)}_{\text{supplied power}}$$

The previous equation means that the derivative of the total energy is equal to the supplied power.

Matrix M is always positive definite (indeed, the kinetic energy is $E_c = \frac{1}{2}\dot{\vartheta}^T(t)M\dot{\vartheta}(t) > 0$ if the system is in motion) and the matrix K is semidefinite (because the potential energy $E_P = \frac{1}{2}\vartheta^T(t)K\vartheta(t) \geq 0$). In the particular case of the building under seismic action, matrix K is positive definite.

To have a state representation we pre-multiply equation (9.32) by M^{-1} to get:

$$\ddot{\vartheta}(t) = -M^{-1}K\vartheta(t) + M^{-1}Ru(t) \quad (9.33)$$

We can choose as state vector $x(t) = \begin{bmatrix} \vartheta(t) & \dot{\vartheta}(t) \end{bmatrix}^T$. Then

$$\frac{d}{dt}\begin{bmatrix} \vartheta(t) \\ \dot{\vartheta}(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ -M^{-1}K & 0 \end{bmatrix}\begin{bmatrix} \vartheta(t) \\ \dot{\vartheta}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ M^{-1}R \end{bmatrix} u(t) \quad (9.34)$$

Denoting by $S = M^{-1}K$, matrix A is given by:

$$A = \begin{bmatrix} 0 & I \\ -S & 0 \end{bmatrix} \quad (9.35)$$

Let us study the modes of the system. The eigenvalues and eigenvectors satisfy

$$A\overline{x} = \lambda\overline{x} \Rightarrow \begin{bmatrix} 0 & I \\ -S & 0 \end{bmatrix} \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \end{bmatrix} = \lambda \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \end{bmatrix} \tag{9.36}$$

From the first equation, we have $\overline{x}_2 = \lambda\overline{x}_1$. This tells us that the generic eigenvector associated with matrix A it is of the type $\overline{x} = \begin{bmatrix} \overline{x}_1^T & \lambda\overline{x}_1^T \end{bmatrix}^T$. Using the second equation

$$-S\overline{x}_1 = \lambda\overline{x}_2 = \lambda(\lambda\overline{x}_1) \quad \Rightarrow \quad S\overline{x}_1 = -\lambda^2\overline{x}_1 \tag{9.37}$$

Let us now consider the expression of $S = M^{-1}K$ and let $T = M^{1/2}$. Then, by pre-multiplying by $T = M^{1/2}$, we have

$$\left[M^{-1/2}KM^{-1/2}\right]M^{1/2}\overline{x}_1 = -\lambda^2 M^{1/2}\overline{x}_1 \Rightarrow \left[M^{-1/2}KM^{-1/2}\right]y \doteq \hat{S}y = -\lambda^2 y$$

Matrix $\hat{S} = M^{-1/2}KM^{-1/2}$ is positive definite and its eigenvalues are real and positive. Moreover $\hat{S} = M^{-1/2}KM^{-1/2} = T^{-1}ST$, hence its eigenvalues are those of S.

Denote the m eigenvalues of S as $\omega_1^2, \omega_2^2, \dots, \omega_m^2$, then the $n = 2m$ eigenvalues associated with A are

$$\omega_k^2 = -\lambda_k^2 \quad \Rightarrow \quad \lambda_k = \pm j\omega_k \tag{9.38}$$

This means that vibrating systems without damping have in general oscillating modes $e^{\pm j\omega t}$, namely,

$$\sin(\omega_k t) \quad \text{and} \quad \cos(\omega_k t) \tag{9.39}$$

which are undamped oscillations. The coefficients ω_k are named **proper pulsations** or **resonance pulsations**. The numbers $f_k = \omega_k/(2\pi)$ are the proper frequencies.

Proposition 9.2. *In an undamped vibrating system, the modes are of the form* (9.39)*, where the proper pulsations are ω_k and ω_k^2 are the eigenvalues of $M^{-1}K$.*

Consider for instance the third floor when $u(t) = 0$. Its displacement $\vartheta_3(t)$ has the following general expression:

$$\vartheta_3(t) = \alpha_1\cos(\omega_1 t + \varphi_1) + \alpha_2\cos(\omega_2 t + \varphi_2) + \alpha_3\cos(\omega_3 t + \varphi_3)$$

where the terms α_i and ϕ_i depend on the initial conditions of the system.

To explain the significance of the eigenvectors, consider initial conditions with $\dot{\vartheta}(0) = 0$ and $\vartheta(0) = \overline{\vartheta}_1$, where $\overline{\vartheta}_1$ is an eigenvector of the matrix. The equation that governs the free response of the system is

$$\begin{aligned} \ddot{\vartheta}(t) &= -S\vartheta(t) \qquad \vartheta(0) = \overline{\vartheta}_1 \\ \Rightarrow -\omega_1^2\overline{\vartheta}_1\cos(\omega_1 t) &= -S\,\overline{\vartheta}_1\cos(\omega_1 t) \\ \Rightarrow S\,\overline{\vartheta}_1 &= \omega_1^2\overline{\vartheta}_1 \end{aligned}$$

(just compute the second derivative of $\overline{\vartheta}_1\cos(\omega_1 t)$ to check the expression and take into account that $\overline{\vartheta}_1$ is an eigenvector of S). This means that, if we initialise the building floors in positions equal to the components of $\overline{\vartheta}_1$, with $\dot{\vartheta}(0) = 0$, the system shows only the oscillation mode corresponding to an harmonic component with pulsation ω_1. Clearly any elastic motion (guitar string, vibrating shells, acoustic instruments, vocal sound) is the sum of all harmonic components.

Finally, assume that there is friction to be taken into account. A typical model is

$$M\ddot{\vartheta}(t) = -K\vartheta(t) - H\dot{\vartheta}(t)Ru(t), \tag{9.40}$$

where H is a positive semidefinite matrix including the friction coefficients. In this case the eigenvalues are not imaginary, but have a negative real part. The energy balance is now

$$\underbrace{\frac{d}{dt}\left(\frac{1}{2}\dot{\vartheta}^{\top}(t)M\dot{\vartheta}(t) + \frac{1}{2}\vartheta^{\top}(t)K\vartheta(t)\right)}_{\text{energy=kinetic+elastic}} + \underbrace{\dot{\vartheta}^{\top}(t)H\dot{\vartheta}(t)}_{\text{dissipated power}} = \underbrace{\dot{\vartheta}^{\top}(t)Ru(t)}_{\text{supplied power}}$$

9.6 Linear electrical circuit

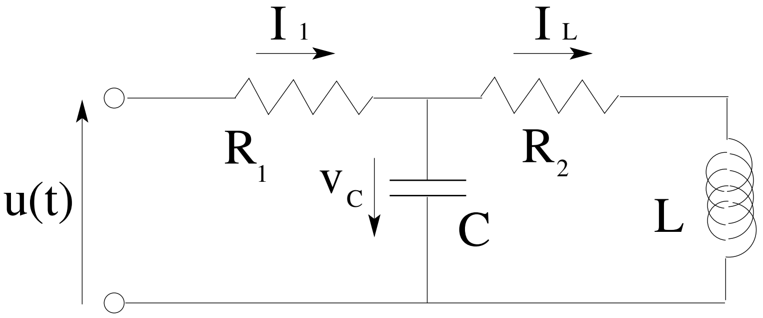

Figure 9.9: Example of a linear circuit.

An important class of linear systems is represented by linear electrical circuits. An electrical circuit is represented by linear equations if it includes only resistive, capacitive and inductive components, along with voltage and current generators. The dipole laws associated with these components are the following.

- Resistors: $v(t) = Ri(t)$
- Inductors: $L\dot{i}(t) = v(t)$
- Capacitors: $i(t) = C\dot{v}(t)$

An electrical circuit is the connection of the previous components and of voltage and current generators. The equations can be written by using Kirchoff laws. Typically, the inputs of these systems are provided by the generators. As state variables, we can take

- the capacitor voltages;
- the inductors currents.

Consider, for example, Figure 9.9. Applying Kirchhoff loop law to the two loops, we obtain the following two equations:

$$\begin{aligned} v &= R_1 i_1 + v_C \\ v_C &= R_2 i_L + L\dot{i}_L \end{aligned} \tag{9.41}$$

Conversely, applying Kirchhoff node law to the central node, we have the current balance

$$i_1 = i_L + i_C = i_L + C\dot{v}_C, \tag{9.42}$$

where we have assumed the orientation in such a way that i_1 is incoming while i_L and i_C are outgoing. Replacing i_1 with this expression, we get two first order differential equations

$$\begin{cases} \dot{v}_C = -\frac{1}{R_1 C} v_C - \frac{1}{C} i_L + \frac{1}{R_1 C} v \\ \dot{i}_L = \frac{1}{L} v_C - \frac{R_2}{L} i_L \end{cases} \tag{9.43}$$

Denoting by $u(t) = v(t)$, $x_1(t) = v_C(t)$, $x_2(t) = i_L(t)$, we obtain the matrix representation

$$\frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} -\frac{1}{R_1 C} & -\frac{1}{C} \\ \frac{1}{L} & -\frac{R_2}{L} \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{R_1 C} \\ 0 \end{bmatrix} u(t) \tag{9.44}$$

The choice of the output of the system depends on the problem. For example, if we want to describe this circuit as an admittance, the output is $i_1(t)$:

$$y = i_L + C\dot{v}_C = i_L + C\left(-\frac{1}{R_1C}v_C - \frac{1}{C}i_L + \frac{1}{R_1C}v\right) = -\frac{1}{R_1}v_C + \frac{1}{R_1}v \tag{9.45}$$

Note that y depends on the derivative $\dot{v}_c$ of v_c, and this would lead to a non-proper system. The equation can be modified by replacing $\dot{v}_c$ with its expression (from the state equation) in order to have a dependence on state variables only, and not on their derivatives. The output is then

$$y(t) = \left[\begin{array}{cc} -\frac{1}{R_1} & 0 \end{array}\right]\left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array}\right] + \left[\begin{array}{c} \frac{1}{R_1} \end{array}\right] v(t) \tag{9.46}$$

This system is proper, but not strictly proper.

We stress that some circuits might have some singularities. This is the case in which there are cut-sets of inductors (three-phases systems) in which $I_1 + I_2 + \dots I_m = 0$ or circuits of capacitors in which $v_1 + v_2 + \dots v_m = 0$. Moreover, for some choices of inputs and outputs, a circuit can be non-proper. For instance, in a capacitor, if we consider the voltage as an input and the current as an output, we have

$$i_c(t) = \frac{d}{dt}v_c(t) \text{ with transfer function } \quad i_c(s) = [s]v_c(s),$$

which is not proper.

Going back to the example, let us consider the characteristic polynomial of matrix A:

$$\det(sI - A) = s^2 + \left(\frac{1}{R_1C} + \frac{R_2}{L}\right)s + \frac{1}{LC}\left(\frac{R_2}{R_1} + 1\right) \tag{9.47}$$

Note that the term that multiplies s is the trace of $-A$, while the term that multiplies s^0 is the determinant of A. The roots of this polynomial have negative real part, since it is a second order polynomial with positive coefficients. This is an expected result, because the circuit is dissipative.

The study of the circuit can also be done using the Laplace transform. Assuming $x(0) = 0$ for simplicity, the dipole laws described earlier become:

- $v(s) = Ri(s)$ (resistors);
- $v(s) = sLi(s)$ (inductors);
- $i(s) = sCv(s)$ (capacitors).

Note that, in the special case where we assume $s = j\omega$, we are representing the dipoles in the sinusoidal regime, according to the theory of frequency analysis for electrical circuits.

These relationships are linear in the Laplace transform domain and can be used as generalised impedances. Then we can apply standard properties of series-parallel connections between components. In the case of the circuit under test, the passive network can be considered as an impedance having this expression:

$$Z(s) = R_1 + \frac{1}{sC} \parallel (R_2 + sL) = R_1 + \frac{1}{sC + \frac{1}{R_2+sL}} = \frac{r_1LCs^2 + (CR_1R_2 + L)s + (R_1 + R_2)}{LCS^2 + R_2Cs + 1} \tag{9.48}$$

It is easily verifiable that the impedance is the inverse of transfer function $W(s) = C(sI - A)-1B = Z^{-1}(s)$, as expected.

The representation in terms of state description may not be useful for circuits of the dimension considered here, because such calculations can be performed quickly using the Laplace transform. However, for more complex circuits, the state representation can be more useful, because it can be easily implemented on a computer.

9.7 Operational amplifiers

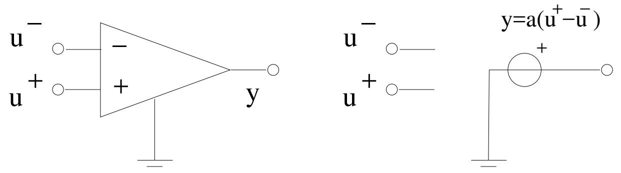

Figure 9.10: Operational amplifier.

Operational amplifiers are components whose symbol and equivalent circuit are shown in Figure 9.10. The factor a is a very large number, ideally tending to infinity. This electronic component is practically only used in feedback as, for example, in Figure 9.11. The Kirchoff equation for the "–' node is, in terms of Laplace transform (omitting the dependence on s)

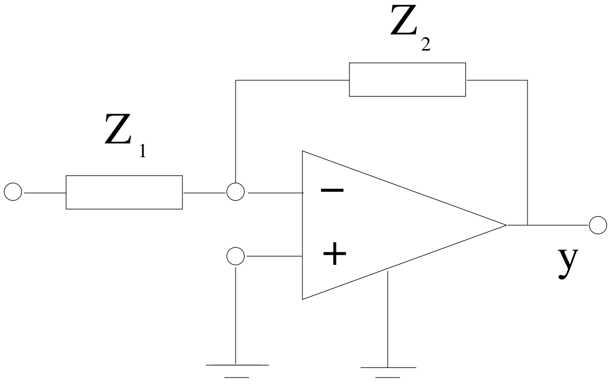

Figure 9.11: Operational amplifier in inverting feedback configuration.

$$\frac{y-u^-}{Z_2}+\frac{u-u^-}{Z_1}=i^-,$$

the current entering the "-" terminal. Hence

$$\frac{y}{Z_2}+\frac{u}{Z_1}=\frac{u^-}{Z_2}+\frac{u^-}{Z_1}+i^-=\frac{1}{a}\left(\frac{y}{Z_2}+\frac{y}{Z_1}\right)+i^-\approx 0,$$

where we have used the fact that the amplification is roughly infinite, $a\approx\infty$, and the input impedance is infinite, hence $i^-\approx 0$. Then the approximate law is

$$\frac{y(s)}{Z_2(s)}+\frac{u(s)}{Z_1(s)}=0, \tag{9.49}$$

where we have restored the dependence on s.

