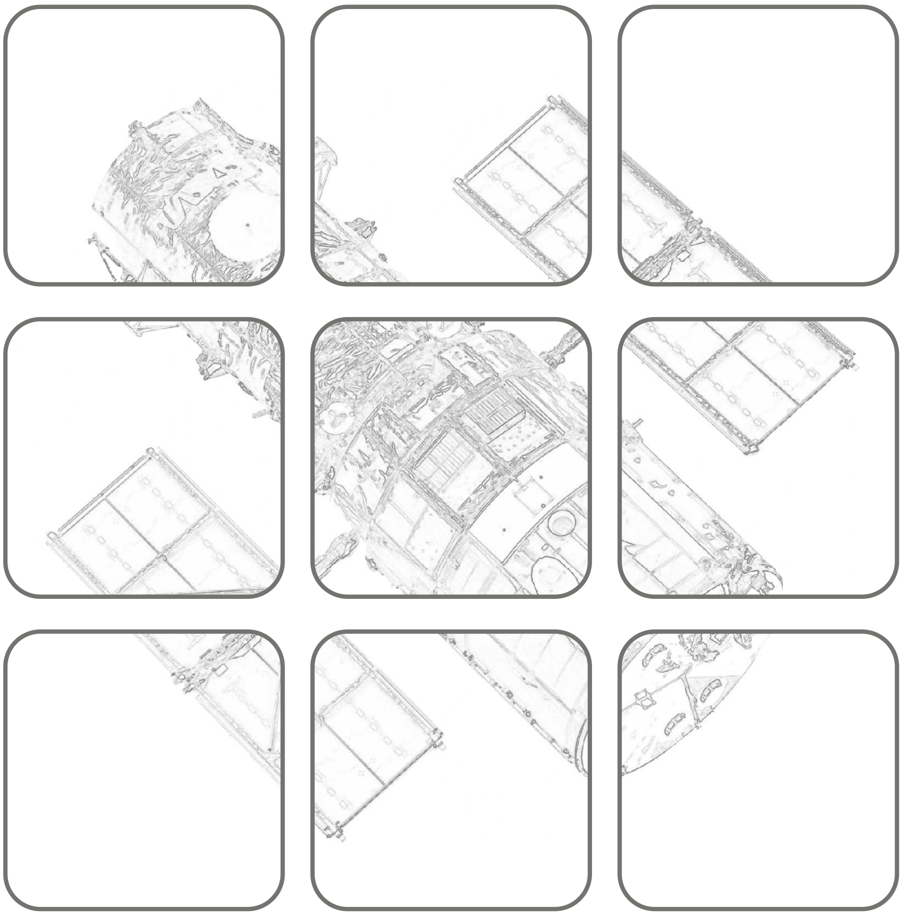
TECHNISCHE
UNIVERSITÄT
WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS-
& REGELUNGSTECHNIK

CONTROL IN QUANTUM SCIENCE
THEORETICAL CONCEPTS

Lecture
SS 2025

Ass. Prof. Dr. techn. Andreas DEUTSCHMANN-OLEK
Dr. techn. Lukas TARRA



Control in Quantum Science: Theoretical Concepts

Lecture
SS 2025

Ass. Prof. Dr. techn. Andreas DEUTSCHMANN-OLEK
Dr. techn. Lukas TARRA

TU Wien
Institut für Automatisierungs- und Regelungstechnik
Gruppe für komplexe dynamische Systeme

Gußhausstraße 27–29
1040 Wien
Telefon: +43 1 58801 – 37615
Internet: https://www.acin.tuwien.ac.at

Contents

List of Figures

1 Nonlinear Systems

The analysis and design methods for controlling linear systems are by far the most advanced. This is due to the superposition principle, which significantly simplifies the mathematical treatment of this class of dynamical systems. However, physical laws often contain significant nonlinearities. When these can no longer be neglected, one must resort to the methods of nonlinear control engineering.

Due to the *superposition principle*, *local* and *global* properties coincide in linear systems. This is no longer the case for *nonlinear dynamical systems.* If one restricts oneself to local properties in nonlinear systems, often linear methods can still be used by linearizing the system equations. However, if one is interested in global properties, the full nonlinear mathematical model must be examined.

A large class of nonlinear dynamical systems can be described by mathematical models of first-order nonlinear differential equations. For these models, there is no simple tool available for input-output description as in the case of Laplace transformation in linear systems. Therefore, the analysis of such systems is preferably done in state space.

1.1 Linear and Nonlinear Systems

The relationship

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{1.1}$$

describes a linear, time-invariant, autonomous system of n-th order with lumped parameters. Besides the superposition principle, the system can be characterized by additional properties.

The equilibrium points $\mathbf{x}_R$ of (1.1) are solutions to the linear system of equations

$$\mathbf{0} = \mathbf{A}\mathbf{x}_R \ . \tag{1.2}$$

In the case where $\det(\mathbf{A}) \neq 0$, the system has exactly one equilibrium point, namely $\mathbf{x}_R = \mathbf{0}$; otherwise, it has infinitely many equilibrium points.

> *Exercise* 1.1. Provide a second-order system (1.1) with infinitely many equilibrium points.

With the transition matrix

$$\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t} = \mathbf{E} + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2} + \ldots + \mathbf{A}^n\frac{t^n}{n!} + \ldots \tag{1.3}$$

the solution of the initial value problem is

$$\mathbf{x}(t) = \mathbf{\Phi}(t)\mathbf{x}_0 \ . \tag{1.4}$$

It is easy to see that $\mathbf{x}(t)$ satisfies the inequality

$$a_1 \mathrm{e}^{-\alpha_1 t} \leq \|\mathbf{x}(t)\| \leq a_2 \mathrm{e}^{\alpha_2 t} \tag{1.5}$$

with real numbers $a_1, a_2, \alpha_1, \alpha_2 > 0$. That is, a trajectory $\mathbf{x}(t)$ of the system (1.1) cannot *converge to the equilibrium* $\mathbf{x}_R = \mathbf{0}$ in finite time nor *grow beyond all bounds* in finite time.

These properties do not necessarily hold for a nonlinear, autonomous system of n-th order

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \ . \tag{1.6}$$

The equilibrium points of this system are now solutions to the nonlinear system of equations

$$\mathbf{0} = \mathbf{f}(\mathbf{x}_R) \ . \tag{1.7}$$

No general statement can be made about the solution set $\mathcal{X}_R$ of (1.7). Thus, $\mathcal{X}_R$ can consist of exactly one element, a finite number of elements, or an infinite number of elements.

> *Exercise* 1.2. Provide a first-order system (1.6) with exactly three equilibrium points.

Nonlinear systems can also converge to the equilibrium state in finite time. Consider the equation

$$\dot{x} = -\sqrt{x}, \qquad x_0 > 0 \ . \tag{1.8}$$

For the solution of the above system, we have

$$x(t) = \begin{cases} \left(\sqrt{x_0} - \frac{t}{2}\right)^2 & \text{for} \quad 0 \leq t \leq 2\sqrt{x_0} \\ 0 & \text{otherwise} \ . \end{cases} \tag{1.9}$$

The solution of a nonlinear system can also grow beyond bounds in finite time. For example, consider the system

$$\dot{x} = 1 + x^2, \qquad x_0 = 0 \tag{1.10}$$

with the solution given by

$$x(t) = \tan(t), \qquad 0 \leq t < \frac{\pi}{2} \ . \tag{1.11}$$

There is no solution for $t \geq \frac{\pi}{2}$.

1.2 Satellite Control

Figure 1.1 shows a communication satellite. If the satellite is considered as a rigid body, its rotational motion can be described by the relationship

$$\boldsymbol{\Theta}\dot{\mathbf{w}} = -\mathbf{w} \times (\boldsymbol{\Theta}\mathbf{w}) + \mathbf{M} \tag{1.12}$$

with

$$\mathbf{w} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}, \tag{1.13a}$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \Theta_{11} & \Theta_{12} & \Theta_{13} \\ \Theta_{12} & \Theta_{22} & \Theta_{23} \\ \Theta_{13} & \Theta_{23} & \Theta_{33} \end{bmatrix}, \tag{1.13b}$$

$$\mathbf{M} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} \tag{1.13c}$$

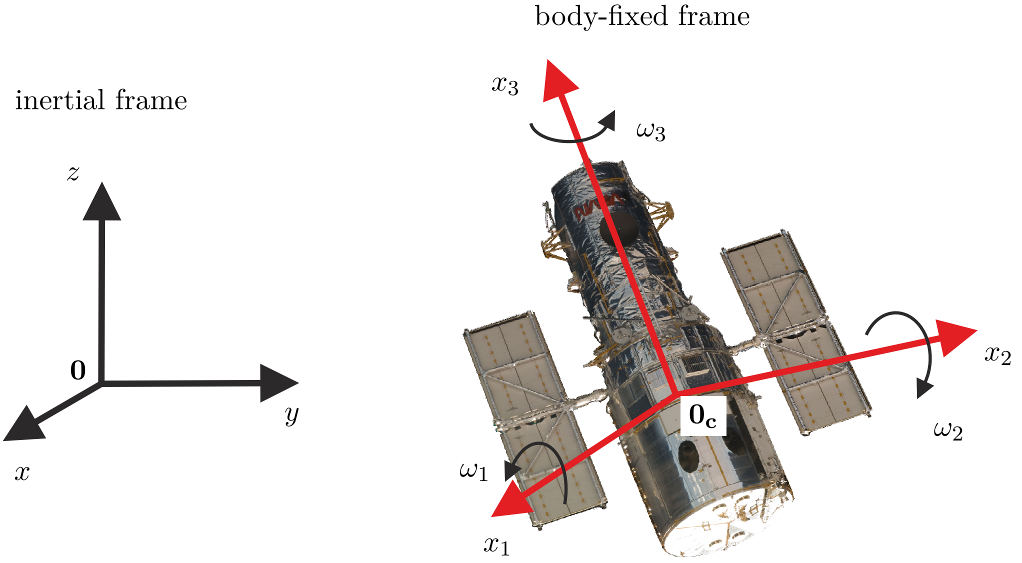


Figure 1.1: Rotational motion of a satellite.

Here, $\mathbf{w}$ denotes the vector of angular velocities, $\boldsymbol{\Theta}$ the inertia matrix, and $\mathbf{M}$ the vector of torques. The quantities $\mathbf{w}$, $\boldsymbol{\Theta}$, and $\mathbf{M}$ are referred to the satellite-fixed coordinate frame $(0_C, x_1, x_2, x_3)$ at the center of mass 0_C. If the coordinate frame $(0_C, x_1, x_2, x_3)$ is aligned with the principal axes of inertia of the satellite, we have

$$\boldsymbol{\Theta} = \begin{bmatrix} \Theta_{11} & 0 & 0 \\ 0 & \Theta_{22} & 0 \\ 0 & 0 & \Theta_{33} \end{bmatrix}, \tag{1.14}$$

which simplifies the above system to

$$\Theta_{11}\dot{\omega}_1 = -(\Theta_{33} - \Theta_{22})\omega_2\omega_3 + M_1 \tag{1.15a}$$
$$\Theta_{22}\dot{\omega}_2 = -(\Theta_{11} - \Theta_{33})\omega_1\omega_3 + M_2 \tag{1.15b}$$
$$\Theta_{33}\dot{\omega}_3 = -(\Theta_{22} - \Theta_{11})\omega_1\omega_2 + M_3 \tag{1.15c}$$

> *Exercise* 1.3. How many fundamentally different equilibrium states can you specify for the satellite (1.15) when $\mathbf{M} = \mathbf{0}$?

