B15 Linear Dynamic Systems and Optimal Control

Michaelmas Term 2024
University of Oxford

Kostas Margellos
kostas.margellos@eng.ox.ac.uk

Syllabus

Transforming continuous time linear systems (linear differential equations) to state space form. Existence and uniqueness of solutions to linear systems. Time domain solution of state space equations and connections with stability. Connections between state space and transfer functions. Controllability and observability properties and their interpretation. Kalman decomposition. Minimum energy control. State feedback control design via pole placement. State observer design. Separation principle. Output feedback control design. Linear Quadratic Regulator (LQR) and its interpretation. Riccati equations.

Learning outcomes:

- Familiarization with state space representation as a modelling formalism of differential equations.
- Linearization of nonlinear differential equations.
- Derivation of the zero input and zero state solution of linear systems; computation of the state transition matrix.
- Understand the implications of the solution form to the stability of the system.
- Understand connections between state space equations and transfer functions.

- Analyze structural linear system properties: controllability and observability.
- Derivation of minimum energy control laws.
- Design state feedback controllers via pole placement.
- Design linear state observers.
- Understand the separation principle to perform output feedback control.
- Linear Quadratic Regulator (LQR) control design.

Lecture notes

These lecture notes are made available on Canvas.

Any comments or corrections shall be sent to `kostas.margellos@eng.ox.ac.uk`

Recommended text

- F Callier & C Desoer *Linear System Theory* Springer Science, 1991.
- J Lygeros & F Ramponi *Lecture Notes on Linear System Theory* ETH Zurich, 2013.
- K Astrom & R Murray *Feedback Systems: An Introduction for Scientists and Engineers* Princeton U.P., 2008.

Other reading

- J Hespanha *Linear Systems* Princeton U.P., 2009.
- D Liberzon *Calculus of Variations and Optimal Control Theory: A Concise Introduction* Princeton U.P., 2012.

The course follows the recommended texts which, however, provide a more general treatment of the theory targeting at a graduate level. It should be acknowledged that the exposition and worked out examples in these notes have been inspired significantly, and follow in parts the handouts of the course "Signal and System Theory II", taught by Prof. John Lygeros at ETH Zurich. Special thanks to Licio Romao for proof-reading an earlier version of these notes and for providing several constructive comments.

Contents

1 Introduction to linear systems and state space form

Optimal control involves regulating and improving – as far as a particular performance criterion is concerned – the behaviour of a given plant. To this end, we typically exploit information from sensors (measurements) to design intelligent actuator commands. To achieve this a model of the underlying plant needs to be developed, which is then employed for control design purposes. Such a model captures the physics governing the behaviour of the plant, e.g., equations of motion for ground or aerial vehicles, Kirchhoff's current and voltage laws for electric circuits, etc. High fidelity models, which usually involve non-linear differential equations, capture accurately the physical description and evolution of the system, however, impose difficulties when it comes to designing controllers. We thus very often abstract the behaviour of such systems to simpler ones, facilitating the control design procedure. In these notes we will first consider *linear systems* in continuous time as a modelling abstraction of the actual plant dynamics, and then provide *optimal control* methodologies to regulate their behaviour.

In this realm, the main objectives of the notes are as follows.

1. *Linear systems (Chapters 1–4):* We aim at introducing linear systems and analyze them in a rigorous mathematical manner in the time domain. As such, our analysis complements traditional control design methodologies based on transfer functions, that are typically performed in the frequency domain. In particular, we will represent linear systems in the so called state-space form, which comprises a linear ordinary differential equation and an algebraic output equation, and we will answer existence and uniqueness questions for solutions to these equations. We will show how to construct solutions to a certain class of linear systems in closed form, and study their properties as far as stability is concerned. We will also analyze structural properties that pertain their behaviour, termed controllability and observability. Such properties allow us to decide whether it is possible to design controllers to steer the system trajectories to a desired location of the space, and whether it is possible to infer their initial condition by inspecting/measuring the output of the system.

2. *Optimal control (Chapters 5–8):* We will build on the linear systems analysis to

design controllers for such systems. We will introduce different design methodologies that are optimal with respect to a given criterion. In particular, we will discuss minimum energy controllers that are open loop, however, correspond to the minimum effort controller to steer the system to a given location. We will then move to feedback controllers, and provide a systematic way of designing the feedback control gains so that we achieve a prescribed closed loop performance. Finally, we will focus on designing feedback controllers while optimizing a certain criterion. We will pose this as a general optimal control problem and characterize its solution.

Beyond their importance in engineering and optimal control, linear systems have particularly appealing properties from a mathematical point of view. Understanding their behaviour requires a wide range of mathematical tools from linear algebra and analysis, however, they are simple enough to derive their properties in closed form, and as such serve as an introduction to mathematical proofs and formal logic. To render the notes self-contained, we provide a condensed summary of selected topics from linear algebra and analysis in *Chapter 9*.

Basic notation: We will be using $\mathbb{R}$ and $\mathbb{C}$ to denote the set of real and complex numbers, respectively. We denote by $\mathbb{R}^n$ the set of n-dimensional vectors, and by $\mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) the set of $m \times n$ matrices with real (complex) entries (for arbitrary m and n)*. We will use $t \in \mathbb{R}$ for the continuous time variable. By $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ we denote a function that takes as input a vector in $\mathbb{R}^n$ and returns a vector in $\mathbb{R}^m$.

1.1 Modelling examples

We first consider a couple of examples and detail their physical description. Our analysis will be entirely in continuous time.

1.1.1 Pendulum motion

Consider the pendulum illustrated in Figure 1, i.e., a mass m is hanging from a weightless spring with length l. Initially the pendulum creates an angle with the

*It should be noted that we will assume throughout that all linear spaces involved are finite dimensional; a more general treatment involving infinite dimensional spaces (spaces of functions) is outside the scope of these notes.

vertical axis and has some angular velocity. If it is released from its initial position this angle will change as a function of time; we will denote it by $\theta(t)$.

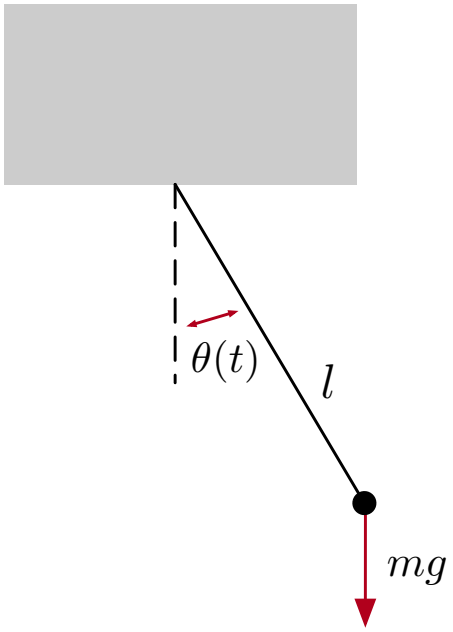


Figure 1: Pendulum with mass m.

We would like to describe the physics that govern the motion of the pendulum, or in other words the evolution of $\theta(t)$. To this end, note that the pendulum performs a rotational motion with $\dot{\theta}(t)$ being its angular, and $l\dot{\theta}(t)$ its linear velocity. The pendulum mass experiences its weight mg, as well as a friction force $dl\dot{\theta}(t)$ (proportional to the linear velocity) with direction opposing its motion, with d being the friction dissipation constant. By Newton's law of motion we have that

$$ml\ddot{\theta}(t) = -dl\dot{\theta}(t) - mg\sin\theta(t),$$

where $\ddot{\theta}(t)$ is the angular acceleration, and $mg\sin(\theta(t))$ denotes the gravity force component along the direction of motion. This is a second order ordinary differential equation (ODE) with respect to $\theta(t)$. We assume that we can only measure the angular position of the pendulum, hence have access to a measurement equation

$$y(t) = \theta(t).$$

For every time instance t the motion of the pendulum is captured by its angular position and velocity; we can thus set

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \theta(t) \\ \dot{\theta}(t) \end{bmatrix} \in \mathbb{R}^2,$$

and notice that $\dot{x}_1(t) = \dot{\theta}(t) = x_2(t)$ and $\dot{x}_2(t) = -\frac{d}{m}x_2(t) - \frac{g}{l}\sin x_1(t)$. Under these variable assignments, we are able to represent the second order ODE and the measurement equation more compactly as

$$\dot{x}(t) = \begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}x_2(t)\\ -\frac{d}{m}x_2(t) - \frac{g}{l}\sin x_1(t)\end{bmatrix},$$
$$y(t) = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} = \begin{bmatrix}1 & 0\end{bmatrix}x(t).$$

Notice that the first equation is now a first order ODE but for a "lifted" system that involves a two-dimensional vector $x(t)$, while the output equation is an algebraic equation that involves multiplying $x(t)$ with a matrix (row vector in this case) that "selects" the component that can be measured.

The ODE together with the algebraic equation capture the behaviour of the pendulum as they model the underlying physics (ODE) and encode the information that is available to us by means of measurements (algebraic equation). In the considered example this is a nonlinear system due to the presence of the trigonometric function, and it is also unforced/autonomous, as no external force is applied to the pendulum.

1.1.2 Electric circuit

Consider the electric circuit of Figure 2, that involves a resistance R in series with an inductor L and a capacitor C. An external voltage $u(t)$ is applied to the system. We would like to determine an ODE that captures the behaviour of the voltage $v_C(t)$ across the capacitor and the current $i_L(t)$ along the inductor. We assume that both $v_C(t)$ and $i_L(t)$ can be measured.

The voltage across the inductor is given by $v_L(t) = L\frac{di_L(t)}{dt}$, while the current across the inductor equals the current across the capacitor, leading to $i_L(t) = C\frac{dv_C(t)}{dt}$. Therefore, the voltage across the inductor is given by $v_L(t) = LC\frac{d^2v_C(t)}{dt^2}$. Denoting by $v_R(t) = Ri_L(t)$ the voltage across the resistance ($i_L(t)$ flows through R), by Kirchhoff's voltage law we have that

$$v_L(t) + v_R(t) + v_C(t) = u(t) \;\Rightarrow\; LC\frac{d^2v_C(t)}{dt^2} + RC\frac{dv_C(t)}{dt} + v_C(t) = u(t).$$

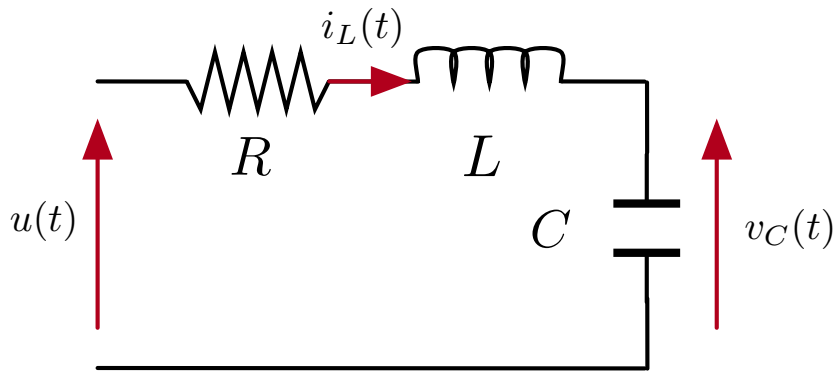


Figure 2: RLC circuit with external input voltage $u(t)$.

This is a second order ordinary differential equation (ODE) with respect to $v_C(t)$. We assume that we can measure both the voltage across C and the current across L, hence we have access to a measurement equation

$$y(t) = \begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix}.$$

The fact that we can measure both $v_C(t)$ and $i_L(t)$ – notice that both these quantities are related to the system's energy – suggests setting

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} \in \mathbb{R}^2.$$

Moreover, $v_L(t) = -v_R(t) - v_C(t) + u(t) \;\Rightarrow\; L\frac{di_L(t)}{dt} = -Ri_L(t) - v_C(t) + u(t)$. Using this fact, under our choice for $x(t)$ we have that

$$i_L(t) = C\frac{dv_C(t)}{dt} \;\Rightarrow\; \dot{x}_1(t) = \frac{1}{C}x_2(t),$$
$$L\frac{di_L(t)}{dt} = -Ri_L(t) - v_C(t) + u(t) \;\Rightarrow\; \dot{x}_2(t) = -\frac{1}{L}x_1(t) - \frac{R}{L}x_2(t) + \frac{1}{L}u(t).$$

Under these variable assignments, we are able to represent the second order ODE and the measurement equation as

$$\dot{x}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{C} \\ -\frac{1}{L} & -\frac{R}{L} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{1}{L} \end{bmatrix} u(t),$$
$$y(t) = \begin{bmatrix} v_C(t) \\ i_L(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \end{bmatrix} u(t).$$

Similarly to the pendulum example, we have transformed a second order ODE to a first order one but for a "lifted" system. Moreover, the resulting ODE and

the algebraic equation capture the behaviour of the capacitor's voltage and the inductor's current (ODE), and indicate the quantities we can measure (algebraic equation). However, in contrast to the pendulum's compact description, the one obtained here is linear with respect to $x(t)$ (and $u(t)$), and is not autonomous due to the presence of the external input $u(t)$.

1.2 State space representation

The two preceding examples illustrate that despite the differences between the underlying physical system (pendulum vs. electric circuit), we are able to capture the system's behaviour by means of a similar mathematical representation. To make this precise we denote by $u(t) \in \mathbb{R}^m$ the input and by $y(t) \in \mathbb{R}^p$ the output of a given system, while we refer to $x(t) \in \mathbb{R}^n$ as the state of the system.

We say that a *nonlinear, time-varying system* is in *state space* form if it can be represented by

$$\dot{x}(t) = f(x(t), u(t), t),$$
$$y(t) = h(x(t), u(t), t),$$

where f and h are nonlinear functions of $x(t)$, $u(t)$ and (possibly) the time t. We refer to the dimension of the state n as the *order* of the system.

The function f is often referred to as dynamics or vector field. The interpretation of $u(t)$ and $y(t)$ is straightforward: they capture the actuation commands and the sensor measurements, respectively. The role of the state $x(t)$, however, might be less obvious. It should be thought of as an "internal" vector, whose elements correspond to physical quantities that change over time, hence their evolution is described by means of ODEs. With reference to the pendulum example, we may not be able to access all these variables in the output of the system, but possibly only a subset of them. In that case, $y(t)$ contains some of the elements of $x(t)$.

If it happens that the vector field f does not depend explicitly on time, we say that the system is time-invariant. If in addition, it depends neither on t, nor on $u(t)$, we say that the system is autonomous. The pendulum example involved a

nonlinear, autonomous system. If the vector field f and the output function h are linear with respect to $x(t)$ and $u(t)$, then we say that the underlying system is linear. In particular we consider the following two classes of linear systems.

Linear, time-varying systems:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),$$
$$y(t) = C(t)x(t) + D(t)u(t),$$

where $A(t), B(t), C(t), D(t)$, are matrices whose entries may depend on time.

Linear, time-invariant systems (LTI):

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

where matrices A, B, C, D, are independent of time.

The electric circuit example involved an LTI system. It was non-autonomous, due to the presence of the external input $u(t)$. For illustration purposes, a block diagram of an LTI system is shown in Figure 3.

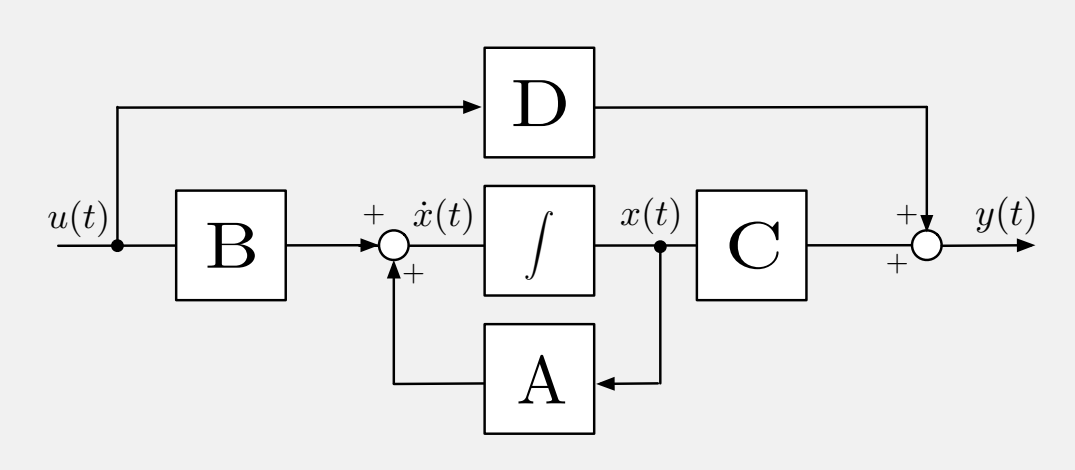


Figure 3: Block diagram of an LTI system.

Deriving the state space form from the physical description of a system requires experience. To make this process more systematic, we provide a sequence of steps that could provide further insight.

1. Determine the input variables $u(t)$. This could include external forces for mechanical systems or voltage/current sources for electrical ones.

2. Determine the output variables $y(t)$. To this end, consider sensor output signals that we can measure.

3. Select the state variables $x(t)$. These are typically variables that:

   - Capture the past and together with inputs give information about the future evolution of the system, namely, $\dot{x}(t)$.
   - For mechanical systems these are typically position and velocity components, while for electrical systems these contain voltages and currents.
   - The state variables are often related to storing energy.

4. Describe the physical evolution of the system (ODE) $\dot{x}(t) = f(x(t), u(t), t)$. This involves taking derivatives of the state variables and exploiting physical laws, like Newton's laws for mechanical systems, and Kirchhoff's laws for electrical ones.

5. Specify the output equation $y(t) = h(x(t), u(t), t)$. This refers to determining the quantities we can measure: either a subset of the state variables, or some of the inputs, or a combination of them.

Not all systems can be written in state space form. As an example, consider a system with delay, with output equation given by $y(t) = u(t - t_d)$ where t_d denotes the delay. To determine the output of the system at time instances after t we need information about the entire input history over the time interval $(t - t_d, t]$ rather than a pointwise estimate. Therefore, the output function h would need to admit as input a functional rather than a vector; such systems are called infinite dimensional.

1.3 Linearization

Often the underlying physical system is nonlinear, as with the pendulum example. Designing controllers for nonlinear systems is in general a difficult task; here we will do so for linear systems only. We show how to abstract/approximate nonlinear

systems (locally) with linear ones. To this end, consider a nonlinear vector field

$$f(x(t),u(t)) = \begin{bmatrix} f_1(x(t),u(t)) \\ \vdots \\ f_n(x(t),u(t)) \end{bmatrix},$$

where f_i is assumed to be sufficiently differentiable and for simplicity we have assumed that it does not depend on time explicitly. Further assume that we have access to a nominal state-input trajectory $(x^\star(\cdot), u^\star(\cdot))$ such that $\dot{x}^\star(t) = f(x^\star(t), u^\star(t))$; notice that this entails a state and an input as functions of time and not just a point. Access to such a trajectory could result from the application of an optimal, albeit open loop controller. As such, when applying $u^\star(t)$ to the real system due to model mismatch and the presence of disturbances, the resulting state trajectory may drift away from $x^\star(t)$. Designing directly a controller that would involve feedback (thus depending on the system state) would alleviate this issue, but is a difficult task for generic nonlinear systems.

We will do this indirectly. To this end, consider a perturbation $u_p(t)$ superimposed to $u^\star(t)$, resulting in a controller $u(t) = u^\star(t) + u_p(t)$. Under the input $u(t)$, the state becomes $x(t) = x^\star(t) + x_p(t)$, where $x_p(t) = x(t) - x^\star(t)$ encodes the perturbation from the nominal state trajectory. As such we have

$$x(t) = x^\star(t) + x_p(t) \quad \text{and} \quad u(t) = u^\star(t) + u_p(t).$$

We suppose that the perturbations $x_p(t), u_p(t)$ are small enough. We will employ $u_p(t)$ to ensure that if we start with $x_p(0)$ small (close to the nominal initial state), $x_p(t)$ remains small, or in other words the state trajectory track the nominal one. In particular, we will show that $\dot{x}_p(t)$ is (approximately) a linear function of $x_p(t)$ and $u_p(t)$, i.e., the perturbations form a linear system. Employing tools from control of linear systems developed in subsequent chapters we would be able to design $u_p(t)$ as a function of $x_p(t) = x(t) - x^\star(t)$ (thus the input will depend on the system state) to "regulate" the perturbation system. However, the validity of this design will only be local, as we have assumed the perturbations are small enough.

To achieve this, by the Taylor series expansion around $(x^\star(t), u^\star(t))$ we have that

$$\begin{aligned} \dot{x}(t) &= f(x(t),u(t)) = f(x^\star(t),u^\star(t)) \\ &\quad + \frac{\partial f}{\partial x}(x^\star(t),u^\star(t))\ x_p(t) + \frac{\partial f}{\partial u}(x^\star(t),u^\star(t))\ u_p(t) + \text{ higher order terms}, \end{aligned}$$

where recall that $x_p(t) = x(t) - x^\star(t)$ and $u_p(t) = u(t) - u^\star(t)$. Neglecting the higher order terms as the perturbations have been assumed to be small enough, and since $f(x^\star(t), u^\star(t)) = \dot{x}^\star(t)$, we have that

$$\begin{aligned}\dot{x}(t) - \dot{x}^\star(t) &\approx \frac{\partial f}{\partial x}(x^\star(t), u^\star(t))\ x_p(t) + \frac{\partial f}{\partial u}(x^\star(t), u^\star(t))\ u_p(t),\\ \Leftrightarrow \dot{x}_p(t) &\approx \frac{\partial f}{\partial x}(x^\star(t), u^\star(t))\ x_p(t) + \frac{\partial f}{\partial u}(x^\star(t), u^\star(t))\ u_p(t),\end{aligned}$$

where we used the fact that $\dot{x}_p(t) = \dot{x}(t) - \dot{x}^\star(t)$. Since f is a vector, the partial derivatives with respect to x and u are matrices given by

$$\frac{\partial f}{\partial x}(x^\star(t), u^\star(t)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^\star(t), u^\star(t)) & \dots & \frac{\partial f_1}{\partial x_n}(x^\star(t), u^\star(t)) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x^\star(t), u^\star(t)) & \dots & \frac{\partial f_n}{\partial x_n}(x^\star(t), u^\star(t)) \end{bmatrix} \in \mathbb{R}^{n\times n},$$

$$\frac{\partial f}{\partial u}(x^\star(t), u^\star(t)) = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(x^\star(t), u^\star(t)) & \dots & \frac{\partial f_1}{\partial u_m}(x^\star(t), u^\star(t)) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1}(x^\star(t), u^\star(t)) & \dots & \frac{\partial f_n}{\partial u_m}(x^\star(t), u^\star(t)) \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

The resulting perturbation dynamics form a linear time-varying system.

Fact 1 (Linearization)**.** *Consider the nonlinear system $\dot{x}(t) = f(x(t), u(t))$, where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and the functions comprising f are differentiable. The linearization of that system around a nominal trajectory $(x^\star(\cdot), u^\star(\cdot))$ is given by the linear, time-varying system*

$$\dot{x}_p(t) = A(t)x_p(t) + B(t)u_p(t),$$

where $A(t) = \frac{\partial f}{\partial x}(x^\star(t), u^\star(t)) \in \mathbb{R}^{n\times n}$ and $B(t) = \frac{\partial f}{\partial u}(x^\star(t), u^\star(t)) \in \mathbb{R}^{n\times m}$. If we are given a nominal operating point rather than a trajectory, i.e., $(x^\star, u^\star)$, then the linearization around this point is given by the LTI system

$$\dot{x}_p(t) = Ax_p(t) + Bu_p(t),$$

where $A = \frac{\partial f}{\partial x}(x^\star, u^\star) \in \mathbb{R}^{n\times n}$ and $B = \frac{\partial f}{\partial u}(x^\star, u^\star) \in \mathbb{R}^{n\times m}$ have now time independent entries.

For an autonomous system the linearization implies that locally, in the vicinity of a nominal trajectory or point, the nonlinear system can be approximated by a linear one. The following example highlights this point.

Example 1. *Consider the state space representation of the pendulum in Figure 1. Linearize the system around the nominal operating points (notice that this is an autonomous system so no input appears)*

$$x^\star = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad x^\star = \begin{bmatrix} \pi \\ 0 \end{bmatrix}.$$

Solution: *By the state space representation of the pendulum we obtain that*

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} x_2 \\ -\frac{d}{m}x_2 - \frac{g}{l}\sin x_1 \end{bmatrix}.$$

The linearization of the system around a nominal operating point $x^\star = [x_1^\star \; x_2^\star]^\top$ is given by $\dot{x}_p(t) = A\, x_p(t)$ (there are no inputs so no B matrix computation), where

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^\star) & \frac{\partial f_1}{\partial x_2}(x^\star) \\ \frac{\partial f_2}{\partial x_1}(x^\star) & \frac{\partial f_2}{\partial x_2}(x^\star) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l}\cos x_1^\star & -\frac{d}{m} \end{bmatrix}.$$

It turns out that the linearized system depends only on $x_1^\star$; for the two different values of $x_1^\star$ according to the operating points we obtain

$$x_1^\star = 0 \;\Rightarrow\; A = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l} & -\frac{d}{m} \end{bmatrix} \quad \text{and} \quad x_1^\star = \pi \;\Rightarrow\; A = \begin{bmatrix} 0 & 1 \\ \frac{g}{l} & -\frac{d}{m} \end{bmatrix}.$$

Since the only nonlinearity in this case was the trigonometric function $\sin x_1$, the outcome of the linearization could have been anticipated by performing a small angle approximation.