In the case of two resistances $Z_1=R_1$ and $Z_2=R_2$, we have $\frac{y}{R_2}+\frac{u}{R_1}=0$, from which:

$$y(t)=-\frac{R_2}{R_1}u(t). \tag{9.50}$$

Then the operational amplifier allows us to have an amplifier with tunable gain. The problem of having a negative sign can be solved by properly connecting two amplifiers (in series).

Consider, for example, $Z_1 = R$ and $Z_2 = \frac{1}{sC}$, i.e., adopting a resistor and a capacitor. We get

$$y(s) = -\frac{1}{sCR}u(s), \tag{9.51}$$

which, up to a multiplicative negative factor, is the transfer function of an integrator block. Similarly, with $Z_1 = \frac{1}{sC}$ and $Z_2 = R$, we obtain the transfer function of a differentiating block.

Another interesting configuration for an operational amplifier is shown in Figure 9.12.

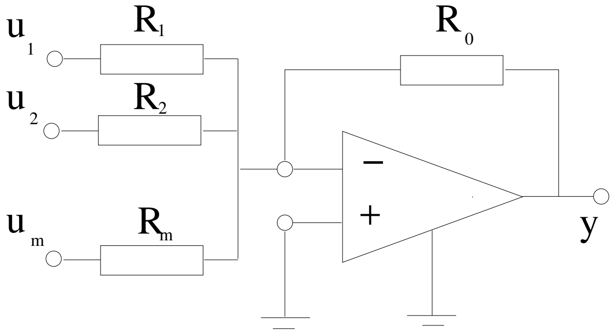

Figure 9.12: Feedback amplifier: adder.

The equation at node u^- is:

$$\frac{u_1 - u^-}{R_1} + \frac{u_2 - u^-}{R_2} + \ldots + \frac{u_m - u^-}{R_m} + \frac{y - u^-}{R_0} = i^- \tag{9.52}$$

Noting that $u^- = y/a \approx 0$ and $i^- \approx 0$, we have

$$y = -\sum_{i=1}^{m} \frac{R_0}{R_i} u_i \tag{9.53}$$

The circuit is a “summing” device that provides as an output a linear combination of the inputs. Again, the minus sign is not really a problem, one can use other devices to invert the sign.

Since we are able to build amplifiers, integrators and summing circuits, we are able to build any transfer function.

There are many different amplifier configurations to realise transfer functions, which are typically optimised so as to use the smallest number of components.

9.8 Two-tank system

Consider Figure 9.13, which represents a system consisting of two tanks connected by a pipe. There is also an outgoing pipe from the lower tanks to a basin. A pump collects the water from the basin and injects it into the upper tank. The balance equations describing the volumetric flow rate and volume variations are:

$$\begin{aligned} \dot{v}_1 &= q_0 - q_1 \\ \dot{v}_2 &= q_1 - q_2 \\ v_1 &= (h_1 - h_{10})S_1 \\ v_2 &= (h_2 - h_{20})S_2 \end{aligned}$$

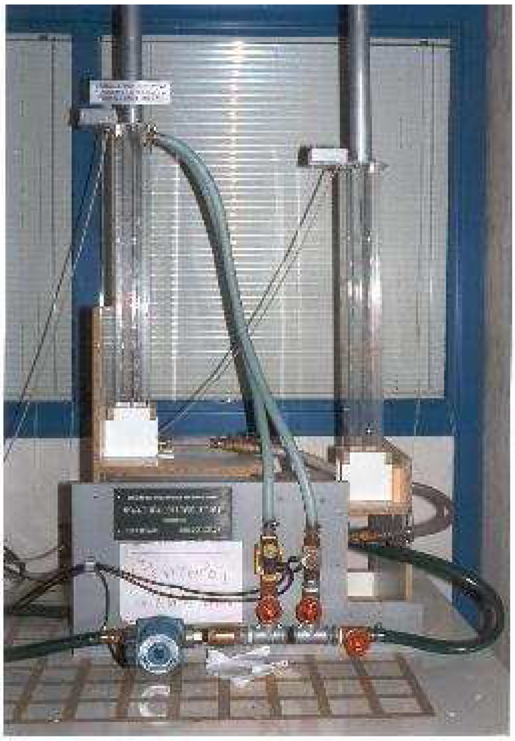

Figure 9.13: Two-tank system.

where q_0, q_1, q_2 are the input flow in the upper tank, the flow between first and second tank, and the flow from the second tank to the basin, respectively. Based on fluid dynamics, we can also consider the the following equations, concerning the dependence between the heights (potential energy) and the flows.

$$\begin{aligned} h_1 - h_2 &= \alpha^2 q_1^2 \\ h_2 &= \beta^2 q_2^2 \end{aligned}$$

From these, we obtain:

$$\begin{aligned} q_1 &= \frac{1}{\alpha}\sqrt{h_1 - h_2} \\ q_2 &= \frac{1}{\beta}\sqrt{h_2} \end{aligned}$$

Note that the inertial effect of the water in the pipes is not considered, because it can be experimentally verified that it is negligible. The derivatives of the volume of water in the tanks are

$$\dot{v}_1 = S_1\dot{h}_1 \quad \dot{v}_2 = S_2\dot{h}_2$$

where S is the tank section surface. The resulting state equations are

$$\begin{cases} \dot{h}_1(t) = u - \frac{1}{\alpha S_1}\sqrt{h_1(t) - h_2(t)} \\ \dot{h}_2(t) = \frac{1}{\alpha S_2}\sqrt{h_1(t) - h_2(t)} - \frac{1}{\beta S_2}\sqrt{h_2(t)} \end{cases}$$

where $u = q_0/S_1$ is the input, while h_1 and h_2 are the state variables. This is a nonlinear system. To study the static problem, write the steady-state conditions $\dot{h}_1(t) = 0$ and $\dot{h}_2(t) = 0$, which implies a flow balance $q_0 = q_1 = q_2$

$$\begin{cases} 0 = \frac{q_0}{S_1} - \frac{1}{\alpha S_1}\sqrt{h_1(t) - h_2(t)} \\ 0 = \frac{1}{\alpha S_2}\sqrt{h_1(t) - h_2(t)} - \frac{1}{\beta S_2}\sqrt{h_2(t)} \end{cases}$$

After simple computations, it turns out that

$$\begin{aligned} \overline{h}_1 &= S_1^2(\alpha^2 + \beta^2)\overline{u}^2 \\ \overline{h}_2 &= S_1^2\beta^2\overline{u}^2 \end{aligned}$$

where $\overline{u} = \overline{q}_0 S_1$ is the constant input flow. Note that $\overline{h}_1 > \overline{h}_2$ because α and β are positive constants. In fact, the proposed model is valid for $h_1(t) > h_2(t)$. If we want to consider the possibility of having (for example, during a transient) $h_2(t) > h_1(t)$, then the model has to be changed as follows

$$\sqrt{h_1(t) - h_2(t)} \Longrightarrow \mathrm{sign}(h_1(t) - h_2(t))\sqrt{|h_1(t) - h_2(t)|}$$

so that the equilibrium pairs $(\overline{h}_1, \overline{h}_2)$ are a straight line in the plane $\overline{h}_1 - \overline{h}_2$.

There are some problems with this system, which comes from a simplified model. First, the levels are measured by floaters, connected to variable inductors. The measurement of $h_1(t)$, returned by the first floater, is highly affected by noise. Second, the pump is not continuously controlled, but it is quantised. The input flow is controlled by two identical switching valves. Therefore there are three possible input values:

$$u \in \{0, \bar{u}, 2\bar{u}\}$$

(if both valves are closed, one is open and the other closed, they are both open). In this case, to achieve an arbitrary average input flow, one can use the pulse-width modulation PWM technique shown in Figure 9.14.

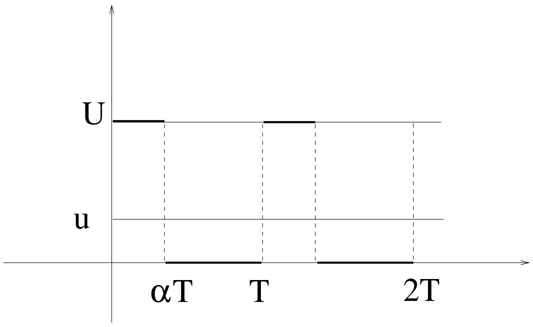

Figure 9.14: Example of segmentation.

Let us consider the problem of achieving a desired value u_{des}, between 0 and U. The technique consists in fixing a sampling time T, and applying the control

$$u_{quant} = \begin{cases} U & \text{for} \quad kT \le t < kT + \alpha T \\ \bar{0} & \text{for} \quad kT + \alpha T \le t < (k+1)T \end{cases}$$

with

$$\alpha = \frac{u_{des}}{U}$$

The average value is the desired value u_{des}. This technique, therefore, allows us to obtain a quantised input with the same mean value of the desired continuous signal. To analyse the error, write this input as

$$u_{quant} = u_{des} + w(t)$$

where, if u_{des} is constant, $w(t)$ is a periodic error of period T. This control works well, as long as the dominating frequency $\omega = 2\pi/T$ of $w(t)$ is considerably greater than the system bandwidth. In the real equipment, we have $\omega \approx 3$, which is greater then the bandwidth.

9.9 Magnetic levitator

Consider the magnetic levitator shown in Figure 9.15. This system is described by a mechanical equation of the form

$$m\ddot{y}(t) = mg - f(t)$$

The orientation is downward: increasing y means lowering the sphere. The term $f(t)$ is the lifting force, where f is a a function of the current and the distance and is approximately

$$f(t) = k\frac{i^2(t)}{y^2(t)}.$$

Then we get

$$\ddot{y}(t) = g - \frac{k}{m}\frac{i^2}{y^2}.$$

If we introduce also the electric equation of the magnet, driven by voltage V, the equations describing the system become:

$$\begin{aligned} L\frac{di(t)}{dt} &= -Ri(t) + V(t) \\ m\ddot{y}(t) &= mg - k\frac{i^2(t)}{y^2(t)} \end{aligned}$$

If we introduce $x_1 = y$, $x_2 = \dot{y}$ and $x_3 = i$ as state variables, we get the system

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = g - \frac{k}{m}\frac{x_3^2(t)}{x_1^2(t)} \\ \dot{x}_3(t) = -\frac{R}{L}x_3(t) + \frac{1}{L}u(t) \end{cases}$$

9.10 Cart–pendulum system

Consider the cart–pendulum system represented in Fig. 9.16 The variables are: y, the distance of the cart from the reference and θ the angle of the pendulum. The cart equation is

$$M\ddot{y}(t) = u - \varphi\dot{y}(t)$$

where M is the mass of the cart, φ is the friction coefficient and $u(t)$ is the applied force. We neglect the reaction of the pole on the cart which is much heavier. The pole equation is

$$mr^2\ddot{\theta}(t) = mgr\sin(\theta(t)) + mr\cos(\theta(t))\ddot{y}(t)$$

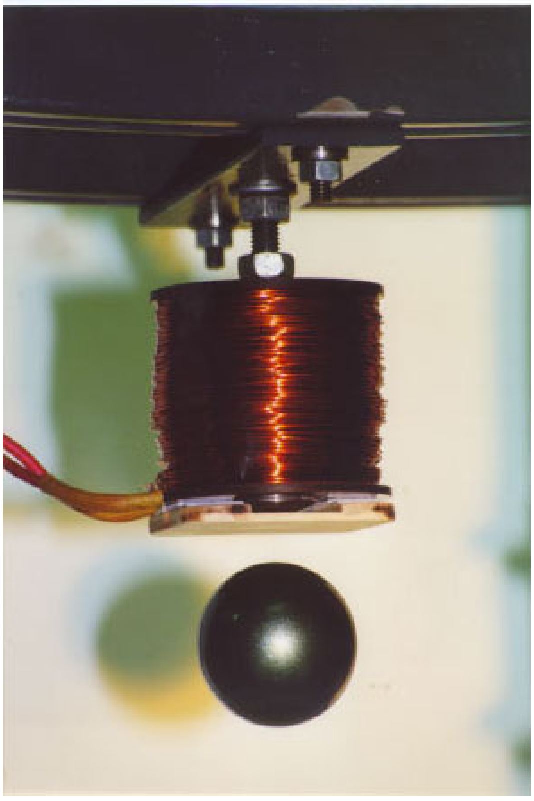

Figure 9.15: The magnetic levitator.

where the last term is due to the fact that the reference on the cart is not inertial. We replace $\ddot{y}(t)$ from the first equation to get

$$\ddot{\theta}(t) = \frac{g}{r} r \sin(\theta(t)) + \frac{1}{Mr} \cos(\theta(t))[u - \varphi \dot{y}(t)]$$

If we apply the linearization at $\theta = 0$ we get $\sin(\theta) \approx \theta$ and $\cos(\theta) \approx 1.$, hence

$$\ddot{\theta}(t) = \frac{g}{r}\theta(t) + \frac{1}{Mr}[u - \varphi \dot{y}(t)]$$

Denoting by $x_1 = \theta$, $x_2 = \dot{\theta}$, $x_3 = y$ and $x_4 = \dot{y}$, we get $\dot{x}(t) = Ax(t) + Bu(t)$, where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \alpha & 0 & 0 & -\mu\varphi \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\nu\varphi \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \mu \\ 0 \\ \nu \end{bmatrix}$$

where $\mu = 1/M$, $\nu = 1/(Mr)$ and $\alpha = g/r$. This system is unstable: its poles are 0, $-\nu\varphi$ and $\pm\sqrt{\alpha}$. It is reachable and it can be stabilized.

Real experiments show that pole assignment design is not suitable for the real problem due to the lack of robustness. Indeed the term φ is uncertain and time–varying. To analyze the situation, write

$$\varphi = \varphi_0 + \Delta$$

where φ_0 is the nominal value of φ and w is uncertain for which we assume the following bound

$$|\Delta| \leq \delta$$

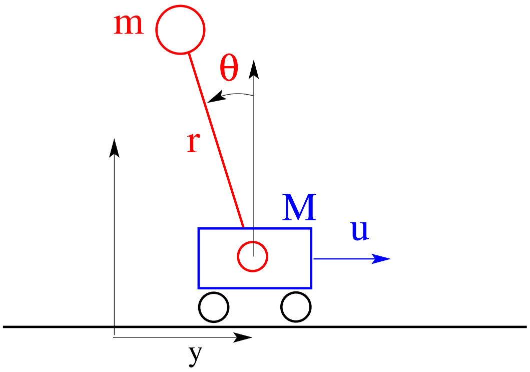

Figure 9.16: The cart–pendulum

Write the model as

$$\dot{x}(t) = A_0 x(t) + BH\Delta + Bu(t)$$

with

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \alpha & 0 & 0 & -\mu\varphi_0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\nu\varphi_0 \end{bmatrix}, \quad \text{and} \quad H = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$$

We consider now an optimal control approach to ensure robustness as follows.