1.3 Ball on Beam

A ball with mass m_K rolls on a pivot-mounted beam (see Figure 1.2). The setup is influenced by applying a moment M at the pivot point of the beam. The geometric relationships hold as follows:

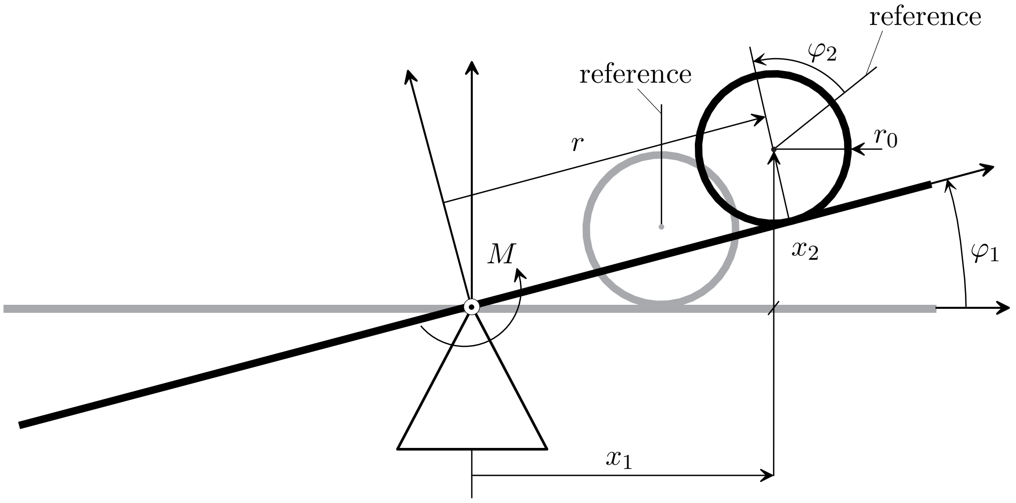


Figure 1.2: Beam with rolling ball.

$$x_1 = r\cos(\varphi_1) - r_0\sin(\varphi_1) \tag{1.16a}$$
$$x_2 = r\sin(\varphi_1) + r_0\cos(\varphi_1) \tag{1.16b}$$

and

$$\dot{r} = -r_0\dot{\varphi}_2\ . \tag{1.17}$$

Neglecting friction forces, the Lagrangian is given by

$$\begin{aligned} L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) = &\underbrace{\frac{1}{2} m_K \left(\dot{x}_1^2(\varphi_1, \dot{\varphi}_1, r, \dot{r}) + \dot{x}_2^2(\varphi_1, \dot{\varphi}_1, r, \dot{r}) \right)}_{\text{translational kinetic energy}} \\ &+ \underbrace{\frac{1}{2} \left(\Theta_B \dot{\varphi}_1^2 + \Theta_K (\dot{\varphi}_1 + \dot{\varphi}_2)^2 \right)}_{\text{rotational kinetic energy}} - \underbrace{m_K g x_2(\varphi_1, r)}_{\text{potential energy}} \end{aligned} \tag{1.18}$$

with the mass of the ball m_K, the moment of inertia of the beam Θ_B, the moment of inertia of the ball $\Theta_K = \frac{2}{5} m_K r_0^2$, and the acceleration due to gravity g.

> *Exercise* 1.4. Show that for the moment of inertia of a homogeneous ball with radius r_0, the following holds:
> $$\Theta_K = \frac{2}{5} m_K r_0^2 \ .$$

Using the generalized coordinates $r(t)$ and $\varphi_1(t)$, the Euler-Lagrange equations yield the system's equations of motion in the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial}{\partial \dot{r}} L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) \right) - \frac{\partial}{\partial r} L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) = 0 \tag{1.19a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial}{\partial \dot{\varphi}_1} L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) \right) - \frac{\partial}{\partial \varphi_1} L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) = M \ . \tag{1.19b}$$

To simplify the results, it is assumed that the ball is a point mass, so $r_0 = 0$ and $\Theta_K = 0$. Thus, the Lagrangian simplifies to

$$L(\varphi_1, \dot{\varphi}_1, r, \dot{r}) = \frac{1}{2} m_K \dot{r}^2 + \frac{1}{2} m_K r^2 \dot{\varphi}_1^2 + \frac{1}{2} \Theta_B \dot{\varphi}_1^2 - m_K g r \sin(\varphi_1) \tag{1.20}$$

and the mathematical model becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \varphi_1 = \frac{1}{m_K r^2 + \Theta_B} (M - 2 m_K r \dot{r} \dot{\varphi}_1 - g m_K r \cos(\varphi_1)) \tag{1.21a}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} r = r \dot{\varphi}_1^2 - g \sin(\varphi_1) \ . \tag{1.21b}$$

The equilibrium positions of this system are given by

$$\varphi_{1,R} = 0 \tag{1.22a}$$

$$M_R = g m_K r_R \tag{1.22b}$$

$$r_R \quad \text{arbitrary.} \tag{1.22c}$$

> *Exercise* 1.5. Replace the rolling ball in Figure 1.2 with a frictionless sliding cube of mass m_2 and edge length l. Provide the Lagrangian function and the equations of motion for this model.

Exercise 1.6. Figure 1.3 shows a crane with a pivot arm. Determine the equations of motion using Lagrangian mechanics. Introduce the generalized coordinates as the angles φ_1 and φ_2. The input variables are the two moments M_1 and M_2.

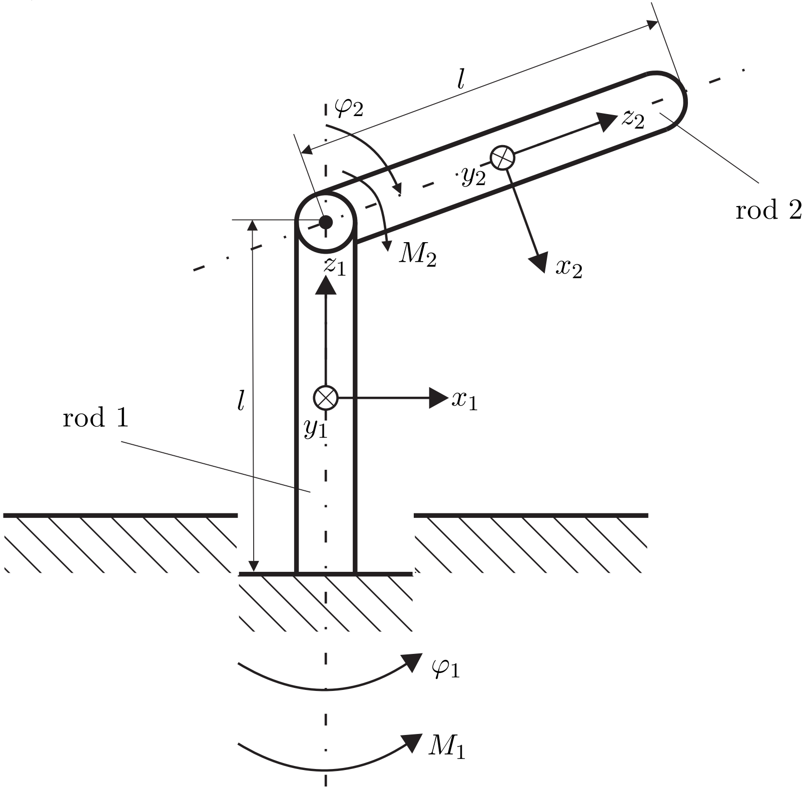


Figure 1.3: Crane with pivot arm.

Exercise 1.7. In Figure 1.4, a simple manipulator consisting of five beam elements is depicted. It is a system with two degrees of freedom, where the quantities q_1 and q_2 are introduced as generalized coordinates. This manipulator has the special property that the system of differential equations decouples when a simple geometric relationship is satisfied. That is, q_1 or q_2 is only influenced by M_1 or M_2. This is particularly convenient for controller design. Such examples are typical mechatronic tasks, as in this case the construction is carried out in such a way that the control task is subsequently simplified. However, knowledge of the mathematical model is required to accomplish this. Manipulators of this type were built, among others, by the company Hitachi under the model designation HPR10II.

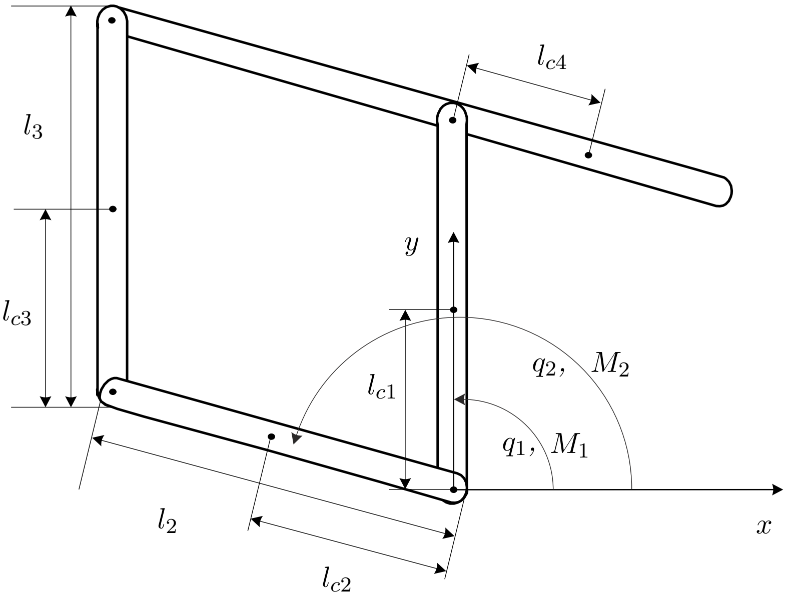


Figure 1.4: Closed kinematic chain.

1.4 Positioning with Static Friction

Figure 1.5 shows a mass m sliding on a rough surface subject to the spring force $F_S = cx$, the friction force F_R, and the input force F_u.

In the friction force model, a distinction is made between *static* and *dynamic models*. In the static model, the friction force F_R is given as a function of the velocity $v = \frac{\mathrm{d}}{\mathrm{d}t}x$.

As shown in Figure 1.6, the friction force generally consists of a *velocity-proportional (viscous) component* $r_v v$, a *Coulomb component (dry friction)* $r_C \mathrm{sign}(v)$, and a *static friction component* described by the parameter r_H. It has also been experimentally observed that the force-velocity curve when entering or leaving the static friction state follows the shape of the dashed curve in Figure 1.6 (*Stribeck effect*). The velocity v_S at which the friction force F_R reaches a minimum is also referred to as the Stribeck velocity. Very often, this behavior is described in the form

$$F_R = r_v v + r_C \,\mathrm{sgn}(v) + (r_H - r_C) \exp\left(-\left(\frac{v}{v_0}\right)^2\right) \mathrm{sgn}(v) \tag{1.23}$$

where a reference velocity v_0 is used for the total friction force. Hence, the mathematical model of Figure 1.5 written relative to the relaxed position of the spring x_0 reads

(1) The static friction condition is satisfied, so $v = 0$ and $|F_u - cx| \leq r_H$,

$$\frac{\mathrm{d}}{\mathrm{d}t}x = 0 \tag{1.24a}$$

$$m\frac{\mathrm{d}}{\mathrm{d}t}v = 0 \tag{1.24b}$$

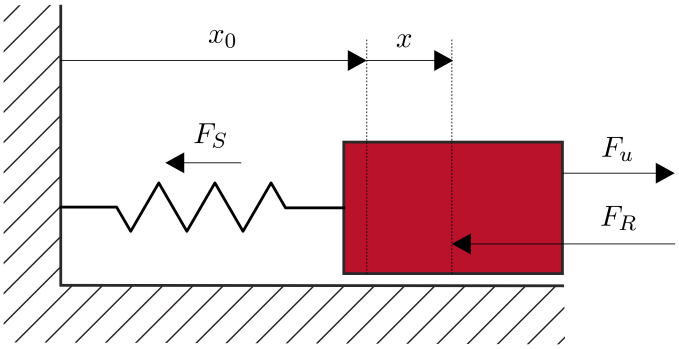


Figure 1.5: Spring-mass system with static friction.