1.4 Summary

This chapter introduced the so called *state space* modelling formalism. In particular, we discussed and illustrated by means of examples the following three main state space descriptions, where $x(t) \in \mathbb{R}^n$ denotes the state, $u(t) \in \mathbb{R}^m$ the input, and $y(t) \in \mathbb{R}^p$ the output of the system.

1. *Nonlinear, time-varying systems:*

$$\dot{x}(t) = f(x(t), u(t), t),$$
$$y(t) = h(x(t), u(t), t),$$

   where f and h are nonlinear functions of $x(t)$, $u(t)$ and (possibly) the time.

2. *Linear, time-varying systems:*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),$$
$$y(t) = C(t)x(t) + D(t)u(t),$$

   where $A(t), B(t), C(t), D(t)$, are matrices whose entries may depend on time.

3. *Linear, time-invariant systems (LTI):*

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

   where matrices A, B, C, D, are independent of time.

We have also detailed a procedure called *linearization* (see Fact 1) that allows approximating locally the behaviour of the system around a nominal trajectory $(x^\star(\cdot), u^\star(\cdot))$ by a time-varying linear system. If rather than a nominal trajectory we have an operating point $(x^\star, u^\star)$, then linearization yields a time-invariant linear system.

2 Solutions of linear time invariant systems

Consider the state space description of an LTI system

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

$$y(t) = Cx(t) + Du(t), \tag{2.2}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^p$ is its output. Given an input trajectory $u(\cdot)$ as a function of time, we are interested in determining a *state solution* $x(\cdot) : \mathbb{R} \to \mathbb{R}^n$ that satisfies (in a sense that would be made precise in the sequel) the ODE in (2.1) and an *output solution* $y(\cdot) : \mathbb{R} \to \mathbb{R}^p$ from (2.2). To perform so, given $u(\cdot)$ as a function of time, and an initial condition (t_0, x_0) (time-state pair), we need to answer the following questions:

1. Do there always exist solutions of the LTI system for every initial condition?
2. If a solution exists, is it unique?

In this chapter we address these questions and show that LTI systems admit a unique state and output solution. Moreover, we will show how these solutions can be computed in closed form.

2.1 Existence and uniqueness of solutions

To address the existence and uniqueness questions, we first focus on a more abstract description of dynamic equations that are possibly nonlinear and are governed by the following ODE:

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

where f is a function of the state $x(t)$, the input $u(t)$, and may also depend explicitly on the time variable t. Given f, an input trajectory $u(\cdot)$ and an initial condition (t_0, x_0), we say that $x(\cdot)$ is a state solution to the ODE if the following two conditions are satisfied

$$x(t_0) = x_0,$$

$$\dot{x}(t) = f(x(t), u(t), t), \text{ for all } t \in \mathbb{R}.$$

Notice that we need to check that both the initial condition and the dynamic equation encoded by f to infer that a given $x(\cdot)$ constitutes a state solution of the ODE. However, for arbitrary choices of f and $u(\cdot)$ a solution does not always exist, and if it exists it is not necessarily unique. To get insight about potential issues, consider the following examples that refer to two autonomous systems.

✎ **Example 2** (No solutions for some initial conditions)**.** *Consider the ODE*

$$\dot{x}(t) = -\text{sgn}(x(t)) = \begin{cases} -1 & \text{if } x(t) \geq 0; \\ 1 & \text{if } x(t) < 0. \end{cases}$$

Consider the initial condition $(t_0, x_0) = (0, 0)$*. The systems starts at* $t = 0$ *and* $\dot{x}(0) = -1$*; however, by the time* $x(t)$ *becomes infinitesimally negative,* $\dot{x}(t)$ *becomes positive, bringing it back to zero state. One may be tempted to say that* $x(t) = 0$ *for all* t*. However, such a candidate solution satisfies the initial condition, but leads to* $\dot{x}(t) = 0$*; clearly this is different from* $\dot{x}(t) = -1$ *which is the ODE branch that includes the case* $x(t) = 0$*. In fact, the system is performing an infinite number of transitions in "zero" time, and the solution is thus undefined for all* $t > 0$*. Such chattering systems are called Zeno.*

✎ **Example 3** (Infinite number of solutions)**.** *Consider the ODE given by* $\dot{x}(t) = 3x(t)^{2/3}$*, and an initial condition given by* $(t_0, x_0) = (0, 0)$*. For* $a \geq 0$*, consider the following candidate solution*

$$x(t) = \begin{cases} (t-a)^3 & \text{if } t \geq a; \\ 0 & \text{if } t < a. \end{cases}$$

Note that since $a \geq t_0 = 0$*, the initial condition* $x(0) = 0$ *is always satisfied. Moreover, if* $t < a$*,* $x(t) = 0$ *for all* t *is a (trivial) solution of the ODE. Consider now the case where* $t \geq a$*, and notice that*

$$\dot{x}(t) = 3(t-a)^2 = 3x(t)^{2/3},$$

where the last equality follows from the definition of $x(t)$ *for* $t \geq a$*. Therefore the candidate* $x(t)$ *is a solution of the given ODE. However, this is the case for any* $a \geq 0$*, hence an infinite number of solutions exist.*

The first example highlights the necessity of f to depend continuously on $x(t)$, while the second one shows that continuity is not sufficient as the function tends to have infinite slope as $x(t)$ tends to zero. Cases like this can be excluded by introducing a Lipschitz continuity (see Definition 11 in the Appendix) requirement on the dependency of f with respect to $x(t)$. The following theorem summarizes the conditions under which we can guarantee existence and uniqueness of solutions to ODEs.

Theorem 1 (Existence & uniqueness of solutions to ODEs)**.** *Let* $\dot{x}(t) = f(x(t), u(t), t)$*, and assume that*

1. *The input* $u(\cdot)$ *and* f *are continuous functions with respect to time* t*.*
2. f *is Lipschitz continuous with respect to its first argument* $x(t)$*.*

We then have that for all initial conditions (t_0, x_0)*, there exists a unique continuous function* $x(\cdot) : \mathbb{R} \to \mathbb{R}^n$ *such that*

$$\begin{aligned} x(t_0) &= x_0, \\ \dot{x}(t) &= f(x(t), u(t), t), \ \textit{for all } t \in \mathbb{R}. \end{aligned}$$

Notice that u being a continuous function of time is rarely the case in practice. In particular, several optimal controllers are "bang-bang", i.e., they involve switching between their extreme values. Therefore, it would have been more realistic if $u(\cdot)$ (and f) are piecewise continuous as functions of time. This would imply that a set of discontinuity points may exist, however, we could still claim the existence and uniqueness conclusions of the theorem* if we further assume that i) within every finite interval of time the number of discontinuity points is finite, ii) left and right limits at discontinuity points are well defined, and iii) the value of u and f is taken to be the one of the right limit. The only difference is that the (unique) solution should satisfy $\dot{x}(t) = f(x(t), u(t), t)$, for all t except from the finite number of discontinuity points, i.e., almost everywhere.

*Note that even if we impose a piecewise continuity assumption this would only be a sufficient condition for existence and uniqueness of solutions. A more relaxed condition would be to consider u and f to be Lebesgue measurable functions of time, however, such developments would require results from measure theory and are outside the scope of these notes.

We will hereafter assume that the input trajectory $u(\cdot)$ is "well-behaved", i.e., it is either a continuous function of time, or a piecewise continuous one (based on the discussion above). We are now ready to return back to the LTI system in (2.1)-(2.2) and show that it admits a unique state $x(\cdot)$ and a unique output solution $y(\cdot)$.

Theorem 2 (Existence & uniqueness of solutions to LTI systems)**.** *The LTI system*

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

admits a unique continuous state $x(\cdot)$ and output solution $y(\cdot)$.

Proof: *To show that the given LTI system admits a unique continuous state solution $x(\cdot)$, it suffices to invoke Theorem 1, and show that $f(x(t), u(t), t) = Ax(t) + Bu(t)$ is Lipschitz continuous with respect to $x(t)$. By the Lipschitz continuity definition, considering two different states x and $\hat{x}$ (we drop the dependence on t for simplicity) we have that*

$$\|(Ax + Bu) - (A\hat{x} + Bu)\|^2 = \|A(x - \hat{x})\|^2 = (x - \hat{x})^\top A^\top A(x - \hat{x}),$$

where the last step follows from the definition of the Euclidean norm. By Fact 16 in the Appendix, we have that $A^\top A \preceq \lambda_{\max}(A^\top A)\ I$, where $\sqrt{\lambda_{\max}(A^\top A)}$ is the maximum singular value of A. By the definition of a positive semidefinite matrix this implies that

$$\begin{aligned}(x - \hat{x})^\top A^\top A(x - \hat{x}) &\leq (x - \hat{x})^\top \lambda_{\max}(A^\top A)\ I(x - \hat{x}) \\ &= \lambda_{\max}(A^\top A)(x - \hat{x})^\top (x - \hat{x}) \\ &= \lambda_{\max}(A^\top A)\|x - \hat{x}\|^2.\end{aligned}$$

Overall, we have shown that (taking square root in both sides)

$$\|(Ax + Bu) - (A\hat{x} + Bu)\| \leq \sqrt{\lambda_{\max}(A^\top A)}\|x - \hat{x}\|,$$

which implies that $Ax(t) + Bu(t)$ is Lipschitz continuous with respect to $x(t)$, with Lipschitz constant $L = \sqrt{\lambda_{\max}(A^\top A)}$, the maximum singular value of A. The fact that the output solution $y(\cdot)$ is unique and continuous follows directly from the fact that the output equation is a linear function of $x(\cdot)$, hence a unique continuous state solution implies a unique continuous output one.

2.2 Characterization of solutions

Theorem 2 shows that the LTI system in (2.1)-(2.2) admits a unique continuous state and output solution. Here, we provide a closed form characterization of these solutions starting at (t_0, x_0), where for simplicity we consider from now on $t_0 = 0$, and verify that they indeed satisfy (2.1)-(2.2).

In particular, the state and output solutions are given by

$$x(t) = \overbrace{\Phi(t)x_0}^{\text{zero input transition}} + \overbrace{\int_0^t \Phi(t-\tau)Bu(\tau)d\tau}^{\text{zero state transition}}$$

$$y(t) = \underbrace{C\Phi(t)x_0}_{\text{zero input response}} + \underbrace{\int_0^t C\Phi(t-\tau)Bu(\tau)d\tau + Du(t),}_{\text{zero state response}}$$

where $\Phi(t) \in \mathbb{R}^{n\times n}$ is called the *state transition matrix* (will be defined below). Notice that once $x(t)$ is computed, then $y(t)$ is directly calculated by means of $y(t) = Cx(t) + Du(t)$. The state solution consists of two parts:

- *Zero input transition:* This is the state solution if the system was autonomous, i.e., if $u(t) = 0$ for all t. Notice that this is a linear function of the initial state x_0. It also justifies the term state transition matrix for $\Phi(t)$, as given x_0, $\Phi(t)x_0$ dictates the state at time t.
- *Zero state transition:* This is a convolution integral, and it is a linear function of the input $u(t)$. To see this, consider the superposition principle of linearity (see Definition 9 in the Appendix), and notice that for $a_1, a_2 \in \mathbb{R}$ and inputs u_1, u_2, we have that

  $$\int_0^t \Phi(t-\tau)B(a_1u_1(\tau) + a_2u_2(\tau))d\tau$$
  $$= a_1\int_0^t \Phi(t-\tau)Bu_1(\tau)d\tau + a_2\int_0^t \Phi(t-\tau)Bu_2(\tau)d\tau.$$

  This was anticipated as the integrand is a linear function of u, and the integral is the continuous analogue of summation.

Similar comments pertain to the output solution $y(t)$ (the system's response) which comprises two terms, namely, the *zero input response* and the *zero state response*.

The state transition matrix is defined below and exhibits certain useful properties.

Fact 2 (State transition matrix and properties)**.** *The state transition matrix is defined as a matrix exponential, i.e., by the Taylor series expansion*

$$\Phi(t) = e^{At} = I + At + \frac{A^2t^2}{2!} + \ldots + \frac{A^kt^k}{k!} + \ldots$$

It satisfies the following properties:

1. $\Phi(0) = I$.
2. $\frac{d}{dt}\Phi(t) = A\Phi(t)$.
3. *It is invertible and its inverse is* $\Phi(-t)$*, i.e.,* $\Phi(t)\Phi(-t) = \Phi(-t)\Phi(t) = I$.
4. $\Phi(t_1 + t_2) = \Phi(t_1)\Phi(t_2)$ *for any* $t_1, t_2 \in \mathbb{R}$.

Note that calculating the integral or the derivative of a matrix exponential involves calculating the integral or the derivative of every entry of the matrix involved. However, e^{At} is **not** equal to a matrix where each of the entries is raised to the exponent; we discuss ways to compute it in the next section. Moreover, note that for arbitrary matrices A and B, we have that

$$e^{(A+B)t} \neq e^{At}e^{Bt},$$

except if A and B commute, i.e., if $AB = BA$ (see Fact 14 in the Appendix).

We are now ready to show that $x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau$ is indeed a state solution of the LTI system. The fact that $y(t) = C\Phi(t)x_0 + \int_0^t C\Phi(t-\tau)Bu(\tau)d\tau + Du(t)$ is the output solution follows then from $y(t) = Cx(t) + Du(t)$. By means of Theorem 2, these solutions will also have to be unique.

Proof that $x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau$ **is a state solution.** We have assumed that $u(\cdot)$ is "well-behaved" and we have shown in Theorem 2 that $Ax(t) + Bu(t)$ is Lipschitz continuous with respect to $x(t)$. Therefore, to show that the candidate expression for $x(t)$ is indeed a solution, it suffices to show that it satisfies

the initial condition and the ODE:

1. *Initial condition satisfaction:*

$$x(0) = \Phi(0)x_0 + \int_0^0 \Phi(0-\tau)Bu(\tau)d\tau = I\ x_0 = x_0,$$

where the second equality is due to the fact that $\Phi(0) = I$ from Fact 2, and the fact that the integral vanishes as the integration limits are the same.

2. *ODE satisfaction:* We will show that the candidate $x(t)$ satisfies $\dot{x}(t) = Ax(t) + Bu(t)$ for all t. To this end, since the integral in the expression of $x(t)$ involves t both at the integration limits and at the integrand, we will employ Leibniz rule below for differentiating integrals:

$$\frac{d}{dt}\int_{g_1(t)}^{g_2(t)} p(t,\tau)d\tau$$
$$= p(t,g_2(t))\frac{d}{dt}g_2(t) - p(t,g_1(t))\frac{d}{dt}g_1(t) + \int_{g_1(t)}^{g_2(t)} \frac{\partial}{\partial t}p(t,\tau)d\tau.$$

Setting $g_1(t) = 0$, $g_2(t) = t$ and $p(t,\tau) = \Phi(t-\tau)Bu(\tau)$, we obtain that

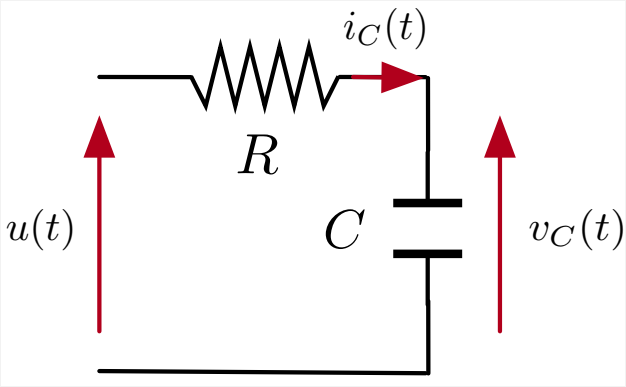
$$\dot{x}(t) = \frac{d}{dt}\Phi(t)x_0$$
$$+ \Phi(t-t)Bu(t)\overbrace{\frac{d}{dt}t}^{1} - \Phi(t-0)Bu(0)\overbrace{\frac{d}{dt}0}^{0} + \int_0^t \frac{d}{dt}\Phi(t-\tau)Bu(\tau)d\tau,$$

where given the cancellations shown above and the first two properties of Fact 2 (notice that by the chain rule $\frac{d}{dt}\Phi(t-\tau) = A\Phi(t-\tau)$), we have that

$$\begin{aligned}\dot{x}(t) &= A\Phi(t)x_0 + Bu(t) + A\int_0^t \Phi(t-\tau)Bu(\tau)d\tau \\ &= A(\Phi(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau) + Bu(t) \\ &= Ax(t) + Bu(t),\end{aligned}$$

thus concluding the proof.

✎ **Example 4** (Solution of an RC circuit)**.** *Consider the RC circuit in the figure below, where $u(t)$ is the input voltage. Derive the unit step response of the system if the output is taken to be the voltage across the capacitor.*



Solution: *The system is described by*

- *Current that flows through the capacitor:* $i_C(t) = C\frac{dv_C(t)}{dt}$.
- *Kirchhoff's voltage law:* $u(t) = Ri_C(t) + v_C(t)$.

Substituting the first equation in the second one, we obtain

$$\frac{d}{dt}v_C(t) = -\frac{1}{RC}v_C(t) + \frac{1}{RC}u(t)$$
$$y(t) = v_C(t),$$

where the second equation corresponds to the output of the system. Take the state to be $x(t) = v_C(t)$*, i.e., the voltage across the capacitor, and denote the initial condition by* $x_0 = v_C(0)$*. The physical description corresponds then to an LTI system; however, all quantities are scalar, so with reference to the general LTI system description,*

$$A = -\frac{1}{RC},\ B = \frac{1}{RC},\ C = 1,\ D = 0, \quad \textit{[all scalars]}.$$

Matrix C *should not be confused with the capacitance symbol. The state transition matrix (also scalar) is thus given by* $\Phi(t) = e^{At} = e^{-\frac{t}{RC}}$*. For a unit step response we have* $u(t) = 1$ *for all* $t \geq 0$*. Hence, the step response of the system (coincides in this case with the state solution as* $C = 1$*) is given by*

$$\begin{aligned} y(t) = x(t) &= e^{-\frac{t}{RC}}x_0 + \int_0^t e^{-\frac{t-\tau}{RC}}\frac{1}{RC}u(\tau)d\tau \\ &= e^{-\frac{t}{RC}}x_0 + \int_0^t e^{-\frac{t-\tau}{RC}}\frac{1}{RC}d\tau, \qquad [u(\tau) = 1 \textit{ for all } \tau \geq 0] \\ &= e^{-\frac{t}{RC}}x_0 + (1 - e^{-\frac{t}{RC}}). \end{aligned}$$

If $u(t) = 0$ *(capacitor* C *discharging over* R*), verify that this solution could be obtained by the solving the first order ODE from first principles.*

2.3 Computation of the state transition matrix

To determine the state and output solutions the state transition matrix $\Phi(t)$ needs to be computed. In the previous example this was straightforward as all quantities were scalars; in general this is more difficult as its definition involves an infinite Taylor series expansion of the matrix exponential. In the sequel we show how this can be performed systematically for diagonalizable and non-diagonalizable matrices.

2.3.1 Diagonalizable matrices

A matrix is called diagonalizable if its eigenvectors are linearly independent (see Definition 6 in the Appendix). If a matrix A is diagonalizable, it can be decomposed as (see also Fact 15 in the Appendix)

$$A = W\Lambda W^{-1},$$

where W is a matrix whose columns are the eigenvectors of A (invertible since the eigenvectors are linearly independent), and Λ is a diagonal matrix whose diagonal entries correspond to the eigenvalues of A. For diagonalizable matrices the state transition matrix can be computed in an efficient way that does not involve infinite series.

Fact 3 (State transition matrix for diagonalizable matrices)**.** *Consider an LTI system with $A \in \mathbb{R}^{n\times n}$ being diagonalizable, admitting a decomposition $A = W\Lambda W^{-1}$. Its state transition matrix is given by*

$$\Phi(t) = e^{At} = We^{\Lambda t}W^{-1},$$

where $e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{bmatrix}$*, with λ_i being the i-th eigenvalue of A.*

Proof: *We first show that for an arbitrary integer k, $A^k = W\Lambda^k W^{-1}$. We show this by means of induction:*

1. *Base case ($k = 0$): We have that $W\Lambda^0 W^{-1} = WW^{-1} = I = A^0$, hence the base case is trivially satisfied.*

2. *Induction hypothesis: Assume the statement holds true for an arbitrary* k*, i.e.,* $A^k = W\Lambda^k W^{-1}$.

3. *Show the claim for the* $(k+1)$*-th case: By the induction hypothesis we have that* $A^k = W\Lambda^k W^{-1}$*. Hence,*

$$A^{k+1} = A^k A = W\Lambda^k \cancelto{I}{W^{-1}W} \Lambda W^{-1} = W\Lambda^{k+1}W^{-1}.$$

We will now show that $e^{At} = We^{\Lambda t}W^{-1}$*. By the Taylor series expansion of the matrix exponential we have that*

$$\begin{aligned} e^{At} &= I + At + \frac{A^2t^2}{2!} + \ldots + \frac{A^k t^k}{k!} + \ldots \\ &= WW^{-1} + W\Lambda t W^{-1} + W\frac{\Lambda^2 t^2}{2!}W^{-1} + \ldots W\frac{\Lambda^k t^k}{k!}W^{-1} + \ldots \\ &= W\Big(I + \Lambda t + \frac{\Lambda^2 t^2}{2!} + \ldots + \frac{\Lambda^k t^k}{k!} + \ldots\Big)W^{-1} \\ &= We^{\Lambda t}W^{-1}, \end{aligned}$$

where in the second equality we used the fact that $A^k = W\Lambda^k W^{-1}$ *and that* t^k *is a scalar so it can be moved inside the triple matrix product. The last equality is due to the fact that term in the parenthesis is the Taylor series expansion of* $e^{\Lambda t}$.

We illustrate this fact by means of the following example.

✎ **Example 5.** *Consider the LTI system corresponding to the RLC circuit of Figure 2 with* $R = 3$*,* $L = 1$ *and* $C = 0.5$*. Compute the system's state transition matrix.*

Solution: *Under the given numerical values we have that* $A = \begin{bmatrix} 0 & 2 \\ -1 & -3 \end{bmatrix}$*. Computing the eigenvalues and eigenvectors of* A *we obtain* $\lambda_1 = -1$ *and* $\lambda_2 = -2$*, and*

$$w_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \textit{and} \quad w_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \;\Rightarrow\; W = \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix}.$$

Notice that the eigenvectors are linearly independent (W is invertible) hence A is diagonalizable. The state transition matrix can be then computed by

$$\begin{aligned}\Phi(t) = e^{At} &= We^{\Lambda t}W^{-1} \\ &= \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \\ &= \begin{bmatrix} 2e^{-t} - e^{-2t} & 2e^{-t} - 2e^{-2t} \\ -e^{-t} + e^{-2t} & -e^{-t} + 2e^{-2t} \end{bmatrix}.\end{aligned}$$

Notice that the matrix exponential e^{At} is different from the matrix that would emanate if the entries of A are raised to the exponent.

2.3.2 Non-diagonalizable matrices with particular structure

In this section we consider LTI systems with a matrix $A \in \mathbb{R}^{n\times n}$ which is not necessarily diagonalizable. In particular, we have that (see also Section 9.1.4)

$$\begin{aligned} & n \text{ distinct eigenvalues } \Rightarrow \ A \text{ diagonalizable} \\ \iff \quad & A \text{ non-diagonalizable } \Rightarrow \text{ not all eigenvalues are distinct.}\end{aligned}$$

The difficulty, however, with non-diagonalizable (or else defective) matrices stems not from the fact that their eigenvalues are not distinct, but from the fact that the number of linearly independent eigenvectors is also strictly smaller than n, hence the matrix of eigenvectors W is no longer invertible. We denote the number of linearly independent eigenvectors by $k < n$. We define the algebraic multiplicity of an eigenvalue as the number of times it appears in the spectrum of A, while we define its geometric multiplicity as the number of linearly independent eigenvectors corresponding to this eigenvalue. For non-diagonalizable matrices, the geometric multiplicity of some eigenvalue is strictly less than its algebraic one.

Example 6. *Consider the matrices*

$$A_1 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}, \ A_2 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \ A_3 = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix},$$

where λ is real. Which of these matrices are diagonalizable? For each case compute the algebraic and geometric eigenvalue multiplicity.

Solution: *By a direct computation (the eigenvalues are the diagonal entries) it follows that all three matrices have $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$ as a repeated eigenvalue with algebraic multiplicity equal to 3. The linearly independent eigenvectors of each matrix are given by*

1. *A_1 has geometric multiplicity 1* $\Rightarrow$ $\begin{bmatrix}1\\0\\0\end{bmatrix}$.

2. *A_2 has geometric multiplicity 2* $\Rightarrow$ $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$.

3. *A_3 has geometric multiplicity 3* $\Rightarrow$ $\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$.

Therefore, A_3 is a diagonalizable matrix (its algebraic and geometric multiplicities coincide), while A_1 and A_2 are non-diagonalizable. Note that A_3 is diagonalizable, despite the fact that all its eigenvalues are equal. This shows that the reverse implication of the statement in the beginning of this section does not hold.