- Fix Q positive definite and $R = 1$ and solve the Riccati equation

  $$A^\top P + PA - PBB^\top P + Q$$

  and consider the optimal control

  $$u_{opt} = -B^\top Px$$

- Apply the control $u = u_{opt} + u_r$, where u_r is a robustifying term that will be specified soon, to get the system

  $$\dot{x}(t) = [A_0 - B^\top P]x(t) + BH\Delta(t) + Bu_r(t)$$

  namely

  $$\dot{x}(t) = \bar{A}x(t) + BH\Delta(t) + Bu_r(t)$$

  Note that the Lyapunov equation holds

  $$\bar{A}^\top P + P\bar{A} = -[Q + PBB^\top P] = -\bar{Q}$$

- Take

  $$u_r(t) = -\gamma B^\top Px$$

  where γ is selected as explained in the Lyapunov redesign section.

It turns out the the optimal control stabilizes the cart-pendulum system.

Appendix A

Elements of Mathematics

In this section we will briefly summarise some basic concepts of mathematics that are required for the study of dynamical systems. In particular, the concepts of linear algebra will be reported, since it is a fundamental language for the formalisation of important concepts. Once again, this manuscript is not a text book, but a guideline: this appendix is essentially listing the preliminary notions that are necessary to understand the course (which should have been thoroughly studied and learned in previous courses).

Mathematics in engineering has three fundamental roles:

- it is a language to describe models;
- it is a language to express laws and properties;
- it provides computational tools.

In this course we need to keep in mind all of these aspects.

A.1 Vector spaces

Linear algebra, including vector space algebra and matrix theory, is essential in system and control theory. Without linear algebra it is impossible to formulate and to solve fundamental problems in these fields.

A vector space X defined over a field C is a set in which the operations of multiplication by a scalar and of sum are defined and have (among others) the following properties:

1. if $x \in X$ and $\lambda \in C$, then $\lambda x \in X$;

2. if $x_1,\ x_2 \in X$, then $x_1 + x_2 \in X$.

We will consider vector spaces defined over the field $\mathbb{R}$. As is known, there are some axioms that must be satisfied.[1]

Example A.1. *The following sets are vector spaces.*

- *The points (geometric vectors) in the plane or in the space.*
- *The points on a 2D-plane in the 3D-space including the origin.*
- *The set of polynomials of degree $\leq n$.*
- *The set of n-tuple $(\alpha_1, \dots, \alpha_n)$.*

[1] We briefly remind these axioms. For $u,\ v \in X$ and $\alpha,\ \beta \in C$: 1) $u + (v + w) = (u + v) + w$; 2) $u + v = v + u$; 3) (zero element) $\exists\, 0 \in X$: $v + 0 = v$ for all $v \in X$; 4) (inverse element) $\forall\, v \in X\ \exists\ -v \in X$: $v + (-v) = 0$; 5) $\alpha(\beta\, v) = (\alpha\beta)\, v$; 6) $1\ v = v$, where 1 is the multiplicative identity in C, 7) $\alpha(u + v) = \alpha u + \alpha v$.

- *The set of solutions of a linear homogeneous differential equation* $a\ddot{y}(t) + b\dot{y}(t) + cy(t) = 0$.
- *The set of polynomials of degree n.*
- *The set of polynomials of degree n for which* $p(0) = 0$.

The following sets are not vector spaces.

- *The points on a 2D-plane in the 3D-space not including the origin.*
- *The set of solutions of a linear non-homogeneous differential equation* $a\ddot{y}(t)+b\dot{y}(t)+cy(t) = 1$.
- *The set of polynomials of degree n for which* $p(0) = 1$.

Definition A.1. ***(Generator set.)*** *A set of vectors* $x_1, x_2, \dots, x_m \in X$ *is said to be a* ***generator set*** *if any element* $x \in X$ *can be written as a linear combination of these vectors, that is,*

$$x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_m x_m$$

for some coefficients $\lambda_i \in \mathbb{R}$.

Definition A.2. ***(Set of linearly independent vectors.)*** *Vectors* $x_1, x_2, \dots, x_n \in X$ *are said* ***linearly independent*** *if the condition*

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_m x_m = 0$$

implies that $\lambda_i = 0$ *for each i (in simple words, if the linear combination with all coefficients equal to zero is the only one that gives the* 0 *vector).*

Definition A.3. ***(Basis.)*** *A set of vectors* $x_1, x_2, \dots, x_n \in X$ *is called* ***basis*** *if it is a set of generators that are linearly independent.*

Theorem A.1. *Given a basis,* $x_1, x_2, \dots, x_n \in X$, *each vector x can be uniquely written as*

$$x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n,$$

namely, the coefficients λ_i *are uniquely determined by x.*

Given a vector space X, its basis is not unique. The following fundamental property holds.

Theorem A.2. *All bases of X have the same number of elements.*

This number is invariant: we can replace a basis with another one having the same number of elements. This has a considerable importance.

Definition A.4. *The cardinality (the number of elements) of any basis of X is called* **dimension** *of the vector space X:*

$$Dim(X).$$

Example A.2. ***(Polynomials of degree*** **4*.)*** *Consider the set of polynomials of degree* 4, *having the form:*

$$p(s) = a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s^1 + a_0.$$

The 0 *polynomial is* $p(s) \equiv 0$, *with all zero coefficients. The polynomials s and* $s^3 + 1$ *are linearly independent. The polynomials* s^4, s^3 , s^2, s^1, 1, $s^3 + 1$ *are a set of generators, the set* s^4, s^3 , s^2, s^1, 1 *is a basis. We can also take* $(s + 1)^4$, $(s + 1)^3$, $(s + 1)^2$, $(s + 1)^1$, 1, *which is a different basis, but has the same number of elements. The reader should prove the above statements as an exercise.*

Example A.3. ***(Euclidean space.)*** *Consider the space of vectors in $\mathbb{R}^3$, with 3 real components, of the type $x = [\alpha,\ \beta,\ \gamma]^\top$. This is a vector space of dimension 3. A possible basis is*

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

which is called canonical basis.

The set $X = \{0\}$, including the 0 vector only, is a vector space, and it is named **trivial space**. It has dimension 0.

Definition A.5. ***Subspace*** *A subset X_1 of X which is also a vector space is called a* ***subspace****.*

Example A.4. ***(Subspaces.)*** *If X is the space of polynomials of degree 4, then the polynomials of degree 2 form a subspace. In the 3D-space, any plane including the origin is a subspace. Precisely, in the space of vectors (x_1, x_2, x_3) with coordinates x_1, x_2 and x_3, the subset $\{(x_1, x_2, x_3) :\ 2x_1 - x_2 + x_3 = 0\}$ is a subspace.*

Given two subspaces X_1 and X_2, the following operations are defined and provide subspaces.

- Intersection:

$$X_1 \bigcap X_2 = \{x :\ \ x \in X_1,\ \text{and}\ x \in X_2\}$$

- Sum:

$$X_1 + X_2 = \{x = x_1 + x_2,\ \ x_1 \in X_1,\ \text{and}\ x_2 \in X_2\}$$

Definition A.6. ***(Direct sum.)*** *The sum X of two subspaces is said direct if any element $x \in X$ can be written as $x = x_1 + x_2$, where $x_1 \in X_1$ and $x_2 \in X_2$ are* ***uniquely determined****.*

It can be shown that the sum is direct if and only if

$$X_1 \bigcap X_2 = \{0\}.$$

When the sum is direct we use the notation

$$X = X_1 \bigoplus X_2.$$

Given a basis $v_1, v_2, \ldots, v_n$, each vector of the vector space can be written as

$$v = x_1 v_1 + x_2 v_2 + \cdots + x_n v_n,$$

where the numbers $x_1,\ x_2,\ \ldots, x_n$ are called the **components** of v. For a fixed basis, they are uniquely determined. Therefore, if the basis is fixed, there is a correspondence between the vectors v and n-tuples of numbers $(x_1, \ldots, x_n)$:

$$v \leftrightarrow \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

where $\mathbb{R}^n$ is, by definition, the set of all n-tuples. For example, in the 4-degree polynomial space, with basis s^4, s^3, s^2, s, 1, the polynomial $x^4 + 5x + 3$ is associated with the tuple $(1, 0, 0, 5, 3)$.

Therefore, without loss of generality we can consider the space $\mathbb{R}^n$ of n-tuples. When necessary, we will consider $\mathbb{C}^n$, the space of complex n-tuples.

A.2 Matrices

Given two spaces X and Y, we say that a map (or operator, or application) $L : X \to Y$ is **linear** if for all $x_a, x_b \in X$ and $\alpha, \beta \in \mathbb{R}$, (or $\mathbb{C}$) we have

$$L(\alpha x_a + \beta x_b) = \alpha L(x_a) + \beta L(x_b).$$

Examples of linear maps are: from the 4th degree polynomial space to the 3rd degree polynomial space, the derivative; between 3D spaces, a rigid rotation (with respect to an axis that includes the origin).

We have seen that, for a fixed basis, any space of dimension n is isomorphic to $\mathbb{R}^n$. Hence, we can focus on operators from the space $\mathbb{R}^n$ to the space $\mathbb{R}^m$. A linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ is represented by a matrix A of size $m \times n$. Let $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$. Then, we have

$$y = Ax,$$

where

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Hence, the coefficients y_i are computed according to the rule

$$y_i = \sum_{j=1}^{n} a_{ij} \, x_j.$$

Example A.5. *The derivative as an operator from the* 3*rd degree polynomial space to the* 2*nd degree polynomial space, with basis* 1, s, s^2, s^3 *and* 1, s, s^2 *respectively, is represented by the matrix*

$$D = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

$$\begin{tikzcd} \mathrm{Y} & \mathrm{Z} \arrow[l, "\mathrm{A}"'] & \mathrm{X} \arrow[l, "\mathrm{B}"'] \end{tikzcd}$$

Figure A.1: Composition of linear maps.

The composition of two linear operators represented by matrices A and B

$$\mathbb{R}^n \underbrace{\to}_{A} \mathbb{R}^m \underbrace{\to}_{B} \mathbb{R}^p$$

is a linear operator. Given $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^p$, if $y = Ax$ and $z = By$, then

$$z = By = B(Ax) = (BA)x = Cx.$$

If B has size $p \times m$ and A has size $m \times n$, matrix C has size $p \times n$ and its generic element is given by

$$[C]_{ij} = c_{ij} = \sum_{k=1}^{m} b_{ik} \, a_{kj}, \quad i = 1, 2, \dots, p, \quad j = 1, 2, \dots, n.$$

Definition A.7. ***(Image.)*** *The image (or range) of a $m \times n$ matrix is the set*

$$Ra(A) = \{y : y = Ax, \quad x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m.$$

Such a set is a subspace of $\mathbb{R}^m$. Given a matrix M, **its columns form a set of generators for its image**, though not necessarily a basis.

Definition A.8. ***(Kernel.)*** *The* ***kernel*** *(or nullspace) of a $m \times n$ matrix is the set*

$$Ker(A) = \{x : \ Ax = 0\} \subseteq \mathbb{R}^n.$$

Such a set is a subspace of $\mathbb{R}^n$.

Example A.6. *Consider the matrix*

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

Its columns are generators for the space $Ra(M)$. We notice that, if we subtract the second column from the first one, we get the third column. Therefore there are two linearly independent columns only (we choose for instance the first two), which are a basis of $Ra(M)$ (a subspace of the 4D space $\mathbb{R}^4$). The kernel of matrix M is given by all the vectors of the form

$$\bar{x} = \alpha \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \qquad \alpha \in \mathbb{R},$$

and is a subspace of $\mathbb{R}^3$.

Example A.7. *Let us compute the kernel of the following matrix:*

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

We have to find all the solutions x of $Mx = 0$, the null space. To get a basis of such a subspace, we can subtract the first row from the second and the first row multiplied by two from the third

$$\begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & -3 & 1 & 0 \\ 0 & -3 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = 0$$

and then subtract the second row from the third

$$\left[\begin{array}{cc|cc} 1 & 2 & 0 & 1 \\ 0 & -3 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = 0$$

The solution is unchanged. The third row is neglected, since it is zero. We parameterise $x_3 = \alpha$ and $x_4 = \beta$ and we derive $x_2 = \alpha/3$ and $x_1 = -\beta - 2x_2 = -\beta - 2\alpha/3$. Finally we have that any vector in the kernel can be written as

$$x_K = \begin{bmatrix} \beta - \frac{2}{3}\alpha \\ \frac{1}{3}\alpha \\ \alpha \\ -\beta \end{bmatrix} = \alpha \begin{bmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} = \alpha v_1 + \beta v_2$$

The two vectors v_1 and v_2 form a basis for the kernel of matrix M.

Definition A.9. ***(Transposed matrix.)*** *Given the $m \times n$ matrix A, its transposed matrix $A^\top$ is the $n \times m$ matrix whose elements are*

$$[A^\top]_{ij} = a_{ji}.$$

The transpose of a product is the product of the transposed matrices in reverse order:

$$(A\,B\,C\ldots)^\top = (\ldots C^\top\,B^\top\,A^\top).$$

A.2.1 Determinant

For a square $n \times n$ matrix, the *determinant* is a function defined as follows. Given the fundamental tuple of integers $(1, 2, \ldots, n)$ and the generic permutation

$$p = (p_1, p_2, \ldots, p_n)$$

we say that p is of even class (class = 2) if the number of element swaps to get the fundamental is even, of odd class (class = 1) in the opposite case.

Let P be the set of all permutations of the fundamental and $cl(p) \in \{1, 2\}$ the class of a permutation p. Then the determinant (according to the Leibniz formula) is defined as

$$\det(A) = \sum_{p \in P} (-1)^{cl(p)}\; a_{1p_1} a_{2p_2} \ldots a_{np_n}.$$

It turns out that

$$\det(A) = \det(A^\top) = \sum_{p \in P} (-1)^{cl(p)}\; a_{p_1 1} a_{p_2 2} \ldots a_{p_n n}.$$

The following formulae (Laplace formulae) can be used to compute the determinant of a matrix, by rows or by columns. Denote by $\hat{A}_{ij}$ the complement matrix to entry (i, j), namely, the square submatrix obtained from A by deleting ith row and and the jth column. We have that

$$\begin{aligned} \det(A) &= \sum_{j=1}^{n} (-1)^{i+j}\; a_{ij}\; \det(\hat{A}_{ij}), \quad \text{for each } i \\ \det(A) &= \sum_{i=1}^{n} (-1)^{i+j}\; a_{ij}\; \det(\hat{A}_{ij}), \quad \text{for each } j \end{aligned}$$

A square matrix having zero (non-zero) determinant is called **singular** (**non-singular**).

Fundamental properties of the determinant are the following (A, B are $n \times n$ matrices):

- $\det(AB) = \det(A)\det(B)$;
- $\det(I) = 1$;
- Let $\tilde{A}_\lambda$ be the matrix achieved from A by multiplying a row or a column by λ: then, $\det(\tilde{A}_\lambda) = \lambda \det(A)$;
- $\det(\mu A) = \mu^n \det(A)$;
- $\det(A) \neq 0$ if and only if the row (columns) are a basis of $\mathbb{R}^n$.