(2) The adhesion condition is not fulfilled

$$\frac{\mathrm{d}}{\mathrm{d}t}x = v \tag{1.25a}$$

$$m\frac{\mathrm{d}}{\mathrm{d}t}v = F_u - F_R - cx \tag{1.25b}$$

with the friction force F_R according to (1.23).

When implementing the mathematical model (1.24) and (1.25) in a numerical simulation program like MATLAB/SIMULINK, it must be ensured that the *structural switching* between (1.24) and (1.25) is correctly implemented. For example, SIMULINK offers dedicated blocks to detect zero-crossings of variables and implement the switching of states using the STATEFLOW TOOLBOX.

Combining static friction with an integral controller generally leads to undesirable limit cycles. To demonstrate this, in the next step, a PI controller will be designed as a

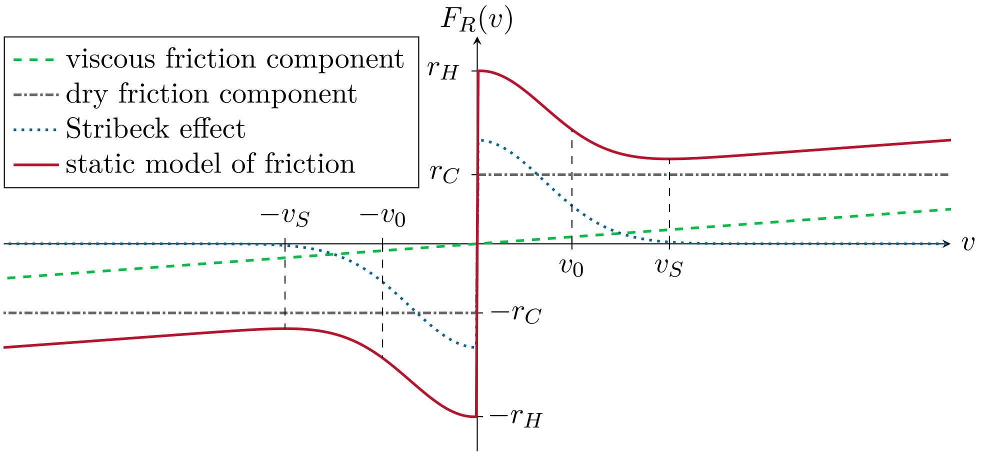


Figure 1.6: Static friction model.

position controller for the spring-mass system shown in Figure 1.5 with the input force F_u. For the design of the PI controller, it is common practice to neglect the Coulomb friction component and the static friction component, i.e., $r_H = r_C = 0$. This results in a simple linear system with position x as the output and force F_u as the input, with the corresponding transfer function

$$G(s) = \frac{\hat{x}}{\hat{F}_u} = \frac{1}{ms^2 + r_v s + c} \tag{1.26}$$

If the parameters are chosen as $c = 2$, $m = 1$, $r_C = 1$, $r_v = 3$, $r_H = 4$, and $v_0 = 0.01$, then the PI controller $R(s) = 4\frac{s+1}{s}$ for the linear system (1.26) leads to the step response of the closed loop shown in Figure 1.7.

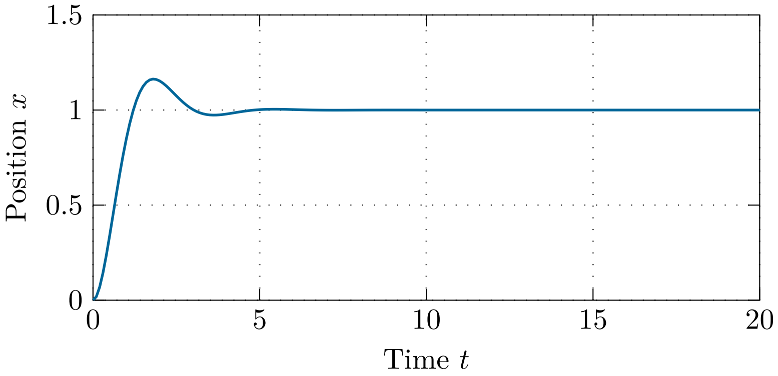


Figure 1.7: Step response of the linear system.

Implementing the PI controller on the original model (1.24) and (1.25), we obtain the position and velocity profiles shown in Figure 1.8.

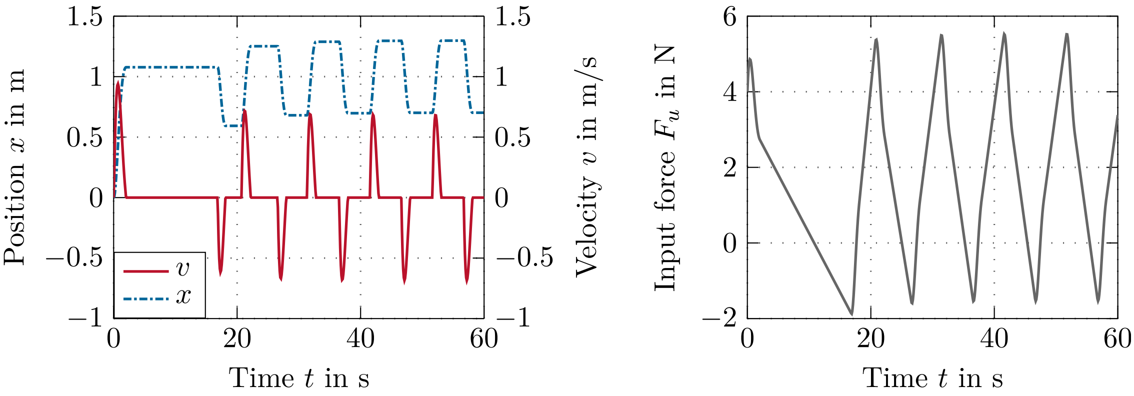


Figure 1.8: Position control of a spring-mass system with static friction using a PI controller.

> *Exercise* 1.8. Try to replicate the results of Figure 1.8 in Matlab/Simulink. Consider measures to prevent limit cycles (Dead Zone, Integrator with switchable I component, Dithering, etc.).

> *Exercise* 1.9. Determine the Stribeck velocity v_S for the friction model approach (1.23) with the parameters $r_C = 1$, $r_v = 3$, $r_H = 4$, and $v_0 = 0.01$.

In addition to static friction models, various dynamic models can be found in the literature. Many of these models are essentially based on a brush-like contact model of two rough surfaces. In the so-called *LuGre model*, the friction force is calculated in the form

$$F_R = \sigma_0 z + \sigma_1 \frac{\mathrm{d}}{\mathrm{d}t} z + \sigma_2 \Delta v \ , \tag{1.27}$$

with the relative velocity Δv of the two contact surfaces. The average deflection of the brushes z satisfies the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} z = \Delta v - \frac{|\Delta v|}{\chi} \sigma_0 z \tag{1.28}$$

with

$$\chi = r_C + (r_H - r_C) \exp\left(- \left(\frac{\Delta v}{v_0} \right)^2 \right) \ . \tag{1.29}$$

Analogous to the static friction model (see (1.23)), r_C denotes the coefficient of Coulomb friction, r_H denotes the static friction, and v_0 denotes a reference velocity. The coefficients σ_0, σ_1, and σ_2 allow parameterization of the friction force model using measurement data. For a constant relative velocity Δv, the static friction force ($\frac{\mathrm{d}}{\mathrm{d}t} z = 0$) is calculated as

$$F_R = \sigma_2 \Delta v + r_C \operatorname{sgn}(\Delta v) + (r_H - r_C) \exp\left(- \left(\frac{\Delta v}{v_0} \right)^2 \right) \operatorname{sgn}(\Delta v) \ . \tag{1.30}$$

It can be seen that (1.30) corresponds to the relationship in (1.23). Therefore, the parameter σ_2 in (1.27) corresponds to the parameter r_v of the viscous friction component in (1.23). The advantage of the dynamic friction model is that no structural switching is required for simulation. However, in general, the entire differential equation system becomes *very stiff*, requiring the use of special integration algorithms.

1.5 Linear and Nonlinear Oscillator

The simplest linear oscillator with an angular frequency of ω_0 is described by a differential equation system of the form

$$\dot{x}_1 = -\omega_0 x_2 \tag{1.31a}$$
$$\dot{x}_2 = \omega_0 x_1 \tag{1.31b}$$

with the output variable x_1. A fundamental disadvantage of this oscillator is that disturbances can change the amplitude (see Figure 1.9 left). It is obvious to extend the linear oscillator in a way that the amplitude is "‘stabilized"’. One possibility is shown by the following system

$$\dot{x}_1 = -\omega_0 x_2 - x_1\Big(x_1^2 + x_2^2 - 1\Big) \tag{1.32a}$$
$$\dot{x}_2 = \omega_0 x_1 - x_2\Big(x_1^2 + x_2^2 - 1\Big)\ . \tag{1.32b}$$

The influence of the nonlinear terms can be seen in Figure 1.9 (right).

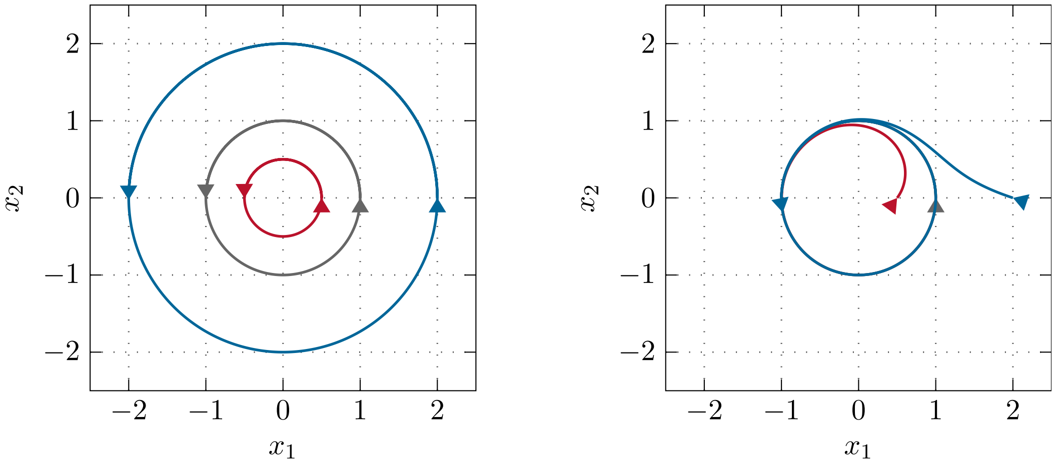


Figure 1.9: Nonlinear and linear oscillator.

> *Exercise* 1.10. Calculate the general solution for the nonlinear oscillator (1.32). Use the transformed variables
>
> $$x_1(t) = r(t)\cos(\varphi(t)) \tag{1.33a}$$
> $$x_2(t) = r(t)\sin(\varphi(t)) \ . \tag{1.33b}$$

1.6 Vehicle Maneuvers

Figure 1.10 shows a drastically simplified model of a vehicle maneuver. The control variables considered are the rolling speed u_1 and the rotational speed u_2 of the axle.

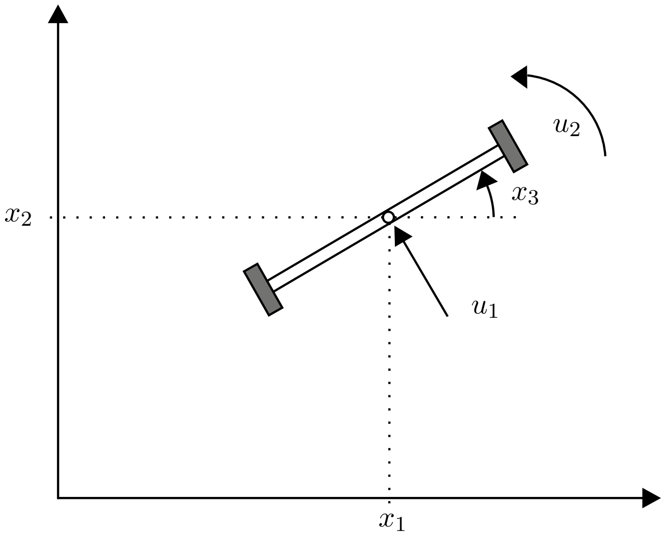


Figure 1.10: Simple vehicle model.