We show that for a particular class of non-diagonalizable matrices the state transition matrix can still be computed efficiently. We first illustrate this by means of a couple of examples.

Example 7. *Consider an LTI system with*

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

Compute the state transition matrix $\Phi(t) = e^{At}$.

Solution: *Note first that A is non-diagonalizable; it has a repeated eigenvalue at 0 (algebraic multiplicity $= 2$), and one linearly independent eigenvector $[1\ 0]^\top$ (geometric multiplicity $= 1$). Notice that $A^k = 0$ for all $k \geq 2$. By the definition of $\Phi(t)$ by means of the Taylor series expansion we have that*

$$\begin{aligned}\Phi(t) = e^{At} &= I + At + \frac{A^2t^2}{2!} + \ldots + \frac{A^kt^k}{k!} + \ldots \\ &= I + At = \begin{bmatrix} 1 & 3t \\ 0 & 1 \end{bmatrix}.\end{aligned}$$

Matrices such that $A^k = 0$ for some integer k are called *nilpotent*. For such matrices it becomes easy to compute the matrix exponential, as the infinite series is truncated after a certain term.

Example 8. *Consider an LTI system with*

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

Compute the state transition matrix $\Phi(t) = e^{At}$.

Solution: *Note first that A is non-diagonalizable; it has a repeated eigenvalue at 2 (algebraic multiplicity $= 2$), and one linearly independent eigenvector $[1\ 0]^\top$ (geometric multiplicity $= 1$). Notice that A can be decomposed as the sum of an identity (modulo a proportionality constant) and a nilpotent matrix, namely,*

$$A = A_1 + A_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}.$$

However, as computed in the previous example $e^{A_2t} = \begin{bmatrix} 1 & 3t \\ 0 & 1 \end{bmatrix}$, while

$$e^{A_1t} = \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{2t} \end{bmatrix},$$

since A_1 is diagonal and the exponential of a diagonal matrix is a matrix whose entries include the diagonal terms raised to the exponent.

Notice now that A_1 and A_2 commute, i.e., $A_1A_2 = A_2A_1$. Therefore,

$$\Phi(t) = e^{At} = e^{(A_1+A_2)t} = e^{A_1t}e^{A_2t}$$
$$= \begin{bmatrix} e^{2t} & 0 \\ 0 & e^{2t} \end{bmatrix} \begin{bmatrix} 1 & 3t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e^{2t} & 3te^{2t} \\ 0 & e^{2t} \end{bmatrix}.$$

The preceding example illustrates that even if a matrix is not nilpotent it is sometimes still possible to compute the state transition matrix efficiently if we can decompose it as the sum of an identity matrix (modulo a proportionality constant) and a nilpotent one. It turns out that this decomposition is more general.

Fact 4 (Sum of nilpotent and identity matrices)**.** *Consider an LTI system with $A = \lambda\, I + N$, where λ is a real scalar, N is a Nilpotent matrix, and I is an identity matrix of appropriate dimension. The state transition matrix is given by*

$$\Phi(t) = e^{At} = e^{\lambda It}e^{Nt},$$

where $e^{\lambda It}$ is a diagonal matrix with its diagonal entries being equal to $e^{\lambda t}$.

Note that Example 8 involved a repeated eigenvalue with algebraic multiplicity of 2, and resulted in a state transition matrix containing a term proportional to t. For matrices of the same structure but higher dimension, one would expect terms $t, t^2, \ldots, t^{r-1}$, where r denotes the algebraic multiplicity of the repeated eigenvalue.

2.3.3 Non-diagonalizable matrices with generic structure

Here we consider generic non-diagonalizable matrices, that do not necessarily exhibit the structure of Fact 4. Since there are only $k < n$ linearly independent eigenvectors, we define a procedure to append to them $n - k$ additional (linearly independent) vectors so that we create an invertible matrix. The resulting family of vectors is termed *generalized eigenvectors*.

Procedure to construct generalized eigenvectors.
For $i = 1, \ldots, k$ take w_i (the i-th linearly independent eigenvector of A), and repeat the following steps:

1. Set $w_i^1 = w_i$, and let λ be the eigenvalue corresponding to w_i.
2. For $j = 1, \ldots, \mu_i$ construct vectors recursively by

$$(\lambda I - A) w_i^{j+1} = w_i^j,$$

where μ_i is such that there is no other vector w with $(\lambda I - A)w = w_i^{\mu_i}$ that is linearly independent with $\{w_i^1, w_i^2, \ldots, w_i^{\mu_i}\}$.

Output of the procedure: A set of μ_i generalized eigenvectors $\{w_i^1, w_i^2, \ldots, w_i^{\mu_i}\}$ for each $i = 1, \ldots, k$.

It can be shown that the generalized eigenvectors constructed above are linearly independent. Moreover, since w_i^1 is an eigenvector of A, we have that $Aw_i^1 = \lambda w_i^1 \Leftrightarrow (\lambda I - A)w_i^1 = 0$. By the recursive construction of the generalized eigenvectors one can show that* $(\lambda I - A)^j w_i^j = 0$ for all $j = 1, \ldots, \mu_i$.

Since we have $k < n$ linearly independent eigenvectors the main steps of the procedure outlined above are repeated for each of those. Each time we get in return a different set of generalized eigenvectors (possibly also with a different cardinality); stacking the k vector families as blocks next to each other we obtain

$$T = \left[w_1^1,\ w_1^2,\ \ldots,\ w_1^{\mu_1},\ w_2^1,\ w_2^2,\ \ldots,\ w_2^{\mu_2},\ \ldots\ldots,\ w_k^1,\ w_k^2,\ \ldots,\ w_k^{\mu_k}\right].$$

The number of generalized eigenvector blocks corresponding to the same eigenvalue is equal to the geometric multiplicity of that eigenvalue. It can be shown that T is an invertible matrix with $\sum_{i=1}^k \mu_i = n$, containing the constructed generalized eigenvectors as its columns.

*Although outside the scope of these notes, based on this property one can show that the subspace spanned by the generalized eigenvectors exhibits an invariance property under the mapping A, i.e., if $(\lambda I - A)^j x = 0$ then $(\lambda I - A)^j Ax = 0$; hence, the particular procedure to construct generalized eigenvectors not only guarantees that they are linearly independent but has also certain geometric implications.

✎ **Example 9.** *Consider the non-diagonalizable matrices A_1 and A_2 of Example 6. For each case compute the generalized eigenvectors matrix.*

Solution:

Generalized eigenvectors for A_1: *We have $k = 1$ linearly independent eigenvectors so we only have to run the main steps of the generalized eigenvectors procedure once. To this end, set $w_1^1 = [1\ 0\ 0]^\top$ and notice that this corresponds to an eigenvalue λ. We have that*

$$(\lambda I - A)w_1^2 = w_1^1 \ \Rightarrow\ w_1^2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \text{and } (\lambda I - A)w_1^3 = w_1^2 \ \Rightarrow\ w_1^3 = \begin{bmatrix}0\\0\\1\end{bmatrix}.$$

We thus have $\mu_1 = 3$ and after the second iteration we can no longer find a linearly independent eigenvector. Hence, the generalized eigenvectors are given by $T = \begin{bmatrix} w_1^1 & w_1^2 & w_1^3 \end{bmatrix}$ and comprises of one block with three vectors.

Generalized eigenvectors for A_2: *We have $k = 2$ linearly independent eigenvectors so we have to repeat the main steps of the generalized eigenvectors procedure twice. To this end, in the first run set $w_1^1 = [1\ 0\ 0]^\top$ and notice that it corresponds to an eigenvalue λ. We have that*

$$(\lambda I - A)w_1^2 = w_1^1 \ \Rightarrow\ w_1^2 = \begin{bmatrix}0\\1\\0\end{bmatrix},$$

while after this the first procedure run terminates with $\mu_1 = 2$ as we cannot find another linearly independent vector w such that $(\lambda I - A_2)w = w_1^2$. To see this, notice that this results in the system of equations $\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} w = \begin{bmatrix}0\\1\\0\end{bmatrix}$ which does not admit a solution.

In the second run set $w_2^1 = [0\ 0\ 1]^\top$ (the second linearly independent eigenvector of A_2) and notice that it also corresponds to an eigenvalue λ. As we cannot find another linearly independent vector w with $(\lambda I - A_2)w = w_2^1$ the procedure terminates with $\mu_2 = 1$.

Hence, the generalized eigenvectors matrix is given by $T = \begin{bmatrix} w_1^1 & w_1^2 & w_2^1 \end{bmatrix}$ and comprises of two blocks, with two and one vectors, respectively.

Fact 5 (State transition matrix for non-diagonalizable matrices – Jordan canonical form)**.** *Consider an LTI system with a possibly non-diagonalizable matrix A that has k linearly independent eigenvectors. A can be written in the so called Jordan canonical form $A = TJT^{-1}$, where T is a matrix whose columns are the generalized eigenvectors, and*

$$J = \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & J_k \end{bmatrix} \in \mathbb{C}^{n \times n}, \text{ where } J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_i & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{\mu_i \times \mu_i},$$

where for $i = 1, \dots, k$, λ_i is the eigenvalue corresponding to the i-th linearly independent eigenvector.
The state transition matrix can be then computed by $\Phi(t) = Te^{Jt}T^{-1}$, where

$$e^{Jt} = \begin{bmatrix} e^{J_1 t} & 0 & \dots & 0 \\ 0 & e^{J_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{J_k t} \end{bmatrix} \text{ and } e^{J_i t} = \begin{bmatrix} e^{\lambda_i t} & te^{\lambda_i t} & \frac{t^2}{2!}e^{\lambda_i t} & \dots & \frac{t^{\mu_i - 1}}{(\mu_i - 1)!}e^{\lambda_i t} \\ 0 & e^{\lambda_i t} & te^{\lambda_i t} & \dots & \frac{t^{\mu_i - 2}}{(\mu_i - 2)!}e^{\lambda_i t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_i t} \end{bmatrix}.$$

Note that J_i can be decomposed as the sum of $\lambda_i I$ plus a Nilpotent matrix. Hence, by means of Fact 4, $e^{J_i t}$ can be efficiently computed in closed form as shown above. Moreover, if it happens that A is diagonalizable, then $k = n$, T is equal to the eigenvector matrix W, and $J = \Lambda$, i.e., Fact 5 reduces to Fact 3.

✎ **Example 10.** *Consider two different LTI systems, whose matrix A is equal to A_1 and A_2, respectively, as these are defined in Example 6. For each case compute the state transition matrix.*

Solution:

Matrix A_1***:*** *For this case we determined* $k = 1$ *linearly independent eigenvector and one block of generalized eigenvectors with* $\mu_1 = 3$*. Hence,*

$$J = J_1, \text{ where } J_1 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Since T *is in this case the identity matrix (see computation in previous example), the state transition matrix is given by*

$$\Phi(t) = Te^{Jt}T^{-1} = e^{J_1 t} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} \\ 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & 0 & e^{\lambda t} \end{bmatrix}$$

Matrix A_2***:*** *For this case we determined* $k = 2$ *linearly independent eigenvector and two blocks of generalized eigenvectors, one with* $\mu_1 = 2$ *and another one with* $\mu_2 = 1$*. Hence,*

$$J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}, \text{ where } J_1 = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \text{ and } J_2 = \lambda.$$

The matrix exponentials corresponding to J_1 *and* J_2 *are then given by*

$$e^{J_1 t} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix} \text{ and } e^{J_2 t} = e^{\lambda t}.$$

Since T *is the identity matrix, the state transition matrix is given by*

$$\Phi(t) = Te^{Jt}T^{-1} = e^{Jt} = \begin{bmatrix} e^{J_1 t} & 0 \\ 0 & e^{J_2 t} \end{bmatrix} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & 0 \\ 0 & e^{\lambda t} & 0 \\ 0 & 0 & e^{\lambda t} \end{bmatrix}.$$

2.4 Summary

This chapter studied the computation of state and output solutions to LTI systems. The main learning outcomes of the chapter can be summarized as follows:

1. LTI systems admit a unique continuous solution (Theorem 2).

2. The state and output solutions of LTI systems are given by

$$x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau,$$
$$y(t) = C\Phi(t)x_0 + \int_0^t C\Phi(t-\tau)Bu(\tau)d\tau + Du(t),$$

   where $\Phi(t) = e^{At}$ is the state transition matrix.

3. The state transition matrix, necessary for the computation of the state and output solutions, can be calculated by means of three different ways according to the structure of the A matrix.

   (a) A is a diagonalizable matrix, i.e., its eigenvectors are linearly independent. The state transition matrix can be then computed by means of

$$e^{At} = We^{\Lambda t}W^{-1},$$

   where W is a matrix whose columns are the eigenvectors of A and Λ is a diagonal matrix whose diagonal elements are its eigenvalues (Fact 3).

   (b) A is a non-diagonalizable matrix, i.e., its eigenvectors are not all linearly independent, that can be written as $A = \lambda I + N$, where λ is a scalar and N is a non-zero Nilpotent matrix. The state transition matrix can be then computed by

$$e^{At} = e^{\lambda It}e^{Nt},$$

   where e^{Nt} is calculated efficiently by means of the (truncated in this case) Taylor series expansion of the matrix exponential (Fact 4).

   (c) A is a non-diagonalizable matrix with generic structure. The state transition matrix can be then computed by

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

   where T is a matrix whose columns are the so called generalized eigenvectors of A and J is a block-diagonal matrix. Such a decomposition is called Jordan canonical form. (Fact 5).

3 Stability and connections with transfer functions

3.1 Stability

In the previous chapter we showed how to determine solutions for the state $x(t)$ and the output $y(t)$ of LTI systems. These solutions depend on the so called state transition matrix $\Phi(t)$, which in turn depends on the eigenvalues and eigenvectors of the system's A matrix (both for diagonalizable and non-diagonalizable matrices). This structure imposes the following questions:

1. Can we anticipate the system's evolution, in particular the "long-run" behaviour as $t \to \infty$, by looking at its eigenvalues and eigenvectors?

2. If yes, what are the implications on the stability of the system?

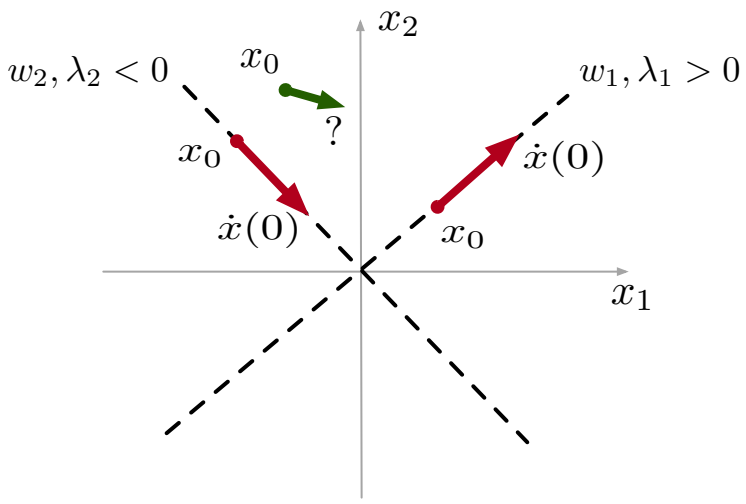


Figure 4: Schematic diagram showing that for autonomous systems $\dot{x}(t) = Ax$ starting on an eigenvector solutions stay on that eigenvector. The norm of the solution increases or decreases according to the associated eigenvalue.

To gain some intuition on the qualitative characteristics of the state (similarly for the output) solution, consider the schematic diagram of Figure 4. This diagram illustrates a two-state autonomous system governed by $\dot{x}(t) = Ax(t)$. The dashed lines represent the two eigenvectors w_1 and w_2 (assumed to be real for the sake of this illustration), which correspond to the real eigenvalues $\lambda_1 > 0$ and $\lambda_2 < 0$, respectively. If the initial condition x_0 happens to be on one of these eigenvectors, then $x(t)$ will stay on that eigenvector for all t. To see this recall that if w_i is an

eigenvector associated to eigenvalue λ_i, $i = 1, 2$, then if $x_0 = w_i$ (starting on that eigenvector),

$$Aw_i = \lambda_i w_i \;\Rightarrow\; \dot{x}(0) = Ax_0 = Aw_i = \lambda_i w_i,$$

i.e., the direction $\dot{x}(0)$ the state will move is aligned with w_i, however, it gets rescaled by λ_i. Therefore, whether $\|x(t)\|$ will increase or decrease depends on the sign of the associated eigenvalue. For the particular example if we start on eigenvector w_2 then the system's state will tend to the origin ($\lambda_2 < 0$), while if we start on w_1 ($\lambda_1 > 0$) the state will grow towards infinity, thus having certain implications about stability. However, it still appears unclear how the system will evolve if we start from an initial condition that is not lying on any of the eigenvectors ("green" dot in Figure 4).

We will now formally address this question for autonomous LTI systems, i.e., systems governed by

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

thus we will analyze the so called zero input transition of the state solution, namely $x(t) = \Phi(t)x_0$. We will do this separately for diagonalizable and non-diagonalizable A matrices.

3.1.1 Diagonalizable matrices

For the developments of this subsection we assume that matrix A of an LTI system state space description is diagonalizable. We start by illustrating the influence of eigenvalues and eigenvectors on the state response by means of an example.

Example 11. *Consider the numerical values for the RLC circuit given in Example 5, and the computed state transition matrix and eigenvectors*

$$\Phi(t) = \begin{bmatrix} 2e^{-t} - e^{-2t} & 2e^{-t} - 2e^{-2t} \\ -e^{-t} + e^{-2t} & -e^{-t} + 2e^{-2t} \end{bmatrix}, \; w_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \text{and} \quad w_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Compute the zero input transition of the solution $x(t)$, and determine its behaviour as $t \to \infty$, if $x_0 = a_1 w_1 + a_2 w_2$ for some scalars a_1, a_2.

Solution: *As matrix A is diagonalizable, these eigenvectors are linearly in-*

dependent. Therefore, an arbitrary initial condition x_0 can always be written as $x_0 = a_1 w_1 + a_2 w_2$ for some scalars a_1, a_2. The state solution (zero input transition in this case) is then given by

$$\begin{aligned} x(t) = \Phi(t)x_0 &= \Phi(t)(a_1 w_1 + a_2 w_2) \\ &= a_1\Phi(t)w_1 + a_2\Phi(t)w_2 \\ &= a_1 \begin{bmatrix} 2e^{-t} \\ -e^{-t} \end{bmatrix} + a_2 \begin{bmatrix} e^{-2t} \\ -e^{-2t} \end{bmatrix} \\ &= a_1 e^{-t} w_1 + a_2 e^{-2t} w_2. \end{aligned}$$

Notice that if $a_1 = 0$ or if $a_2 = 0$ then the solution starts at an eigenvector and stays on that eigenvector. Taking the limit as $t \to \infty$, and since the eigenvalues are bot negative (-1 and -2), we see that

$$\lim_{t\to\infty} x(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The previous example suggests that starting at an arbitrary initial condition, the zero input transition $x(t) = \Phi(t)x_0$ is a linear combination of the eigenvectors (notice that these are independent as matrix A is diagonalizable). The coefficients of this linear combination are time-dependent and rely on the eigenvalues, whose sign indicates the effect on $\|x(t)\|$. In the previous example they gave rise to a decaying exponential behaviour; however, according to the eigenvalues, different terms may appear in the solution.

	$\lambda = 0$	$\lambda = \sigma$		$\lambda = j\omega$	$\lambda = \sigma + j\omega$	
		$\sigma < 0$	$\sigma > 0$		$\sigma < 0,\ \omega \neq 0$	$\sigma > 0,\ \omega \neq 0$
Terms in solution	1	$e^{\sigma t}$		$\sin\omega t, \cos\omega t$	$e^{\sigma t}\sin\omega t, e^{\sigma t}\cos\omega t$	
Limit as $t \to \infty$	constant	0	∞	periodic	0	∞

Table 1: Classification of the different terms that may appear in the solution of LTI systems with diagonalizable A matrix, and their asymptotic behaviour.

Since eigenvalues are in general complex, we denote $\lambda = \sigma + j\omega$. Each eigenvalue

contributes to the state transition matrix $\Phi(t)$ by certain terms, and as a result to the zero input transition as this is a linear combination of the terms that appear in $\Phi(t)$ (the coefficients of this combination will depend on the eigenvectors). Recalling that $e^{\lambda t} = e^{(\sigma+j\omega)t} = e^{\sigma t}(\cos \omega t + j \sin \omega t)$, the contribution of each eigenvalue to the solution according to the different values of σ and ω is summarized in Table 1.

Investigating the limiting behaviour of these terms allows us to determine the asymptotic behaviour of the system as $t \to \infty$. In particular, the limiting behaviour of $x(t)$ is closely related to the stability properties of systems in the form of $\dot{x}(t) = Ax(t)$ where $x(t) = 0$ is an equilibrium solution as it results in $\dot{x}(t) = 0$. We consider the following notions of stability (we only provide an informal definition):

1. *Stability:* A system is called stable* if we can stay arbitrarily close enough to 0 if we start sufficiently close to it.

2. *Asymptotic stability:* A system is called asymptotically stable if it is stable, and approaches 0 as time tends to infinity, i.e., $\lim_{t\to\infty} \|x(t)\| = 0$. In other words, not only we stay close to 0, but also converge to it.

We then say that a system that is not stable is unstable. Denote by $\lambda_i = \sigma_i + j\omega_i$, $i = 1, \ldots, n$ the eigenvalues of $A \in \mathbb{R}^{n\times n}$. From Table 1 it can be observed that

- If $\mathrm{Re}(\lambda_i) \neq 0$ for all i, then $x(t)$ is a linear combination of $e^{\sigma_i t}, e^{\sigma_i t} \sin \omega_i t$, $e^{\sigma_i t} \cos \omega_i t$. As a result $\lim_{t\to\infty} \|x(t)\| = 0$ if $\sigma_i < 0$ for all i, otherwise $\lim_{t\to\infty} \|x(t)\| = \infty$ (for some initial conditions) if there exists i with $\sigma_i > 0$.

- If $\mathrm{Re}(\lambda_i) \leq 0$ for all i (we allow eigenvalues to have zero real part), then $x(t)$ is a linear combination of terms $e^{\sigma_i t}, e^{\sigma_i t} \sin \omega_i t$, $e^{\sigma_i t} \cos \omega_i t$ but also $1, \sin \omega_i t, \cos \omega_i t$. The latter terms do not vanish as $t \to \infty$, hence the solution of the system is constant or periodic.

The relationship of these observations to stability are summarized in the fact below.

*Formally, a system is called stable if for all $\epsilon > 0$, there exists $\delta > 0$ such that if $\|x(0)\| < \delta$ then $\|x(t)\| < \epsilon$ for all $t \geq 0$.

Fact 6 (Stability of $\dot{x}(t) = Ax(t)$ with diagonalizable A)**.** *Consider an autonomous LTI system $\dot{x}(t) = Ax(t)$ with $A \in \mathbb{R}^{n\times n}$ diagonalizable. Let λ_i, $i = 1, \ldots, n$, be the eigenvalues of A. We then have that the system is:*

- *Stable (for all initial conditions) if and only if $\mathrm{Re}(\lambda_i) \leq 0$ for all $i = 1, \ldots, n$.*
- *Asymptotically stable (for all initial conditions) if and only if $\mathrm{Re}(\lambda_i) < 0$ for all $i = 1, \ldots, n$.*
- *Unstable (for some initial conditions) if and only if there exists i such that the corresponding eigenvalue has $\mathrm{Re}(\lambda_i) > 0$.*

Note that if the system is asymptotically stable, then the magnitude of the real part of the eigenvalues carries information about the rate with which the state decays towards zero. Figure 5 illustrates some zero input transition solution patterns for LTI systems with diagonalizable A matrix. The different snapshots correspond to different eigenvalue locations in the complex plane.

3.1.2 Non-diagonalizable matrices

For the developments of this subsection we assume that matrix A of an LTI system state space description is non-diagonalizable. Since A is not diagonalizable, its eigenvectors are no longer linearly independent, and at least one of its eigenvalues is repeated; say we have one such eigenvalue with algebraic multiplicity $r > 1$.

	$\lambda = 0$	$\lambda = \sigma$		$\lambda = j\omega$	$\lambda = \sigma + j\omega$	
		$\sigma < 0$	$\sigma > 0$		$\sigma < 0,\ \omega \neq 0$	$\sigma > 0,\ \omega \neq 0$
Terms in solution	$1, t, t^2, \ldots, t^{r-1}$	$e^{\sigma t}, te^{\sigma t}, \ldots, t^{r-1}e^{\sigma t}$		$\sin\omega t, \ldots, t^{r-1}\sin\omega t,$ $\cos\omega t, \ldots, t^{r-1}\cos\omega t$	$e^{\sigma t}\sin\omega t, \ldots, t^{r-1}e^{\sigma t}\sin\omega t,$ $e^{\sigma t}\cos\omega t, \ldots, t^{r-1}e^{\sigma t}\cos\omega t$	
Limit as $t \to \infty$	∞	0	∞	∞	0	∞

Table 2: Classification of the different terms that may appear in the solution of LTI systems with non-diagonalizable A matrix that has an eigenvalue repeated $r > 1$ times, and their asymptotic behaviour.