A.2.2 Rank

The columns or the rows of a matrix are vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$. The following fundamental theorem holds.

Theorem A.3. *Given a generic matrix $A \in \mathbb{R}^{m \times n}$, the maximum number $\hat{m}$ of linearly independent rows is equal to the maximum number $\hat{n}$ of linearly independent columns.*

Definition A.10. ***Rank**. The number $\hat{n} = \hat{m} = \hat{r}$ is called **rank** of matrix A and is denoted by*

$$Rank(A).$$

It turns out that $Rank(A)$ is equal to the dimension of the largest square submatrix of A having nonzero determinant.

If the matrix $m \times n$ has rank equal to the maximum possible (which is the minimum of m and n, number of rows and columns) is said to be a *full rank* matrix, or to have full rank. In particular, a square matrix has full rank if and only if its determinant is non-zero.

The following relations are useful when determining the rank of a matrix or when computing its image or its kernel.

Theorem A.4. *For a $m \times n$ matrix A, the following relations apply:*

$$\begin{aligned} n &= Rank(A) + Dim(Ker(A)) \\ m &= Rank(A) + Dim(Ker(A^\top)). \end{aligned}$$

A.2.3 Linear systems of equations

Given a $m \times n$ matrix A and a vector $b \in \mathbb{R}^m$, consider the system of equations

$$Ax = b.$$

Theorem A.5. *Given the system $Ax = b$, a necessary and sufficient condition for the existence of a solution is that*

$$Rank([A|b]) = Rank(A),$$

namely, the rank does not increase if b is added as a new column to A.

If a solution exists, then the set of all solutions is

$$\bar{x} + \tilde{x},$$

where $\bar{x}$ is any solution and $\tilde{x} \in Ker(A)$ is an element of the kernel.

In the case of a square matrix A ($n \times n$) and a vector $b \in \mathbb{R}^n$, consider the system of equations

$$Ax = b.$$

Since A is square, the following properties are equivalent:

- the system admits a unique solution for all b;
- A has rank n;
- A has a trivial kernel (the kernel is $\{0\}$);
- $\det(A) \neq 0$.

If $\det(A) = 0$, there can be either no solutions or infinite solutions, depending on b. The system $Ax = 0$ admits non-zero solutions if and only if $\det(A) = 0$.

Example A.8. *Consider the following matrix and vector:*

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Then a solution exists if and only if $b_1 = b_2$. If (x_1, x_2) is a solution, then $(x_1 + \alpha, x_2 - \alpha)$ is a solution as well.

Definition A.11. ***(Inverse matrix.)*** *Given a square matrix A such that* $\det(A) \neq 0$*, the matrix $B = A^{-1}$ is its inverse matrix if for every x*

$$y = Ax, \quad \Rightarrow \quad x = By.$$

The inverse matrix has then the property

$$AB = BA = I$$

where I is the identity matrix. The following formula holds to compute the elements B_{ij} of $B = A^{-1}$:

$$B_{ij} = (-1)^{i+j} \; \frac{\det(\hat{A}_{ji})}{\det A}.$$

where $\hat{A}_{ji}$ is the complement of a_{ij}. Note that i and j are exchanged in A_{ji} in the formula!!

The inverse of a product is the product of the inverses in reverse order:

$$(A\,B\,C\dots)^{-1} = (\dots C^{-1}\,B^{-1}\,A^{-1}).$$

Finally, we have

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

A.2.4 Partitioned matrices

In many occasions we have to deal with matrices of matrices, namely matrices whose elements are themselves matrices or vectors. For example, a matrix A can be thought as a "row" of column vectors

$$A = [\; \bar{a}_1 \; \bar{a}_2 \; \dots \bar{a}_m \;]$$

or as a column of row vectors

$$B = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_n \end{bmatrix}.$$

More in general, a matrix can be of the type

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}$$

where A_{ij} are submatrices. Of a particular interest is the product rule: given two partitioned matrices

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1m} \\ B_{21} & B_{22} & \dots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{p1} & B_{p2} & \dots & B_{pm} \end{bmatrix}$$

under the assumption that the number of rows in each submatrix A_{kj} matches the number of columns in the submatrix B_{ik}, the matrix product is

$$C = BA = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{p1} & C_{p2} & \dots & C_{pn} \end{bmatrix}$$

where the same product rule holds

$$C_{ij} = \sum_{k=1}^{m} B_{ik} A_{kj}, \quad i = 1, 2, \dots, p, \quad j = 1, 2, \dots, n.$$

A.2.5 Elementary row and column tranformations

Row and column operations are fundamental to determine rank, kernel, image space of matrices, as well as solutions of linear systems of equations.

Given a $n \times m$ matrix, we call elementary row or column transformation[2] one of the following operations:

- multiplication of one row (column) by a non-zero scalar;
- permutation of two rows (columns);
- adding one row (column) to another row (column) multiplied by a scalar.

Row operations have the following properties

- they leave the rank unchanged;
- they leave the kernel unchanged;
- they leave the set of solutions unchanged if the matrix $M = [A\ b]$ represents a system of linear equations $Ax = b$.

Column operations have the following properties

- they leave the rank unchanged;
- they leave the image space unchanged.

A $m \times n$ matrix M is an upper-staircase matrix if row i has v_i non-zero elements in the first positions and

$$\text{if } v_k < n, \text{ then } v_{k+1} > v_k,$$

[2] This type of transformations is **not** the similarity transformation presented in Subsection A.2.6.

while if $v_k = n$, then $v_h = n$, for $h \geq k$. For instance

$$\begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

is an upper-staircase matrix. A $m \times n$ matrix M is row-reduced if

- it is an upper staircase matrix;
- the first non-zero element is equal to a 1, namely $M_{i,v_i+1} = 1$, unless the row is zero;
- in each column where a pivot (unitary) element exists, all other elements are 0.

For instance the following matrix is row-reduced.

$$\begin{bmatrix} \underline{1} & 0 & 3 & 0 & 1 \\ 0 & \underline{1} & 1 & 0 & 1 \\ 0 & 0 & 0 & \underline{1} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Each $m \times n$ matrix can be row-reduced by means of elementary row operations. The procedure works as follows.

Procedure A.1.

1. *Set $i := 1,\ j := 1$.*
2. *If, in the jth column, there exists a non-zero element in position $k \geq i$, (including the ith element) take it in position i (by row permutation). Otherwise consider the next column: $j := j + 1$.*
3. *Multiply row i by $1/M_{ij}$. Now the element i, j has become a pivot (it is equal to 1).*
4. *Add row i to all other rows $k \neq i$ after multiplication by $-M_{kj}$, to set to zero all elements of column j (excluding the pivot $M_{ij} = 1$).*
5. *Set $i := i + 1,\ j := j + 1$ and go to Step 2 to repeat the procedure until either $i = m$ or all elements $M_{k,h}$ are 0 for $k \geq i$ and $h \geq j$.*

Example A.9. *Given the matrix M below on the left, set $i = 1$ and $j = 1$. $M_{11} = 1$ becomes pivot. perform the following operations: $Row(2) := Row(2) - Row(1)$, $Row(3) := Row(3) - 2Row(1)$, achieving the matrix M' on the right,*

$$M = \begin{bmatrix} 1 & 1 & -1 & 1 & 0 & -1 \\ 1 & 1 & 2 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 2 \\ 2 & 2 & 0 & 0 & 1 & 2 \end{bmatrix} \quad \Rightarrow \quad M' = \begin{bmatrix} 1 & 1 & -1 & 1 & 0 & -1 \\ 0 & 0 & 3 & -2 & 0 & 2 \\ 0 & 0 & -1 & 0 & 1 & 2 \\ 0 & 0 & 2 & -2 & 1 & 4 \end{bmatrix}$$

Set $i = 2$ and $j = 2$. The element M'_{22} is 0 as all the other elements of the second column. Then consider the third column, $j = 3$. The element M'_{23} is non-zero. Make it unitary (pivot) by division of the row $Row(2) := Row(2)/M'_{23}$, achieving the matrix M'' below on the left. Then *operate: $Row(1) := Row(1) - Row(2) \cdot M''_{13}$, $Row(3) := Row(3) - Row(2) \cdot M''_{33}$, $Row(4) := Row(4) - Row(2) \cdot M''_{43}$, achieving the matrix M''' below on the right.*

$$M'' = \begin{bmatrix} 1 & 1 & -1 & 1 & 0 & -1 \\ 0 & 0 & 1 & -2/3 & 0 & 2/3 \\ 0 & 0 & -1 & 0 & 1 & 2 \\ 0 & 0 & 2 & -2 & 1 & 4 \end{bmatrix} \Rightarrow M''' = \begin{bmatrix} 1 & 1 & 0 & 1/3 & 0 & -1/3 \\ 0 & 0 & 1 & -2/3 & 0 & 2/3 \\ 0 & 0 & 0 & -2/3 & 1 & 8/3 \\ 0 & 0 & 0 & -2/3 & 1 & 8/3 \end{bmatrix}$$

In the third column $M'''_{23} = 1$ is non-zero (pivot). Let $i = 3$ and $j = 4$. The element $M'''_{34} \neq 0$. Operate: $Row(3) := Row(3)/M'''_{34}$, so the element becomes 1 (pivot), achieving the matrix M'''' below on the left. Operate the following: $Row(1) := Row(1) - Row(3) \cdot M''''_{14}$, $Row(2) := Row(2) - Row(3) \cdot M''''_{24}$, $Row(4) := Row(4) - Row(3) \cdot M''''_{44}$, achieving the matrix M''''' below on the right.

$$M'''' = \begin{bmatrix} 1 & 1 & 0 & 1/3 & 0 & -1/3 \\ 0 & 0 & 1 & -2/3 & 0 & 2/3 \\ 0 & 0 & 0 & 1 & -3/2 & -8/2 \\ 0 & 0 & 0 & -2/3 & 1 & 8/3 \end{bmatrix} \Rightarrow M''''' = \begin{bmatrix} \underline{1} & 1 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & \underline{1} & 0 & -1 & -2 \\ 0 & 0 & 0 & \underline{1} & -3/2 & -8/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Setting $i = 4$ and $j = 5$, $M'''''_{45} = 0$ and there are only zero elements corresponding to greater or equal indices: STOP.
The matrix is row-reduced. The underlined elements are pivot. The number of nonzero rows is the rank, in this case equal to 3.

The procedure is useful to find the kernel and the solution of a system of linear equations.

Example A.10. *Consider the system $Ax = b$, with*

$$Ax = \begin{bmatrix} 1 & 1 & -1 & 1 & 0 \\ 1 & 1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 2 & 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 2 \\ 2 \end{bmatrix} = b$$

Consider the complete system matrix

$$[A|b] = \left[\begin{array}{ccccc|c} 1 & 1 & -1 & 1 & 0 & -1 \\ 1 & 1 & 2 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 2 \\ 2 & 2 & 0 & 0 & 1 & 2 \end{array}\right]$$

A necessary and sufficient condition for the existence of solutions is

$$rank(A) = rank([A|b])$$

equivalent to

$$rank(\hat{A}) = rank([\hat{A}|\hat{b}])$$

where $[\hat{A}|\hat{b}]$ is the transformed matrix after row-reduction. Checking the existence of solutions is easy. Indeed we arrive to a form of the type

$$[\hat{A}|\hat{b}] = \left[\begin{array}{c|c} \hat{A}_1 & \hat{b}_1 \\ \hline \hat{0} & \hat{b}_2 \end{array}\right]$$

and the rank equality is equivalent to $\hat{b}_2 = 0$, which is then a necessary and sufficient condition for the existence of solutions. In the example we get

$$\left[\begin{array}{ccccr|c}
x_1 & x_2 & x_3 & x_4 & x_5 & \hat{b} \\ \hline
\underline{1} & 1 & 0 & 0 & 1/2 & 1 \\
0 & 0 & \underline{1} & 0 & -1 & -2 \\
0 & 0 & 0 & \underline{1} & -3/2 & -4 \\ \hline
0 & 0 & 0 & 0 & 0 & [0]
\end{array}\right]$$

and the 0 *evidenced in brackets means that a solution exists.*

How can we find all solutions of $Ax = b$? We remind that, if $\bar{x}$ is any solution ($A\bar{x} = b$), the solution set is

$$Sol(A, b) = \{x = \bar{x} + \tilde{x}, \ \ where\ \tilde{x} \in Ker(A)\}.$$

- *A particular solution $\bar{x}$ is obtained by setting to* 0 *all elements x_i which do not correspond to columns with pivots and computing the remaining, corresponding to pivots, which are equal to the corresponding entry of the last column.*

- *The kernel basis is the solution we get after setting to* 0 *the last column (the known terms). This is achieved as follows. We consider the columns without pivots: each of them is associated with a vector of the kernel basis. For each non-pivot column, we set to* 1 *the corresponding component x_i and to* 0 *all the elements corresponding to other non-pivot columns. We then simply derive the elements corresponding to pivot columns.*

In the specific case, the pivots are in the positions 1, 3 *and* 4. *The non-pivot positions are* 2 *and* 5.

To get a particular solution we put $x_2 = 0$ and $x_5 = 0$, then $x_1 = 1$, $x_3 = -2$ and $x_4 = -4$, namely

$$\bar{x}^\top = \left[\begin{array}{ccccc} 1 & 0 & -2 & -4 & 0 \end{array}\right]$$

A kernel basis is achieved by setting to 0 *the last column (zeroing the elements of column $\hat{b}$). The first vector of the basis is achieved by setting $x_2 = 1$ and $x_5 = 0$ and computing the other terms*

$$v_1^\top = \left[\begin{array}{ccccc} 1 & -1 & 0 & 0 & 0 \end{array}\right]$$

The second vector, achieved by setting $x_2 = 0$ and $x_5 = 1$, is

$$v_2^\top = \left[\begin{array}{ccccc} -1/2 & 0 & 1 & 3/2 & 1 \end{array}\right].$$

The two vectors v_1 and v_2 form a basis for the 2-dimensional kernel.

In the case of square non-singular matrices, row reduction is useful to compute the inverse as well. Let P be a square invertible matrix. Its inverse is achieved as follows.

Procedure A.2. ***Inverse computation***

1. *Form the matrix*
$$M = [P|I]$$
where I is the identity of the proper size.

2. *Reducing the matrix by row we get*
$$\hat{M} = [I|Q]$$
(since P is invertible, there will be no zero rows).

3. Then $P^{-1} = Q$.

Example A.11. *Let*

$$M = [P|I] = \left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

After row-reduction we get

$$[I|Q] = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 2 & 1 & -1 \end{array}\right]$$

where $Q = P^{-1}$.

The procedure can be applied to columns. Matrix M is column-reduced if $M^\top$ is row-reduced. A basis of the image space is given by the non-zero columns of the column-reduced matrix.