The corresponding mathematical model is given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}-\sin(x_3)\\ \cos(x_3)\\ 0\end{bmatrix} u_1 + \begin{bmatrix}0\\0\\1\end{bmatrix} u_2 \ . \tag{1.34}$$

Linearizing the model around an equilibrium point

$$\mathbf{x}_R = \begin{bmatrix}x_{1,R}\\ x_{2,R}\\ x_{3,R}\end{bmatrix}, \qquad \mathbf{u}_R = \begin{bmatrix}0\\0\end{bmatrix} , \tag{1.35}$$

results in

$$\Delta\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Delta\mathbf{x} + \begin{bmatrix} -\sin(x_{3,R}) \\ \cos(x_{3,R}) \\ 0 \end{bmatrix} \Delta u_1 + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Delta u_2 \ . \tag{1.36}$$

It can be easily verified that the controllability matrix

$$\mathcal{R}(\mathbf{A}, \mathbf{B}) = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} \end{bmatrix} \tag{1.37}$$

has rank two. Therefore, every linearized model of the vehicle maneuver around an equilibrium point is uncontrollable. However, from experience, it is known that this may not hold for the original system (or what is your experience with parking?).

1.7 Direct Current (DC) Machines

Figure 1.11 shows the equivalent circuit diagram of a separately excited DC machine. The

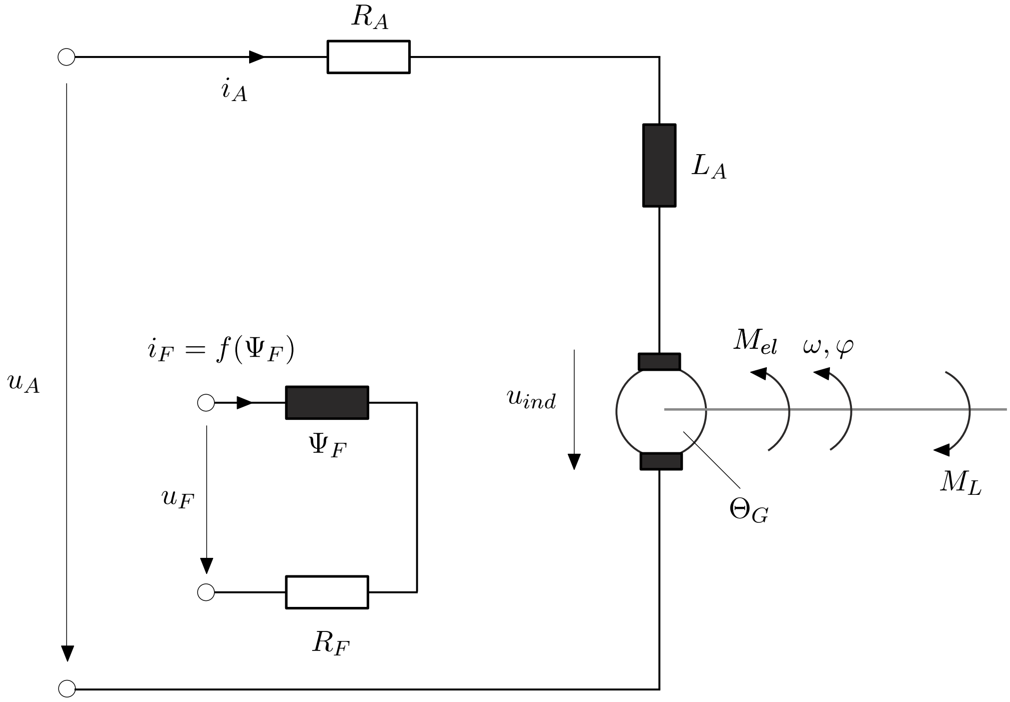


Figure 1.11: Equivalent circuit diagram of a separately excited DC machine.

corresponding mathematical model can be formulated in the form

$$
\begin{align}
L_A \frac{\mathrm{d}}{\mathrm{d}t} i_A &= u_A - R_A i_A - \underbrace{k\psi_F\omega}_{u_{ind}} \tag{1.38a}\\
\frac{\mathrm{d}}{\mathrm{d}t}\psi_F &= u_F - R_F i_F \tag{1.38b}\\
\Theta_G \frac{\mathrm{d}}{\mathrm{d}t}\omega &= \underbrace{k\psi_F i_A}_{M_{el}} - M_L \tag{1.38c}
\end{align}
$$

where L_A is the armature inductance, R_A is the armature resistance, $i_F = f(\psi_F)$ is the field current, R_F is the field circuit resistance, Θ_G is the moment of inertia of the DC machine and all rigidly flanged components, and k is the armature circuit constant. The state variables in this case are the armature current i_A, the linked field flux ψ_F, and the angular velocity ω, while the control variables are the armature voltage u_A, the field voltage u_F, and the load torque M_L acts as a disturbance on the system. This description of the separately excited DC machine already assumes that the following model assumptions have been taken into account:

- The spatially distributed windings can be modeled as concentrated inductances in their respective winding axes,
- the inductances in the armature and field circuits twisted by 90° against each other already indicate a complete decoupling between the armature and field,
- the resistances in the armature and field circuits are constant,
- no iron losses are considered,
- there are no saturation effects in the armature circuit, and
- commutation is assumed to be ideal (no torque ripple).

To classify the steady-state behavior of the DC machine independently of the specific machine parameters, a normalization of (1.38) to dimensionless quantities is carried out. Using the reference values of the nominal angular velocity ω_0, the nominal linked field flux $\psi_{F,0}$, and

$$
\begin{align}
u_{A,0} &= u_{ind,0} = k\psi_{F,0}\omega_0 \ , \tag{1.39a}\\
i_{A,0} &= \frac{u_{A,0}}{R_A} \ , \tag{1.39b}\\
M_{el,0} &= k\psi_{F,0} i_{A,0} \ , \tag{1.39c}\\
u_{F,0} &= R_F i_{F,0} \tag{1.39d}
\end{align}
$$

(1.38) is then transformed into dimensionless form as

$$\frac{L_A}{R_A}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{i_A}{i_{A,0}}\right) = \frac{u_A}{u_{A,0}} - \frac{i_A}{i_{A,0}} - \frac{\psi_F}{\psi_{F,0}}\frac{\omega}{\omega_0} \tag{1.40a}$$

$$\frac{\psi_{F,0}}{u_{F,0}}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\psi_F}{\psi_{F,0}}\right) = \frac{u_F}{u_{F,0}} - \tilde{f}\left(\frac{\psi_F}{\psi_{F,0}}\right) \tag{1.40b}$$

$$\frac{\Theta_G\omega_0}{M_{el,0}}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\omega}{\omega_0}\right) = \frac{\psi_F}{\psi_{F,0}}\frac{i_A}{i_{A,0}} - \frac{M_L}{M_{el,0}} \ , \tag{1.40c}$$

where $\frac{i_F}{i_{F,0}} = \frac{f(\psi_F)}{i_{F,0}} = \tilde{f}\left(\frac{\psi_F}{\psi_{F,0}}\right)$. Due to the larger air gap in the armature transverse direction, $\frac{L_A}{R_A} \ll \frac{\psi_{F,0}}{u_{F,0}}$ and magnetic saturation effects in the armature circuit can generally be neglected. For simplification of notation, all normalized quantities $\frac{x}{x_0}$ are denoted in the form $\frac{x}{x_0} = \tilde{x}$ in the following.

For constant input quantities u_A, u_F, and M_L, the equations for the steady state from (1.40) are given by

$$0 = \tilde{u}_A - \tilde{i}_A - \tilde{\psi}_F\tilde{\omega} \tag{1.41a}$$

$$0 = \tilde{u}_F - \tilde{f}\left(\tilde{\psi}_F\right) \tag{1.41b}$$

$$0 = \tilde{\psi}_F\tilde{i}_A - \tilde{M}_L \ . \tag{1.41c}$$

Considering the normalized flux $\tilde{\psi}_F$ as an independent input quantity - which can always be calculated from $\tilde{u}_F$ via (1.41b) in the steady state - the following relationships can be specified for the steady state of the separately excited DC machine

$$\tilde{i}_A = \frac{1}{\tilde{\psi}_F}\tilde{M}_L \ , \tag{1.42a}$$

$$\tilde{\omega} = \frac{1}{\tilde{\psi}_F}\tilde{u}_A - \frac{1}{\tilde{\psi}_F^2}\tilde{M}_L \tag{1.42b}$$

It should be noted that the flux ψ_F is limited by iron saturation in the stator circuit, which is why $\psi_{F,0}$ can always be set in such a way that approximately in the entire operating range the following holds

$$\tilde{\psi}_F = \frac{\psi_F}{\psi_{F,0}} \leq 1 \ . \tag{1.43}$$

> *Exercise* 1.11. Show that in the case of a constant excitation DC machine $\psi_F = \psi_{F,0}$ the mathematical model (1.38) is linear.

There is a distinction between armature control and field control in separately excited DC machines. In armature control, the excitation flux is set as in the case of a constant excitation DC machine $\psi_F = \psi_{F,0}$, and the control of the angular velocity ω is done through the armature circuit voltage u_A.

> *Exercise* 1.12. Draw the steady-state characteristics of (1.42) for $\tilde{\psi}_F = 1$ with $\tilde{u}_A$ as a parameter ($\tilde{u}_A = -1.0, \ -0.5, 0.5, 1.0$) in the range $-0.5 \leq \tilde{M}_L \leq 0.5$.

In contrast, in field control, the armature voltage is operated at the nominal value $u_A = \pm u_{A,0}$, and the speed control is done through the excitation voltage u_F by weakening the excitation flux in the range $\tilde{\psi}_{F,\min} \leq \tilde{\psi}_F \leq 1$. Setting $\tilde{u}_A = 1$ in (1.42), the steady-state characteristics shown in Figure 1.12 are obtained. The maximum achievable angular velocity $\tilde{\omega}_{\max}$ for a constant load torque $\tilde{M}_L$ is obtained from (1.42) with $\tilde{u}_A = 1$ through the relationship

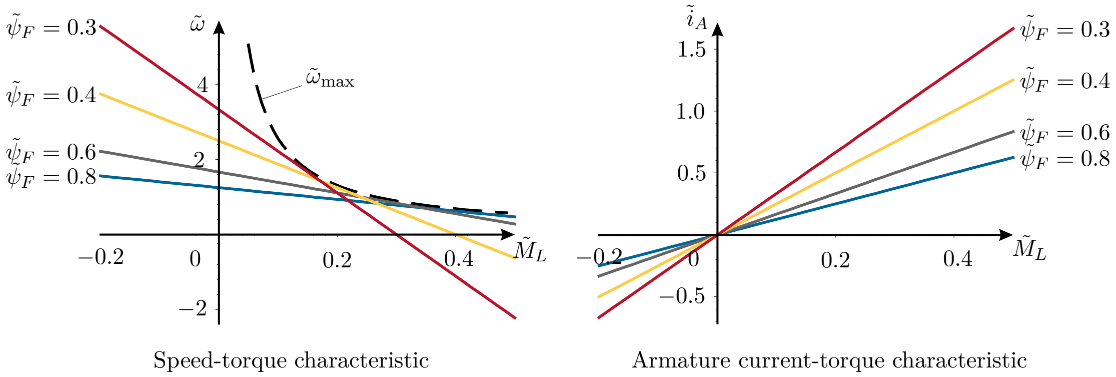
$$\frac{\mathrm{d}\tilde{\omega}}{\mathrm{d}\tilde{\psi}_F} = -\frac{1}{\tilde{\psi}_F^2}\left(1 - \frac{2}{\tilde{\psi}_F}\tilde{M}_L\right) = 0 \tag{1.44}$$

in the form

$$\tilde{\psi}_{F,\min} = 2\tilde{M}_L \ , \tag{1.45a}$$

$$\tilde{\omega}_{\max} = \frac{1}{4\tilde{M}_L} \ . \tag{1.45b}$$

It can be seen from (1.45) that for a given constant load torque $\tilde{M}_L$, the lower limit of the flux is given by $\tilde{\psi}_{F,\min} = 2\tilde{M}_L$.