By inspection of Facts 4 & 5, we can construct the state transition matrix $\Phi(t)$ for non-diagonalizable matrices which, however, may now include terms that involve

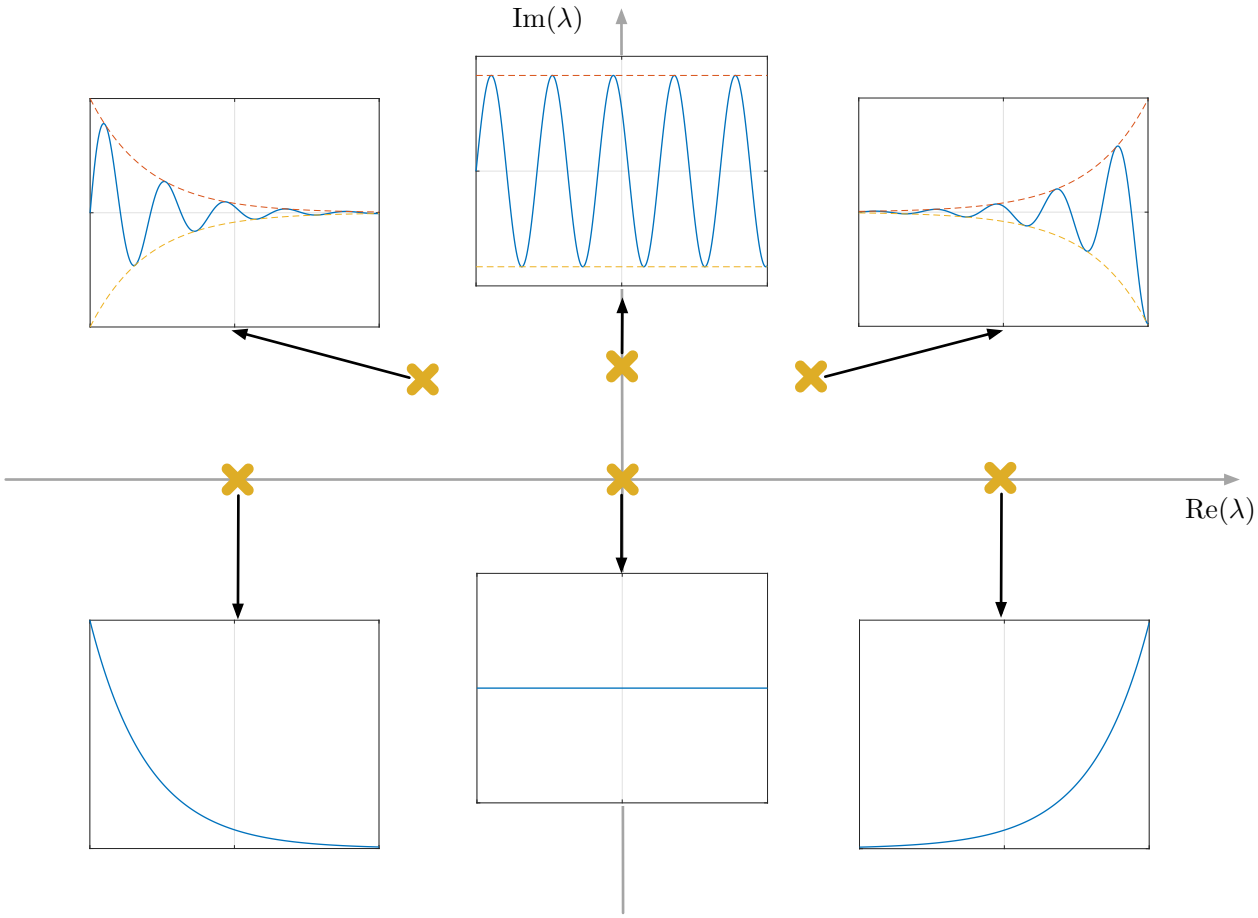


Figure 5: Zero input transition solution patterns for LTI systems with diagonalizable A matrix.

$t, t^2, \ldots, t^{r-1}$. The zero input transition is in turn a linear combination of the terms appearing in $\Phi(t)$; as a result, it may include additional terms with respect to the case where A is diagonalizable. We list the terms with which each eigenvalue of the form $\lambda = \sigma + j\omega$ contributes to the solution in Table 2.

If all eigenvalues have non-zero real part, a similar observation with the diagonalizable case pertains. However, in contrast to the diagonalizable case, if an eigenvalue has zero real part ($\lambda = 0$ or if $\lambda = j\omega$), then its contribution in the solution is not necessarily constant or periodic and may include terms

$$t, t^2, \ldots, t^{r-1} \quad \text{and/or} \quad t\sin\omega t, t\cos\omega t, \ldots, t^{r-1}\sin\omega t, t^{r-1}\cos\omega t.$$

These terms tend to infinity as $t \to \infty$, however, their linear combination may lead to cancellations so in certain occasions the solution may still be bounded. As this depends on the linear combination coefficients, this is now dictated by the eigenvectors (and how initial conditions relate to them), which for diagonalizable matrices did not play a role in assessing the behaviour of the system. To emphasize

the fact that the individual terms tend to infinity but their linear combination may still be bounded, we have highlighted the corresponding entries in Table 2.

Fact 7 (Stability of $\dot{x}(t) = Ax(t)$ with non-diagonalizable A)**.** *Consider an autonomous LTI system $\dot{x}(t) = Ax(t)$ with $A \in \mathbb{R}^{n\times n}$ non-diagonalizable. Let λ_i, $i = 1, \ldots, n$, be the eigenvalues of A (at least one of them would be repeated). We then have that the system is:*

- *Asymptotically stable (for all initial conditions) if and only if $\mathrm{Re}(\lambda_i) < 0$ for all $i = 1, \ldots, n$.*
- *Unstable (for some initial conditions) if there exists i such that the corresponding eigenvalue has $\mathrm{Re}(\lambda_i) > 0$.*

The unstable case is no longer "if and only" as in the diagonalizable case, as the system can potentially be also unstable if all eigenvalues have non-positive real part but at least one has zero real part. Note that if the repeated eigenvalues of a non-diagonalizable A matrix do not have zero real part, then the stronger results of Fact 6 would hold for the non-diagonalizable case as well.

We have analyzed the stability of zero input transitions; in case inputs are present, then the zero state solution becomes relevant to assess the stability of $x(t)$. Some of the obtained stability results extend to such cases. As an example, if all eigenvalues of the A matrix involved have negative real part, then if the input remains bounded the state $x(t)$ remains bounded as well.

3.2 Connections with transfer functions

Consider the state space description of an LTI system

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

and recall that $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^p$ is the output of the system. We have already discussed that the state $x(t)$ acts as an internal variable; we will thus treat the system from an input-output point of view and investigate connections with transfer functions in the frequency domain.

To this end, given the complex variable s denote by $\mathcal{L}\{x(t)\} = X(s)$ the Laplace transform of $x(t)$. Similarly, let $U(s)$ and $Y(s)$ denote the Laplace transforms of $u(t)$ and $y(t)$, respectively.

3.2.1 From state space to transfer functions

Given matrices (A, B, C, D) that encode the state space description of an LTI system, we first show how to obtain a transfer function $G(s) = \frac{Y(s)}{U(s)}$. To achieve this, we take the Laplace transform in both sides of the state space equation:

$$\begin{aligned} sX(s) - x_0 &= AX(s) + BU(s) \\ Y(s) &= CX(s) + DU(s), \end{aligned}$$

where $sX(s) - x_0$ is the Laplace transform of $\dot{x}(t)$, with x_0 denoting the initial condition of the state. We can solve with respect $X(s)$ in the first equation (note that all quantities are matrices or vectors so the multiplication order becomes important) to obtain $X(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}BU(s)$. Substituting it in the second equation we get

$$Y(s) = C(sI - A)^{-1}x_0 + C(sI - A)^{-1}BU(s) + DU(s).$$

We will now assume $x_0 = 0$. Assuming zero initial conditions stems from our objective to determine the transfer function of the system, i.e., a relationship between input and output. In other words it is as if we concentrate only on the zero state response. Notice that once a transfer function is determined we could take the inverse Laplace transform to obtain the zero state response; in case of a non-zero initial condition we could then add to the resulting solution the term corresponding to the zero input response $C\Phi(t)x_0$ to obtain the system's output solution.

Fact 8 (Transfer function given (A, B, C, D).)**.** *Assuming $x_0 = 0$, given matrices (A, B, C, D) the transfer function corresponding to an LTI system is given by*

$$G(s) = \frac{Y(s)}{U(s)} = C(sI - A)^{-1}B + D \in \mathbb{C}^{p\times m}.$$

This follows from the expression of $Y(s)$, taking $x_0 = 0$ and dividing with $U(s)$. Notice that the transfer function is rather a transfer matrix, whose elements are

complex (as they depend on s), and its dimension is $p \times m$, i.e., number of outputs times number of inputs. We could represent it as

$$G(s) = \begin{bmatrix} \frac{n_{11}(s)}{d_{11}(s)} & \frac{n_{12}(s)}{d_{12}(s)} & \cdots & \frac{n_{1m}(s)}{d_{1m}(s)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{n_{p1}(s)}{d_{p1}(s)} & \frac{n_{p2}(s)}{d_{p2}(s)} & \cdots & \frac{n_{pm}(s)}{d_{pm}(s)} \end{bmatrix} \in \mathbb{C}^{p \times m}.$$

The (i, j)-th element is the ratio between two polynomials of s, namely, $n_{ij}(s)$ and $d_{ij}(s)$, and captures the transfer function (scalar) between input j and a particular output i. In case $p = m = 1$, then we only have one input and one output in the system, and $G(s)$ reduces to transfer function of the scalar case.

Example 12. *Consider the LTI system corresponding to the RLC circuit of Figure 2 with $R = 3$, $L = 1$ and $C = 0.5$. Compute the transfer function of the system.*

Solution: *Under the given numerical values we have*

$$A = \begin{bmatrix} 0 & 2 \\ -1 & -3 \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } D = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The system has two outputs and one input, so the transfer function $G(s)$ will be a column vector with two elements. These elements encode the transfer function from the input to the first and the second output, respectively. Notice also that

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

By means of Fact 8, we then have that

$$\begin{aligned} G(s) &= C(sI - A)^{-1}B + D \\ &= \frac{1}{(s+1)(s+2)} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s+3 & 2 \\ -1 & s \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{2}{(s+1)(s+2)} \\ \frac{s}{(s+1)(s+2)} \end{bmatrix}. \end{aligned}$$

It should be remarked that, by taking the inverse Laplace transform of $Y(s)$ we

obtain the output solution $y(t)$ as this is defined in the time domain, i.e.,

$$y(t) = \mathcal{L}^{-1}\{Y(s)\} = C\mathcal{L}^{-1}\{(sI-A)^{-1}\}x_0 + \mathcal{L}^{-1}\{C(sI-A)^{-1}BU(s) + DU(s)\}$$
$$= C\Phi(t)x_0 + \int_0^t C\Phi(t-\tau)Bu(\tau)d\tau + Du(t).$$

By direct comparison between these two expressions we obtain that the state transition matrix is given by

$$\Phi(t) = \mathcal{L}^{-1}\{(sI-A)^{-1}\},$$

which in turn suggests yet another way of computing the state transition matrix, this time by taking the inverse Laplace transform of $(sI-A)^{-1}$.

3.2.2 From transfer functions to state space

We now investigate whether we could follow the opposite route and given a transfer function $G(s)$ obtain a state space description, i.e., matrices (A, B, C, D) such that $G(s) = C(sI-A)^{-1}B+D$. It turns out that this is possible, however, the resulting quadruple of matrices (A, B, C, D) is not unique and multiple choices exist. We refer to each choice (A, B, C, D) that results in the same transfer function $G(s)$ as a realization of $G(s)$.

To see that multiple realizations of $G(s)$ may exist, let $T \in \mathbb{R}^{n\times n}$ be any invertible matrix and consider the following coordinate trnasformation

$$\hat{x}(t) = Tx(t) \;\Rightarrow\; x(t) = T^{-1}\hat{x}(t).$$

Under this change of variables we obtain a new LTI system description

$$\dot{\hat{x}}(t) = T\dot{x}(t) = TAx(t) + TBu(t) \;\Rightarrow\; \dot{\hat{x}}(t) = TAT^{-1}\hat{x}(t) + TBu(t),$$
$$y(t) = Cx(t) + Du(t) \;\Rightarrow\; y(t) = CT^{-1}\hat{x}(t) + Du(t).$$

The new description has matrices $(\hat{A}, \hat{B}, \hat{C}, \hat{D}) = (TAT^{-1}, TB, CT^{-1}, D)$. Notice now that the transfer function $\hat{G}(s)$ of the new description is given by

$$\begin{aligned}
\hat{G}(s) &= \hat{C}(sI-\hat{A})^{-1}\hat{B} + \hat{D} \\
&= CT^{-1}(sI - TAT^{-1})^{-1}TB + D \\
&= CT^{-1}(sTT^{-1} - TAT^{-1})^{-1}TB + D \qquad \text{[by expressing } I = TT^{-1}\text{]} \\
&= C\cancel{T^{-1}}\cancel{T}^{I}(sI-A)^{-1}\cancel{T^{-1}}\cancel{T}^{I}B + D = G(s),
\end{aligned}$$

where in the last equality we used the fact that the inverse of the product of invertible matrices is the product of the inverses with reverse order. The new system description has the same transfer function with the original one. As this holds true for any coordinate transformation T, we have multiple realizations of $G(s)$. However, notice that all of them have the same D matrix.

Among those realizations we will provide a specific one for systems with a single input and a single output (hence the transfer function would be scalar). To this end, if after any pole-zero cancellations the transfer function is given by (notice that this corresponds to a system with $D = 0$ as this is strictly proper)

$$G(s) = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \ldots + b_n}{s^n + a_1 s^{n-1} + \ldots + a_n},$$

then the following system corresponds to a realization of $G(s)$:

$$\begin{aligned}
\dot{x}(t) &= \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \ldots & -a_1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(t) \\
y(t) &= \begin{bmatrix} b_n & b_{n-1} & b_{n-2} & \ldots & b_1 \end{bmatrix} x(t).
\end{aligned}$$

This realization is known as controllable canonical form; we will revisit it in the sequel and the term "controllable" will become clear.

3.2.3 Eigenvalues vs. Poles

By the definition of the matrix inverse in Section 9.1.2, we can equivalently rewrite the transfer function as

$$G(s) = C \, \frac{\operatorname{adj}(sI - A)}{\det(sI - A)} \, B + D,$$

where $\operatorname{adj}(sI - A)$ is the adjoint matrix and $\det(sI - A)$ the determinant of $sI - A$. Assuming no pole-zero cancellations, notice that all elements (transfer functions) in $G(s)$ would have the same denominator, which is in turn equal to $\det(sI - A)$.

As such, the poles of $G(s)$ are given by the roots of this determinant when equated with zero, i.e.,

$$\text{poles of } G(s): \quad \text{roots of } \det(sI - A) = 0.$$

This is clearly related to the characteristic polynomial of A, and as a result poles are related with the eigenvalues of A. This relation is summarized below.

Fact 9 (Eigenvalues vs. Poles)**.** *If there are no pole-zero cancellations, the poles of $G(s)$ coincide with the eigenvalues of A.*

The previous fact suggests that if there are no pole-zero cancellations, eigenvalues and poles contain the same information about a system. In particular, the established stability results hold true, but rather than looking at the eigenvalues of A, one could inspect the poles of $G(s)$. The distinction between diagonalizability and non-diaginalizability of A refers to whether the poles are distinct or repeated.

Overall, state space analysis and transfer functions exhibit several similarities but also differences. Their respective advantages are summarized below:

1. Advantages of transfer functions over state space.

    - Lead to algebraic manipulations rather than solving ODEs.
    - Same transfer function for all coordinate transformations.
    - Easier to compute zero state reposponses (no need to compute a convolution integral).
    - We could have transfer functions for systems that do not admit a state space description (e.g., presence of delays).

2. Advantages of state space over transfer functions.

    - Preserves physical intuition about the underlying system.
    - Contains information about parts of the system that may be lost in a transfer function (pole-zero cancellations). These correspond to what we will refer to as uncontrollable and unobservable parts in the next chapter.

3.3 Summary

This chapter provided the means to assess the stability of LTI systems and discussed the connections between state space representations and transfer functions. The main learning outcomes of the chapter can be summarized as follows:

1. Stability of autonomous LTI systems of the form $\dot{x}(t) = Ax(t)$. If matrix A is diagonalizable, then the system is said to be (see Fact 6)

   - *Stable* if and only if all eigenvalues of A have non-positive real part (we allow them to be zero).
   - *Asymptotically stable* if and only if all eigenvalues of A have negative real part.
   - *Unstable* if and only if there exists at least one eigenvalue of A with positive real part.

   If A is non-diagonalizable, then the asymptotic stability condition above remains unaltered, while the instability statement becomes an "if" condition. The stability result can no longer be claimed: whether the system will be stable or whether the state will grow to infinity if the system has eigenvalues with zero real part would depend in this case on the eigenvectors (see Fact 7).

2. Connections between state space and transfer functions.

   - *From state space to transfer functions.* Given an LTI system encoded by (A, B, C, D), the transfer function (matrix) of the system can be computed by means of (see Fact 8)
   $$G(s) = \frac{Y(s)}{U(s)} = C(sI - A)^{-1}B + D,$$
   where the number of rows of $G(s)$ corresponds to the number of outputs, and the number of columns to the number of inputs of the system.
   - *From transfer functions to state space.* Given a transfer function $G(s)$ there exist multiple realizations (A, B, C, D) of LTI systems that are captured from an input-output point of view by the same transfer functions.
   - *Eigenvalues vs. Poles.* If there are no pole-zero cancellations, then the poles of $G(s)$ coincide with the eigenvalues of A (see Fact 9).

4 Structural properties of linear systems

In this chapter we address the following two fundamental questions:

1. Does there exist an input so that we can steer the system from an initial state to any given final one? The answer to this question is related to the notion of *controllability*.
2. Could we infer the state of the system if we only have access to the output (measurements)? The answer to this question is related to the notion of *observability*.

In the next sections we analyze the controllability and observability properties for LTI systems. In particular, we show that the notion of controllability depends on matrices A and B (that capture the relationship between input and state), while the notion of observability depends on matrices A and C (that capture the relationship between state and output) of the LTI system's state space description.

4.1 Controllability

We consider an LTI system with a state space description governed by matrices (A, B, C, D). For simplicity we assume again that the initial time is zero, i.e., $t_0 = 0$. We define controllability as stated below.

Definition 1 (Controllability)**.** *Consider an arbitrary t. We say that a system is controllable over the time interval $[0, t]$ if one of the following statements holds:*

1. *For any given initial state x_0 and terminal state x_1, there exists an input function $u(\cdot) : [0, t] \to \mathbb{R}^m$ such that $x(t) = x_1$.*
2. *For any given terminal state x_1, there exists an input function $u(\cdot) : [0, t] \to \mathbb{R}^m$ such that $x(t) = x_1$, starting at $x_0 = 0$.*

The second statement follows directly from the first one. To see this, notice that the "only if" part is a direct consequence of the fact that if the system is controllable,

then by the first statement there exists an input function such that the system can reach any terminal state x_1, from any initial condition x_0, hence also from an initial state $x_0 = 0$. To show the "if" part it suffices to show that if there exists an input to drive the system from $x(0) = 0$ to any arbitrary terminal state, then we can also steer the system from $x_0 \neq 0$ to any terminal state x_1. Fix any such $x_0 \neq 0$. If there exists an input to steer the system from $x(0) = 0$ to the terminal state $x_1 - \Phi(t)x_0$, then the same input would drive the system from x_0 to x_1.

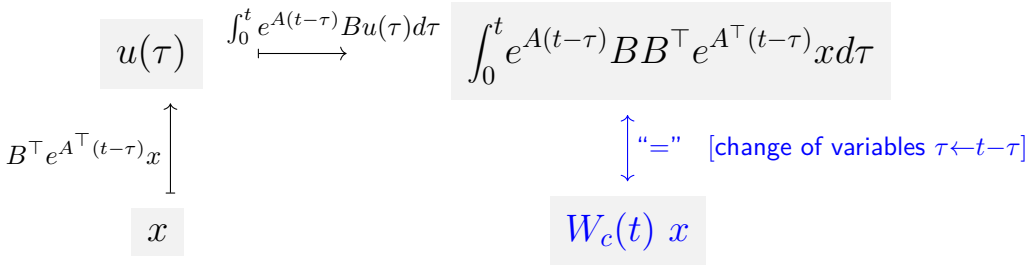
Definition 1 is natural, however, it cannot be easily checked for arbitrary systems. To this end, the following fact provides an alternative condition.

Fact 10 (Controllability gramian)**.** *A system is controllable over the time interval* $[0, t]$ *if and only if the so called controllability gramian*

$$W_c(t) = \int_0^t e^{A\tau} B B^\top e^{A^\top \tau} d\tau \in \mathbb{R}^{n\times n},$$

is invertible.

It can be shown that $W_c(t) = W_c(t)^\top \succeq 0$, and also that $W_c(t)$ is invertible for some $t > 0$ if and only if it is invertible for all $t > 0$. This implies that time is not important as far as controllability is concerned. We will not provide a formal proof for Fact 10, but an informal one providing some intuition behind the definition of the gramian $W_c(t)$, and the reasons why this is related to controllability. To this end, fix an arbitrary $x \in \mathbb{R}^n$ and consider the diagram below.



The controllability gramian $W_c(t)$ can be constructed as the matrix representation emanating from the composition of two mappings: i) one from x to u, namely, $u(\tau) = B^\top e^{A^\top (t-\tau)} x$; and, ii) one from u to $\int_0^t e^{A(t-\tau)} B u(\tau) d\tau$. This composition can be thus written as

$$\int_0^t e^{A(t-\tau)} B B^\top e^{A^\top (t-\tau)} x d\tau = \int_0^t e^{A\tau} B B^\top e^{A^\top \tau} x d\tau = W_c(t) x,$$

where the first equality is due to the change of variables $\tau \leftarrow t - \tau$, and the second one follows from the definition of the controllability gramian $W_c(t)$. The following observations are in order:

- By the second part of Definition 1, to assess the controllability of a system, we can assume without loss of generality that the initial condition is $x_0 = 0$. Therefore, the second mapping in the composition defining $W_c(t)$ is the zero state transition (see 2.2). It follows then that varying x (hence also the control input as this is generated by the first mapping), $W_c(t)x$ corresponds to the states that can be reached from zero.
- The system is controllable if we could reach an arbitrary terminal state x_1 starting from zero. For any fixed x_1, if $W_c(t)$ is invertible, taking $x = W_c(t)^{-1}x_1$, our choice for u (see diagram) would result in steering the system to $W_c(t)x = x_1$. This justifies why invertibility of $W_c(t)$ is related to controllability.
- We have seen in the first point above that the set of states that can be reached (using some control input) is given by $W_c(t)x$, i.e., by the range space of $W_c(t)$. If $W_c(t)$ is invertible, we can reach any state, i.e.,

$$\text{range}(W_c(t)) = \mathbb{R}^n \;\Leftrightarrow\; \text{null}(W_c(t)) = \{0\}.$$

The controllability gramian condition is more general, and could be employed for linear time-varying systems as well. However, despite being useful in identifying the set of states that can be reached from the origin, the condition of Fact 10 is not easy to check in general as it involves computation of the matrix exponential and integration. However, it can be employed to develop an alternative condition for LTI systems, which is easier to check.

Fact 11 (Controllability matrix)**.** *A system is controllable over the time interval* $[0, t]$ *if and only if the so called controllability matrix*

$$P = \begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n \times nm},$$

is full rank, i.e., $\text{rank}(P) = n$ *(notice that the rank of* P *is at most equal to the number of rows* n*, as it has fewer rows than columns). We equivalently say that* (A, B) *is in this case controllable.*

Proof of Fact 11. By Fact 10 the system is controllable if $W_c(t)$ is invertible. We have already seen (third observation above) that this is equivalent to $\mathrm{null}(W_c(t)) = \{0\}$. By the definition of the null space, the latter can also be written as

$$W_c(t)x = 0 \;\Leftrightarrow\; x = 0.$$

We will show that for the controllability matrix

$$W_c(t)x = 0 \;\Leftrightarrow\; P^\top x = 0.$$

This implies then that $W_c(t)$ is invertible if and only if P has rank n, which by Fact 10 establishes that the system is controllable.