Remark A.1. *If the row and column operations so far described are applied to square matrices, they* **do not** *preserve eigenvalues, eigenvectors and characteristic polynomial. The similarity transformations presented in Section A.2.6 have this property.*

A.2.6 Basis transformation

In representing the space $\mathbb{R}^n$, we implicitly refer to the canonical basis

$$e_k = \begin{bmatrix} 0 & \dots & 0 & \underbrace{1}_{k-th\,position} & 0 \dots & 0 \end{bmatrix}^\top$$

(e_k has all 0 elements except the kth, which is equal to 1). Indeed we have

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{j=1}^{n} x_j e_j.$$

Now suppose we want to write the vector x as the convex combination of vectors forming a new basis, $t_1,\ t_2,\ \dots\ , t_n$, grouped in the matrix

$$T = [t_1\ t_2\ \dots\ t_n].$$

We must have that

$$x = \sum_{j=1}^{n} t_j\, \hat{x}_j = \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_n \end{bmatrix} = T\hat{x},$$

where $\hat{x}_j$ are the components with respect to the new basis. These components, collected in the vector $\hat{x}$, can be determined as follows:

$$x = T\hat{x} \Rightarrow \hat{x} = T^{-1}x. \qquad \text{(A.1)}$$

Given a linear map $y = Ax$, the basis changes $x = T\hat{x}$ and $y = S\hat{y}$ lead to the transformation

$$\hat{y} = S^{-1}AT\, \hat{x}$$

We consider mainly[3] applications of a space to itself, for which the transformation is represented by a square matrix A. In this case the new basis should be the same ($x = T\hat{x}$ and $y = T\hat{y}$). The linear application is then transformed as follows:

$$\hat{y} = T^{-1}AT\, \hat{x} = \hat{A}\hat{x}. \tag{A.2}$$

In this case the transformation is called *similarity transformation*. A similarity transformation has fundamental property, we will show later. In particular we have the following. Any power A^k of $\hat{A}$, with k positive integer, is transformed as A

$$\hat{A}^k = \underbrace{T^{-1}AT\, T^{-1}AT\, \dots\, T^{-1}AT}_{k \text{ times}} = T^{-1}A^kT.$$

If A is invertible, then the property applies also for negative k. This property allows us to define analytic functions of matrices.

A.2.7 Eigenvalues and eigenvectors

Definition A.12. ***(Eigenvalues and Eigenvectors.)*** *Given a square matrix A, a vector $x \neq 0$ and a complex scalar λ such that*

$$Ax = \lambda x,$$

are called respectively ***eigenvector*** *and* ***eigenvalue*** *of A.*

The previous equation can be written in the form

$$(A - \lambda I)x = 0.$$

For this equation to admit a nontrivial solution it is necessary and sufficient that the matrix $(A - \lambda I)$ is singular (equivalently, does not have full rank):

$$\det(\lambda I - A) = 0.$$

This is a polynomial function of λ,

$$\det(\lambda I - A) = p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda^1 + a_0,$$

called **characteristic polynomial** of A. It is a monic polynomial, namely, the coefficient of the highest degree is equal to 1. The eigenvalues are therefore the roots (possibly complex) of this polynomial. Note that complex roots can be present even if A, hence $p(\lambda)$, are real.

The eigenvector $x \neq 0$ that satisfies the relation is also called *right eigenvector*. A vector $z \neq 0$ such that

$$z^\top(A - \lambda I) = 0$$

is called **left eigenvector**. If we transpose the relationship, we have

$$(A^\top - \lambda I)z = 0,$$

hence left eigenvectors are the right eigenvectors of the transpose. Note that the eigenvalues of $A^\top$ and A are the same, because $\det(\lambda I - A) = \det((\lambda I - A)^\top) = \det(\lambda I - A^\top)$.

The set of the eigenvalues of A, which is the set of the roots of $p(\lambda)$, is called **spectrum** of A and is denoted by

$$\sigma(A) = \{\lambda \in C : \;\; p(\lambda) = 0\}.$$

[3] Indeed always, with the exception of malicious exercises given in the tests.

Example A.12. *Let*

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

Then

$$\det(\lambda I - A) = \lambda^2 + 5\lambda + 6$$

The eigenvalues are $\lambda_1 = -2$ and $\lambda_2 = -3$. The corresponding eigenvectors are the solution of

$$(A - \lambda_1)x = \begin{bmatrix} +1 & 2 \\ -1 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \quad \textit{and} \quad (A - \lambda_2)x = \begin{bmatrix} +2 & 2 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

Then possible eigenvectors (there are infinite of them) are

$$t_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \textit{and} \quad t_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Example A.13. *The matrix*

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

is real but its eigenvalues are $\pm j2$, hence complex (imaginary). The eigenvectors are $[1 \quad \pm j]^\top$. Note that the components of the two vectors are conjugate.

Remark A.2. *The eigenvectors are defined up to multiplication by a scalar. If $x \neq 0$ is an eigenvector, $\alpha x \neq 0$ is an eigenvector as well.*

Remark A.3. *If A is real, its eigenvalues have the conjugate property: $\lambda \in \sigma(A)$ implies that the conjugate $\lambda^* \in \sigma(A)$. This is trivial if λ is real. If, for instance $-1 + 2j \in \sigma(A)$, then $-1 - 2j \in \sigma(A)$. Finding a complex eigenvector is harder, but once the eigenvector v of λ is known, the eigenvector associated with λ^* is v^* and comes for free.*

Assume, for the moment, that the matrix A admits n distinct eigenvalues, namely that the roots of $p(\lambda)$ are distinct

$$\sigma(A) = \{\, \lambda_1,\ \lambda_2\ ,\ \dots\ ,\ \lambda_n \,\}.$$

In this case the corresponding eigenvectors t_j are linearly independent. Call T the matrix whose columns are the eigenvectors

$$T = \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix}.$$

Then the following relation holds:

$$A \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix} = \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}.$$

If we define $S \doteq T^{-1}$ we have the following relations

$$\begin{aligned} AT &= T\Lambda \\ SA &= \Lambda S \\ A &= T\Lambda S \\ \Lambda &= SAT. \end{aligned}$$

Note that partitioning S by rows

$$S = \begin{bmatrix} s_1^\top \\ s_2^\top \\ \vdots \\ s_n^\top \end{bmatrix}$$

the relation $SA = \Lambda S$ tells us that the rows of S are left eigenvectors.

The case in which matrix A has no distinct eigenvalues is much more complicated. It can be shown that there is a matrix $T \in \mathbb{R}^{n\times n}$ such that $A = TJT^{-1}$ is a block-diagonal matrix of the form

$$J = diag(J_1, \dots, J_r),$$

where the k-th block J_k, having size $\mu_k \times \mu_k$, is of the form

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

Matrix J is said to be in *Jordan form*, or a *Jordan matrix*.

So in general we may transform A as

$$A = [T_1 T_2 \dots T_r] \begin{bmatrix} J_1 & 0 & 0 & 0 & \dots \\ 0 & J_2 & 0 & 0 & \dots \\ 0 & 0 & J_3 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & J_r \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_r \end{bmatrix} \tag{A.3}$$

where $T_k \in \mathbb{R}^{n\times\mu_k}$ and $S_k^\top \in \mathbb{R}^{\mu_k\times n}$. The Jordan form is unique up to permutations of the blocks. The number of blocks and their dimensions are a characteristic of the matrix and cannot be changed. For each eigenvalue λ, three basic parameters can be defined.

Algebraic Multiplicity: multiplicity of λ as a root of the characteristic polynomial.

Geometric Multiplicity: the number of blocks associated with λ in the Jordan form.

Ascent: the size of the largest block associated with λ.

Remark A.4. *For simple eigenvalues all of these numbers are equal to* 1.

The ascent of an eigenvalue is particularly important. There is a simple way to calculate it *without passing through the Jordan form.* Define the following numbers:

$$\begin{aligned} \rho_1 &= rank(\lambda I - A)^1 \\ \rho_2 &= rank(\lambda I - A)^2 \\ &\vdots \\ \rho_g &= rank(\lambda I - A)^g \end{aligned}$$

It can be shown that $\rho_i \geq \rho_{i+1}$. The computation is carried out until equality occurs:

$$\rho_1 > \rho_2 > \dots \rho_g = \rho_{g+1}.$$

The first index for which the equality holds is the ascent g of λ ($g = \min_i : \rho_i = \rho_{i+1}$).

Example A.14. *Let*

$$A_1 = \begin{bmatrix} -2 & 1 & 0 \\ -1 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -2 & -1 & -1 \\ -1 & -2 & -1 \\ -1 & -1 & -2 \end{bmatrix}$$

Matrix A_1 has the eigenvalue -1 with algebraic multiplicity 3. Computing the rank of the powers $(A_1 + 1I)^k$, we get $\rho_1 = 2$, $\rho_2 = 1$, $\rho_3 = 0$ and $\rho_4 = 0$. The ascent of the eigenvalue -1 is therefore 3. Matrix A_2 has a simple eigenvalue $\lambda = -4$ (hence of ascent 1) and another eigenvalue $\lambda = -1$. For the latter eigenvalue, $\rho_1 = 1$ and $\rho_2 = 1$, hence the ascent is 1.

The following properties concerning the eigenvalues are fundamental.

Proposition A.1. *A similitude transformation preserves the characteristic polynomial, hence the eigenvalues. In fact, if $\hat{A} = T^{-1}AT$, then*

$$\begin{aligned} \det(sI - \hat{A}) &= \det(sT^{-1}T - T^{-1}AT) = \det[T^{-1}(sI - A)T] = \\ &= \det(T^{-1})\det(sI - A)\det(T) = \det(sI - A). \end{aligned}$$

The eigenvectors are transformed as any other vector.

A.2.8 Norms and inner product

Given a vector space X, for $x \in X$ we can define a function $\|x\|$, called norm, having the following properties:

- $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$;
- $\|\lambda x\| = |\lambda| \|x\|$;
- $\|x + y\| \leq \|x\| + \|y\|$.

The norm represents the "size" of x. A famous norm is the geometric length, or 2-norm, or Euclidean norm

$$\|x\| \doteq \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Other norms are the 1-norm

$$\|x\|_1 \doteq \sum_{i=1}^{n} |x_i|$$

and the ∞-norm

$$\|x\|_\infty \doteq \max_i |x_i|.$$

The Euclidean and the 1-norm are special cases of the p-norm

$$\|x\| \doteq \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

There are many other norms. In general $\|Fx\|$ is a norm of $x \in \mathbb{R}^n$, if F has rank n. All norms in $\mathbb{R}^n$ are equivalent. Precisely, given two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, there exist positive constants α and β such that

$$\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1.$$

We will mainly use the Euclidean norm.

Given a vector x in $\mathbb{C}^n$ we define the Hermitian conjugate x^H as the vector obtained by transposing x and conjugating the elements.

$$\begin{bmatrix} 2+j \\ 3-j4 \end{bmatrix}^H = \begin{bmatrix} 2-j & 3+j4 \end{bmatrix}$$

Clearly for real vectors in $\mathbb{R}^n$ the Hermitian conjugate is the transposed: $x^H = x^\top$. Given two vectors x, y in $\mathbb{C}^n$, we define the scalar product as

$$(x, y) \doteq x^H y$$

For vectors in $\mathbb{R}^n$,

$$(x, y) \doteq x^\top y$$

As a particular case, when $x = y$ we have that

$$(x, x) = ||x||^2 = \sum_{i=1}^{n} x_i^2.$$

This is the square of the length, or Euclidean norm, of the vector.

Two vectors x and y are **orthogonal** if their scalar product is zero:

$$(x, y) = 0.$$

In three dimensions, the inner product is equal to the product of the lengths and the cosine of the angle θ formed by the vectors

$$(x, y) = ||x||\, ||y|| \cos(\theta).$$

This enables us to define "angles" between vectors in dimension greater than 3 as

$$\cos(\theta) = \frac{(x, y)}{||x||\, ||y||}.$$

To give a further interpretation, consider a unit vector v with $||v|| = 1$ (note that, given a vector $x \neq 0$, it is always possible to normalise it and get a unit vector $v = \frac{x}{||x||}$). Then the quantity (y, v) represents the component y along the direction v.

Definition A.13. *Given a subspace X of $\mathbb{R}^n$, the set of all orthogonal vectors is a subspace and it is called orthogonal complement*

$$X^\perp = \{y : \ (x, y) = 0\}$$

Definition A.14. *A square $n \times n$ matrix $Q = [q_1\ q_2 \ldots q_n]$ is said* ***orthogonal*** *if its columns are non-zero and pairwise orthogonal: $q_i^\top q_j = 0$ if $i \neq j$. It is said* ***orthonormal*** *if its columns are pairwise orthogonal and of unit length:*

$$q_i^\top q_j = \begin{cases} 0 & if \quad i \neq j \\ 1 & if \quad i = j \end{cases} \tag{A.4}$$

The columns of an orthonormal matrix form an orthonormal basis. Orthonormal bases are very convenient in computation, because the new components of a vector x are just computed as $Q^\top x$. Indeed, an orthonormal matrix has an inverse equal to the transpose

$$Q^{-1} = Q^\top$$

because $Q^\top Q$ has components $[Q^\top Q]_{ij}$ as in (A.4), hence is the identity.

Transformations via an orthonormal matrix preserve the scalar product and the Euclidean norm:

$$(Qx, Qy) = x^\top Q^\top Qy = x^\top y = (x, y).$$

Given a matrix M, we can define an induced norm, which is the maximum stretching of the vector on which the matrix operates:

$$\|M\| \doteq \sup_{\|x\|=1} \|Mx\| = \sup_{\|x\|\neq 0} \frac{\|Mx\|}{\|x\|}$$

The induced norm depends of the chosen vector norm. If we take the Euclidean norm,

$$\|M\| = \max \sqrt{\sigma(M^\top M)}$$

is the maximum square root of the eigenvalues of $M^\top M$, which are all real and nonnegative, as we will shown soon.

A.2.9 Symmetric matrices

We remind that, given a matrix $A \in \mathbb{C}^{m\times n}$ with elements $\left[a_{ij}\right]$, its **transpose** is defined as the matrix $A^\top$ of dimension $n \times m$ with elements $\left[a_{ji}\right]$. For complex matrices we define the **hermitian conjugate** as the matrix A^H of dimension $n \times m$ with elements $\left[a^*_{ji}\right]$ (transpose and conjugate).

Example A.15. *For instance we have*

$$\begin{bmatrix} 2+j & 1 & 1+3j \\ 3-j4 & 2-2j & 4j \end{bmatrix}^H = \begin{bmatrix} 2-j & 3+j4 \\ 1 & 2+2j \\ 1-3j & -4j \end{bmatrix}$$

We can easily verify the following properties:

- $(AB)^\top = B^\top A^\top$;
- $(AB)^H = B^H A^H$.

A matrix is said **Hermitian** if $P = P^H$. In the case of real matrices an Hermitian matrix $P \in \mathbb{R}^{n\times n}$ is called **symmetric**: $P = P^\top$. We can prove the following.