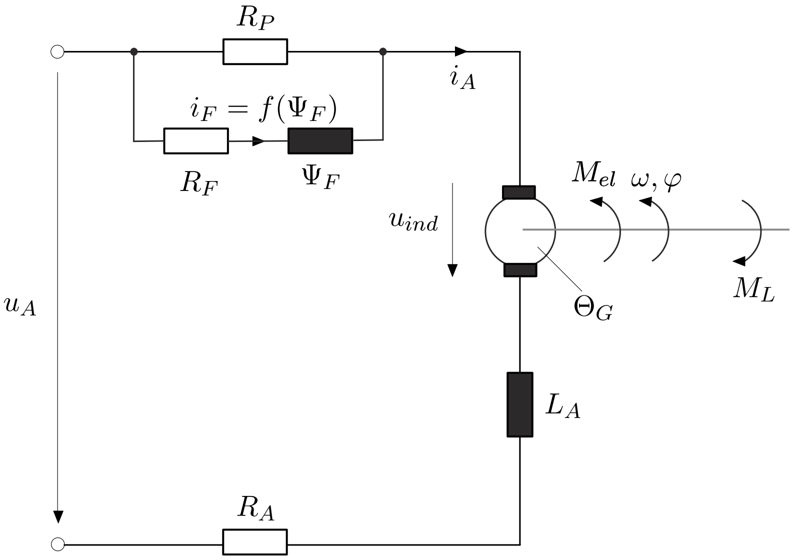
Speed-torque characteristic

Armature current-torque characteristic

Figure 1.12: Characteristic curves for DC machines.

The left image of Figure 1.12 shows, among other things, that reducing the flux $\tilde{\psi}_F$ depending on the load torque $\tilde{M}_L$ does not necessarily lead to an increase in the angular velocity $\tilde{\omega}$. Therefore, in practice, a combination of armature and field control is usually chosen - namely, in a way that the angular velocity is controlled by the armature voltage u_A up to the nominal value of angular velocity ω_0 and the excitation flux ψ_F is maintained at its nominal value $\psi_{F,0}$, and only when the armature voltage $u_{A,0}$ is reached, further increase in angular velocity is achieved through field weakening.

> *Exercise* 1.13. Figure 1.13 shows the equivalent circuit diagram of a series-wound machine, which is very commonly used in traction drives.Any external resistances in the armature circuit are added to the armature resistance R_A, and the adjustable

> resistance R_P is used for field weakening. Provide a mathematical model of the series-wound machine and consider how the resistance R_P affects the steady-state behavior.
>
> 
>
>
> Figure 1.13: Equivalent circuit diagram of a series-wound machine.

1.8 Hydraulic Actuator (Double Rod Cylinder)

Figure 1.14 shows a double rod cylinder controlled by a 3/4-way valve with zero overlap. It should be noted that this configuration also includes the very common case of a double-acting cylinder with a single piston rod (differential cylinder). Here, x_k denotes the piston position, $V_{0,1}$ and $V_{0,2}$ are the volumes of the two cylinder chambers for $x_k = 0$, A_1 and A_2 describe the effective piston areas, m_k is the sum of all moving masses, q_1 and q_2 denote the flow from the control valve to the cylinder and from the cylinder to the control valve, respectively, q_{int} is the internal leakage oil flow, and $q_{ext,1}$ and $q_{ext,2}$ describe the external leakage oil flows. In general, the density of oil ρ_{oil} is a function of pressure p and temperature T. The temperature influence will be neglected further, and the isothermal bulk modulus β_T will be used as a constitutive equation with

$$\frac{1}{\beta_T} = \frac{1}{\rho_{oil}} \left(\frac{\partial \rho_{oil}}{\partial p} \right)_{T\,=\,\mathrm{const.}} \tag{1.46}$$

The continuity equations for the two cylinder chambers are

$$\frac{\mathrm{d}}{\mathrm{d}t}(\rho_{oil}(p_1)(V_{0,1} + A_1 x_k)) = \rho_{oil}(p_1)(q_1 - q_{int} - q_{ext,1}) \tag{1.47a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\rho_{oil}(p_2)(V_{0,2} - A_2 x_k)) = \rho_{oil}(p_2)(q_{int} - q_{ext,2} - q_2) \tag{1.47b}$$

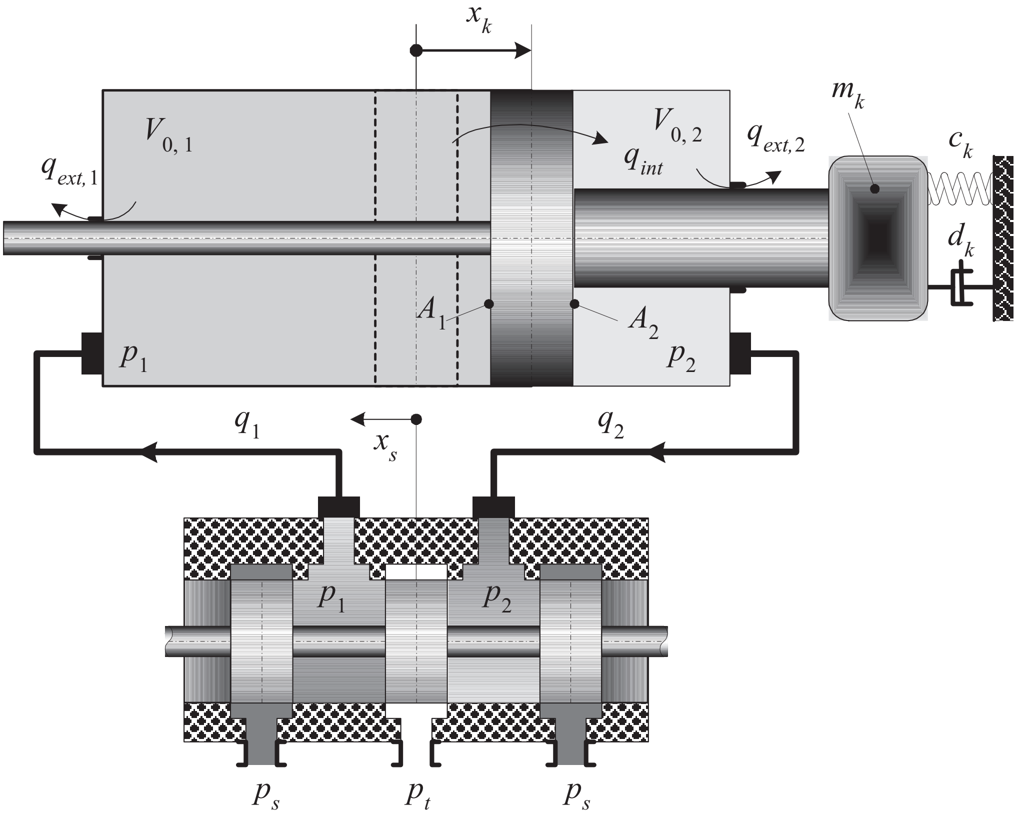


Figure 1.14: Double rod cylinder with 3/4-way valve.

with the cylinder pressures p_1 and p_2. Since the internal and external leakage oil flows q_{int}, $q_{ext,1}$, and $q_{ext,2}$ are generally laminar, there is a linear relationship between leakage oil flow and pressure drop. Using relation (1.46), equation (1.47) simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}t}p_1 = \frac{\beta_T}{(V_{0,1}+A_1x_k)}\left(q_1 - A_1\frac{\mathrm{d}}{\mathrm{d}t}x_k - C_{int}(p_1-p_2) - C_{ext,1}p_1\right) \tag{1.48a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_2 = \frac{\beta_T}{(V_{0,2}-A_2x_k)}\left(-q_2 + A_2\frac{\mathrm{d}}{\mathrm{d}t}x_k + C_{int}(p_1-p_2) - C_{ext,2}p_2\right) \tag{1.48b}$$

with the laminar leakage coefficients C_{int}, $C_{ext,1}$, and $C_{ext,2}$. For a 3/4-way valve with zero overlap, the flows q_1 and q_2 are calculated as

$$q_1 = K_{v,1}\sqrt{p_S-p_1}\,\mathrm{sg}(x_s) - K_{v,2}\sqrt{p_1-p_T}\,\mathrm{sg}(-x_s) \tag{1.49a}$$

$$q_2 = K_{v,2}\sqrt{p_2-p_T}\,\mathrm{sg}(x_s) - K_{v,1}\sqrt{p_S-p_2}\,\mathrm{sg}(-x_s) \tag{1.49b}$$

with the tank pressure p_T, the supply pressure p_S, the control spool position x_s, the function $\mathrm{sg}(x_s) = x_s$ for $x_s \geq 0$ and $\mathrm{sg}(x_s) = 0$ for $x_s < 0$, and the valve coefficients $K_{v,i} = C_d A_{v,i}\sqrt{2/\rho_{oil}}$, $i = 1, 2$. Here, the term $A_{v,i}x_s$ denotes the orifice area and C_d denotes the flow coefficient ($C_d \approx 0.6 - 0.8$, depending on the geometry of the control edge, Reynolds number, flow direction, etc).

Neglecting the dynamics of the control valve and considering the control valve position x_s as an input to the system, a mathematical model for Figure 1.14 is obtained in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}p_1 = \frac{\beta_T}{(V_{0,1} + A_1 x_k)}(q_1 - A_1 v_k - C_{int}(p_1 - p_2) - C_{ext,1}p_1) \tag{1.50a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_2 = \frac{\beta_T}{(V_{0,2} - A_2 x_k)}(-q_2 + A_2 v_k + C_{int}(p_1 - p_2) - C_{ext,2}p_2) \tag{1.50b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}x_k = v_k \tag{1.50c}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}v_k = \frac{1}{m_k}(A_1 p_1 - A_2 p_2 - d_k v_k - c_k x_k) \tag{1.50d}$$

with q_1 and q_2 from (1.49).

1.9 Literatur

[1.1] C. Canudas de Wit, H. Olsson, K. J. Åström, and P. Lischinsky, "A New Model for Control of Systems with Friction," *IEEE Transactions on Automatic Control*, vol. 40, no. 3, pp. 419–425, Mar. 1995.

[1.2] W. Leonhard, *Control of Electrical Drives*. Springer, Berlin: Dover Publications, 1990.

[1.3] H. E. Merritt, *Hydraulic Control Systems*. New York, USA: John Wiley & Sons, 1967.

[1.4] H. Murrenhoff, *Grundlagen der Fluidtechnik*. Aachen, Germany: Shaker, 2001.

[1.5] G. Pfaff, *Regelung elektrischer Antriebe I*. München: Oldenbourg, 1990.

[1.6] M. W. Spong, *Robot Dynamics and Control*. New York: John Wiley & Sons, 1989.