To prove the last equivalence, notice first that*

$$W_c(t)x = 0 \;\Leftrightarrow\; B^\top e^{A^\top \tau} x = 0, \text{ for all } \tau \in [0, t].$$

By the Taylor series expansion around $\tau = 0$ we obtain that

$$\begin{aligned} B^\top e^{A^\top \tau} x &= B^\top e^{A^\top \tau} x\big|_{\tau=0} + \frac{d}{d\tau} B^\top e^{A^\top \tau} x\big|_{\tau=0} \tau + \dots \\ &\qquad + \frac{d^{n-1}}{d\tau^{n-1}} B^\top e^{A^\top \tau} x\big|_{\tau=0} \frac{\tau^{n-1}}{(n-1)!} + \dots \\ &= B^\top x + B^\top A^\top x \tau + \dots + B^\top (A^{n-1})^\top x \frac{\tau^{n-1}}{(n-1)!} + \dots \end{aligned}$$

Therefore, $B^\top e^{A^\top \tau} x = 0$ is equivalent to

$$\begin{aligned} &B^\top x = 0, B^\top A^\top x = 0, \dots, B^\top (A^{n-1})^\top x = 0 \\ \Leftrightarrow\; &\begin{bmatrix} B^\top \\ B^\top A^\top \\ \vdots \\ B^\top (A^{n-1})^\top \end{bmatrix} x = 0 \;\Leftrightarrow\; P^\top x = 0. \end{aligned}$$

Notice that we consider only the first $n-1$ terms of the Taylor series expansion to be identically equal to zero, as by the Cayley-Hamilton theorem (see Theorem 8), taking the transpose and multiplying from the left with $B^\top$, we obtain

$$B^\top (A^n)^\top = -a_1 B^\top (A^{n-1})^\top - \dots - a_{n-1} B^\top A^\top - a_n B^\top,$$

*To see this notice that $W_c(t) = 0$ is equivalent to $e^{A\tau} B B^\top e^{A^\top \tau} x = 0$ for all $\tau \in [0, t]$, which in turn implies that $B^\top e^{A^\top \tau} x$ is in the null space of $e^{A\tau} B$. As such, it will be orthogonal to the range space of its transpose (rowspace), i.e., to $B^\top e^{A^\top \tau} x$. However, the only way it can happen that a vector is orthogonal to itself is if this vector is zero, i.e., $B^\top e^{A^\top \tau} x = 0$.

where $a_1, \ldots, a_n$ are the coefficients of the characteristic polynomial of A. Hence, if the first $n-1$ terms are zero, $B^\top (A^n)^\top$ (and subsequently all higher order terms) will be zero as well. We have thus shown that $W_c(t)x = 0$ is equivalent to $P^\top x = 0$, which in turn implies that P has to be full rank, thus concluding the proof.

A direct consequence of the proof of Fact 11, is that the set of states we can reach is given by $\mathrm{range}(W_c(t)) = \mathrm{range}(P)$. Overall, to decide whether a given LTI system is controllable results in checking the controllability matrix condition of Fact 11. We illustrate this by means of an example.

Example 13. *Consider the LTI system corresponding to the RLC circuit of Figure 2 with $R = 3$, $L = 1$ and $C = 0.5$. Check whether the system is controllable.*

Solution: *Under the given numerical values we have*

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

Since this is a second order system, i.e., $n = 2$, the controllability matrix is thus given by

$$P = \begin{bmatrix} B & AB \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & -3 \end{bmatrix}.$$

The rows of P are linearly independent (here P is a square matrix, so we could equivalently check that $\det(P) \neq 0$), hence P is a full rank matrix. As a result, the system is controllable.

4.2 Observability

We consider again an LTI system with a state space description governed by matrices (A, B, C, D), and assume that the initial time is zero, i.e., $t_0 = 0$. We define observability as stated below.

Definition 2 (Observability)**.** *Consider an arbitrary t. We say that a system is observable over the time interval $[0, t]$ if one of the following statements holds:*

1. *Given an input function $u(\cdot) : [0, t] \to \mathbb{R}^m$, having access to the output function $y(\cdot) : [0, t] \to \mathbb{R}^p$ allows us to uniquely determine the system state $x(\tau)$, for all $\tau \in [0, t]$.*
2. *Given an input function $u(\cdot) : [0, t] \to \mathbb{R}^m$, having access to the output function $y(\cdot) : [0, t] \to \mathbb{R}^p$ allows us to uniquely determine the initial state x_0.*

The second statement follows directly from the first one. To see this, recall that the state of the system is given by $x(t) = \Phi(t)x_0 + \int_0^t \Phi(t-\tau)Bu(\tau)d\tau$. It can be thus observed that given $u(\cdot)$, to infer $x(\tau)$ for all $\tau \in [0, t]$, it suffices to infer the initial condition x_0.

In a sense dual to controllability, we provide the following two conditions to check whether the system is observable. As with controllability, the second one is straightforward to check once we have access to the system's state space description.

Fact 12 (Observability gramian & matrix)**.** *A system is observable over the time interval $[0, t]$ if and only if*

1. *The so called observability gramian*

$$W_o(t) = \int_0^t e^{A^\top \tau} C^\top C e^{A\tau} d\tau \in \mathbb{R}^{n\times n},$$

   is invertible.

2. *The so called observability matrix*

$$Q = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np\times n},$$

is full rank, i.e., $\mathrm{rank}(Q) = n$ *(notice that the rank of* Q *is at most equal to the number of columns* n*, as it has fewer columns than rows). We equivalently say that* (A, C) *is in this case observable.*

We will not provide a formal proof for this fact, but the following interpretations of the observability matrix condition:

- Assume that under the same input $u(\cdot)$, two different initial conditions $x_0 \neq \hat{x}_0$ lead to the same output $y(\cdot)$. Since the initial condition cannot be uniquely determined from the output, then this implies that the system is unobservable. By the definition of the output solution (and recalling that $\Phi(t) = e^{At}$) this is equivalent to

$$Ce^{A\tau}x_0 + \int_0^\tau Ce^{A(\tau-s)}Bu(s)ds + Du = Ce^{A\tau}\hat{x}_0 + \int_0^\tau Ce^{A(\tau-s)}Bu(s)ds + Du$$
$$\Leftrightarrow \; Ce^{A\tau}(x_0 - \hat{x}_0) = 0, \text{ for all } \tau \in [0, t].$$

  The last statement implies that the system is unobservable (recall we cannot distinguish between x_0 and $\hat{x}_0$) if and only if $\mathrm{null}(Ce^{A\tau}) \neq \{0\}$. Equivalently the system is observable if and only if $\mathrm{null}(Ce^{A\tau}) = \{0\}$, i.e.,

$$Ce^{A\tau}x = 0 \;\Leftrightarrow\; x = 0.$$

  This provides some insight on why the term $Ce^{A\tau}$ appears in the observability gramian condition. Moreover, by performing a Taylor series expansion of $Ce^{A\tau}x$ around $\tau = 0$ as in the proof of Fact 11, then we obtain that $Cx = CAx = \ldots = CA^{n-1}x = 0$. Stacking these conditions one under the other gives rise to the observability matrix condition of Fact 12.

- Consider differentiating the output $y(t) = Cx(t) + Du(t)$, using the fact that $\dot{x}(t) = Ax(t) + Bu(t)$. This leads to (consider all derivatives well defined)

$$\begin{aligned}
y(t) &= Cx(t) + Du(t) \\
\dot{y}(t) &= C\dot{x}(t) + D\dot{u}(t) = CAx(t) + CBu(t) + D\dot{u}(t) \\
\ddot{y}(t) &= CA^2x(t) + CABu(t) + CB\dot{u}(t) + D\ddot{u}(t) \\
&\ldots \text{ derivatives up to order } n-1.
\end{aligned}$$

Stacking these equations one under the other and setting $t = 0$ we obtain

$$\underbrace{\begin{bmatrix} y(0) \\ \dot{y}(0) \\ \vdots \\ y^{(n-1)}(0) \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}}_{Q} x_0 + \underbrace{\begin{bmatrix} D & 0 & \dots & 0 \\ CB & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{n-2}B & CA^{n-1}B & \dots & D \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} u(0) \\ \dot{u}(0) \\ \vdots \\ u^{(n-1)}(0) \end{bmatrix}}_{U},$$

where Y, U and M are of appropriate dimension, and Q is the observability matrix. This is a system of linear equations with respect to x_0. Therefore, if we have access to the input and output functions $u(\cdot)$ and $y(\cdot)$ (and hence also their derivatives), we can infer x_0. In particular, this system has more equations than unknowns, while if we only have one output ($p = 1$), then Q is a square matrix, hence

$$x_0 = Q^{-1}(Y - MU).$$

This derivation provides an additional intuition on why the observability matrix should be full rank for the system to be observable. However, that way of inferring x_0 is not practical, as measurements are typically affected by noise, and taking derivatives of the output (measurements) is likely to amplify that noise. We will provide efficient ways to construct an estimate of the state when we will discuss about linear state observers in the sequel.

✎ **Example 14.** *Consider the the equations of motion of the pendulum of Figure 1, and its linearization around the origin as this was derived in Example 1. Let $g = 10$, $l = 10$, $d = 1$ and $m = 1$. Check whether the system is observable.*

Solution: *Under the given numerical values we have*

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

Since this is a second order system, i.e., $n = 2$, the observability matrix is thus given by

$$Q = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The columns of Q are linearly independent (here Q is the identity matrix), hence Q is a full rank matrix. As a result, the system is observable.

4.3 Summary

This chapter provided the means to decide about certain structural properties of linear systems, namely *controllability and observability*. The main learning outcomes of the chapter can be summarized as follows:

1. *Controllability:* Informally, controllability refers to the ability to steer the system using some input from any initial state to any desired terminal state. For LTI systems to decide whether the system is controllable it suffices to check the following condition:

   Controllable LTI system if and only if

   $$P = \begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n \times nm} \text{ is full rank.}$$

2. *Observability:* Informally, observability refers to the ability to infer the state of the system if we have access to its input and output (measurements). For LTI systems to decide whether the system is observable it suffices to check the following condition:

   Observable LTI system if and only if

   $$Q = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np \times n} \text{ is full rank.}$$

5 Minimum energy control & Kalman decomposition

5.1 Minimum energy control

We have already seen that if the system is controllable, then there exists a control input $u(\cdot)$ as a function of time to steer the system from the initial condition $x_0 = 0$ (note that due to the second part of Definition 1 this choice is without loss of generality) to an arbitrary terminal state x_1. However, this is an existential statement. The following questions remain still.

1. Can we construct an input $u(\cdot)$ that can drive the system from $x_0 = 0$ to an arbitrary x_1?
2. If yes, can we do this in a minimum effort fashion?

We will show that for controllable systems indeed it is possible to design controllers that can lead the system to the desired terminal state. The controller that we will develop in this chapter, however, we will be open loop; we will construct feedback controllers in the subsequent chapters. The constructed controller will though be a minimum effort one (in a way that would be made precise below; we will call such controllers *minimum energy controllers* – their relationship with energy will be discussed in the sequel.

5.1.1 Controller design

We consider the energy of an input signal to be given by

$$\textit{energy:} \quad \int_0^t u(\tau)^\top u(\tau) d\tau = \int_0^t \|u(\tau)\|^2 d\tau.$$

We first provide a closed form expression of the minimum energy input, and then provide a couple of interpretations on what we mean by the term "energy".

Fact 13. *Consider an LTI system and assume that it is controllable. The control input that steers the system from $x_0 = 0$ to $x(t) = x_1$, and has the minimum energy, is given by*

$$u(\tau) = B^\top e^{A^\top (t-\tau)} W_c(t)^{-1} x_1 \text{ for all } \tau \in [0, t],$$

where $W_c(t)$ is the controllability gramian.

Note that $W_c(t)$ is invertible, as the system is assumed to be controllable (see Fact 10). We now provide a proof of this fact.

Proof of Fact 13. We first show that the candidate control input can indeed steer the system from $x_0 = 0$ to a given x_1. To see this notice that by the schematic diagram below Fact 10, and since $W_c(t)$ is invertible, we can

$$\begin{array}{rlll} & \text{reach} & W_c(t)x & \text{using} \quad u(\tau) = B^\top e^{A^\top(t-\tau)}x \\ \underset{x=W_c(t)^{-1}x_1}{\Longleftrightarrow} & \text{reach} & x_1 & \text{using} \quad u(\tau) = B^\top e^{A^\top(t-\tau)}W_c(t)^{-1}x_1. \end{array}$$

The latter is the candidate input, hence we have shown that it is indeed possible to reach x_1 with that input.

We will now show that the candidate input has the minimum energy. To this end, consider any other input that steers the system from $x_0 = 0$ to x_1. Represent such an input by $u(\tau) + \hat{u}(\tau)$, i.e., as the sum of the candidate input plus a perturbation component $\hat{u}(\tau)$. By the definition of the solution we have that

$$\begin{aligned} x_1 &= e^{At}\cancel{x_0}^{0} + \int_0^t e^{A(t-\tau)}B(u(t) + \hat{u}(t))d\tau \\ &= \underbrace{\int_0^t e^{A(t-\tau)}Bu(t)d\tau}_{x_1} + \int_0^t e^{A(t-\tau)}B\hat{u}(t)d\tau, \end{aligned}$$

where the first term in the last step is equal to x_1 as the candidate input u was also shown to steer the system from $x_0 = 0$ to x_1. We thus have that

$$\int_0^t e^{A(t-\tau)}B\hat{u}(t)d\tau = 0.$$

Denote the energy of the candidate input u by $E(t) = \int_0^t u(\tau)^\top u(\tau)d\tau$, and the energy of $u(\tau) + \hat{u}(\tau)$ by $\hat{E}(t)$. Note that this is given by

$$\begin{aligned} \hat{E}(t) &= \int_0^t (u(\tau) + \hat{u}(\tau))^\top (u(\tau) + \hat{u}(\tau))d\tau \\ &= \underbrace{\int_0^t u(\tau)^\top u(\tau)d\tau}_{E(t)} + \int_0^t \hat{u}(\tau)^\top u(\tau)d\tau + \int_0^t u(\tau)^\top \hat{u}(\tau)d\tau + \int_0^t \hat{u}(\tau)^\top \hat{u}(\tau)d\tau \\ &= E(t) + 2\int_0^t u(\tau)^\top \hat{u}(\tau)d\tau + \int_0^t \hat{u}(\tau)^\top \hat{u}(\tau)d\tau, \end{aligned}$$

where the last equality is due to the fact $u(\tau)^\top \hat{u}(\tau)$ is a scalar so it coincides with its transpose. By substituting the representation of the candidate u, the second term that appears in the right-hand side of the expression for $\hat{E}(t)$ becomes (recall that $W_c(t) = W_c(t)^\top$)

$$\int_0^t u(\tau)^\top \hat{u}(\tau) d\tau = \int_0^t x_1^\top W_c(t)^{-1} e^{A(t-\tau)} B\hat{u}(\tau) d\tau = 0,$$

since we have shown that $\int_0^t e^{A(t-\tau)} B\hat{u}(t) d\tau = 0$. We then have that

$$\hat{E}(t) = E(t) + \underbrace{\int_0^t \hat{u}(\tau)^\top \hat{u}(\tau) d\tau}_{\geq 0} \;\Rightarrow\; \hat{E}(t) \geq E(t).$$

This implies that the candidate u results in lower energy from any other controller that can steer the state to x_1, hence it is the minimum energy controller.

✎ **Example 15.** *Consider the LTI system corresponding to the RC circuit of Example 4 with $R = 1$ and $C = 1$. Determine the minimum energy controller to steer the system from $x_0 = 0$ to $x_1 = 0.5$ at time t.*

Solution: *Under the given numerical values notice that*

$$A = -\frac{1}{RC} = -1 \;\; \textit{and} \;\; B = \frac{1}{RC} = 1 \quad \textit{[scalars]}.$$

The matrix exponential is in this case $e^{A\tau} = e^{-\tau}$, hence the controllability gramian $W_c(t)$ is given by

$$\begin{aligned} W_c(t) &= \int_0^t e^{A\tau} B B^\top e^{A^\top \tau} d\tau \\ &= \int_0^t e^{-2\tau} d\tau = \frac{1 - e^{-2t}}{2}. \end{aligned}$$

For any $\tau \in [0, t]$, the minimum energy controller is then given by

$$\begin{aligned} u(\tau) &= B^\top e^{A^\top (t-\tau)} W_c(t)^{-1} x_1 \\ &= 1 \cdot e^{-(t-\tau)} \cdot \left(\frac{1 - e^{-2t}}{2}\right)^{-1} \cdot 0.5 = \frac{e^{\tau - t}}{1 - e^{-2t}}. \end{aligned}$$

Notice that the control input tends to infinity as t tends to zero, implying that we need infinite control effort if we want to steer our system in "zero" time.

5.1.2 Energy interpretation

Consider first the mathematical expression for the minimum energy $E(t)$ associated with the input signal

$$u(\tau) = B^\top e^{A^\top(t-\tau)} W_c(t)^{-1} x_1 \text{ for all } \tau \in [0,t],$$

that steers the system to x_1. This is given by

$$\begin{aligned} E(t) &= \int_0^t u(\tau)^\top u(\tau) d\tau \\ &= \int_0^t x_1^\top W_c(t)^{-1} e^{A(t-\tau)} B B^\top e^{A^\top(t-\tau)} W_c(t)^{-1} x_1 d\tau \\ &= x_1^\top W_c(t)^{-1} \underbrace{\int_0^t e^{A(t-\tau)} B B^\top e^{A^\top(t-\tau)} d\tau}_{W_c(t)} W_c(t)^{-1} x_1 \\ &= x_1^\top W_c(t)^{-1} x_1. \end{aligned}$$

The following remarks are in order.

- The energy expression is quadratic in x_1. This implies that the further away we want to steer a system (hence the higher $\|x_1\|$), the more energy we need to expend. This is in line with our intuition.
- The energy expression is also proportional with respect to $W_c(t)^{-1}$, which reflects how controllable a system is. In particular, if a system is controllable this is well defined since $W_c(t)$ is invertible; however, the further away a system is from being controllable, the closest $W_c(t)$ is to become singular. This in turn implies that the closest a system is to be uncontrollable, the more energy we would need to steer it to a given terminal state x_1. In the limiting case we would need infinite energy.
- Very often the quantity $E(t) = \int_0^t u(\tau)^\top u(\tau) d\tau$ is related to physical energy. For example, if the input $u(\tau)$ represents voltage across a certain component in a circuit, then $E(t)$ is related to the energy of the system modulo some proportionality constant (e.g., related to some resistance).

It should be noted that the minimum energy controller is optimal with respect to a certain performance criterion, namely, the energy $\int_0^t u(\tau)^\top u(\tau) d\tau$ which depends solely on the input. We will see in Chapter 8 how to construct optimal controllers for more general performance criteria that involve the state as well.

5.2 Kalman decomposition

Controllability and observability carry important information about a system. In fact, we have already seen that controllability can be exploited to design controllers – minimum energy ones – and we will see alternative control design procedures in the sequel. Moreover, we will also discuss how observability can be exploited to design state estimators and infer the state of the system.

However, not all systems are controllable and/or observable. It is thus useful if we can decompose a systems into subsystems that exhibit these properties. In fact, the so called *Kalman decomposition* provides the means to achieve this. We summarize this below.

Theorem 3 (Kalman decomposition)**.** *An LTI system can be brought, by means of an appropriate coordinate transformation, to the following form*

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\\ \dot{x}_3(t)\\ \dot{x}_4(t)\end{bmatrix} = \overbrace{\begin{bmatrix}A_{11} & 0 & A_{13} & 0\\ A_{21} & A_{22} & A_{23} & A_{24}\\ 0 & 0 & A_{33} & 0\\ 0 & 0 & A_{43} & A_{44}\end{bmatrix}}^{A}\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\\ x_4(t)\end{bmatrix} + \overbrace{\begin{bmatrix}B_1\\ B_2\\ 0\\ 0\end{bmatrix}}^{B} u(t)$$

$$y(t) = \underbrace{\begin{bmatrix}C_1 & 0 & C_3 & 0\end{bmatrix}}_{C} x(t) + Du(t),$$

where the state is partitioned into

$$x(t) = \begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\\ x_4(t)\end{bmatrix} \begin{matrix}\textit{controllable + observable}\\ \textit{controllable + unobservable}\\ \textit{uncontrollable + observable}\\ \textit{uncontrollable + unobservable}\end{matrix}$$

Morover, the eigenvalues of the new A matrix are the eigenvalues of the matrices A_{11}, A_{22}, A_{33}, and A_{44}.

Note that under the Kalman decomposition states $x_1(t)$ and $x_2(t)$ correspond to a controllable subsystem, while $x_1(t)$ and $x_3(t)$ to an observable one. As such, the corresponding pairs of matrices satisfy the controllability and observability matrix

condition, respectively.

$$\textit{controllable:} \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \textit{observable:} \begin{bmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{bmatrix}, \begin{bmatrix} C_1 & C_3 \end{bmatrix}.$$

Figure 6 provides a schematic illustration of Kalman decomposition. Note that the arrows do not represent input-output signals, but show pictorially interdependencies among the different subsystems. For example, the three arrows entering the second subsystem imply that the evolution of $x_2(t)$ depends on all other three states, each of them being weighted according to the submatrices A_{21}, A_{23} and A_{24}. Effectively, the arrows entering a subsystem correspond to the non-zero contribution from other subsystems in the specific matrix row of the Kalman decomposition.

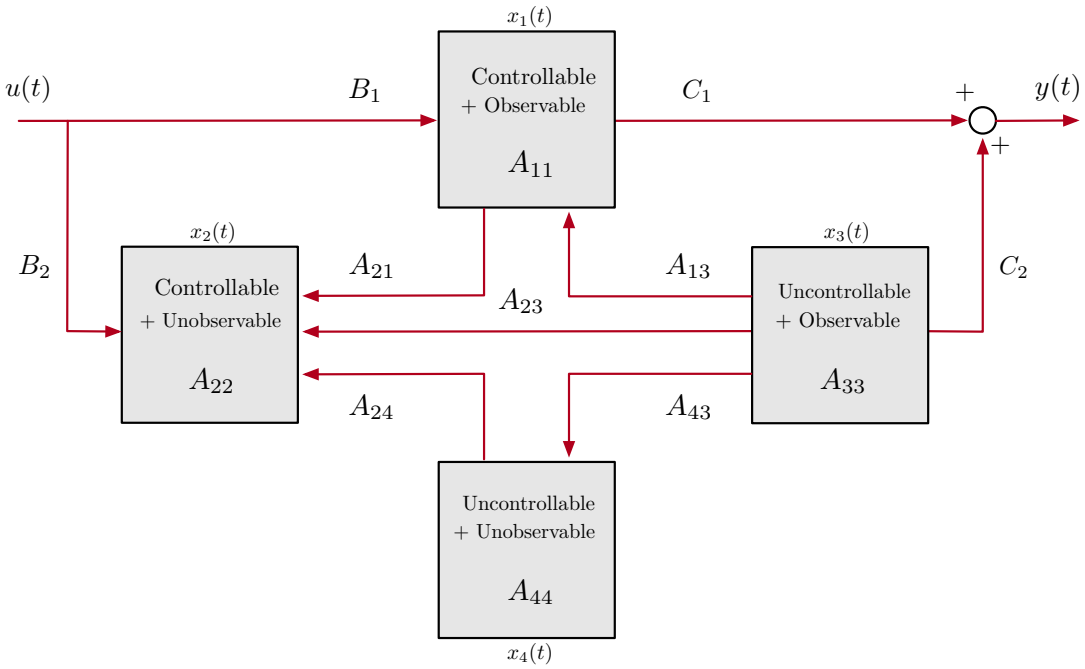


Figure 6: Schematic diagram for Kalman decomposition. Note that the arrows do not represent input-output signals, but show pictorially interdependencies among the different subsystems.

Notice that some subsystems are uncontrollable and/or unobservable. In this case we cannot control (steer arbitrarily) or observe (estimate) that part of the state; however, we would like these parts to be stable. Since stability is dictated by the eigenvalues of the system, and these correspond to the eigenvalues of the diagonal blocks, we would like:

1. The eigenvalues of A_{33} and A_{44} (uncontrollable part) to have negative real part. In this case the system is called *stabilizable*.

2. The eigenvalues of A_{22} and A_{44} (unobservable part) to have negative real part. In this case the system is called *detectable*.

Designing a controller for stabilizable systems (or an observer as we will see in the sequel for detectable ones) will ensure that the overall system is asymptotically stable as the part that we cannot control is asymptotically stable as well. Finally, it should be mentioned that the uncontrollable and unobservable parts lead to pole zero cancellations in the associated transfer function.

✎ **Example 16.** *Consider an LTI system with $x(t) \in \mathbb{R}^2$, written in state-space form as*

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

Comment on which states are (un)controllable and (un)observable.

Solution: *Note that the system is already in the Kalman decomposition form. To see this, consider Theorem 3, and select the subsystem corresponding to the first and third row and column of the associated state-space matrices. This results in the sub-matrices*

$$\begin{bmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{bmatrix}, \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \text{ and } \begin{bmatrix} C_1 & C_3 \end{bmatrix}.$$

By inspecting this subsystem, and setting

$$A_{11} = -1,\ A_{13} = 1,\ A_{33} = 1,\ B_1 = 1,\ C_1 = 1,\ \text{and } C_3 = 0,$$

we observe that the given system has state-space matrices that exhibit this form. In particular, the two states of the given system correspond to states $x_1(t)$ and $x_2(t)$ in Theorem 3. As such, the first state is controllable and observable, while the second state is observable but uncontrollable.
Notice that this is expected as the second state is not affected by the input directly, nor indirectly through the other state, and evolves autonomously. In fact, the system is not even stabilizable as A_{33} is non-negative. The fact

that the system is uncontrollable implies that there would be some pole-zero cancellation in the associated transfer function. To verify this, notice that

$$G(s) = C(sI - A)^{-1}B + \cancelto{0}{D}$$
$$= \frac{1}{(s+1)(s-1)}\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}s-1 & 1\\ 0 & s+1\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \frac{\cancel{(s-1)}}{(s+1)\cancel{(s-1)}} = \frac{1}{s+1}.$$

5.3 Summary

This chapter showed how minimum energy controllers can be designed for controllable systems, and discussed Kalman decomposition as a state partitioning to (un)controllable and (un)observable subsystems. The main learning outcomes of the chapter can be summarized as follows:

- *Minimum energy control.* Minimum energy controllers are designed for controllable systems, and are able to steer the system from $x_0 = 0$ to an arbitrary terminal state x_1, while minimizing the energy (or control effort) $\int_0^t u(\tau)^\top u(\tau) d\tau$. Such controllers and their energy are given by

  $$\text{control input:}\quad u(\tau) = B^\top e^{A^\top(t-\tau)} W_c(t)^{-1} x_1 \text{ for all } \tau \in [0, t],$$
  $$\text{energy:}\quad E(t) = x_1^\top W_c(t)^{-1} x_1.$$

  Minimum energy controllers are related to energy: the further away a system is from being controllable, the closest the controllability gramian $W_c(t)$ is to become singular. This in turn implies that controller's energy will grow, implying that we need more energy/effort to steer the system to x_1.