Theorem A.6. *An Hermitian matrix (symmetric if real) has real eigenvalues.*

Proof Consider the expressions

$$\begin{aligned} Px &= \lambda x \\ P^* x^* &= \lambda^* x^* \end{aligned}$$

pre-multiply the first by x^H, and transpose the second, thus obtaining

$$\begin{aligned} x^H Px &= \lambda x^H x \\ x^{*T} P^{*T} &= x^{*T} \lambda^* \quad \Rightarrow \quad x^H P^H = x^H \lambda^* \end{aligned}$$

Post-multiply the second equation by x, to get

$$\begin{aligned} x^H Px &= \lambda x^H x \\ x^H P^H x &= \lambda^* x^H x \end{aligned}$$

Since P is Hermitian, *i.e.*, $P = P^H$, these two expressions are the same:

$$\lambda x^H x = \lambda^* x^H x$$

Since $x^H x = \sum_{i=1}^n x_i^* x_i = \sum_{i=1}^n |x_i|^2$ is always greater than zero because $x \neq 0$, then it must necessarily happen that $\lambda = \lambda^*$, that is, the eigenvalues of the matrix P are real. □

It can also be shown the following.

Theorem A.7. *For a symmetric matrix, the eigenvectors q_i and q_j associated with two distinct eigenvalues are orthogonal:*

$$q_i^\top q_j = q_j^\top q_i = 0.$$

In general (even if there are repeated eigenvalues) a symmetric matrix admits n eigenvectors that form an orthonornal basis.

Indeed, we can derive eigenvectors that are mutually orthogonal and that can be normalised to find a basis of orthonormal eigenvectors $Q = [\ q_1\ q_2\ \dots q_n\ ]$. In this case, (A.4) holds and then $Q^\top Q = I$: Q is **orthonormal**.

The previous properties imply that, for a symmetric matrix P, the diagonalising similarity transformation is orthonormal

$$\begin{aligned} PQ &= Q^\top \Lambda \\ Q^\top P &= \Lambda Q^\top \\ P &= Q \Lambda Q^\top \\ \Lambda &= Q^\top A Q. \end{aligned}$$

A.2.10 Semidefinite and definite matrices

A symmetric matrix P is said **positive semidefinite** (or **negative semidefinite**) if:

$$x^\top P x \geq 0 \quad (\leq 0) \quad \forall\, x.$$

A symmetric matrix P is said **positive definite** (or **negative definite**) if:

$$x^\top P x > 0 \quad (< 0) \quad \forall\, x \neq 0.$$

In all other cases, when $x^\top P x$ can assume both negative and positive signs, we call P **indefinite**. Note that P is a negative (semi)definite matrix if and only if $-P$ is positive (semi)definite.

If P is positive definite, in two dimensions the level surfaces of the function $V(x) = x^\top P x$ are ellipses, in three dimension are ellipsoids. The graph of $V(x) = x^\top P x$ with x two dimensional is shown in Figure A.2 and its level curves are represented in Figure A.3.

A fundamental property of these matrices is the following.

Theorem A.8. *Denote as $\lambda_1, \dots, \lambda_n$ the real eigenvalues of the symmetric matrix P.*

- *P is positive (negative) semidefinite if and only if, for all i, $\lambda_i \geq 0$ ($\lambda_i \leq 0$);*
- *P is positive (negative) definite if and only if, for all i, $\lambda_i > 0$ ($\lambda_i < 0$).*

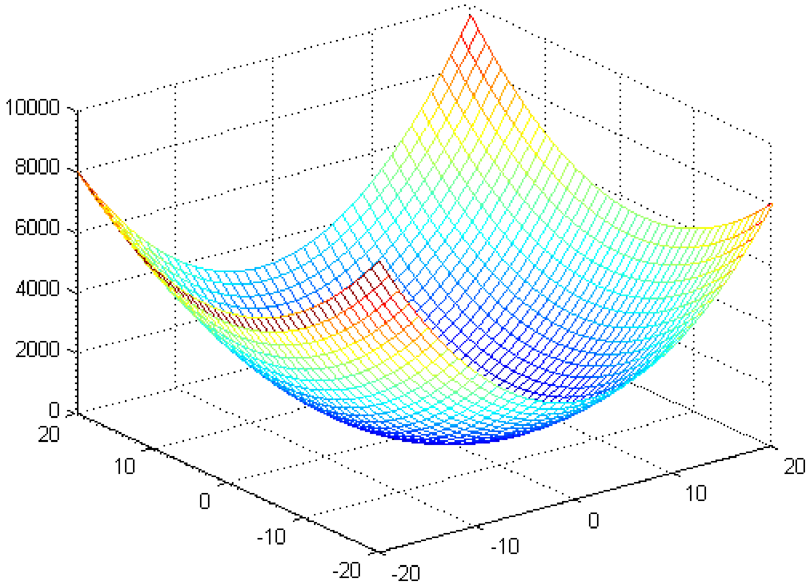

Figure A.2: Graph of the quadratic function $V(x)$.

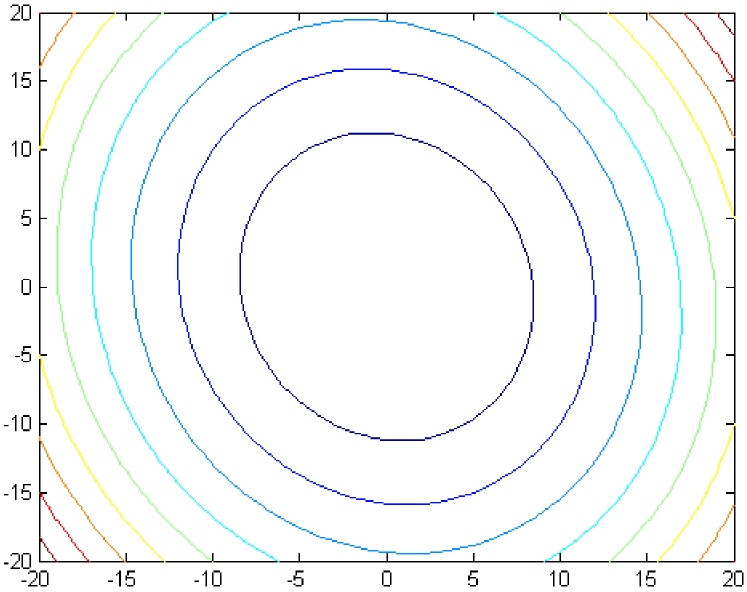

Figure A.3: Level curves of the quadratic function $V(x)$.

Proof Matrix P is diagonalisable by similarity transformation according to the formula $P = Q\Lambda Q^\top$, where Q is orthonormal. Then, denoting by $z = Qx$,

$$x^\top Px = x^\top Q^\top \Lambda Qx = z^\top \Lambda z = \sum_i \lambda_i z_i^2$$

This function is positive if and only if $\lambda_i > 0$ for all i (and so on).

An additional method to check whether a matrix P is positive definite is the **Sylvester criterion**. Denote by P_k the principal sub-matrix formed by considering the first k rows and k columns.

Proposition A.2. *P is positive definite if and only if* $\det P_k > 0$ $\forall\, k = 1, 2, \dots, n$.

Example A.16. *Matrix P, with*

$$P = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} = P_3, \quad P_2 = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}, \quad P_1 = [3],$$

is positive definite because the three first leading determinants determinants $\det(P_1)$, $\det(P_2)$ *and* $\det(P_3) = \det(P)$ *are positive (check it).*

A.3 Polynomials and power series

A polynomial in the complex variable s is a function of the form

$$p(s) = p_0 + p_1 s + p_2 s^2 + \cdots + p_n s^n.$$

The largest power n is the degree of the polynomial. A polynomial of degree n has always n roots in the complex field: $\lambda_1, \lambda_2, \dots, \lambda_n$ (which are not necessarily distinct). Then the polynomial can be written as

$$p(s) = (s - \lambda_1)(s - \lambda_2)\dots(s - \lambda_n).$$

If the coefficients are real, the roots are in general complex numbers; however, the set of the roots is symmetric with respect to the real axis: if z is a root, then also z^* is a root.

We call *power series* a polynomial of "infinite" degree:

$$f(s) = f_0 + f_1 s + f_2 s^2 + \cdots + f_h s^h + \cdots = \sum_{h=0}^{\infty} f_h\, s^h.$$

A function $f(s)$ of this type is defined in a certain domain of convergence, having the form

$$D = \{s : \ |s| < \rho\},$$

where $\rho \geq 0$ is called *convergence radius*. It depends on the coefficients $\{f_h\}$. For instance

$$1 + s + s^2 + s^3 + s^4 \dots$$

has convergence radius $\rho = 1$. The series

$$1 + \frac{s}{1!} + \frac{s^2}{2!} + \frac{s^3}{3!} + \frac{s^4}{4!} + \dots$$

has convergence radius $\rho = \infty$. These two series, where convergent, are equal to $\frac{1}{1-s}$ and e^s respectively. A polynomial is a particular series for which all coefficients are 0 after a finite n and its convergence radius is infinite.

Functions that can be expressed as power series are called *analytic* and have several significant properties (most of which are not reported here, for brevity). The following principle applies.

Theorem A.9. ***(Identity principle of power series.)*** *Let l be a continuous curve in the complex plane having one extremum at the origin and containing other points different from* 0*. Then, given two series $f(s) = \sum_{h=0}^{\infty} f_h\, s^h$ and $g(s) = \sum_{h=0}^{\infty} g_h\, s^h$, whose convergence disks include l, we have that $f(s) = g(s)$ for each point of l if and only if $g_h = f_h$ for each $h \geq 0$.*

It is obvious that $g_h = f_h$ implies that $f(s) = g(s)$ for each point of l. The fundamental point of the theorem is that, if $f(s) = g(s)$ on the curve, then f and g have the same coefficients, hence the same convergence disk, hence they are the same function.

By means of power series, we can define functions of matrices. Let $f(s)$ be an analytic function and M a square matrix. Assume that the set of the eigenvalues of M is contained in the convergence disk of f.[4] Then we can define the function of M associated with $f(s)$ as

$$f(M) \doteq \sum_{h=0}^{\infty} f_h\, M^h.$$

Note that $M^0 = I$ as a postulate. As an explanation, consider the case of M diagonalisable so that

$$M^h = T\Lambda^h T^{-1},$$

with $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ diagonal. Then

$$f(M) = \sum_{h=0}^{\infty} f_h\, T\Lambda^h T^{-1} = T\left[\sum_{h=0}^{\infty} f_h \Lambda^h\right] T^{-1}$$

where

$$\left[\sum_{h=0}^{\infty} f_h \Lambda^h\right] = diag\left\{\sum_{h=0}^{\infty} f_h \lambda_1^h, \sum_{h=0}^{\infty} f_h \lambda_2^h, \ldots, \sum_{h=0}^{\infty} f_h \lambda_n^h\right\} = diag\{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)\}.$$

Then it becomes apparent why we need to assume that the eigenvalues are in the convergence disk.

Example A.17. *We can define*

$$e^M \doteq I + \frac{M}{1!} + \frac{M^2}{2!} + \frac{M^3}{3!} + \frac{M^4}{4!} + \ldots$$

$$\cos(M) \doteq I - \frac{M^2}{2!} + \frac{M^4}{4!} + \ldots$$

and

$$\sin(M) \doteq \frac{M}{1!} - \frac{M^3}{3!} + \frac{M^5}{5!} + \ldots$$

Polynomials are particular power series. Then we can define the polynomial of a matrix. Given $p(s) = p_0 + p_1 s + p_2 s^2 + \cdots + p_n s^n$ we just define

$$p(M) = p_0 I + p_1 M + p_2 M^2 + \ldots p_n M^n.$$

The following property is important

Proposition A.3. ***(Cayley-Hamilton identity.)*** *Let A be a square matrix and*

$$\det(sI - A) = p(s) = a_0 + a_1 s + a_2 s^2 + \cdots + s^n$$

its characteristic polynomial. Then, the corresponding matrix polynomial calculated in A is zero:

$$p(A) = a_0 I + a_1 A + a_2 A^2 + \cdots + A^n = 0.$$

[4] This hypothesis can be removed thanks to the analytic extension technique.

The proof is simple in the case of A diagonalisable, $A = T\Lambda T^{-1}$, because, as we have seen,

$$P(A) = T diag\{p(\lambda_1), p(\lambda_2), \dots, p(\lambda_n)\}T^{-1} = 0$$

since $p(\lambda_i) = 0$ if λ_i is an eigenvalue.

Cayley-Hamilton identity implies the following fact.

Proposition A.4. ***(Dependence of the powers of*** **A.)** *Each power A^k of A, with $k \geq 0$, is a linear combination of the first n powers*

$$I \ , \ A \ , \ A^2 \ , \dots , \ A^{n-1}.$$

This property is trivial for the first powers $0, 1, \dots, n-1$. To prove it for the following powers, just write Cayley-Hamilton identity as

$$A^n = -a_{n-1}A^{n-1} - \dots - a_2A^2 - a_1A - a_0I.$$

This means that A^n is a linear combination of the first n powers. By induction, we assume that this is true for a given $k \geq n$

$$A^k = -\alpha_{n-1}A^{n-1} - \dots - \alpha_2A^2 - \alpha_1A - \alpha_0I,$$

and then we multiply by A getting

$$\begin{aligned} A^{k+1} &= -\alpha_{n-1}A^n - \dots - \alpha_2A^3 - \alpha_1A^2 - \alpha_0A = \\ &= -\alpha_{n-1}[-\alpha_{n-1}A^{n-1} - \dots - \alpha_2A^2 - \alpha_1A - \alpha_0I] \\ & \quad - \dots - \alpha_2A^3 - \alpha_1A^2 - \alpha_0A = \\ &= -\alpha^*_{n-1}A^{n-1} - \dots - \alpha^*_2A^2 - \alpha^*_1A^1 - \alpha^*_0I \end{aligned}$$

(where α^*_j are easily computable), hence the property is true for $k+1$.

Remark A.5. *If A is invertible, the property is also true for negative powers, A^k with $k < 0$.*

A.4 The impulse response

The concept of impulse is easy to grasp with intuition, but it does not have a simple mathematical description. Intuitively, an impulse is a phenomenon with high intensity and a very short duration. To represent it mathematically, we can consider a function $\delta_\epsilon(t)$ defined as follows

$$\delta_\epsilon(t) = \begin{cases} 0 & \text{if} \quad t < -\frac{\epsilon}{2} \\ \frac{1}{\epsilon} & \text{if} \quad -\frac{\epsilon}{2} \leq t \leq \frac{\epsilon}{2} \\ 0 & \text{if} \quad t > \frac{\epsilon}{2} \end{cases}$$

For any ϵ, the support of this function (namely, the interval where the function is non-zero), is precisely $[-\epsilon/2, \epsilon/2]$ and becomes increasingly small for $\epsilon \to 0$, while the amplitude of the function in this interval, $1/\epsilon$, becomes increasingly large. Note that the integral of this function on the real axis is always equal to 1. Let t_0 be a point inside the interval $[a, b]$ and let ϵ be a value such that the support of function $\delta_\epsilon(t - t_0)$ is included in $[a, b]$ ($t_0 - \epsilon/2$ and $t_0 + \epsilon/2$ must both belong to $[a, b]$).