2 Dynamical Systems

A dynamical system (without input) allows the description of the change of certain points (elements of a suitable set $\mathcal{X}$) in time t. In control engineering, these points are given by the state $\mathbf{x}(t)$ of the system. If we choose the set of states as $\mathcal{X} = \mathbb{R}^n$, then an autonomous dynamical system is a mapping

$$\boldsymbol{\Phi}_t(\mathbf{x}) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \tag{2.1}$$

with

$$\mathbf{x}(t) = \boldsymbol{\Phi}_t(\mathbf{x}_0) \ . \tag{2.2}$$

From the relationship

$$\mathbf{x}_0 = \boldsymbol{\Phi}_0(\mathbf{x}_0) \tag{2.3}$$

it follows that $\boldsymbol{\Phi}_0$ must be the identity mapping $\mathbf{I}$ with $\mathbf{x} = \mathbf{I}(\mathbf{x})$. From the relationships

$$\begin{aligned} \mathbf{x}(t) &= \boldsymbol{\Phi}_t(\mathbf{x}_0) && \text{(2.4a)} \\ \mathbf{x}(s+t) &= \boldsymbol{\Phi}_s(\mathbf{x}(t)) && \text{(2.4b)} \\ \mathbf{x}(s+t) &= \boldsymbol{\Phi}_{s+t}(\mathbf{x}_0) && \text{(2.4c)} \end{aligned}$$

we now have

$$\mathbf{x}(s+t) = \boldsymbol{\Phi}_s(\boldsymbol{\Phi}_t(\mathbf{x}_0)) = \boldsymbol{\Phi}_{s+t}(\mathbf{x}_0) \tag{2.5}$$

or

$$\boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_{s+t} \ , \tag{2.6}$$

where $\circ$ denotes the composition of the mappings $\boldsymbol{\Phi}_s$ and $\boldsymbol{\Phi}_t$. By exchanging the order in the above considerations, we obtain

$$\boldsymbol{\Phi}_{s+t} = \boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_t \circ \boldsymbol{\Phi}_s \ , \tag{2.7}$$

justifying the notation $\boldsymbol{\Phi}_{s+t}$.

> *Exercise* 2.1. Let $\mathbf{a}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{b}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ be two linear mappings from $\mathbb{R}^n$ to itself. Is the composition $(\mathbf{a} \circ \mathbf{b})(\mathbf{x}) = \mathbf{a}(\mathbf{b}(\mathbf{x}))$ again a linear mapping? Does $\mathbf{a} \circ \mathbf{b} = \mathbf{b} \circ \mathbf{a}$ hold?

> In other words, are linear mappings commutative with respect to composition? The linear mappings $\mathbf{a}$ and $\mathbf{b}$ are given by the matrices $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{y} = \mathbf{A}\mathbf{x}$ and $\mathbf{y} = \mathbf{B}\mathbf{x}$. What are the matrix representations of the above compositions?

Furthermore, it is assumed that $\boldsymbol{\Phi}_t(\mathbf{x})$ is a (continuously) differentiable mapping with respect to $\mathbf{x}$.

Definition 2.1 (Dynamical System)**.** A *(autonomous) dynamical system* is a C^1 (continuously differentiable) mapping

$$\boldsymbol{\Phi}_t(\mathbf{x}) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \ , \tag{2.8}$$

that satisfies the following conditions:

(1) $\boldsymbol{\Phi}_0$ is the identity mapping $\mathbf{I}$, and

(2) the composition $\boldsymbol{\Phi}_s(\boldsymbol{\Phi}_t(\mathbf{x}))$ satisfies the relations

$$\boldsymbol{\Phi}_{s+t} = \boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_t \circ \boldsymbol{\Phi}_s \tag{2.9}$$

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

Note that from the above definition, it immediately follows

$$\boldsymbol{\Phi}_{-s}(\boldsymbol{\Phi}_s(\mathbf{x}_0)) = \boldsymbol{\Phi}_0(\mathbf{x}_0) = \left(\boldsymbol{\Phi}_s^{-1} \circ \boldsymbol{\Phi}_s\right)(\mathbf{x}_0) = \mathbf{x}_0 \tag{2.10}$$

The mapping $\boldsymbol{\Phi}_t$ thus satisfies the following conditions:

(1) $\boldsymbol{\Phi}_0 = \mathbf{I}$,

(2) $\boldsymbol{\Phi}_{s+t} = \boldsymbol{\Phi}_s \circ \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_t \circ \boldsymbol{\Phi}_s$, and

(3) $\boldsymbol{\Phi}_s^{-1} = \boldsymbol{\Phi}_{-s}$.

A dynamical system according to Definition 2.1 is closely related to a system of differential equations. From

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t}(\boldsymbol{\Phi}_{t+\Delta t}(\mathbf{x}_0) - \boldsymbol{\Phi}_t(\mathbf{x}_0)) \\
&= \left(\lim_{\Delta t \to 0} \frac{1}{\Delta t}(\boldsymbol{\Phi}_{\Delta t} - \mathbf{I})\right) \circ \boldsymbol{\Phi}_t(\mathbf{x}_0) \\
&= \frac{\partial}{\partial t}\boldsymbol{\Phi}_t\bigg|_{t=0} \circ \boldsymbol{\Phi}_t(\mathbf{x}_0) \\
&= \frac{\partial}{\partial t}\boldsymbol{\Phi}_t\bigg|_{t=0}(\mathbf{x}(t))
\end{aligned} \tag{2.11}$$

it follows

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)), \qquad \mathbf{f}(\mathbf{x}(t)) = \frac{\partial}{\partial t}\boldsymbol{\Phi}_t\bigg|_{t=0}(\mathbf{x}(t)) \ . \tag{2.12}$$

Thus, a dynamical system also satisfies the relationship

(4) $\frac{\partial}{\partial t}\mathbf{\Phi}_t\Big|_{t=0}(\mathbf{x}(t)) = \mathbf{f}(\mathbf{x}(t))$ with $\mathbf{x}(t) = \mathbf{\Phi}_t(\mathbf{x}_0)$. The mapping $\mathbf{\Phi}_t$ is also called the *flow* of the differential equation system (2.12).

> *Exercise* 2.2. Choose the specific dynamical system $\mathbf{x}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{x}_0$ or $\mathbf{\Phi}_t(\mathbf{x}) = \mathrm{e}^{\mathbf{A}t}\mathbf{x}$. Now interpret the properties of the transition matrix according to points (1) - (3) of a dynamical system. What does the corresponding differential equation system look like?

As an example, the motion of a point $\mathbf{x}_0 \in \mathbb{R}^3$ on a unit sphere with the origin as the center is considered (see Figure 2.1). As an approach for a (continuous) transformation that maps points on the unit sphere back to themselves, the form

$$\mathbf{x}(t) = \mathbf{D}(t, \mathbf{x}_0)\mathbf{x}_0 = \mathbf{\Phi}_t(\mathbf{x}_0) \tag{2.13}$$

is chosen with a (3×3) matrix $\mathbf{D}$. Due to $\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_0 = \mathbf{x}^{\mathrm{T}}(t)\mathbf{x}(t) = 1$, the conditions

$$\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{I} \tag{2.14}$$

must be satisfied.

> *Exercise* 2.3. Show the validity of (2.14).

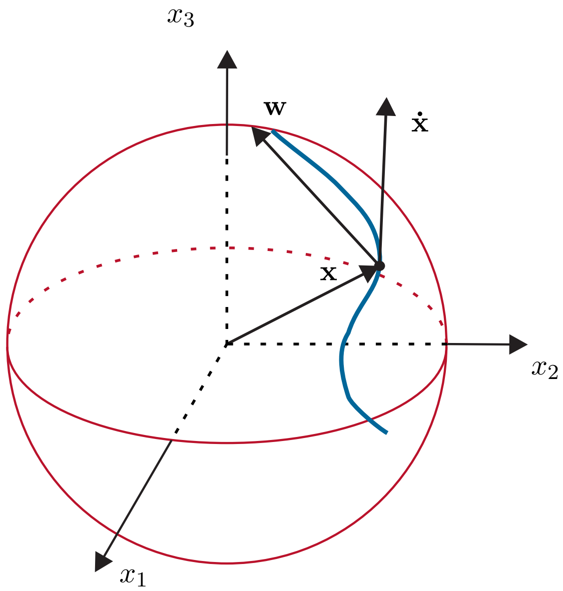


Figure 2.1: Motion on a sphere.

For the mapping in Figure 2.1 to describe a dynamical system, the conditions

(1) $\mathbf{D}(0, \mathbf{x}) = \mathbf{I}$ and

(2) $\mathbf{D}(s + t, \mathbf{x}) = \mathbf{D}(s, \mathbf{D}(t, \mathbf{x})\mathbf{x})\mathbf{D}(t, \mathbf{x}) = \mathbf{D}(t, \mathbf{D}(s, \mathbf{x})\mathbf{x})\mathbf{D}(s, \mathbf{x})$

must hold. Furthermore, it is known that a dynamical system is associated with a system of differential equations of the form

$$\dot{\mathbf{x}} = \frac{\partial}{\partial t}(\mathbf{D}(t,\mathbf{x})\mathbf{x})\bigg|_{t=0} = \frac{\partial}{\partial t}\mathbf{D}(t,\mathbf{x})\bigg|_{t=0}\mathbf{x} \tag{2.15}$$

Additionally, the relationship

$$\begin{aligned}\mathbf{W} &= \left(\frac{\partial}{\partial t}\mathbf{D}(t,\mathbf{x}_0)\right)\mathbf{D}^{\mathrm{T}}(t,\mathbf{x}_0)\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{D}(t+\Delta t,\mathbf{x}_0)-\mathbf{D}(t,\mathbf{x}_0))\mathbf{D}^{\mathrm{T}}(t,\mathbf{x}_0)\\ &\quad\text{using condition (2):}\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{D}(\Delta t,\mathbf{D}(t,\mathbf{x}_0)\mathbf{x}_0)\mathbf{D}(t,\mathbf{x}_0)-\mathbf{D}(t,\mathbf{x}_0))\mathbf{D}^{\mathrm{T}}(t,\mathbf{x}_0)\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{D}(\Delta t,\mathbf{D}(t,\mathbf{x}_0)\mathbf{x}_0)-\mathbf{I})\mathbf{D}(t,\mathbf{x}_0)\mathbf{D}^{\mathrm{T}}(t,\mathbf{x}_0)\\ &= \frac{\partial}{\partial t}\mathbf{D}(t,\mathbf{x})\bigg|_{t=0}.\end{aligned} \tag{2.16}$$

holds. By using (2.14), it is immediately clear that $\mathbf{W}$ is skew-symmetric, because

$$\frac{\partial}{\partial t}\left(\mathbf{D}\mathbf{D}^{\mathrm{T}}\right) = \left(\frac{\partial}{\partial t}\mathbf{D}\right)\mathbf{D}^{\mathrm{T}} + \mathbf{D}\left(\frac{\partial}{\partial t}\mathbf{D}^{\mathrm{T}}\right) = \mathbf{0} \tag{2.17}$$

or

$$\left(\frac{\partial}{\partial t}\mathbf{D}\right)\mathbf{D}^{\mathrm{T}} = -\mathbf{D}\left(\frac{\partial}{\partial t}\mathbf{D}^{\mathrm{T}}\right). \tag{2.18}$$

A skew-symmetric matrix $\mathbf{W}$ generally has the form

$$\mathbf{W}(\mathbf{x}) = \begin{bmatrix} 0 & -\omega_3(\mathbf{x}) & \omega_2(\mathbf{x})\\ \omega_3(\mathbf{x}) & 0 & -\omega_1(\mathbf{x})\\ -\omega_2(\mathbf{x}) & \omega_1(\mathbf{x}) & 0\end{bmatrix} \tag{2.19}$$

and thus the differential equation (2.15) can be written as follows

$$\dot{\mathbf{x}} = \mathbf{W}\mathbf{x} = \mathbf{w}(\mathbf{x})\times\mathbf{x} \tag{2.20}$$

with $\mathbf{w}^{\mathrm{T}}(\mathbf{x}) = [\omega_1(\mathbf{x}),\, \omega_2(\mathbf{x}),\, \omega_3(\mathbf{x})]$. This means that when a dynamical system describes the motion of a point on a sphere, the differential notation yields the cross product.

2.1 Differential Equations

By a dynamical system according to Definition 2.1, a system of differential equations is defined. The investigation of when a differential equation of the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{2.21}$$

describes a dynamical system in the above sense will be examined subsequently. However, in a first step, some basic concepts will be explained.