- *Kalman decomposition.* It shows that through an appropriate coordinate transformation a system can be decomposed into four subsystems, namely,

  1. $x_1(t)$: controllable + observable;
  2. $x_2(t)$: controllable + unobservable;
  3. $x_3(t)$: uncontrollable + observable;
  4. $x_4(t)$: uncontrollable + unobservable.

6 State feedback control

In the previous chapter we considered a minimum effort control design, however, the resulting controller was open loop. Here we introduce a feedback control design for the case where the entire state is available in the form of measurements in the system output, i.e., when $y = x$. Therefore, for the results of this chapter the output equation becomes irrelevant. In this setting, the following questions naturally arise:

1. Is it possible design a state feedback control input so that we steer an LTI system to a desired state?
2. If yes, how do we construct such a state feedback controller?

The first of these questions is related to controllability. We have already seen that if the system is controllable then we can design minimum energy controllers; here, we will show that controllability is equivalent to the existence of a state feedback controller. In particular, we will show that such a state feedback controller can be constructed by means of a methodology called pole placement, hence it is possible for controllable systems to select the feedback control gains appropriately so that the resulting closed loop system is driven to a desired state.

6.1 Closed loop system

We consider an LTI system in state space form, i.e.,

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

Note that under a full state feedback regime $y(t) = x(t)$, which implies that $C = I$ and $D = 0$. With reference to the state feedback control architecture schematically shown in Figure 7, we have that

$$u(t) = Kx(t) + r(t),$$

where $r(t) \in \mathbb{R}^m$ is an external input vector that we consider to be fixed, while $K \in \mathbb{R}^{m \times n}$ is the *feedback gain matrix* whose elements are the control gains that

we need to tune. Note that K is a matrix, as we may have multiple inputs and outputs (in this setting outputs coincide with states). In the single input, single output case, the resulting gain would be a scalar.

Under this state feedback controller, the ODE governing the behaviour of the LTI system, i.e., $\dot{x}(t) = Ax(t) + Bu(t)$, becomes

$$\dot{x}(t) = (A + BK)x(t) + Br(t).$$

Such a system is referred to as *closed loop system* as its evolution does not depend on the input $u(t)$, but only on the state $x(t)$ and some fixed, external input $r(t)$. Matrix $A + BK$ will be hereafter referred to as closed loop matrix.

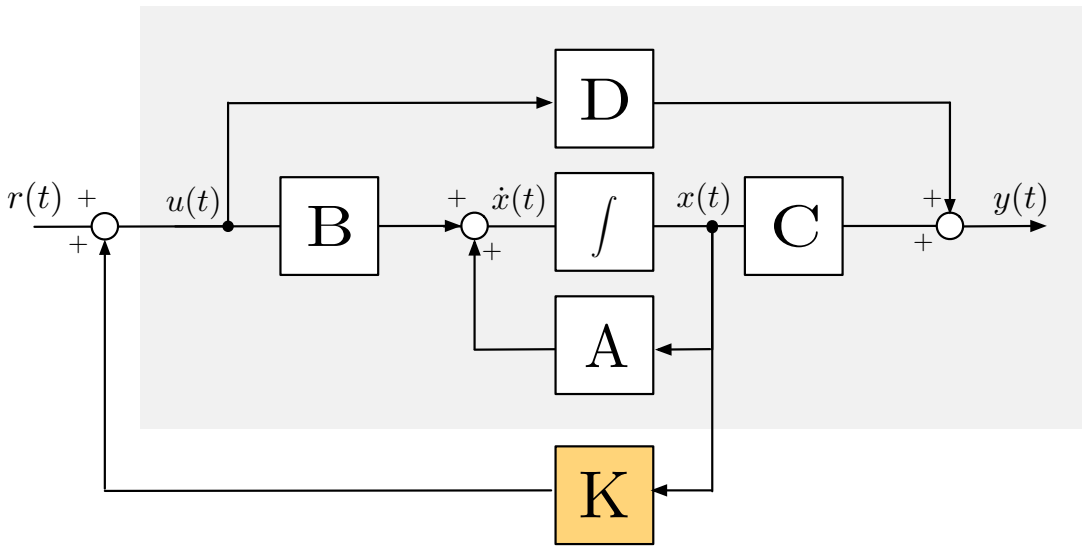


Figure 7: Block diagram for the state feedback controller.

We aim at designing the feedback gain matrix K so that the closed loop system evolves in a prescribed manner. In particular, we will show that deciding on whether a given LTI system is controllable is related to the ability of placing the eigenvalues of $A+BK$ at any desired location, and we will provide a procedure to do so referred to as *pole placement*. Due to the relationship between poles and eigenvalues this is also known as eigenvalue placement or eigenvalue assignment.

6.2 Pole placement

6.2.1 Single input systems

We first consider LTI systems with a single input, i.e., $u(t) \in \mathbb{R}$ ($m = 1$), and hence

$$K = \begin{bmatrix} k_1 & \dots & k_n \end{bmatrix} \in \mathbb{R}^{1\times n},$$

is a row vector. We aim at designing K so that we steer the system at a desired state; naturally, this is related to controllability (see Definition 1). To build some intuition, the most practically relevant situation is the case where this desired state is finite and we would like to converge to it asymptotically (possibly also sufficiently fast), hence the closed loop system exhibits a stable behaviour. However, we have already seen that stability depends on the eigenvalues of the A matrix that appears in the state space system representation. For the closed loop system to be stable, we are thus interested in the eigenvalues of $A + BK$. To this end, informally we aim at building the following equivalence:

$$\text{Controllability} \quad \Leftrightarrow \quad \text{Setting the eigenvalues of } A + BK \text{ to any } \{\lambda_1, \dots, \lambda_n\},$$

where $\{\lambda_1, \dots, \lambda_n\}$ is a target set of n (possibly complex) eigenvalues. It is to be understood that if one of the eigenvalues in the target set is complex, then its complex conjugate will have to be in that set as well, as complex eigenvalues appear in conjugate pairs.

Changing the target set of eigenvalues we thus change the evolution of the closed loop system. In fact, if we are able to select K so that we move the eigenvalues (and hence the closed loop system) at any desired target set then we make sure that through feedback our system can be driven to any state, hence it is controllable. We formalize this in the following theorem.

Theorem 4 (Controllability & eigenvalues of $A+BK$)**.** *A single input LTI system is controllable if and only if for any target set of eigenvalues* $\{\lambda_1, \dots, \lambda_n\}$*,*

$$\text{there exists } K \text{ such that: } \textit{eigenvalues of } A + BK = \{\lambda_1, \dots, \lambda_n\}.$$

The previous theorem provides a link between the ability to place the eigenvalues

of $A+BK$ and controllability, hence it answers the first question set forth at the beginning of the chapter. However, it does not state how we can design such a K. The design procedure for K to achieve this eigenvalue placement is most commonly referred to as *pole placement*, and we will detail it below.

To this end, let $\{\lambda_1, \ldots, \lambda_n\}$ be any target set of eigenvalues for the closed loop matrix $A+BK$. These eigenvalues will then be the roots of the characteristic polynomial of $A+BK$, i.e.,

target characteristic polynomial:

$$\begin{aligned}\det(\lambda I - (A+BK)) &= (\lambda-\lambda_1)\cdots(\lambda-\lambda_n)\\ &= \lambda^n + d_1\lambda^{n-1} + \ldots + d_{n-1}\lambda + d_n,\end{aligned}$$

where $d_1, \ldots, d_n$ are real coefficients, uniquely determined by the particular choice of the target set of eigenvalues.

At the same time, it can be shown that if a system is controllable then there exists an invertible matrix T (in fact this is an "if and only if" condition) such that the coordinate transformation $\hat{x}(t) = Tx(t)$ brings the system in controllable canonical form *, i.e., $\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{B}u(t)$, where

$$\hat{A} = TAT^{-1} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0\\ 0 & 0 & 1 & \ldots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \ldots & 1\\ -a_n & -a_{n-1} & -a_{n-2} & \ldots & -a_1 \end{bmatrix} \text{ and } \hat{B} = TB = \begin{bmatrix} 0\\ 0\\ \vdots\\ 0\\ 1 \end{bmatrix},$$

where $a_1, \ldots, a_n$ are the coefficients of the characteristic polynomial of A, which is given by $\lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n$. In the new coordinates, the state feedback control input can be written as

$$\begin{aligned}u(t) = Kx(t) + r(t) &= KT^{-1}\hat{x}(t) + r(t)\\ &= \hat{K}\hat{x}(t) + r(t),\end{aligned}$$

*Note that we have already introduced the controllable canonical form in Section 3.2.2, as one possible realization of the system's transfer function. The term "controllable" shall now be clear due to the fact that any controllable system admits such a representation via an appropriate coordinate transformation T. In fact, even though we will not provide this in these notes, it can be shown that for single input systems T^{-1} involves the controllability matrix (will be a square matrix in the case of a single input). Note that an observable canonical form also exists.

where $\hat{K} = KT^{-1} = \begin{bmatrix} \hat{k}_1 & \dots & \hat{k}_n \end{bmatrix}$ is a row vector including the new control gains. Application of this feedback controller results in the following closed loop system in the new coordinates:

$$\begin{aligned}\dot{\hat{x}}(t) &= (\hat{A} + \hat{B}\hat{K})\hat{x}(t) + \hat{B}r(t) \\ &= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \hat{k}_1 - a_n & \hat{k}_2 - a_{n-1} & \hat{k}_3 - a_{n-2} & \dots & \hat{k}_n - a_1 \end{bmatrix} \hat{x}(t) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} r(t),\end{aligned}$$

where this particular form emanates from the structure of $\hat{A}$ and $\hat{B}$ that are in controllable canonical form. The closed loop system is thus still in controllable canonical form, hence the entries in the last row correspond to the coefficients of the characteristic polynomial of $\hat{A} + \hat{B}\hat{K}$, i.e.,

characteristic polynomial in new coordinates:

$$\begin{aligned}&\det(\lambda I - (\hat{A} + \hat{B}\hat{K})) \\ &\qquad = \lambda^n - (\hat{k}_n - a_1)\lambda^{n-1} - \dots - (\hat{k}_2 - a_{n-1})\lambda - (\hat{k}_1 - a_n).\end{aligned}$$

The eigenvalues remain unaffected by a coordinate transformation (since $\hat{A}+\hat{B}\hat{K} = T(A + BK)T^{-1}$), hence $A + BK$ and $\hat{A} + \hat{B}\hat{K}$ have the same eigenvalues. As a result, the target characteristic polynomial and the characteristic polynomial in new coordinates should coincide. This implies that their coefficients should be the same, or in other words,

target characteristic polynomial = *characteristic polynomial in new coordinates*

$$\Rightarrow \ \hat{k}_1 = a_n - d_n, \ \hat{k}_2 = a_{n-1} - d_{n-1}, \ \dots, \ \hat{k}_n = a_1 - d_1.$$

Therefore, we can construct $\hat{K} = \begin{bmatrix} a_n - d_n & \dots & a_1 - d_1 \end{bmatrix}$ so that we place the eigenvalues of $\hat{A} + \hat{B}\hat{K}$ to the target set of eigenvalues. Equivalently, since $\hat{K} = KT^{-1}$, through $K = \hat{K}T$ we can construct the control input

$$u(t) = \hat{K}Tx(t) + r(t) = Kx(t) + r(t),$$

so that the eigenvalues of the closed loop matrix $A+BK$ in the original coordinates are also equal to the target set.

Effectively, the derivation above constitutes a proof for the "only if" part of Theorem 4, providing a constructive way to place the eigenvalues of $A+BK$, thus answering the second question set forth at the beginning of the chapter. However, to design K we used a transformation in the so called controllable canonical form. In practice this is not necessary, and we can achieve pole placement by means of the following procedure.

Pole placement procedure. Determine $K = \begin{bmatrix} k_1 & \dots & k_n \end{bmatrix}$ by means of the following steps:

Step 1: Select a target set of eigenvalues $\{\lambda_1, \dots, \lambda_n\}$.

Step 2: Construct the target characteristic polynomial

$$(\lambda - \lambda_1) \cdots (\lambda - \lambda_n) = \lambda^n + d_1\lambda^{n-1} + \dots + d_{n-1}\lambda + d_n.$$

Step 3: Construct the characteristic polynomial of $A + BK$, i.e.,

$$\det(\lambda I - (A + BK)).$$

The coefficients will be linear functions of the gains in K, namely, $k_1, \dots, k_n$.

Step 4: Equate coefficients between the polynomials of Step 2 and Step 3,

$\Rightarrow$ Solve a system of n equations with n unknowns, namely, $k_1, \dots, k_n$.

Therefore, designing a state feedback controller involves implementing the pole placement procedure outlined above, which results in solving a linear system of equations with an equal number of unknowns. To solve that system of equations we distinguish the following cases:

1. The system is controllable. Then the system of equations admits a unique solution; in fact, existence of a solution for controllable systems is guaranteed by means of Theorem 4. Naturally, if we have a controllable system, we would like the closed loop system to be asymptotically stable (see Chapter 3). Hence the target eigenvalues should all have negative real part.

2. The system is uncontrollable and at least one eigenvalue corresponding to the uncontrollable part has positive real part. By Theorem 3 this will be one of the eigenvalues of A_{33} and A_{44} if Kalman decomposition is performed. In that case the system of equations does not have a solution. This corresponds to a case where the system is unstable, and the unstable part is uncontrollable, so we can not achieve a stable closed loop system by means of state feedback.

3. The system is uncontrollable but all eigenvalues corresponding to the uncontrollable part have negative real part. In that case the system of equations admits a solution as long as we include the eigenvalues of the uncontrollable part in the target set of eigenvalues. However, the solution might not necessarily be unique. This corresponds to the case of an uncontrollable but stabilizable system.

We illustrate the pole placement procedure by means of an example.

Example 17. *Consider an LTI system with matrices*

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

Design a state feedback controller so that the closed loop system has eigenvalues at -1 and -2.

Solution: *The system is of order $n = 2$, so we seek a control gain matrix of the form $K = \begin{bmatrix} k_1 & k_2 \end{bmatrix}$. The closed loop matrix is thus given by*

$$A + BK = \begin{bmatrix} 0 & 1 \\ -1 + k_1 & -1 + k_2 \end{bmatrix}.$$

To determine k_1 and k_2 so that the closed loop system has eigenvalues at -1 and -2, we apply the pole placement procedure. We thus have:

Step 1. The target set of eigenvalues is provided in this case, and we denote it as $\{\lambda_1, \lambda_2\} = \{-1, -2\}$.

Step 2. *The target characteristic polynomial is given by*

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = (\lambda + 1)(\lambda + 2) = \lambda^2 + 3\lambda + 2.$$

Step 3. *The characteristic polynomial of $A + BK$ is given by*

$$\det(\lambda I - (A + BK)) = \lambda^2 + (1 - k_2)\lambda + (1 - k_1).$$

Step 4. *Equating the coefficients of the polynomials of Step 2 and Step 3 we obtain*

$$\begin{aligned} 1 - k_1 = 2 \;&\Rightarrow\; k_1 = -1 \\ 1 - k_2 = 3 \;&\Rightarrow\; k_2 = -2. \end{aligned}$$

The resulting system of equations admitted a unique solution, as the given system was controllable. This can be verified by checking that the controllability matrix is full rank, i.e.,

$$\operatorname{rank}([B \; AB]) = \operatorname{rank}\left(\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \right) = 2.$$

6.2.2 Multiple input systems

Up to this point the analysis refers to single input systems. In fact this is only used when resorting to the controllable canonical form which refers to single input systems. Even if the system has multiple inputs ($m > 1$) though, we could still extend the result of Theorem 4 and place the eigenvalues of the closed loop system at target locations following the pole placement procedure. However, for multiple input systems, the system of equations in the last step of the pole placement procedure will involve n equations with nm unknowns. Hence, we will have more unknowns than equations, implying that the system will not admit a unique solution even if the original LTI system is controllable. As a result, there will be multiple choices of K that will lead to the same eigenvalues for the closed loop system (in

fact, the admissible choices for K will lie on a subspace with dimension $n(m-1)$)*.

To obtain a unique K from the resulting system of equations, we could either impose additional considerations, e.g., choose a K that not only results in a target set of eigenvalues but at the same time minimizes a certain performance criterion, or force a certain number of elements of K to be zero. This would lead to a sparse feedback gain matrix, thus simplifying some computations, leading to a feedback structure which is easier to implement in practice.

6.3 Summary

This chapter discussed state feedback control, or in other words, it provided a procedure to design feedback gains if the entire state is available. The main learning outcomes of the chapter can be summarized as follows:

- *Closed loop matrix.* We showed that under state feedback the evolution of the closed loop system is captured by the matrix $A + BK$.

- *Single input systems.*

    1. *Controllability and eigenvalues of $A + BK$:* Theorem 4 shows that an LTI system is controllable if and only if for any target set of eigenvalues, there exists a choice for the gain matrix K such that the eigenvalues of the closed loop matrix $A + BK$ become equal to this target set.

    2. *Pole placement:* The important implication of Theorem 4 is that controllability ensures the existence of a state feedback controller. Pole placement is a procedure to construct such a controller. It involves equating the coefficients of the target characteristic polynomial, with those of the characteristic polynomial of $A + BK$, i.e.,

*To build some intuition on why multiple solutions exist, note that similarly to the way the controllable canonical form was employed for single input systems, for multiple input systems the so called Brunowski normal form can be used. However, this form is not unique as it involves a coordinate transformation where we select n columns of the controllability matrix which is now of dimension $n \times nm$, and there are several permutations of the selected columns.

target characteristic polynomial
$$= \text{characteristic polynomial of } A + BK = \det(\lambda I - (A + BK)).$$

Note that the coefficients of $\det(\lambda I - (A + BK))$ are linear functions of the entries of K, hence the equality above results in a system of n equations with n unknowns. *If the original LTI system is controllable, then this system admits a unique solution for the gains in K.*

- *Multiple input systems.* Pole placement is also applicable, however, the resulting set of equations will involve more unknowns than equations. As a result the solution is not unique even if the system is controllable, hence, there will be multiple choices of K that will lead to the same eigenvalues for the closed loop system.

7 Observers & Output feedback control

The pole placement procedure outlined in the previous chapter provides an efficient methodology to design a state feedback controller. However, having access to the entire system state is often unrealistic in practice, as typically only some of the states are available via measurements. Therefore, it would be more practical if an output feedback control design methodology were available, where the feedback was only a function of the current and past outputs of the system.

In this chapter we will address this problem, and provide an output feedback control design methodology. In particular, we will:

- Exploit the state feedback control design procedure, and construct a control input that is not feedback of the actual system state (which is not available), but rather feedback of an estimate of the state. A procedure that generates an estimate of the state using past and present inputs and outputs is known as *state observer* or else state estimator.

- Design a controller which is feedback of the state estimated by means of the observer. This is implicitly an *output feedback controller*, as it will depend on the estimated state, which in turn depends on the system's output and not on the state.

A block diagram summary of the output feedback controller including an observer is provided in Figure 8. We will be referring to this figure when introducing its various elements in the sequel.

7.1 Linear state observers

7.1.1 Observer estimation error

Analogously to the state feedback control design analysis, we will assume that we have an LTI system (see upper shaded block in Figure 8) with a single output

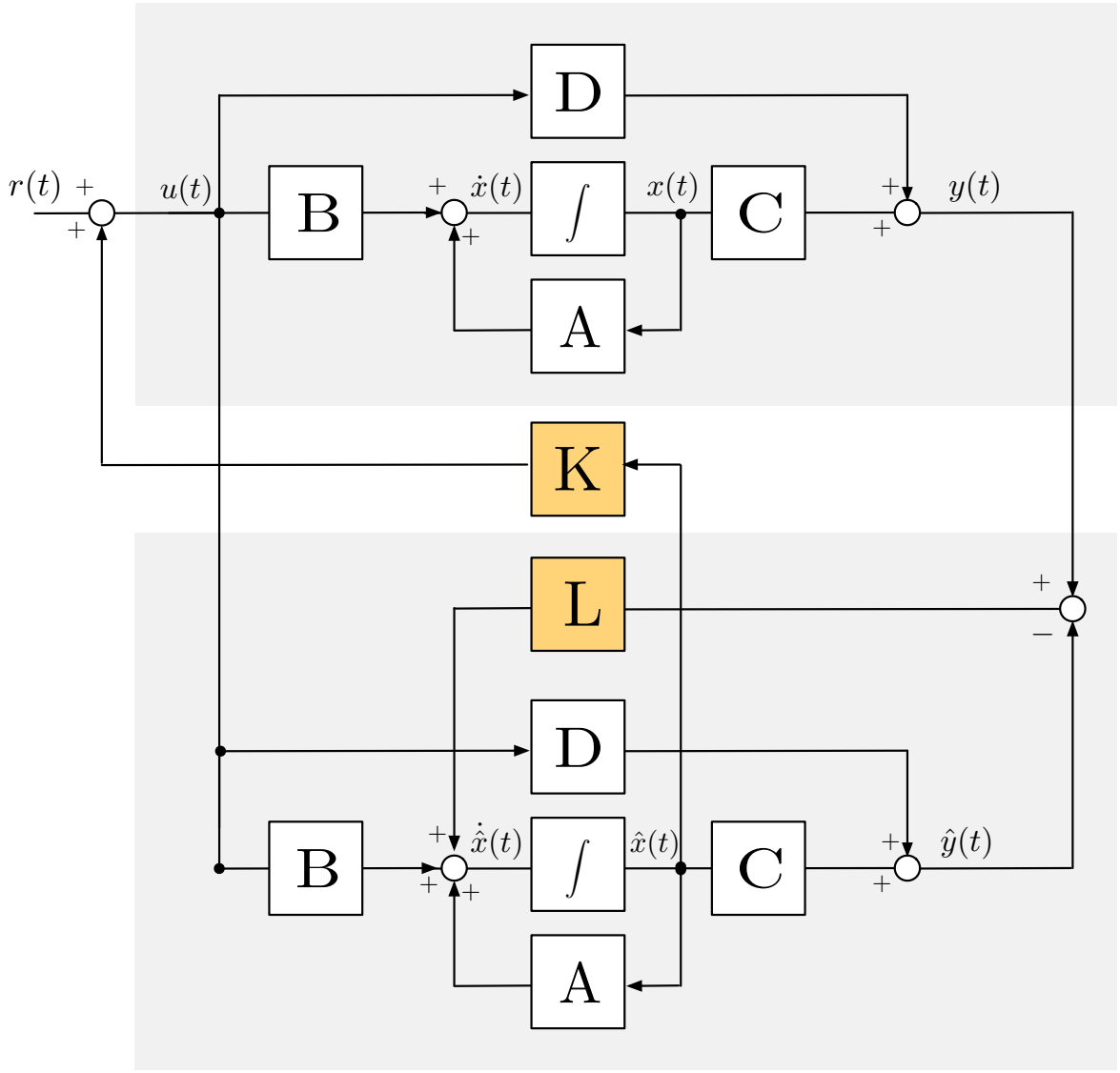


Figure 8: Block diagram for the output feedback controller, including a linear state observer.

$(p = 1)$, i.e.,

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

where $C \in \mathbb{R}^{1\times n}$ and $D \in \mathbb{R}^{1\times m}$ are row vectors. We will discuss at the end of the section the necessary modifications in case of multiple outputs.

It turns out that for linear systems a linear observer is sufficient to construct an estimate of the state. In particular, for any time instance t, the observer will produce an estimate $\hat{x}(t)$ for the actual state $x(t)$, using only current and past inputs and outputs (information available through measurements), i.e., $u(\tau)$ and $y(\tau)$ for $\tau \in [0, t]$. To achieve this, the observer is yet another LTI system, which

acts as a replica of the actual system (in the sense that we employ the state space matrices (A, B, C, D)) described by

Linear state observer:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t))$$
$$\hat{y}(t) = C\hat{x}(t) + Du(t).$$

At time t, the observer uses only its current state and output estimates $\hat{x}(t)$ and $\hat{y}(t)$, respectively, the input $u(t)$ and the actual output $y(t)$ that is available by means of measurements to update its state estimate. Notice that the actual state $x(t)$ is not used anywhere in the observer's equations. The derivative of the state estimate $\hat{x}(t)$ will thus depend on the current input and output of the actual system, while $\hat{x}(t)$ depends implicitly (see the solution form for LTI systems) on past inputs and outputs as well. See the lower shaded block of Figure 8 for a block diagram representation of a linear observer.