Intuitively, the "impulse" in t_0, $\delta(t - t_0)$, is the "limit" for $\epsilon \to 0$ of the function $\delta_\epsilon(t - t_0)$. Note that $\delta(t - t_0)$ is not a function, but a **distribution**. A formal treatment is omitted for simplicity: the following is an intuitive explanation. Given f, a continuous function, consider the integral

$$\begin{aligned} \int_a^b f(t)\, \delta_\epsilon(t - t_0)dt &= \int_{t_0-\frac{\epsilon}{2}}^{t_0+\frac{\epsilon}{2}} f(t)\, \delta_\epsilon(t - t_0)dt = \\ &= \int_{t_0-\frac{\epsilon}{2}}^{t_0+\frac{\epsilon}{2}} f(t)\, \frac{1}{\epsilon}dt = \frac{1}{\epsilon}f(\tau)\epsilon = f(\tau), \end{aligned}$$

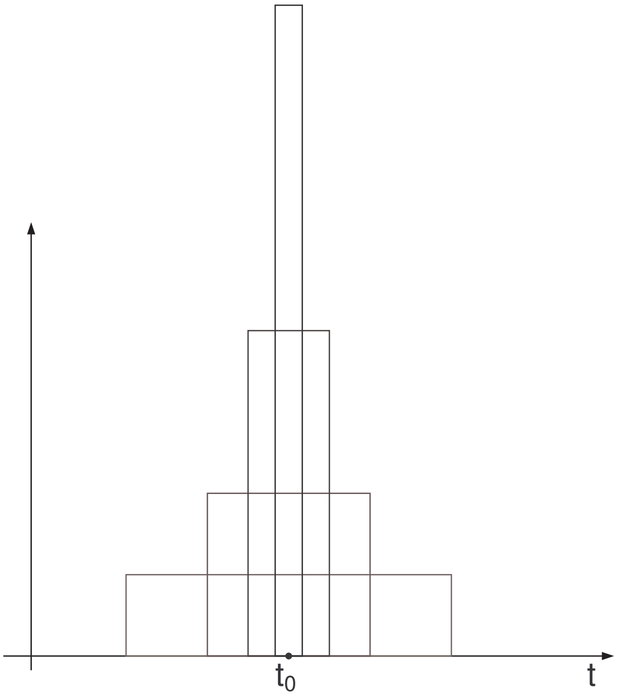

Figure A.4: The impulse can be seen as the limit of a sequence of functions.

where τ is an internal point of the interval $[t_0 - \frac{\epsilon}{2}, t_0 + \frac{\epsilon}{2}]$, which exists for the mean value theorem. Being f continuous, when $\epsilon \to 0$ we have

$$\int_a^b f(t)\,\delta_\epsilon(t - t_0)dt \to f(t_0).$$

Then the limit function $\delta(t - t_0)$ is defined as the "function" (in fact, it is a distribution) having the property that, for every continuous function f, defined on $[a, b]$ having t_0 as an interior point, we have that

$$\int_a^b f(t)\,\delta(t - t_0)dt = f(t_0). \tag{A.5}$$

We can apply this concept to the case of the forced response of a single-input-single-output linear system, given by the convolution

$$y(t) = \int_0^\top W(t - \sigma)u(\sigma)d\sigma.$$

If we take as an input the impulse function $\delta(\sigma - t_0)$ for $\sigma \geq 0$ we have that

$$y_{imp}(t) = \int_0^\top W(t - \sigma)\delta(\sigma - t_0)d\sigma = W(t - t_0),$$

for $t > t_0$ (while $y_{imp}(t) = 0$ for $t < t_0$). Finally, note that, in order to consider the solution due to an impulse at instant 0, we must consider the convolution from 0^- to include the origin in the interior:

$$y(t) = \int_{0^-}^\top W(t - \sigma)u(\sigma)d\sigma.$$

A.5 Laplace transforms

The Laplace transform is of great utility for the study of linear and time-invariant systems. We summarise here its fundamental properties.

Given a function $f(t)$ with **positive support** (namely, which is zero for $t < 0$), we define the *Laplace transform* (if it exists) as the function

$$F(s) = \mathcal{L}[f(t)] = \int_0^\infty f(t)e^{-st}dt, \tag{A.6}$$

where s is a complex variable. If the function F is well defined, we say that f is Laplace-transformable. The Laplace transform gives a one-to-one correspondence: $F(s) = G(s)$ if and only if $f(t) = g(t)$.

With a certain abuse of notation (but with great convenience of representation), it is often accepted to represent the function and its Laplace transform with the same letter

$$f(s) = \int_0^\infty f(t)e^{-st}dt.$$

The expression (A.6) is defined in a suitable domain of s, of the type

$$Dom_f = \{s : \ \Re(s) > c\}.$$

In particular, the smallest value c_f of c for which convergence is ensured, namely for which

$$f(s) = \int_0^\infty |f(t)|e^{-st}dt$$

converges, is called *abscissa of convergence* (relatively to f). Then $Dom_f = \{s : \ \Re(s) > c_f\}$ is a suitable domain.

We assume now that the functions f and g have positive support and are continuous and differentiable for $t > 0$. We consider the common domain of convergence $Dom = Dom_f \bigcap Dom_g$. We denote by $f(s) = \mathcal{L}[f(t)]$ and $g(s) = \mathcal{L}[g(t)]$, respectively. The following fundamental properties holds.

Linearity

$$\mathcal{L}[\alpha f(t) + \beta g(t)] = \alpha f(s) + \beta g(s).$$

Transform of the derivative

$$\mathcal{L}[f'(t)] = sf(s) - f(0).$$

Final limit theorem Assuming that the limits below *exist and are both finite* we have that

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sf(s).$$

Initial limit theorem Assuming that the limits below *exist and are both finite* we have that

$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} sf(s).$$

Convolution theorem

$$\mathcal{L}\left[\int_0^\top g(t-\sigma)f(\sigma)d\sigma\right] = \mathcal{L}\left[\int_0^\top g(\sigma)f(t-\sigma)d\sigma\right] = f(s)g(s).$$

Shifting theorem

$$\mathcal{L}[f(t-\tau)] = e^{-s\tau}F(s).$$

Multiplication by t

$$\mathcal{L}[tf(t)] = -\frac{d}{ds}F(s).$$

Multiplication by $e^{\xi t}$

$$\mathcal{L}[e^{\xi t}f(t)] = F(s-\xi).$$

A.5.1 Transforms of rational functions

In this subsection we consider functions of the type $t^k e^{\xi t}\cos(\omega t)$ and $t^k e^{\xi t}\sin(\omega t)$ and all special cases ($\xi = 0$, $\omega = 0$, $k = 0$). Let us start from the simplest case

$$\mathcal{L}\left[e^{\lambda t}\right] = \frac{1}{s-\lambda}.$$

The property above applies in the case of real and complex λ.

Consider the transform of $\mathcal{L}[\cos(\omega t)]$ and $\mathcal{L}[\sin(\omega t)]$. The simplest way to compute it (simultaneously) is to consider the function $e^{j\omega t} = \cos(\omega t) + j\sin(\omega t)$. By linearity, real and imaginary parts will provide us the transforms of the functions $\cos(\omega t)$ and $\sin(\omega t)$. We get

$$\begin{aligned}
\mathcal{L}\left[e^{j\omega t}\right] &= \frac{1}{s-j\omega} = \frac{s+j\omega}{s+j\omega}\frac{1}{s-j\omega} = \\
&= \frac{s+j\omega}{s^2+\omega^2} = \underbrace{\frac{s}{s^2+\omega^2}}_{\mathcal{L}[\cos(\omega t)]} + j\underbrace{\frac{\omega}{s^2+\omega^2}}_{\mathcal{L}[\cos(\omega t)]}
\end{aligned}$$

Then

$$\begin{aligned}
\mathcal{L}[\cos(\omega t)] &= \frac{s}{s^2+\omega^2} \\
\mathcal{L}[\sin(\omega t)] &= \frac{\omega}{s^2+\omega^2}.
\end{aligned}$$

Using the same fact and setting $\lambda = \xi + j\omega$

$$e^{\lambda t} = e^{(\xi+j\omega)t} = e^{\xi t}(\cos(\omega t) + j\sin(\omega t))$$

since

$$\mathcal{L}\left[e^{(\xi+j\omega)t}\right] = \frac{1}{s-\xi+j\omega}$$

we get, taking real and imaginary parts separately,

$$\begin{aligned}
\mathcal{L}[e^{\xi t}\cos(\omega t)] &= \frac{s-\xi}{(s-\xi)^2+\omega^2} \\
\mathcal{L}[e^{\xi t}\sin(\omega t)] &= \frac{\omega}{(s-\xi)^2+\omega^2}.
\end{aligned}$$

Now consider the Laplace transform of the function $te^{\lambda t}$. This is obtained by considering the theorem of multiplication by t:

$$\mathcal{L}[te^{\lambda t}] = -\frac{d}{ds}\frac{1}{s-\lambda} = \frac{1}{(s-\lambda)^2}.$$

From the trivial fact that $t^k e^{\lambda t} = t(t^{k-1}e^{\lambda t})$ we can recursively demonstrate that

$$\mathcal{L}[t^k e^{\lambda t}] = \frac{k!}{(s-\lambda)^{k+1}}.$$

The transform of the impulse $\delta(t-\tau)$ is important:

$$\mathcal{L}[\delta(t-\tau)] = \int_0^\infty \delta(t-\tau)e^{-st}dt = e^{-s\tau}$$

In particular for $\tau = 0$ this means that the impulse function in the origin is just

$$\mathcal{L}[\delta(t)] = 1.$$

The Heaviside step function in τ is

$$gr(t-\tau) = \begin{cases} 0 & \text{for} \quad t \le \tau \\ 1 & \text{for} \quad t > \tau \end{cases}$$

and its transform can be obtained as follows:

$$\begin{aligned} \mathcal{L}[gr(t-\tau)] &= \int_{t=0}^{t=\infty} gr(t-\tau)e^{-st}dt = \int_{\sigma=-\tau}^{\sigma=\infty} gr(\sigma)e^{-s(\sigma+\tau)}d\sigma = \\ &= e^{-s\tau}\int_{\sigma=0}^{\sigma=\infty} e^{-s\sigma}d\sigma = \frac{e^{-s\tau}}{s}. \end{aligned}$$

In particular, the transform of the step in $t = 0$ is

$$\mathcal{L}[gr(t)] = \frac{1}{s}.$$

Finally note that the forced response of a linear system has transform

$$\mathcal{L}\left[\int_0^t W(t-\sigma)u(\sigma)d\sigma\right] = W(s)u(s).$$

The transforms which have been considered so far are rational. There is an interesting case of **non-rational transform**. Consider the T-delay operator

$$y(t) = DL_T(u) \doteq u(t_T)$$

This operator is linear. Note that if u has positive support, then $u(t_T) = 0$ as $t < T$. In terms of transform we get

$$y(s) = e^{-Ts}u(s)$$

The proof is simple and left to the reader.

A.5.2 Rational functions and their inverse transform

We have seen in the previous section some of the main transforms. In this section, we deal with the inverse transform problem. In particular, we consider the inverse transform of proper rational functions[5] of the type

$$f(s) = \frac{n_0 + n_1 s + n_2 s^2 + \cdots + n_\nu s^\nu}{d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu}.$$

[5] A rational function is proper if the degree of the denominator is greater than or equal to the degree of the numerator.

Note that the coefficient of s^ν at the denominator is $d_\nu = 1$: this is not restrictive because, if $d_\nu \neq 1$, we can always divide both numerator and denominator by d_ν (which is non-zero because we consider proper functions). A proper rational function $f(s)$ can be always be transformed into the sum of a constant and a function strictly proper rational function[6]

$$\begin{aligned} f(s) &= \frac{n_0 + n_1 s + n_2 s^2 + \cdots + n_\nu s^\nu}{d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu} = \frac{n_\nu(d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu)}{d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu} + \\ &+ \frac{(n_0 - n_\nu d_0) + (n_1 - n_\nu d_1)s + \cdots + (n_{\nu-1} - n_\nu d_{\nu-1})s^{\nu-1}}{d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu} \\ &= n_\nu + \frac{\tilde{n}_0 + \tilde{n}_1 s + \tilde{n}_2 s^2 + \cdots + \tilde{n}_{\nu-1} s^{\nu-1}}{d_0 + d_1 s + d_2 s^2 + \cdots + s^\nu} = n_\nu + \tilde{f}(s) \end{aligned}$$

where $\tilde{f}(s)$ is strictly proper. Note that

$$\mathcal{L}^{-1}[f(s)] = n_\nu \delta(t) + \mathcal{L}^{-1}[\tilde{f}(s)],$$

Then we just need to determine the inverse transform of the strictly proper part.

Consider the strictly proper rational function $f(s)$. If f has distinct poles (roots of the denominator) $\lambda_1, \lambda_2, \ldots, \lambda_\nu$, it can always be written in the form

$$f(s) = \sum_{i=1}^{\nu} \frac{r_i}{s - \lambda_i}, \tag{A.7}$$

where the coefficients r_i are called *residuals* and are computable as the limit

$$r_i = \lim_{s \to \lambda_i} (s - \lambda_i) f(s),$$

as we can see from (A.7). If we factorise the denominator writing f as

$$f(s) = \frac{n(s)}{(s - \lambda_1)(s - \lambda_2) \ldots (s - \lambda_\nu)}$$

the formula reduces to

$$r_i = \left. \frac{n(s)}{\prod_{j \neq i}(s - \lambda_j)} \right|_{s = \lambda_i}.$$

As an alternative, we can solve the problem by means of a system of linear equations. Take the expression (A.7) and consider the common denominator

$$f(s) = \frac{n(s)}{d(s)} = \frac{\sum_{i=1}^{\nu} r_i \prod_{j \neq i}(s - \lambda_j)}{(s - \lambda_1)(s - \lambda_2) \ldots (s - \lambda_\nu)} = \frac{\sum_{i=1}^{\nu} r_i \Psi_i(s)}{d(s)}$$

The identity

$$n(s) = \sum_{i=1}^{\nu} r_i \Psi_i(s),$$

must hold. Both sides are polynomials of degree ν in s. By applying the identity principle for polynomials, the coefficients r_i are determined by a linear system (obtained by equating the coefficients of the terms with the same degree).

[6] A rational function is strictly proper if the degree of the numerator is smaller than that of the denominator.