Definition 2.2 (Linear Vector Space)**.** A non-empty set $\mathcal{X}$ is called a linear vector space over a (scalar) field K with the binary operations $+ : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ (addition) and $\cdot : K \times \mathcal{X} \to \mathcal{X}$ (scalar multiplication), if the following vector space axioms are satisfied:

(1) The set $\mathcal{X}$ with the operation $+$ forms a commutative group, i.e., for $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{X}$, the following holds:

$$(1) \quad \mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \qquad \text{Commutativity} \tag{2.22}$$
$$(2) \quad \mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z} \qquad \text{Associativity} \tag{2.23}$$
$$(3) \quad \mathbf{0} + \mathbf{x} = \mathbf{x} \qquad \text{Identity element} \tag{2.24}$$
$$(4) \quad \mathbf{x} + (-\mathbf{x}) = \mathbf{0} \qquad \text{Inverse element} \tag{2.25}$$

(2) The multiplication $\cdot$ by a scalar a, $b \in K$ satisfies the laws:

$$(1) \quad a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y} \qquad \text{Distributivity} \tag{2.26}$$
$$(2) \quad (a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x} \qquad \text{Distrivutivity} \tag{2.27}$$
$$(3) \quad (ab)\mathbf{x} = a(b\mathbf{x}) \qquad \text{Compativility} \tag{2.28}$$
$$(4) \quad 1\mathbf{x} = \mathbf{x}, \quad 0\mathbf{x} = \mathbf{0} \tag{2.29}$$

Definition 2.3 (Linear Subspace)**.** If $\mathcal{X}$ is a linear vector space over the field K, then a subset $\mathcal{S}$ of $\mathcal{X}$ is a linear subspace if $\mathbf{x}$, $\mathbf{y} \in \mathcal{S} \Rightarrow a\mathbf{x} + b\mathbf{y} \in \mathcal{S}$ for all scalars a, $b \in K$.

An expression of the form

$$\sum_{j=1}^{n} a_j \mathbf{x}_j = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_n\mathbf{x}_n \tag{2.30}$$

with $\mathcal{X} \ni \mathbf{x}_j$, $j = 1, \ldots, n$ and scalars $K \ni a_j$, $j = 1, \ldots, n$ is called a *linear combination* of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$. If there exist scalars a_j, $j = 1, \ldots, n$, not all identically zero, such that the linear combination $\sum_{j=1}^{n} a_j \mathbf{x}_j = \mathbf{0}$ holds, then the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ are *linearly dependent*. If apart from the trivial solution $a_j = 0$, $j = 1, \ldots, n$, no scalars exist that satisfy this condition, then the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ are called *linearly independent*. For the set of all linear combinations of vectors in a non-empty subset $\mathcal{M}$ of $\mathcal{X}$, we denote $\operatorname{span}(\mathcal{M})$. The *subspace spanned by* $\mathcal{M}$ (also known as linear hull) is the smallest subspace according to Definition 2.3 that contains $\mathcal{M}$, i.e., all its elements can be represented as linear combinations of elements from $\mathcal{M}$.

If a linear vector space $\mathcal{X}$ is spanned by a finite number n of linearly independent vectors, then $\mathcal{X}$ has dimension n and is called *finite-dimensional*. If no finite number exists, $\mathcal{X}$ is *infinite-dimensional*.

2.1.1 The Concept of Norms

Examples of linear vector spaces include vectors in $\mathbb{R}^n$, $n \times m$-dimensional real-valued matrices, or complex numbers, each with the scalar field $\mathbb{R}$.

Definition 2.4 (Normed Linear Vector Space)**.** A normed linear vector space is a vector space $\mathcal{X}$ over a scalar field K with a real-valued function $\|\mathbf{x}\| : \mathcal{X} \rightarrow \mathbb{R}_+$ that assigns to each $\mathbf{x} \in \mathcal{X}$ a real number $\|\mathbf{x}\|$, called the norm of $\mathbf{x}$, and satisfies the following norm axioms:

$$(1) \|\mathbf{x}\| \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{X} \qquad \text{Non-negativity} \tag{2.31}$$

$$(2) \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0} \tag{2.32}$$

$$(3) \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad \text{Triangle Inequality} \tag{2.33}$$

$$(4) \|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\| \quad \text{for all } \mathbf{x} \in \mathcal{X} \text{ and all } \alpha \in K \tag{2.34}$$

Exercise 2.4. Show that from the norm axioms it follows that $\|\mathbf{x} - \mathbf{y}\| \geq \|\mathbf{x}\| - \|\mathbf{y}\|$.

Next, we consider some classical normed vector spaces, distinguishing between finite and infinite-dimensional vector spaces. The p-norm, $1 \leq p < \infty$, of a finite-dimensinal vector $\mathbf{x}^T = [x_1, \dots, x_n]$ is defined as

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{2.35}$$

and for $p = \infty$ we have

$$\|\mathbf{x}\|_\infty = \max_i |x_i| \ . \tag{2.36}$$

In addition to the ∞-*norm* ("infinity norm") according to (2.36), the most commonly used norms on $\mathbb{R}^n$ are the 1-*norm* ("one norm")

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{2.37}$$

and the 2-*norm* ("square norm" or "Euclidean norm")

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \ . \tag{2.38}$$

The following inequalities hold:

Theorem 2.1 (Hölder's Inequality)**.** *If the relationship*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{2.39}$$

holds for positive numbers $1 \leq p \leq \infty$ *and* $1 \leq q \leq \infty$*, then for* $\mathbf{x}^T = [x_1, \dots, x_n]$ *and* $\mathbf{y}^T = [y_1, \dots, y_n]$*, the inequality*

$$\sum_{i=1}^{n} |x_i y_i| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q \ . \tag{2.40}$$

follows.

Theorem 2.2 (Minkowski's Inequality). *For* $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$, $1 \leq p \leq \infty$, *we have*

$$\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \ . \tag{2.41}$$

The equality in (2.41) holds if and only if $a\mathbf{x} = b\mathbf{y}$ *for positive constants* a *and* b.

Note that Minkowski's inequality corresponds to the triangle inequality (3) for norms in Definition 2.4.

In a finite-dimensional normed vector space, all norms are *equivalent*. This means that if $\| \ \|_\alpha$ and $\| \ \|_\beta$ denote two different norms, there always exist two constants $0 < c_1, c_2 < \infty$ such that

$$c_1\| \ \|_\alpha \leq \| \ \|_\beta \leq c_2\| \ \|_\alpha \tag{2.42}$$

holds.

Exercise 2.5. Prove the statement that in a finite-dimensional vector space, all p-norms are *equivalent*.

Exercise 2.6. Show that the equivalence of norms ($\| \ \|_\alpha \sim \| \ \|_\beta$) is an *equivalence relation*.

Tip: You need to prove the properties of *reflexivity* ($\| \ \|_\alpha \sim \| \ \|_\alpha$), *symmetry* ($\| \ \|_\alpha \sim \| \ \|_\beta \Rightarrow \| \ \|_\beta \sim \| \ \|_\alpha$), and *transitivity* ($\| \ \|_\alpha \sim \| \ \|_\beta$ and $\| \ \|_\beta \sim \| \ \|_\gamma \Rightarrow \| \ \|_\alpha \sim \| \ \|_\gamma$).

Exercise 2.7. Draw in the (x_1, x_2)-plane the sets $\mathcal{M}_1 = \{\mathbf{x} \in \mathbb{R}^2 | \|\mathbf{x}\|_1 \leq 1\}$, $\mathcal{M}_2 = \{\mathbf{x} \in \mathbb{R}^2 | \|\mathbf{x}\|_2 \leq 1\}$, and $\mathcal{M}_\infty = \{\mathbf{x} \in \mathbb{R}^2 | \|\mathbf{x}\|_\infty \leq 1\}$. Verify the inequality

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{2}\|\mathbf{x}\|_2 \tag{2.43}$$

using the image and find suitable positive constants c_1 and c_2 for the inequality

$$c_1\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_\infty \leq c_2\|\mathbf{x}\|_2 \ . \tag{2.44}$$

The equivalence of norms does not hold for infinite-dimensional normed vector spaces. In the *infinite-dimensional* vector space $L_p[t_0, t_1]$, $1 \leq p < \infty$, all real-valued functions $x(t)$ in the interval $[t_0, t_1]$ are considered, satisfying

$$\|x\|_p = \left(\int_{t_0}^{t_1} |x(t)|^p \, \mathrm{d}t\right)^{1/p} < \infty \ . \tag{2.45}$$

It is important to note that in the vector space $L_p[t_0, t_1]$, functions that are *almost everywhere* equal, meaning they differ only on a countable set of points, are considered identical. This is the reason why the norm $\|x\|_p$ in (2.45) satisfies condition (2) of

Definition 2.4. The vector space $L_\infty[t_0, t_1]$ describes all real-valued functions $x(t)$ that are essentially bounded on the interval $[t_0, t_1]$, i.e., bounded except on a countable set of points. The corresponding norm is then $\|x\|_\infty = \operatorname{ess\,sup}_{t_0 \le t \le t_1} |x(t)|$. Hölder's inequality for the L_p spaces is as follows (see Theorem 2.1):

Theorem 2.3 (Hölder's Inequality for L_p Spaces)**.** *For $x(t) \in L_p[t_0, t_1]$ and $y(t) \in L_q[t_0, t_1]$ with $p > 1$,*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{2.46}$$

holds

$$\int_{t_0}^{t_1} |x(t)y(t)| \,\mathrm{d}t \le \|x\|_p \|y\|_q \ . \tag{2.47}$$

The *Minkowski Inequality* for L_p Spaces corresponds to the triangle inequality (3) according to the norm definition 2.4 and is therefore not repeated here.

The common norms here are the L_1, L_2, and the L_∞ norms and are briefly summarized below.

$$\|x\|_1 = \int_{t_0}^{t_1} |x(t)| \,\mathrm{d}t \ , \tag{2.48a}$$

$$\|x\|_2 = \sqrt{\int_{t_0}^{t_1} x^2(t) \,\mathrm{d}t} \ , \tag{2.48b}$$

$$\|x\|_\infty = \operatorname{ess} \sup_{t_0 \le t \le t_1} |x(t)| \ . \tag{2.48c}$$

It is easy to see that for the function

$$x(t) = \begin{cases} 1/t & \text{for } t \ge 1 \\ 0 & \text{for } t < 1 \end{cases} \tag{2.49}$$

the L_1, L_2, and L_∞ norms can be calculated as follows

$$\|x\|_1 = \infty \ , \tag{2.50a}$$

$$\|x\|_2 = 1 \ , \tag{2.50b}$$

$$\|x\|_\infty = 1 \tag{2.50c}$$

and thus the existence of one norm does not imply the existence of other norms.

> *Exercise* 2.8. Calculate the L_1, L_2, and L_∞ norms for the time functions $x(t) = \sin(t)$, $x(t) = 1 - \exp(-t)$, and $x(t) = 1/\sqrt[3]{t}$ for $0 \le t \le \infty$.

Regarding the equivalence of norms, the following definition of topologically equivalent normed vector spaces should be mentioned:

Definition 2.5. Let $(\mathcal{X}, \| \ \|_{\mathcal{X}})$ and $\left(\mathcal{Y}, \| \ \|_{\mathcal{Y}}\right)$ be two normed linear vector spaces. Now, $\mathcal{X}$ and $\mathcal{Y}$ are called topologically isomorphic if there exists a bijective linear mapping $\mathbf{T} : \mathcal{X} \to \mathcal{Y}$ and positive real constants c_1 and c_2 such that

$$c_1 \|\mathbf{x}\|_{\mathcal{X}} \leq \|\mathbf{T}\mathbf{x}\|_{\mathcal{Y}} \leq c_2 \|\mathbf{x}\|_{\mathcal{X}} \tag{2.51}$$

for all $\mathbf{x} \in \mathcal{X}$. The norms $\| \ \|_{\mathcal{X}}$ and $\| \ \|_{\mathcal{Y}}$ are then also called equivalent.