The linear term $L(y(t)-\hat{y}(t))$ acts as a correction proxy, introducing an estimate of the difference between the actual and the estimated output, which we would ideally like to steer to zero. The matrix (column vector in single output systems) $L \in \mathbb{R}^n$ is called the *observer gain matrix*. We would like to select the entries of L so that the observer generates a "good" estimate $\hat{x}(t)$ of $x(t)$. To quantify this, we study the evolution of the so called *estimation error* $e(t) = x(t) - \hat{x}(t)$, i.e.,

$$\begin{aligned}
\dot{e}(t) &= \dot{x}(t) - \dot{\hat{x}}(t) \\
&= Ax(t) + \cancel{Bu(t)} - A\hat{x}(t) - \cancel{Bu(t)} - L(y(t) - \hat{y}(t)) \\
&= A(x(t) - \hat{x}(t)) - L(Cx(t) + \cancel{Du(t)} - C\hat{x}(t) - \cancel{Du(t)}) \\
&= (A - LC)(x(t) - \hat{x}(t)) \\
&= (A - LC)e(t),
\end{aligned}$$

where in the third equation we replaced $y(t)$ and $\hat{y}(t)$ with the output equation of the corresponding LTI description. Therefore, the evolution of the estimation error is an autonomous LTI system. It will be asymptotically stable, hence $\hat{x}(t)$ will converge to $x(t)$, if and only if all eigenvalues of $A - LC$ have negative real part. Otherwise, if the system is stable, given an inaccurate initial estimate, $\hat{x}(t)$ will

remain at some "distance" from the actual estimate, while if the system is unstable then that distance will grow towards infinity, implying that our estimate diverges. Note that matrix B does not influence the behaviour of the error estimate, similarly to the way C does not appear in the state feedback design procedure.

7.1.2 Observer gain selection

It turns out that to design an efficient observer whose estimate will converge to the actual state we need to have the ability to place the eigenvalues of the observer gain matrix $A - LC$ at desired locations, and in particular select the gains in L so that these target eigenvalues have negative real part.

The ability of placing the eigenvalues of $A - LC$ at a target set of eigenvalues is related to observability. This is in some sense dual to the fact that the eigenvalues of $A + BK$ were related to controllability in Theorem 4.

Theorem 5 (Observability & eigenvalues of $A-LC$)**.** *A single output LTI system is observable if and only if for any target set of eigenvalues* $\{\lambda_1, \ldots, \lambda_n\}$*,*

$$\text{there exists } L \text{ such that: } \text{eigenvalues of } A - LC = \{\lambda_1, \ldots, \lambda_n\}.$$

Theorem 5 shows that for observable systems there always exists a choice for L so that we place the eigenvalues of $A-LC$ at any desired set of target locations. Due to the structural similarity between $A-LC$ and $A+BK$, to construct the observer gain matrix L so that we place the eigenvalues of $A - LC$, we can follow exactly the same steps with the pole placement procedure of Chapter 6 with $A - LC$ in place of $A+BK$. This results in equating the coefficients of the target polynomial with the characteristic polynomial of $A - LC$, i.e.,

$$\text{target characteristic polynomial: } \lambda^n + d_1\lambda^{n-1} + \ldots + d_{n-1}\lambda + d_n = \det(\lambda I - (A - LC)).$$

For single output systems this results in solving a linear system of equations with an equal number of unknowns. To solve that system of equations we distinguish

the following cases:

1. The system is observable. Then the system of equations admits a unique solution; in fact, existence of a solution for observable systems is guaranteed by means of Theorem 5. If the target eigenvalues all have negative real part, then we can guarantee that the observer estimation error converges to zero, so asymptotically we tend to construct the entire system state, even if this was unavailable via measurements.
2. The system is unobservable and at least one eigenvalue corresponding to the unobservable part has positive real part. By Theorem 3 this will be one of the eigenvalues of A_{22} and A_{44} if Kalman decomposition is performed. In that case the system of equations does not have a solution. This corresponds to a case where the system is unstable, and the unstable part is unobservable, so we can not achieve a stable observer estimation error.
3. The system is unobservable but all eigenvalues corresponding to the unobservable part have negative real part. The system of equations admits then a solution as long as we include the eigenvalues of the unobservable part in the target set. However, the solution might not necessarily be unique. This corresponds to the case of an unobservable but detectable system.

For systems with multiple outputs ($p > 1$) the previous statements remain valid, however, equating the polynomial coefficients leads to a system of n equations with np unknowns. Hence, we will have more unknowns than equations, implying that the system will not admit a unique solution even if the original LTI system is observable. Note that all these conclusions are in complete symmetry with respect to the ones obtained in the previous chapter for state feedback control design.

Example 18. *Consider the LTI system of Example 14, with matrices*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \quad \textit{and} \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

Design a linear state observer so that the estimation error system has eigenvalues at -1 *and* -2.

Solution: *The system is of order $n = 2$ and has a single output as C is a row vector. We thus seek a control gain matrix of the form $L = \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix}$. The estimation error system involves the matrix*

$$A - LC = \begin{bmatrix} -\ell_1 & 1 \\ -1 - \ell_2 & -1 \end{bmatrix}.$$

To determine ℓ_1 and ℓ_2 so that the eigenvalues of the estimation error system (hence of $A - LC$) become -1 and -2, we apply the pole placement procedure of Chapter 6 with $A - LC$ in place of $A + BK$. We thus have:

Step 1. The target set of eigenvalues is provided in this case, and we denote it as $\{\lambda_1, \lambda_2\} = \{-1, -2\}$.

Step 2. The target characteristic polynomial is given by

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = (\lambda + 1)(\lambda + 2) = \lambda^2 + 3\lambda + 2.$$

Step 3. The characteristic polynomial of $A - LC$ is given by

$$\det(\lambda I - (A - LC)) = \lambda^2 + (1 + \ell_1)\lambda + (1 + \ell_1 + \ell_2).$$

Step 4. Equating the coefficients of the polynomials of Step 2 and Step 3 we obtain

$$\begin{aligned} 1 + \ell_1 = 3 \;\Rightarrow\; \ell_1 = 2 \\ 1 + \ell_1 + \ell_2 = 2 \;\Rightarrow\; \ell_2 = -1. \end{aligned}$$

The resulting system of equations admitted a unique solution, as the given system was single output and it was shown in Example 14 to be observable.

7.2 Output feedback control

7.2.1 Closed loop system

We have already seen that if the state is fully known, then a state feedback controller can be designed. We have also shown that if the state is not fully known then a

linear state observer with gain matrix L can be designed. Combining these designs we come up with the architecture of Figure 8, which serves as a straightforward implementation of output feedback. To see this, notice that it involves feedback of the estimated state $\hat{x}(t)$ with a feedback gain matrix K. The estimated state is the output of the observer, which in turn depends on the output of the actual system $y(t)$. This introduces in an implicit manner output feedback.

However, we still need to show that putting the design of the observer gain matrix L together with that of the (estimated) state feedback gain matrix K leads to the desired closed loop performance. In particular, we would like to ensure that the transient estimation error in the observer will not interfere with the state feedback controller leading to a destabilizing behaviour, which in turn may increase the estimation error and eventually result in an unstable closed loop system.

To address this issue, we will study the behaviour of the closed loop system of Figure 8, and in particular we will compute its eigenvalues. To this end, we have that

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \qquad \rightarrow \textit{ LTI system}$$

$$u(t) = K\hat{x}(t) + r(t) \qquad \rightarrow \textit{ (estimated) state feedback}$$

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) + L(y(t) - \hat{y}(t)) \\ \hat{y}(t) &= C\hat{x}(t) + Du(t). \end{aligned} \qquad \rightarrow \textit{ state observer}$$

Substituting one equation into the other so that we eliminate $\hat{y}(t)$ and $u(t)$, we obtain the following closed loop system description:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + BK\hat{x}(t) + Br(t) \\ \dot{\hat{x}}(t) &= LCx(t) + (A + BK - LC)\hat{x}(t) + Br(t) \\ y(t) &= Cx(t) + DK\hat{x}(t) + Dr(t). \end{aligned}$$

The closed loop system is itself an LTI system with a state vector $\begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} \in \mathbb{R}^{2n}$ (both the actual and the estimated state), an input vector $r(t) \in \mathbb{R}^m$, and an

output vector $y(t) \in \mathbb{R}^p$. We can write it in state space form as

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\hat{x}}(t) \end{bmatrix} = \begin{bmatrix} A & BK \\ LC & A + BK - LC \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} + \begin{bmatrix} B \\ B \end{bmatrix} r(t)$$
$$y(t) = \begin{bmatrix} C & DK \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix} + Dr(t).$$

7.2.2 Separation principle

The state space representation above captures the evolution of the actual and the estimated state. However, we are mainly concerned about the evolution of the actual state $x(t)$ and the estimation error $e(t) = x(t) - \hat{x}(t)$. To this end, replacing $\hat{x}(t) = x(t) - e(t)$, or else considering the (invertible) coordinate transformation

$$\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ x(t) - \hat{x}(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{x}(t) \end{bmatrix},$$

we can render $e(t)$ as one of the states, resulting in the following equivalent state space description of the closed loop system:

$$\begin{bmatrix} \dot{x}(t) \\ \dot{e}(t) \end{bmatrix} = \begin{bmatrix} A + BK & -BK \\ 0 & A - LC \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} r(t)$$
$$y(t) = \begin{bmatrix} C + DK & -DK \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} + Dr(t).$$

The new state space representation has the advantage that not only it involves the state estimation error as one of the system states, but it also involves a block triangular system matrix. However, the determinant of a block triangular matrix coincides with the product of the determinants of the associated blocks. We thus have that

Separation principle:

$$\det\left(\begin{bmatrix} \lambda I - (A + BK) & BK \\ 0 & \lambda I - (A - LC) \end{bmatrix} \right) = \det\big(\lambda I - (A + BK)\big) \det\big(\lambda I - (A - LC)\big).$$

This property of the closed loop system is known as the *separation principle*, as the $2n$ eigenvalues of the closed loop system coincide with the n eigenvalues of the system if full state feedback were available (roots of $\det\big(\lambda I-(A+BK)\big)$), and the n eigenvalues of the state estimation error system (roots of $\det\big(\lambda I-(A-LC)\big)$). Its consequence is that to design output feedback controllers for linear systems, it is sufficient to design the feedback gain matrix K and the observer gain matrix L separately, and then put them them together, while having guarantees about the closed loop system performance. It should be noted that the separation principle does not directly extend to systems that are not linear.

Therefore, if an LTI system is both controllable and observable, Theorems 4 and 5 together with the separation principle imply that the eigenvalues of the closed loop system can be arbitrarily placed by selecting the target eigenvalues of $A+BK$ and $A-LC$ to have a sufficiently negative real part, thus resulting in an arbitrarily fast, asymptotically stable performance. However, the LTI system is typically only an abstraction of the actual system which may exhibit nonlinearities, actuation saturations, etc. Hence, selecting the eigenvalues to have arbitrarily negative real parts may lead to high control gains that in turn may be clipped by input saturations, leading to erroneous closed loop performance. To this end, the choice of the eigenvalues of the closed loop system (and hence the gains of the controller and the observer) usually involves a trade-off between a fast and stable response and other performance considerations. We will investigate an optimal control methodology to achieve such a trade-off in the next chapter.

7.3 Summary

This chapter discussed output feedback control, or in other words, it provided a procedure to design feedback gains if only some components of the state are available in the output of the system in the form of measurements. To construct an estimate of the entire state vector a linear state observer was designed. The main learning outcomes of the chapter can be summarized as follows:

- *Linear state observer.* We showed that linear state observers can be designed to construct an estimate of the actual state (see lower shaded block in Figure

8). In particular, the important implication of Theorem 5 is that observability ensures the existence of an observer such that the state estimation error is under control. Using again the pole placement procedure we can construct such an observer. It involves equating the coefficients of the target characteristic polynomial, with those of the characteristic polynomial of $A - LC$, i.e.,

target characteristic polynomial
$= \text{characteristic polynomial of } A - LC = \det(\lambda I - (A - LC))$.

Note that the coefficients of $\det(\lambda I - (A - LC))$ are linear functions of the entries of L, hence for single output systems the equality above results in a system of n equations with n unknowns. *If the original LTI system is observable, then this system admits a unique solution for the gains in L.*

- *Output feedback.* A feedback controller was designed closing the loop via the estimated state. That state was in turn dependent on the output of the system (through the observer equations), thus resulting in an output feedback implementation (see feedback interconnection between the upper and lower blocks in Figure 8, using a feedback gain matrix K).

- *Separation principle.* The so called separation principle is a property of linear systems, which shows that

eigenvalues of closed loop system
$= \text{eigenvalues of } A + BK \text{ and eigenvalues of } A - LC$.

As a result, we can select K and L separately, each of them by means of pole placement for $A + BK$ and $A - LC$, respectively. For controllable and observable systems, putting them together in the closed loop architecture of Figure 8 can result in a stable performance and a convergent state estimation error.

8 Linear Quadratic Regulator (LQR)

We have seen that minimum energy controllers are optimal in the sense of minimizing the control effort, however, they are open loop. At the same time state and output feedback controllers introduce feedback, however, they are not optimal with respect to a given performance criterion. The following question thus pertains:

- Is it possible to design feedback controllers that are at the same time optimal with respect to a given criterion that involves (possibly) both the state and the input of the system?

It turns out that this is indeed possible. We will assume that the entire state is available by means of measurements, and design a state feedback controller that will optimize a quadratic function of the state and the input. Such a controller is known as Linear Quadratic Regulator (LQR) and is within the realm of optimization based control.

8.1 Finite horizon optimal control problem

8.1.1 Problem statement

Consider a system whose state starts from $x(0) = x_0$ and evolves according to

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

We also consider a finite horizon problem with horizon length T, and aim at designing a control input trajectory $u(\cdot)$ that is optimal with respect to a certain cost criterion over the time interval $[0, T]$, while resulting in a state trajectory that satisfies the ODE above and the initial condition. We take the cost criterion to be a cummulative penalty over the time horizon on the state and the control input, namely,

$$J(u) = \underbrace{\int_0^T \Big(x(t)^\top Q x(t) + u(t)^\top R u(t)\Big) dt}_{\text{running cost}} + \underbrace{x(T)^\top Q_T x(T)}_{\text{terminal cost}},$$

where the integral acts like the continuous analogue of summation, accumulating penalty terms corresponding to the different time instances within our horizon.

Overall, the cost function is the sum of a running cost, which penalizes the state and the input, and a terminal cost, which penalizes the state at the end of the horizon. We can further distinguish three terms:

1. Running state penalty: $x(t)^\top Q x(t)$. This is a quadratic penalty on the state $x(t)$ at any t within the given horizon, with $Q = Q^\top \succeq 0$ (symmetric and positive semidefinite).
2. Running input penalty: $u(t)^\top R u(t)$. This is a quadratic penalty on the state $u(t)$ at any t within the given horizon, with $R = R^\top \succ 0$ (symmetric and positive definite).
3. Terminal state penalty: $x(T)^\top Q_T x(T)$. This is a quadratic penalty on the state $x(T)$ at the end of the horizon, with $Q_T = Q_T^\top \succeq 0$ (symmetric and positive semidefinite).

Notice that if $Q_T = 0$ then we only have a running but no terminal cost, while if $Q = Q_T = 0$ and $R = I$, then we just penalize the input in the cost function, as in the case of minimum energy controllers. It should be noted that the cost function J depends on the control input u that we seek to determine, as well as on the initial state x_0 and the length of the time horizon T (even though the dependency on these parameters is not made explicit). J does not depend on the state $x(t)$, as this is an "internal" variable, and as the ODE evolves then states can be written as functions of the initial condition and past inputs (as also witnessed by the state solution of an LTI system).

The problem of seeking a control input trajectory that is optimal with respect to this quadratic cost, while being compatible with the linear ODE, is called *finite horizon Linear Quadratic Regulator (LQR)* problem. It can be written as

Finite horizon LQR problem: Find $u(\cdot) : [0, T] \to \mathbb{R}^m$ such that we

$$\begin{aligned} \text{minimize} \quad & J(u) = \int_0^T \big(x(t)^\top Q x(t) + u(t)^\top R u(t)\big) dt + x(T)^\top Q_T x(T) \\ \text{subject to} \quad & \dot{x}(t) = Ax(t) + Bu(t), \text{ for all } t \in [0, T], \\ & x(0) = x_0. \end{aligned}$$

8.1.2 Riccati equation and optimal controller

The controller that solves the optimal control problem, or in other words the LQR controller, admits a closed form expression. This is provided in following theorem which establishes the LQR solution and the associated optimal cost.

Theorem 6 (finite horizon LQR controller)**.** *The finite horizon LQR problem can be solved by means of the following steps:*

1. *Solve the so called Riccati differential equation*

$$-\dot{P}(t) = P(t)A + A^\top P(t) + Q - P(t)BR^{-1}B^\top P(t)$$
$$\text{with} \quad P(T) = Q_T,$$

   and denote its (unique) solution by $P(t) \in \mathbb{R}^{n\times n}$. For any t, $P(t)$ is symmetric and positive semidefinite.

2. *The finite horizon optimal LQR controller can be then constructed as*

$$u^\star(t) = K(t)x(t) = -R^{-1}B^\top P(t)x(t),$$

   where $P(t)$ is the solution of the Riccati differential equation over $[0, T]$.

3. *The associated optimal LQR cost is given by*

$$J(u^\star) = x_0^\top P(0)x_0.$$

It should be remarked that:

- To compute $P(t)$ as a function of time we need to solve the Riccati differential equation. This is a differential equation that involves a matrix, so we have as many ODEs as the number of entries of $P(t)$. Existence and uniqueness of solutions to the Riccati equation relies on the fact that $P(t)$ (and hence also the cost) can be shown to remain bounded over finite time horizons.
- The Riccati differential equation is solved backwards in time, as we are given a terminal condition $P(T) = Q_T$, and seek the solution for $t \in [0, T]$. This implies that for a given horizon T we first need to solve the Riccati equation over $[0, T]$, and then construct the optimal controller.

- The finite horizon LQR controller is state feedback, however, the feedback gain matrix is no longer time invariant, but depends on time through $P(t)$. The state feedback gain matrix is then given by $K(t) = -R^{-1}B^\top P(t)$, where R^{-1} is well defined since R is positive definite.
- The optimal cost depends only the initial state x_0 and the initial value of the solution to the Riccati equation $P(0)$.

We first illustrate Theorem 6 by means of an example for a single state system, and then state its proof.

✎ **Example 19.** *Let T be a given finite horizon length, and consider an LTI system whose evolves according to $\dot{x}(t) = u(t)$, starting from $x(0) = x_0$. Design a state feedback control input that minimizes the cost*

$$\int_0^T \big(x(t)^2 + u(t)^2\big)dt.$$

Solution: *By the given ODE we infer that $A = 0$ and $B = 1$, both scalars as the system has $n = 1$ state. From the cost function description we have that*

$$Q = 1,\ R = 1, \quad \text{and} \quad Q_T = 0. \qquad \text{[all scalars]}$$

As a result, $P(t)$ is a scalar that needs to satisfy the Riccati equation, which under the numerical values above becomes:

$$\begin{aligned} -\dot{P}(t) &= P(t)A + A^\top P(t) + Q - P(t)BR^{-1}B^\top P(t) \\ &= 1 - P(t)^2, \end{aligned}$$

with $P(T) = 0$. We solve this ODE by separation of variables. We have that ($P(t) \neq \pm 1$ as this would not be compatible with the terminal condition)

$$\frac{1}{1-P^2}dP = -dt \ \Rightarrow\ \int \frac{1}{1-P^2}dP = -\int dt$$

[using partial fraction expansion]

$$\Leftrightarrow\ \int \Big(\frac{1}{2}\frac{1}{1-P} + \frac{1}{2}\frac{1}{1+P}\Big)dP = -t + \textit{constant}$$

$$\Rightarrow\ -\frac{1}{2}\ln(|1-P|) + \frac{1}{2}\ln(|1+P|) = -t + \textit{constant}$$

[using $P(T) = 0 \Rightarrow$ constant $= T$]

$$\Leftrightarrow\ \ln\Big|\frac{1+P}{1-P}\Big| = 2(T-t)\ \Rightarrow\ P(t) = \frac{e^{2(T-t)} - 1}{e^{2(T-t)} + 1}.$$

Notice that $P(t) \geq 0$ for all $t \in [0,T]$ (we used $\left|\frac{1+P}{1-P}\right| = \frac{1+P}{1-P}$ as the other case is not compatible with the terminal condition). The optimal LQR controller is

$$u^\star(t) = -R^{-1}B^\top P(t)x(t) = -\frac{e^{2(T-t)}-1}{e^{2(T-t)}+1}x(t),$$

resulting in an optimal cost $J(u^\star) = x_0^\top P(0)x_0 = \frac{e^{2T}-1}{e^{2T}+1}x_0^2$.

Proof of Theorem 6. There are several approaches to obtain the LQR controller description, e.g., by means of dynamic programming or via the calculus of variations. Here, we will do this by means of the so called "completing the square" approach; its name will be clear in the sequel. It should be noted that we will not show that the Riccati equation admits a unique solution; we will accept this as a fact and denote this solution by $P(t)$.

The cost criterion that we seek to minimize can be written as

$$\begin{aligned}
J(u) &= \int_0^T \big(x(t)^\top Q x(t) + u(t)^\top R u(t)\big)dt + x(T)^\top Q_T x(T) \\
&\quad \textit{[adding and subtracting } x_0^\top P(0)x_0 \textit{ and since } Q_T = P(T)\textit{]} \\
&= \int_0^T \big(x(t)^\top Q x(t) + u(t)^\top R u(t)\big)dt \\
&\qquad + x(T)^\top P(T)x(T) - x_0^\top P(0)x_0 + x_0^\top P(0)x_0 \\
&\quad \textit{[since } x(T)^\top P(T)x(T) - x_0^\top P(0)x_0 = \int_0^T \frac{d}{dt}\big(x(t)^\top P(t)x(t)\big)dt\textit{]} \\
&= \int_0^T \Big(x(t)^\top Q x(t) + u(t)^\top R u(t) + \frac{d}{dt}\big((x(t)^\top P(t)x(t)\big)\Big)dt + x_0^\top P(0)x_0.
\end{aligned}$$

Differentiating the quadratic expression we obtain that

$$\begin{aligned}
&\frac{d}{dt}\big(x(t)^\top P(t)x(t)\big)dt \\
&= x(t)^\top P(t)\dot{x}(t) + \dot{x}(t)^\top P(t)x(t) + x(t)^\top \dot{P}(t)x(t) \\
&= x(t)^\top P(t)(Ax(t)+Bu(t)) + (Ax(t)+Bu(t))^\top P(t)x(t) + x(t)^\top \dot{P}(t)x(t) \\
&= x(t)^\top \Big(P(t)A + A^\top P(t) + \dot{P}(t)\Big)x(t) + x(t)^\top P(t)Bu(t) + u(t)^\top B^\top P(t)x(t) \\
&= x(t)^\top \Big(P(t)BR^{-1}B^\top P(t) - Q\Big)x(t) + x(t)^\top P(t)Bu(t) + u(t)^\top B^\top P(t)x(t),
\end{aligned}$$

where the second equality is due to the fact that $\dot{x}(t) = Ax(t)+Bu(t)$, and the last one is due to the fact that "blue" terms are equal since $P(t)$ satisfies the Riccati

differential equation. Substituting this derivative within the worked out expression for $J(u)$ (notice that the terms involving Q cancel out) we obtain that

$$\begin{aligned} J(u) &= \int_0^T \Big[x(t)^\top P(t)BR^{-1}B^\top P(t)x(t) + u(t)^\top Ru(t) \\ &\quad + x(t)^\top P(t)Bu(t) + u(t)^\top B^\top P(t)x(t)\Big]dt + x_0^\top P(0)x_0 \\ &= \int_0^T \Big(u(t) + R^{-1}B^\top P(t)x(t)\Big)^\top R\,\Big(u(t) + R^{-1}B^\top P(t)x(t)\Big)dt \\ &\quad + x_0^\top P(0)x_0. \end{aligned}$$

The last step can be shown by direct calculation and using the fact that $P(t)$ is symmetric (to see this expand the product in the second equality and notice that it is identical to the sum in the square brackets in the first equality). The resulting expression involves a quadratic integrand, thus justifying the fact that this procedure is termed "completion of the square". Only the first term in that expression depends on u and, since $R \succ 0$, this quantity is minimized if the integrand is zero. This leads to

$$u^\star(t) = -R^{-1}B^\top P(t)x(t),$$

while the optimal cost becomes $J(u^\star) = x_0^\top P(0)x_0$, thus concluding the proof.

8.2 Infinite horizon optimal control problem

We consider now an optimal control problem with an infinite time horizon. To this end, there is no longer a terminal cost and we let $T \to \infty$, leading to the following infinite horizon LQR problem.

Infinite horizon LQR problem: Find $u(\cdot)$ such that we

$$\begin{aligned} \text{minimize} \quad & J(u) = \int_0^\infty \Big(x(t)^\top Qx(t) + u(t)^\top Ru(t)\Big)dt \\ \text{subject to} \quad & \dot{x}(t) = Ax(t) + Bu(t), \text{ for all } t, \\ & x(0) = x_0. \end{aligned}$$

Notice that the difference with the finite horizon LQR problem is that the cost criterion is now the cumulative penalty over an infinite time horizon. We can

obtain the optimal solution to this problem as established in the following theorem, by taking the limit as $T \to \infty$ of the associated finite horizon problem with zero terminal cost.