Example A.18. *Consider the function*

$$f(s) = \frac{s+4}{(s+1)(s+2)(s+3)} = \frac{r_1}{s+1} + \frac{r_2}{s+2} + \frac{r_3}{s+3}$$

We have that

$$\begin{aligned} r_1 &= \left.\frac{s+4}{(s+2)(s+3)}\right|_{s=-1} = \frac{3}{2} \\ r_2 &= \left.\frac{s+4}{(s+1)(s+3)}\right|_{s=-2} = -2 \\ r_3 &= \left.\frac{s+4}{(s+1)(s+2)}\right|_{s=-2} = \frac{1}{2} \end{aligned}$$

If we proceed with the linear system method, we have that

$$\begin{aligned} (s+4) &= r_1(s+2)(s+3) + r_2(s+1)(s+3) + r_3(s+1)(s+2) \\ &= r_1(s^2+5s+6) + r_2(s^4+4s+3) + r_3(s^2+3s+2) \\ &= (r_1+r_2+r_3)s^2 + (5_r1 + 4r_2 + 3r_3)s + 6r_1 + 3r_2 + 2r_3 \end{aligned}$$

from which we derive the linear system of equations

$$\begin{aligned} r_1 + r_2 + r_3 &= 0 \\ 5_r1 + 4r_2 + 3r_3 &= 1 \\ 6r_1 + 3r_2 + 2r_3 &= 4 \end{aligned}$$

which (as expected) provides the same values for r_1, r_2, r_3.

This procedure can work but "gives troubles" if some poles λ_i are complex. If so, then we can decide to work in a real setting. We factorise the denominator as follows

$$\begin{aligned} d(s) &= \prod_{i=1}^{m}(s-\lambda_i)\prod_{i=1}^{q}(s^2 - 2\xi_i s + \xi_i^2 + \omega_i^2) \\ &= \prod_{i=1}^{m}(s-\lambda_i)\prod_{i=1}^{q}((s-\xi_i)^2 + \omega_i^2) \end{aligned}$$

where the factors $(s-\lambda_i)$ are associated with the m real roots, while the factors $((s-\xi_i)^2+\omega_i^2)$ are associated with the q pairs of complex conjugate roots $\xi_i + j\omega_i$ ($m + 2q = \nu$). Then we have that

$$f(s) = \sum_{i=1}^{m}\frac{r_i}{(s-\lambda_i)} + \sum_{i=1}^{q}\frac{a_i s + b_i}{(s-\xi_i)^2+\omega_i^2}.$$

The coefficients r_i can be determined by means of the limit formula. To compute all of the coefficients r_i, a_i and b_i, we can write the common denominator and calculate the coefficients through a linear system.

Example A.19. *Consider the rational function*

$$f(s) = \frac{2s^3+1}{(s+1)(s+2)((s+1)^2+4)}.$$

This function can be written as

$$f(s) = \frac{n(s)}{d(s)} = \frac{r_1}{s+1} + \frac{r_2}{s+2} + \frac{a_1 s + b_1}{(s^2+2s+5)}.$$

After adopting the common denominator, equating the numerators provides

$$2s^3 + 1 = r_1(s+2)(s^2+2s+5) + r_2(s+1)(s^2+2s+5) + (a_1 s + b_1)(s+1)(s+2)$$

Through the principle of identity, we can determine the coefficients r_1, r_2, a_1 and b_1 of the polynomial by solving the linear system

$$\begin{aligned} r_1 + r_2 + a_1 &= 2 \\ 4r_1 + 3r_2 + 3a_1 + b_1 &= 0 \\ 9r_1 + 7r_2 + 2a_1 + 3b_1 &= 0 \\ 10r_1 + 5r_2 + 2b_1 &= 1 \end{aligned}$$

The solution is

$$\begin{aligned} r_1 &= -\frac{1}{4} \\ r_2 &= 3 \\ a_1 &= -\frac{3}{4} \\ b_1 &= -\frac{23}{4}. \end{aligned}$$

Let us briefly consider now the case of rational functions with multiple poles. The most general case of decomposition is

$$\sum_{i=1}^{m} \sum_{j=1}^{g_i} \frac{r_{ij}}{(s-\lambda_i)^j},$$

where g_i is the multiplicity of the poles λ_i. The r_{ij}'s are obtainable through a system as in the following example.

Example A.20.

$$\begin{aligned} f(s) &= \frac{s^2+s+1}{(s+1)^2(s+2)^3} && \text{(A.8)} \\ &= \frac{r_{11}}{s+1} + \frac{r_{12}}{(s+1)^2} + \frac{r_{21}}{s+2} + \frac{r_{22}}{(s+2)^2} + \frac{r_{23}}{(s+2)^3} && \text{(A.9)} \end{aligned}$$

With the usual technique of the common denominator, we can equate the numerators and get

$$\begin{aligned} s^2+s+1 &= r_{11}(s+1)(s+2)^3 + r_{12}(s+2)^3 + r_{21}(s+1)^2(s+2)^2 + \\ &+ r_{22}(s+1)^2(s+2) + r_{23}(s+1)^2 \end{aligned}$$

from which we can derive the coefficients r_{11}, r_{12}, r_{21}, r_{22} and r_{23}. Equating the coefficients of the corresponding terms, we get the linear system

$$\begin{aligned} r_{11} + r_{21} &= 0 \\ 7r_{11} + r_{12} + 6r_{21} + r_{22} &= 0 \\ 18r_{11} + 6r_{12} + 13r_{21} + 4r_{22} + r_{23} &= 1 \\ 20r_{11} + 12r_{12} + 12r_{21} + 5r_{22} + 2r_{23} &= 1 \\ 8r_{11} + 8r_{12} + 4r_{21} + 2r_{22} + r_{23} &= 1 \end{aligned}$$

The solution is

$$\begin{array}{rcl} r_{11} & = & -4 \\ r_{12} & = & 1 \\ r_{21} & = & 4 \\ r_{22} & = & 3 \\ r_{23} & = & 3. \end{array}$$

Once we have calculated the decomposition, the problem of the inverse transform can be solved term by term (in view of linearity). For the terms whose denominator is of the first degree

$$\mathcal{L}^{-1}\left[\frac{r}{s-\lambda}\right] = r\, e^{\lambda t},$$

while for terms with a second degree denominator we can write

$$\frac{as+b}{(s-\xi)^2+\omega^2} = \frac{a(s-\xi)}{(s-\xi)^2+\omega^2} + \frac{(b+a\xi)}{(s-\xi)^2+\omega^2}$$

whereby

$$\mathcal{L}^{-1}\left[\frac{as+b}{(s-\xi)^2+\omega^2}\right] = a\, e^{\xi t}\cos(\omega t) + \frac{b+a\xi}{\omega}\, e^{\xi t}\sin(\omega t).$$

In the case of multiple eigenvalues, we have terms of degree $k+1$. For these we have

$$\mathcal{L}^{-1}\left[\frac{r}{(s-\lambda)^{k+1}}\right] = \frac{r}{k!}\, t^k e^{\lambda t}$$

where λ is either real or complex.

A.6 Zeta transform

The Zeta transform is the counterpart of the Laplace transform for the analysis of discrete-time systems. Given the close analogy, we give only a brief sketch. Given the sequence (or function in discrete-time) $f(k)$, defined for $k \geq 0$, the Z-transform of f is the function of the complex variable z defined as follows

$$Z[f] \doteq \sum_{k=0}^{\infty} f(k)\frac{1}{z^k} = f(z),$$

namely, it is the corresponding series of powers of $\frac{1}{z}$. The convergence domain of this function is a set of the type

$$\mathcal{D} = \{z : \ |z| > \rho_f\}$$

(which is the complement of the disk of radius ρ_f), where the radius ρ_f depends on the considered function f.

The Zeta transform has properties that are quite similar to those of the Laplace transform. We remember only the following properties.

Linearity $Z[\alpha f + \beta g] = \alpha Z[f] + \beta Z[g]$.

Transform of the anticipated f $Z[f(k+1)] = zZ[f] - zf(0)$.

Convolution transform

$$Z[\sum_{h=0}^{k-1} f(k-h)g(k)] = Z[f]Z[g].$$

The utility of the Zeta transform in the study of discrete-time systems is entirely comparable to that of the Laplace transform in the study of continuous-time systems.

Example A.21. (Exercise.) *The sequence λ^k has Zeta transform*

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{z^k} = \frac{z}{z-\lambda}.$$

To find the transforms of $\cos(\omega k)$ *and* $\sin(\omega k)$*, just consider that*

$$e^{j\omega k} = \cos(\omega k) + j\sin(\omega k) = [e^{j\omega}]^k$$

has transform

$$\frac{z}{z-e^{j\omega}} = \frac{z}{z-(\cos(\omega)+j\sin(\omega))} = \dots$$

and then split real and imaginary parts . . .

A.7 Convex and least square optimization

Consider the constrained optimization problem

$$\min l(\theta) \tag{A.10}$$

$$s.t.$$

$$\theta \in \Theta \tag{A.11}$$

where $l(\theta)$ is an objective function of a vector $\theta \in \mathcal{R}^n$ defined on a domain $\theta \in \Theta$. The set Θ is in general represented by equalities and inequalities.

Finding the solution of this type of problems is important in many contexts including control theory. The solution is not in general easy to find. The main problem it that (A.12) (A.11) my have many local minima so we need to find all of these and take the smallest. A remarkable exception is the one in which the problem is convex.

A set Θ is convex if assuming that θ_a and θ_b belong to Θ, then all their convex combinations belong to Θ

$$\theta = \alpha\theta_a + \beta\theta_b \in \Theta, \quad \forall \alpha+\beta = 1, \ \ \alpha,\beta \geq 0$$

A functional $l(\theta)$ defined on a convex set Θ is convex if

$$l(\alpha\theta_a + \beta\theta_b) \leq \alpha l(\theta_a) + \beta l(\theta_b), \quad \forall \alpha+\beta = 1, \ \ \alpha,\beta \geq 0$$

If Θ and l are convex the optimization problem is said convex.

Proposition A.5. *In a convex optimization problem, any local minimum is a global minimum.*

The previous result, is very important and its consequence is that a convex optimization problem is easier than the general optimization problems.

Example A.22. *The following problem*

$$\begin{aligned} & \min h^\top \theta \\ s.t. \quad & M\theta = r \\ & N\theta \leq s \end{aligned}$$

is a linear programming problem and can be solved for systems of dimension 1000 *or more with no difficulties. The following problem*

$$\begin{aligned} & \min \theta^\top P\theta \\ s.t. \quad & M\theta = r \\ & N\theta \leq s \end{aligned}$$

with P semi–positive definite, is a quadratic programming problem and can be also efficiently solved.

Special cases of quadratic problems are the minimum euclidean norm problems with linear constraints. Consider the linear systems

$$Ax = b$$

There are three cases.

1. Matrix A is square invertible, then the solution is unique.
2. There are infinite solutions. In this case one wish to find the one with smallest Euclidean norm.
3. There are no solutions. In this case one wish to find the vector for which the error $e = Ax - b$ has the smallest Euclidean norm.

The previous are called generalized solutions.

For brevity assume A full rank. The first solution can be found as

$$\begin{aligned} & \min \frac{1}{2}||x||^2 \\ s.t. \quad & Ax = b \end{aligned}$$

The Lagrangian is

$$\frac{1}{2}||x||^2 + \lambda^\top [Ax - b]$$

Differentiating with respect to x we get

$$x^\top + \lambda^\top A = 0$$

namely $x = -A^\top \lambda$. We replace this in $Ax = b$ to get $-AA^\top \lambda = b$ or $\lambda = -[AA^\top]^{-1}b$. Then the minimum norm solution is

$$x^* = -A^\top [AA^\top]^{-1} b$$

In the second case we wish to minimize

$$\min ||Ax - b||^2$$

Write the function as

$$f(x) = [Ax - b]^\top [Ax - b] = x^\top A^\top Ax - b^\top Ax - x^\top A^\top b + b^\top b = x^\top A^\top Ax - 2b^\top Ax + b^\top b$$

Differentiating with respect to x we get

$$2x^\top A^\top A - 2b^\top A = 0$$

transposing $A^\top Ax - A^\top b = 0$, then

$$\hat{x} = [A^\top A]^{-1} A^\top b$$

A.7.1 Application to system identification

Assume that a frequency test is performed on a system with transfer function F with inputs

$$\cos(\omega_k t), \quad k = 1, 2, \ldots, m$$

measuring the steady state output

$$|F(j\omega_k)| \cos(\omega_k t + \phi_k)$$

From magnitude and phase, experimentally determined, we can determine

$$F(j\omega_k) = \alpha_k + j\beta_k$$

If $F(s)$ has to be determined, we parameterize it with its coefficient. As an example consider

$$F(s) = \frac{cs + d}{s^2 + as + b}$$

where a, b, c, d have to be determined. We can write at each frequency experiment

$$\frac{cj\omega + d}{-\omega^2 + aj\omega + b} = \alpha + j\beta$$

namely

$$[\alpha + j\beta][-\omega^2 + aj\omega + b] = cj\omega + d$$

We can split real an imaginary part

$$-\alpha\omega^2 + \alpha b - a\beta\omega = d \quad \text{and} \quad -\beta\omega^2 + \omega\alpha a + \beta b = \omega c$$

So we get two equations for the unknown coefficients

$$\begin{bmatrix} -\beta\omega & \alpha & 0 & -1 \\ \omega\alpha & \beta & -\omega & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} \alpha\omega^2 \\ \beta\omega^2 \end{bmatrix}$$

Repeating the measure m times we get

$$\begin{bmatrix} -\beta_1\omega_1 & \alpha_1 & 0 & -1 \\ \omega_1\alpha_1 & \beta_1 & -\omega_1 & 0 \\ -\beta_2\omega_2 & \alpha_2 & 0 & -1 \\ \omega_2\alpha_2 & \beta_2 & -\omega_2 & 0 \\ -\beta_3\omega_3 & \alpha_3 & 0 & -1 \\ \omega_3\alpha_3 & \beta_3 & -\omega_3 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\beta_m\omega_m & \alpha_m & 0 & -1 \\ \omega_m\alpha_m & \beta_m & -\omega_m & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \neq \begin{bmatrix} \alpha_1\omega_1^2 \\ \beta_1\omega_1^2 \\ \alpha_2\omega_2^2 \\ \beta_2\omega_2^2 \\ \alpha_3\omega_3^2 \\ \beta_3\omega_3^2 \\ \vdots \\ \alpha_m\omega_m^2 \\ \beta_m\omega_m^2 \end{bmatrix}$$

This linear system is of the form

$$\Phi\theta = \Gamma$$

and it has no solution in general. So it can be solved by minimizing the error norm $||\Phi\theta - \Gamma||$.

A different technique for identification is based on time–domain measures. A discrete time input $u(k)$ is applied to the systems and the output $y(k)$ is measured. Consider a model of the form

$$y(k) = \sum_{h=1}^{n} a_h y(k-h) + \sum_{h=1}^{n-1} b_h u(k-h), \quad k = 2, 3, 4 \ldots$$

This model is called ARMA model (AutoRegressive-Moving-Average). The identification of its coefficients can be performed via the least square method presented above.