Theorem 7 (infinite horizon LQR controller)**.** *Assume that (A, B) is controllable. The infinite horizon LQR problem can be solved by means of the following steps:*

1. *For an arbitrary T let $Q_T = 0$ and solve the Riccati differential equation introduced in the finite horizon case. Denote its solution by $P(t)$, note that $P(t)$ depends on T, and compute*

$$\bar{P} = \lim_{T \to \infty} P(t).$$

Notice that the limit is with respect to T and not t. It turns out that the limit exists and $\bar{P} \in \mathbb{R}^{n \times n}$ will be a symmetric constant matrix, independent of t, and in particular, it will be a positive semidefinite solution to the algebraic Riccati equation

$$PA + A^\top P + Q - PBR^{-1}B^\top P = 0.$$

2. *The infinite horizon optimal LQR controller can be then constructed as*

$$u^\star(t) = Kx(t) = -R^{-1}B^\top \bar{P}x(t).$$

3. *The optimal LQR cost is given by*

$$J(u^\star) = x_0^\top \bar{P} x_0.$$

It should be remarked that:

- Unlike the finite horizon case, the cost may no longer be bounded. To ensure that the cost does not escape to infinity we impose the assumption that (A, B) is controllable. In fact, we could allow for systems that are not necessary controllable, but stabilizable.
- We will not show formally that $\bar{P}$ is a constant matrix, i.e., independent of t. This relies on the fact that the system under consideration is time invariant.

Moreover, $\bar{P}$ is guaranteed to be a solution of the algebraic Riccati equation (notice that this follows from the differential one as $\dot{P}(t)$ vanishes for constant matrices), however, it is not necessarily a unique one, and other solutions (possibly negative definite) may also exist (see Example 20). For the other solutions, the resulting controller is not guaranteed to be optimal.

- The optimal controller is again state feedback, however, this time the feedback gain matrix $K = -R^{-1}B^\top \bar{P}$ is time invariant. As in the finite horizon case, the optimal cost depends only on the initial state x_0.
- The constructed optimal controller is not guaranteed to result in a stable closed loop system. This is illustrated in Example 21 below. This example provides insight on the underlying issues for such an undesirable behaviour. We will then discuss which additional property our system needs to exhibit so that we overcome this and achieve a stable closed loop performance.

✎ **Example 20.** *Consider the infinite horizon counterpart of Example 19. Determine the optimal infinite horizon LQR controller, and specify all solutions of the algebraic Riccati equation.*

Solution: *By Example 19 (notice that $Q_T = 0$) we have that*

$$P(t) = \frac{e^{2(T-t)} - 1}{e^{2(T-t)} + 1} \;\Rightarrow\; \bar{P} = \lim_{T\to\infty} P(t) = 1,$$

i.e., taking the limit as $T \to \infty$ we obtain $\bar{P}$ as the limit of the solution of the Riccati differential equation. The infinite horizon LQR controller is then given by $u^\star(t) = -x(t)$, which results in the closed loop system $\dot{x}(t) = -x(t)$, which is asymptotically stable (scalar state, eigenvalue equals to -1). The optimal LQR cost is $J(u^\star) = x_0^\top \bar{P} x_0 = x_0^2$.

For the particular system and cost function matrices (all scalars in this case) – see Example 19 for numerical values) – the algebraic Riccati equation becomes

$$\begin{aligned} & PA + A^\top P + Q - PBR^{-1}B^\top P = 0 \\ \Rightarrow\; & 1 - P^2 = 0 \\ \Rightarrow\; & P = 1 \text{ or } P = -1. \end{aligned}$$

We thus observe that the algebraic Riccati equation admits multiple solutions, with $\bar{P}$ being one of them.

Example 21. *Consider an LTI system whose state evolves according to $\dot{x}(t) = x(t) + u(t)$, starting from $x(0) = x_0$. Design a state feedback control input that minimizes the cost*

$$\int_0^\infty u(t)^2 dt.$$

Moreover, specify all solutions of the algebraic Riccati equation.

Solution: *The integrand of the cost function depends only on the input, and in fact it is quadratic in $u(t)$. Therefore, for this case there is no need to determine the solution of the Riccati differential equation; it suffices to notice that the cost is minimized if $u^\star(t) = 0$, which is thus the optimal controller resulting in zero cost. The closed loop system becomes then*

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

For the given system and cost function we have that $A = B = 1$, $Q = 0$ and $R = 1$ (all scalars). The algebraic Riccati equation becomes then

$$\begin{aligned} & PA + A^\top P + Q - PBR^{-1}B^\top P = 0 \\ \Rightarrow\ & 2P - P^2 = 0 \ \Rightarrow\ P(2-P) = 0 \\ \Rightarrow\ & P = 0 \text{ or } P = 2. \end{aligned}$$

We thus observe again that the algebraic Riccati equation admits multiple solutions, both of them in this case being positive semidefinite.

Example 21 illustrates that the infinite horizon LQR controller does not necessarily lead to a stable closed loop system. To see this notice that the solution of the closed loop system $\dot{x}(t) = x(t)$ escapes to infinity. The reason for this is that the original system when no input is applied is unstable (scalar state, eigenvalue is equal to 1), however, controllable as the controllability matrix in this case would be $P = B = 1$ ($A = B = 1$, all of them being scalars). However, the (unstable but

controllable) state is not penalized in the LQR cost that involves only the input. As a result an increase of the system state (as this escapes to infinity) does not result in an increase in the cost function we seek to minimize, hence the resulting controller does not attempt to prevent this increase.

It turns out, that to avoid such cases we need to "see" the unstable parts of the system in the LQR cost. Formally, this can be achieved if matrix Q that penalizes the state in the cost can be written as $C^\top \bar{Q} C$, where $\bar{Q}$ is some positive definite matrix, and the pair (A, C) is observable. In fact, the system does not necessarily need to be observable, but detectable.

Overall, combining this with Theorem 7, if Q admits a representation as outlined above, and the pairs (A, B) and (A, C) are controllable and observable, respectively, then the infinite horizon LQR controller results in a stable (in fact asymptotically stable) closed loop performance. Moreover, in this case $\bar{P}$ is the unique positive semidefinite solution of the algebraic Riccati equation. This suggests that rather than solving the differential Riccati equation and letting $T \to \infty$ to construct $\bar{P}$, we can directly solve the algebraic Riccati equation and keep its (unique this time) positive semidefinite solution. Notice that this can be verified in Example 20 (taking the output to be equal to the single state, i.e., $C = 1$), where a unique positive semidefinite solution exists (the other one is negative definite), while it is not the case in Example 21, as Q is zero thus not admitting the desired representation.

Remark. We provide some further discussion on the assumptions required for an asymptotically stable behaviour of the closed loop system, when the optimal infinite horizon LQR controller is applied. To this end, let the output of the system be in the form $y(t) = Cx(t)$. The following arguments can be thought of as a sketch of the proof for asymptotic stability of the closed loop system.

1. (A, B) controllable: Under $u(t) = Kx(t)$ the closed loop system becomes $\dot{x}(t) = (A + BK)x(t)$. Since (A, B) is controllable, gain choices exist so that the eigenvalues of $A + BK$ have negative real part. The closed loop state solution $x(t) = e^{(A+BK)t}x_0$ (and hence $u(t) = Kx(t)$) is then bounded by decaying exponentials, converging to zero. Under such a gain, the infinite horizon cost would be finite (as the integral of a decaying exponential). To see

this, notice that under such a gain K, we have that $x(t) \leq \mu e^{-\bar{\lambda}t} x_0$, where $\bar{\lambda} > 0$ is real, indicating the magnitude of the smallest real part among the eigenvalues of $A + BK$, and $\mu > 0$ is some positive constant. The infinite horizon cost under such a controller can be expressed as

$$\begin{aligned} J(u) &= \lim_{T\to\infty} \int_0^T \Big(x(t)^\top Q x(t) + u(t)^\top R u(t) \Big) dt \\ &= \lim_{T\to\infty} \int_0^T x(t)^\top (Q + K^\top R K) x(t) dt \\ &= \lim_{T\to\infty} \int_0^T \mu^2 \Big(x_0^\top (Q + K^\top R K) x_0 \Big) e^{-2\bar{\lambda}t} dt, \end{aligned}$$

where the first equality is due to the choice of $u(t)$, and the second one is due to $x(t)$ being bounded by a decaying exponential. Since $Q + K^\top R K \succeq 0$, notice that $c = \mu^2 \Big(x_0^\top (Q + K^\top R K) x_0 \Big) > 0$. We then have that

$$\begin{aligned} J(u) &= \lim_{T\to\infty} \int_0^T c e^{-2\bar{\lambda}t} dt \\ &= \lim_{T\to\infty} \frac{c}{2\bar{\lambda}} (1 - e^{-2\bar{\lambda}T}) = \frac{c}{2\bar{\lambda}} < \infty, \end{aligned}$$

where the inequality is since $c < \infty$ as the initial state x_0 is arbitrary but finite. This shows that the infinite horizon cost is finite if (A, B) is controllable.

Since under controllability we have shown that there exist gain choices for which the cost is finite, the gain K obtained via Theorem 7 would necessarily be one of those, as otherwise we would get an infinite and hence higher cost (recall that we seek to minimize the cost).

2. $Q = C^\top \bar{Q} C$ with $\bar{Q} \succ 0$: Under $Q = C^\top \bar{Q} C$, we have

$$\begin{aligned} J(u) &= \int_0^\infty \Big(x(t)^\top C^\top \bar{Q} C x(t) + u(t)^\top R u(t) \Big) dt \\ &= \int_0^\infty \Big(y(t)^\top \bar{Q} y(t) + u(t)^\top R u(t) \Big) dt. \end{aligned}$$

The fact that $\bar{Q} \succ 0$ implies that we "see" the output in the cost. Therefore, to keep the cost for the optimal input $u^\star(t)$ computed as in Theorem 7 finite, and since the output is present in the cost, we must have $\lim_{t\to\infty} \|y^\star(t)\| = 0$ and $\lim_{t\to\infty} \|u^\star(t)\| = 0$, where $y^\star(t) = C x^\star(t)$ is the output along the optimal trajectory (the cost is evaluates at the optimal input-output).

Notice that if the limiting behaviour of the output and the input was different (recall that $x^\star$, and hence $y^\star, u^\star$ as linear functions of $x^\star$, is in given by a linear combinations of (complex) exponentials) the cost would be infinite.

3. (A, C) observable: Under observability of (A, C), it becomes possible to infer that $\lim_{t\to\infty} \|x^\star(t)\| = 0$ (and hence that the closed loop system is asymptotically stable), from the fact that $\lim_{t\to\infty} \|y^\star(t)\| = 0$ and $\lim_{t\to\infty} \|u^\star(t)\| = 0$. One way to see this is to notice that if (A, C) is observable, then there exists some gain matrix L such that the eigenvalues of $A - LC$ have negative real part. We then have

$$\begin{aligned}\dot{x}^\star(t) &= Ax^\star(t) + Bu^\star(t) \\ &= (A - LC)x^\star(t) + Ly^\star(t) + Bu^\star(t),\end{aligned}$$

where the second equality follows from adding and subtracting $LCx^\star(t)$, and using $y^\star(t) = Cx^\star(t)$. The "input" of that system $Ly^\star(t) + Bu^\star(t)$ vanishes as time tends to infinity (since $\lim_{t\to\infty} \|y^\star(t)\| = 0$ and $\lim_{t\to\infty} \|u^\star(t)\| = 0$), while $A - LC$ has eigenvalues with negative real part. As a result, we have that $\lim_{t\to\infty} \|x^\star(t)\| = 0$, thus implying that the closed loop system (under the optimal infinite horizon LQR controller) is asymptotically stable.

Alternatively, we could follow the derivation of p. 56, however, evaluating the stacked vectors at $t \to \infty$ instead of $t = 0$. This leads to $Y = U = 0$ as $\lim_{t\to\infty} \|y^\star(t)\| = \lim_{t\to\infty} \|u^\star(t)\| = 0$. Setting $x^\star_\infty = \lim_{t\to\infty} x^\star(t)$ we have $Qx^\star_\infty = 0 \Leftrightarrow x^\star_\infty = 0$, as Q (the observability matrix as in p. 56) is full rank.

8.3 Summary

This chapter introduced the so called *Linear Quadratic Regulator (LQR)* both for finite and infinite horizon optimal control problems, and showed that the LQR controller minimizes a cost function which is quadratic with respect to the state and input vector subject to the linear dynamics. The main learning outcomes of the chapter can be summarized as follows:

- *Finite horizon LQR controller:*

  The finite horizon LQR problem is time varying and is given by

  $$u^\star(t) = -R^{-1}B^\top P(t)x(t),$$

  where $P(t)$ is symmetric and is the unique positive semidefinite solution

of the *Riccati differential equation*

$$-\dot{P}(t) = P(t)A + A^\top P(t) + Q - P(t)BR^{-1}B^\top P(t)$$
$$\text{with} \quad P(T) = Q_T.$$

- *Infinite horizon LQR controller:*

  If (A, B) is controllable or stablizable, the infinite horizon LQR problem is time invariant and is given by

  $$u^\star(t) = -R^{-1}B^\top \bar{P}x(t),$$

  where $\bar{P} = \lim_{T\to\infty} P(t)$ is a symmetric constant matrix that results as the limit of the solution $P(t)$ of the Riccati differential equation with $Q_T = 0$. It turns out that $\bar{P}$ is a positive semidefinite solution (not necessarily unique) of the *algebraic Riccati equation*

  $$PA + A^\top P + Q - PBR^{-1}B^\top P = 0.$$

  If in addition $Q = C^\top \bar{Q} C$ with $\bar{Q}$ being positive definite, and (A, C) is observable or detectable, then $\bar{P}$ is the unique positive semidefinite solution of the algebraic Riccati equation and the resulting LQR controller renders the closed loop system (asymptotically) stable.

9 Appendix

In this chapter we list some basic results from linear algebra and analysis that are used throughout the notes. The statements are provided without proofs; they constitute a condensed summary and should not be treated as a detailed exposition of the topic.

9.1 Selected results from linear algebra

9.1.1 Vectors and independence

Definition 3 (Linear independence). *A set of vectors $x_1, \ldots, x_m \in \mathbb{R}^n$ is said to be linearly independent if for scalars $a_1, \ldots, a_n$,*

$$a_1 x_1 + \ldots + a_m x_m = 0 \Leftrightarrow a_1 = \ldots = a_m = 0.$$

Otherwise, they are called linearly dependent.

9.1.2 Matrix properties

Fact 14 (Matrix product). *For given matrices with arbitrary but appropriate dimensions the matrix product satisfies the following properties:*

1. *Associative: $(AB)C = A(BC)$.*
2. *Distributive with respect to addition: $A(B + C) = AB + AC$.*
3. *Non-commutative (in general): $AB \neq BA$.*
4. *Transpose: $(AB)^\top = B^\top A^\top$.*

Notice that despite the fact that matrices in general do not commute, there still exist matrices such that $AB = BA$. As an example, consider the following case:

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

Definition 4 (Range & null space)**.** *The range and null space associated with a matrix $A \in \mathbb{R}^{m\times n}$ are given by*

$$\begin{aligned}\mathrm{range}(A) &= \{y \in \mathbb{R}^m : \ \exists x \in \mathbb{R}^n \textit{ such that } Ax = y\},\\ \mathrm{null}(A) &= \{x \in \mathbb{R}^n \textit{ such that } Ax = 0\}.\end{aligned}$$

Recall that a square matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if $\mathrm{range}(A) = \mathbb{R}^n$, or equivalently if and only if $\mathrm{null}(A) = \{0\}$. For an invertible matrix $A \in \mathbb{R}^{n\times n}$, we can directly compute its inverse by

$$A^{-1} = \frac{\mathrm{adj}(A)}{\det(A)},$$

where $\det(A)$ denotes the determinant of A, and $\mathrm{adj}(A)$ is the so called adjoint matrix, given by the transpose of a matrix whose (i,j)-th entry is given by $(-1)^{i+j}M_{ij}$, where M_{ij} is the determinant of the $(n-1)\times(n-1)$ matrix that emanates if the i-th row and j-th column of A are removed.

9.1.3 Matrix eigenvalues and eigenvectors

Definition 5 (Eigenvalues & eigenvectors)**.** *A nonzero vector $w \in \mathbb{C}$ is called an eigenvector of a matrix $A \in \mathbb{R}^{n\times n}$, if there exists $\lambda \in \mathbb{C}$ such that*

$$Aw = \lambda w.$$

We then call λ an eigenvalue of A.

Recall that even if the entries of A are real, the eigenvalues and eigenvectors could be complex. The eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ can be determined as the n roots of the so called characteristic polynomial, i.e.,

$$\det(\lambda I - A) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n = 0,$$

where I is an identity matrix with appropriate dimensions. Parameters $a_1, \ldots, a_n$ are the coefficients of this n-th order polynomial.

Theorem 8 (Cayley-Hamilton theorem). *A matrix $A \in \mathbb{R}^{n\times n}$ satisfies its characteristic polynomial, i.e.,*

$$A^n + a_1 A^{n-1} + \ldots + a_{n-1} A + a_n I = 0.$$

9.1.4 Matrix decomposition and positive definiteness

Definition 6 (Diagonalizable matrices). *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be diagonalizable if its eigenvectors are linearly independent.*

We then have the following sufficient conditions for a matrix to be diagonalizable, i.e., to have linearly independent eigenvectors:

1. If a matrix $A \in \mathbb{R}^{n\times n}$ has distinct eigenvalues (i.e., $\lambda_i \neq \lambda_j$ for all $i \neq j$), then its eigenvectors are linearly independent.
2. If a matrix is symmetric ($A = A^\top \in \mathbb{R}^{n\times n}$), then i) its eigenvalues are real; ii) its eigenvectors are orthonormal.

Diagonalizable matrices admit the following decomposition.

Fact 15 (Decomposition of diagonalizable matrices). *If $A \in \mathbb{R}^{n\times n}$ is a diagonalizable matrix, then*

$$A = W\Lambda W^{-1},$$

where $W \in \mathbb{C}^{n\times n}$ is a matrix whose columns are the eigenvectors of A, and $\Lambda \in \mathbb{C}^{n\times n}$ is a diagonal matrix whose diagonal entries correspond to the eigenvalues of A.
If A is also symmetric (and hence diagonalizable), i.e., $A = A^\top$, then

$$A = W\Lambda W^\top.$$

Note that even if W and Λ could in general have complex entries, $W\Lambda W^{-1}$ (this is just A) always has real entries. Intuitively this occurs since complex eigenvalues appear in conjugate pairs and cancellations of the imaginary parts occur when multiplying them together.

Definition 7 (Positive definiteness). *A symmetric matrix ($A = A^\top \in \mathbb{R}^{n\times n}$) is positive definite and we write $A \succ 0$ if $x^\top Ax > 0$ for all $x \neq 0$. It is positive semidefinite and we write $A \succeq 0$ if $x^\top Ax \geq 0$ for all $x \neq 0$.*

Notice that $A \preceq B$ (inequality in the positive semidefinite sense) is equivalent to $B - A \succeq 0$, or in other words $x^\top (B-A)x \geq 0$ all for $x \neq 0$. A consequence of this result is that if A is symmetric, then $A \preceq \max_{i=1,\dots,n} \lambda_i(A)\ I = \lambda_{\max}(A)\ I$, where $\lambda_i(A)$ denotes the i-th eigenvalue of A, $\lambda_{\max}(A)$ the largest eigenvalue of A, and I is the identity matrix of appropriate dimension.

Fact 16 (Maximum singular value dominance). *For $A^\top A$ (notice it is symmetric even if A is not, or if $A \in \mathbb{R}^{m\times n}$), we have that*

$$A^\top A \preceq \lambda_{\max}(A^\top A)\ I,$$

where $\sqrt{\lambda_{\max}(A^\top A)}$ is the maximum singular value of A.

It should be noted that the eigenvalues of $A^\top A$ are non-negative so the square root is always well defined.

9.2 Selected results from analysis

9.2.1 Norms

Definition 8 (Norm). *A norm on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ such that*

1. *Triangle inequality: $\|x + \hat{x}\| \leq \|x\| + \|\hat{x}\|$, for all $x, \hat{x} \in \mathbb{R}^n$.*
2. *Scalar multiplication: $\|ax\| = |a|\ \|x\|$, for all $x \in \mathbb{R}^n$, $a \in \mathbb{R}$.*
3. *Zero element: $\|x\| = 0 \Leftrightarrow x = 0$ (zero vector in $\mathbb{R}^n$).*

There are many norms; in these notes we will be using the Euclidean norm (2-norm), i.e., for any $x = [x_1 \dots x_n]^\top \in \mathbb{R}^n$,

$$\|x\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

Notice that for simplicity we indicate it by $\|\cdot\|$, without introducing a subscript. Note that $\|x\| = \sqrt{x^\top x}$, while for a matrix $A \in \mathbb{R}^{m\times n}$ we have that $\|Ax\| = \sqrt{x^\top A^\top Ax}$ (we used the property of the matrix product transpose).

There also exist several matrix norms, as well as induced norms. Induced norms (as the name suggests) are induced by the application of a matrix $A \in \mathbb{R}^{m\times n}$ on a vector $x \in \mathbb{R}^n$, or in other words by the mapping Ax, and depend on the choice of norm for $\mathbb{R}^n$ (where x takes values from) and $\mathbb{R}^m$ (where Ax takes values from). Here, in the occasions where a matrix norm will be used, we will imply

$$\|A\| = \sqrt{\lambda_{\max}(A^\top A)}, \qquad \text{[maximum singular value]},$$

where $\sqrt{\lambda_{\max}(A^\top A)}$ is the maximum singular value of A.

Notice that we use both $\|\cdot\|$ both for the Euclidean norm and the induced norm above, however, the distinction should always be clear from whether the argument is a vector or a matrix, respectively.

9.2.2 Linearity and continuity

Definition 9 (Linearity). *A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ (possibly vector-valued) is called linear if for any $x, \hat{x} \in \mathbb{R}^n$ and scalars $a_1, a_2 \in \mathbb{R}$,*

$$f(a_1 x + a_2 \hat{x}) = a_1 f(x) + a_2 f(\hat{x}).$$

Definition 10 (Continuity). *Consider any $\hat{x} \in \mathbb{R}^n$. A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is said to be continuous at $\hat{x} \in \mathbb{R}^n$ if for all $\epsilon > 0$, there exists $\delta > 0$, such that*

$$\text{for any } x \in \mathbb{R}^n \text{ with } \|x - \hat{x}\| < \delta \;\Rightarrow\; \|f(x) - f(\hat{x})\| < \epsilon.$$

Intuitively, this implies that we can always pick x close enough to a given $\hat{x}$ (δ-close), if we would like $f(x)$ to remain close to $f(\hat{x})$ (ϵ-close). A stronger property than continuity is the so called Lipschitz continuity condition, that encompasses a certain function growth property.

Definition 11 (Lipschitz continuity). *A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if there exists a (finite) scalar $L > 0$ such that for all $x, \hat{x} \in \mathbb{R}^n$,*

$$\|f(x) - f(\hat{x})\| \leq L\|x - \hat{x}\|.$$

L is then called the Lipschitz constant of f.

Note that the Lipschitz continuity definition provided above is often referred to as global Lipschitz continuity, as the same Lipschitz constant exists for all points $x, \hat{x}$. Intuitively, Lipschitz continuity implies that a function cannot grow infinitely steep (think of x^2 as x tends to infinity). Lipschitz continuous functions are continuous but not vice versa (for instance, $f(x) = \sqrt{x}$ is continuous but not Lipschitz continuous). Moreover, a function f does not need to be differentiable to be Lipschitz continuous.

✎ **Example 22.** *Show that the absolute value function $f(x) = |x|$ (non-differentiable at zero) is Lipschitz continuous with Lipschitz constant $L = 1$.*

Solution: *To show this we start by expanding the left-hand side in the Lipschitz continuity definition (note that the norm in the scalar case is just the absolute value). We distinguish different cases according to the sign of $x, \hat{x}$. Consider first the case where $x, \hat{x} \geq 0$. We then have that $|x| = x$ and $|\hat{x}| = \hat{x}$. Hence,*

$$\big||x| - |\hat{x}|\big| = |x - \hat{x}|.$$

Consider now the case where $x \geq 0$ and $\hat{x} \leq 0$. We then have that $|x| = x$ and $|\hat{x}| = -\hat{x}$, which implies that

$$\big||x| - |\hat{x}|\big| = |x + \hat{x}| \leq |x - \hat{x}|,$$

where the last inequality is due to the fact that $\hat{x} \leq 0$. Reversing the roles of x and $\hat{x}$ covers the remaining cases. Therefore, in all cases we obtain $\big||x| - |\hat{x}|\big| \leq |x - \hat{x}|$, thus showing that the absolute value is a Lipschitz continuous function with Lipschitz constant equal to 1.

However, if a function is differentiable with bounded derivatives, then it is also Lipschitz continuous. In general, the following implications hold:

Continuity $\Leftarrow$ Lipschitz continuity $\Leftarrow$ Differentiability with bounded derivatives.