Finally, it should be noted that norms of finite and infinite-dimensional vector spaces can also be combined. For example, consider the vector space $\mathbf{C}^n[t_0, t_1]$, the set of all vector-valued continuous time functions mapping the interval $[t_0, t_1]$ to $\mathbb{R}^n$. If a norm of the form

$$\begin{aligned} \|\mathbf{x}(t)\|_C &= \sup_{t \in [t_0, t_1]} \|\mathbf{x}(t)\|_2 \\ &= \sup_{t \in [t_0, t_1]} \left(\sum_{i=1}^{n} x_i^2(t) \right)^{1/2} , \end{aligned} \tag{2.52}$$

is defined, then $\| \ \|_2$ provides a norm on $\mathbb{R}^n$ with an n-dimensional vector as the argument, while $\| \ \|_C$ denotes the norm on $\mathbf{C}^n[t_0, t_1]$ with a vector-valued time function as the argument.

> *Exercise* 2.9. Prove that $\|\mathbf{x}(t)\|_C$ from (2.50) is a norm.

2.1.2 Induced Matrix Norm

A real-valued $(m \times n)$ matrix $\mathbf{A}$ describes a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$. Assuming $\|\mathbf{x}\|_p$ denotes a valid norm, one defines the so-called *induced* p-norm as follows:

$$\|\mathbf{A}\|_{i,p} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} . \tag{2.53}$$

It is immediately clear that the following inequality holds for $\mathbf{x} \neq \mathbf{0}$:

$$\|\mathbf{A}\mathbf{x}\|_p = \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \|\mathbf{x}\|_p \leq \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \|\mathbf{x}\|_p = \|\mathbf{A}\|_{i,p} \|\mathbf{x}\|_p . \tag{2.54}$$

For $p = 1, 2, \infty$, we have:

$$\underbrace{\|\mathbf{A}\|_{i,1} = \max_j \sum_{i=1}^{m} |a_{ij}|}_{\text{maximum absolute column sum}} \quad , \quad \|\mathbf{A}\|_{i,2} = \sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})} \quad \text{und} \quad \underbrace{\|\mathbf{A}\|_{i,\infty} = \max_i \sum_{j=1}^{n} |a_{ij}|}_{\text{maximum absolute row sum}} , \tag{2.55}$$

where $\lambda_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ (largest singular value of $\mathbf{A}$). For example, if we consider the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 6 & 5 \\ 9 & 7 & 8 \end{bmatrix} , \tag{2.56}$$

the induced norms can be calculated as (in MATLAB using the commands *norm*(A,1), *norm*(A), and *norm*(A,inf):

$$\|\mathbf{A}\|_{i,1} = 16 \ , \tag{2.57a}$$
$$\|\mathbf{A}\|_{i,2} = 16.708 \ , \tag{2.57b}$$
$$\|\mathbf{A}\|_{i,\infty} = 24 \ . \tag{2.57c}$$

> *Exercise* 2.10. Prove that for $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times l}$ with the induced matrix norm $\| \ \|_{i,p}$, the following holds:
>
> $$\|\mathbf{AB}\|_{i,p} \le \|\mathbf{A}\|_{i,p}\|\mathbf{B}\|_{i,p} \ . \tag{2.58}$$

> *Exercise* 2.11. Show that for $\mathbf{A} \in \mathbb{R}^{m\times n}$, the following inequalities hold:
>
> $$\begin{aligned} \|\mathbf{A}\|_{i,2} &\le \sqrt{\|\mathbf{A}\|_{i,1}\|\mathbf{A}\|_{i,\infty}} \\ \frac{1}{\sqrt{n}}\|\mathbf{A}\|_{i,\infty} \le \|\mathbf{A}\|_{i,2} &\le \sqrt{m}\|\mathbf{A}\|_{i,\infty} \\ \frac{1}{\sqrt{m}}\|\mathbf{A}\|_{i,1} \le \|\mathbf{A}\|_{i,2} &\le \sqrt{n}\|\mathbf{A}\|_{i,1} \end{aligned} \tag{2.59}$$

Using the so-called *Rayleigh quotient*, a convenient estimate of quadratic forms can be given. The Rayleigh quotient of a real-valued (complex-valued) $(n \times n)$ matrix $\mathbf{A}$ with any nontrivial vector $\mathbf{x}$ is defined as:

$$R[\mathbf{x}] = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \ . \tag{2.60}$$

It is important to note that in the complex case, $\mathbf{x}^{\mathrm{T}}$ refers to the transposed, complex conjugate. We want to find the vector $\mathbf{x}$ for which the Rayleigh quotient attains extreme values, i.e.,

$$\left(\frac{\partial}{\partial \mathbf{x}} R[\mathbf{x}]\right)^{T} = \frac{2\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} - \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{(\mathbf{x}^{\mathrm{T}}\mathbf{x})^2}2\mathbf{x} = \frac{2}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}(\mathbf{A}\mathbf{x} - R[\mathbf{x}]\mathbf{x}) = \mathbf{0} \ . \tag{2.61}$$

Since the Rayleigh quotient is real, the extremal value problem reduces to solving an eigenvalue problem of the form:

$$(\mathbf{A} - R[\mathbf{x}]\mathbf{I})\mathbf{x} = \mathbf{0} \tag{2.62}$$

with the identity matrix $\mathbf{I}$.

Therefore, the eigenvectors of $\mathbf{A}$ are solutions to the extremal value problem of the Rayleigh quotient (2.61), and with $\mathbf{x}$ as an eigenvector of $\mathbf{A}$, the Rayleigh quotient $R[\mathbf{x}]$ corresponds to the associated eigenvalue λ due to:

$$R[\mathbf{x}] = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \frac{\lambda\mathbf{x}^{\mathrm{T}}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \lambda \tag{2.63}$$

This allows us to provide the following useful estimation for all $\mathbf{x} \in \mathbb{R}^n$:

$$\lambda_{\min}(\mathbf{A})\|\mathbf{x}\|_2^2 \leq \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})\|\mathbf{x}\|_2^2 \tag{2.64}$$

> *Exercise* 2.12. Show that every square matrix $\mathbf{A}$ can be decomposed into a symmetric part $\mathbf{A}_s$ and a skew-symmetric part $\mathbf{A}_{ss}$. Furthermore, show that in the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, the skew-symmetric part of the matrix $\mathbf{A}$ cancels out.

> *Exercise* 2.13. Use the Rayleigh quotient to show that a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has exclusively real eigenvalues and a positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has exclusively positive real eigenvalues.

2.1.3 Banach Space

In the following, we will consider convergence in normed vector spaces.

Definition 2.6 (Convergence)**.** A sequence of points $(\mathbf{x}_k)$ in a normed linear vector space $(\mathcal{X}, \|\ \ \|)$ with $\mathbf{x}_k \in \mathcal{X}$ is called *convergent* to a limit $\mathbf{x} \in \mathcal{X}$ (in compact notation $\mathbf{x}_k \to \mathbf{x}$) if

$$\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}\| = 0 \tag{2.65}$$

holds. Furthermore, for a continuous function $\mathbf{f}(\mathbf{x})$, it holds that if $\mathbf{x}_k \to \mathbf{x}$, then $\mathbf{f}(\mathbf{x}_k) \to \mathbf{f}(\mathbf{x})$.

The above definition allows to investigate whether a given sequence converges to a given limit or not. However, this requires knowledge of the limit, which is generally not available. Therefore, one often resorts to the concept of a *Cauchy sequence*.

Definition 2.7 (Cauchy Sequence)**.** A sequence $(\mathbf{x}_k)$ with $\mathbf{x}_k \in \mathcal{X}$ is called a *Cauchy sequence* if

$$\lim_{n,m\to\infty} \|\mathbf{x}_n - \mathbf{x}_m\| = 0 \tag{2.66}$$

holds.

The relationship between convergent sequences and Cauchy sequences is characterized by the following theorem.

Theorem 2.4 (Cauchy Sequence)**.** *Every convergent sequence is a Cauchy sequence. However, the converse does not generally hold in normed vector spaces.*

To illustrate this theorem, consider $\mathcal{X} = C[0,1]$, i.e., the sequence of continuous functions $\{x_k(t)\}$, $k = 2, 3, \ldots$ in the interval $0 \leq t \leq 1$, of the form

$$x_k(t) = \begin{cases} 0 & \text{for} \quad 0 \leq t \leq \frac{1}{2} - \frac{1}{k} \\ kt - \frac{k}{2} + 1 & \text{for} \quad \frac{1}{2} - \frac{1}{k} < t \leq \frac{1}{2} \\ 1 & \text{for} \quad \frac{1}{2} < t \leq 1 \ . \end{cases} \tag{2.67}$$

Choosing the L_2 norm for $\{x_k(t)\} \subset C[0,1]$,

$$\|x\|_2 = \left(\int\limits_0^1 x^2(t) \, \mathrm{d}t \right)^{1/2} , \tag{2.68}$$

immediately leads to

$$\begin{aligned} \|x_m - x_n\|_2^2 &= \int_{\frac{1}{2}-\frac{1}{m}}^{\frac{1}{2}-\frac{1}{n}} \left(mt - \frac{m}{2} + 1 \right)^2 \mathrm{d}t + \int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}} \left(mt - \frac{m}{2} - nt + \frac{n}{2} \right)^2 \mathrm{d}t \\ &= \frac{(m-n)^2}{3n^2 m} \end{aligned} \tag{2.69}$$

for $n > m$, and

$$\lim_{n,m \to \infty} \|x_m - x_n\|_2^2 = 0 \ . \tag{2.70}$$

Thus, it can be seen that the sequence (2.67) is a Cauchy sequence for the L_2 norm. However, for the limit function, we have

$$\lim_{k \to \infty} x_k(t) = x(t) = \begin{cases} 0 & \text{for} \quad 0 \leq t < \frac{1}{2} \\ 1 & \text{for} \quad \frac{1}{2} < t \leq 1 \ . \end{cases} \tag{2.71}$$

This shows that the limit function $x(t)$ is not continuous and therefore not an element of $C[0,1]$.

> *Exercise* 2.14. Draw a plot of the sequence (2.67).

Since it is generally of interest that the limit of Cauchy sequences in a normed linear vector space also lies in this vector space, the concept of a *Banach space* is introduced.

Definition 2.8 (Banach space)**.** A normed linear vector space $(\mathcal{X}, \| \ \|)$ is called complete if every Cauchy sequence converges to an element $\mathbf{x} \in \mathcal{X}$. A complete, normed vector space is also called a *Banach space*.

Theorem 2.5 (Cauchy convergence criterion). *In a complete, normed vector space, a sequence converges if and only if it is a Cauchy sequence.*

The normed linear vector spaces $(\mathbb{R}^n, \| \|_p)$, $(\mathbb{R}^n, \| \|_\infty)$, $L_p[t_0, t_1]$, and $L_\infty[t_0, t_1]$ are examples of Banach spaces. Furthermore, it can be shown that $C[0, 1]$ with the norm $\| \|_\infty$ is also a Banach space.

For the following, some important definitions are needed:

Definition 2.9 (Closed subset). A subset $\mathcal{S} \subset \mathcal{X}$ is called *closed* if for every convergent sequence $(\mathbf{x}_k)$ with $\mathbf{x}_k \in \mathcal{S}$, the limit also lies in $\mathcal{S}$. If $\mathcal{S}$ is not closed, one can add to $\mathcal{S}$ the set of all possible limits of convergent sequences in $\mathcal{S}$, and this set is called the *closure* of $\mathcal{S}$ denoted by $\bar{\mathcal{S}}$. Thus, $\bar{\mathcal{S}}$ is the smallest closed subset containing $\mathcal{S}$.

Definition 2.10 (Bounded subset). A subset $\mathcal{S} \subset \mathcal{X}$ is *bounded* if

$$\sup_{\mathbf{x} \in \bar{\mathcal{S}}} \|\mathbf{x}\|_{\mathcal{X}} < \infty \ . \tag{2.72}$$

Definition 2.11 (Compact subset). A subset $\mathcal{S} \subset \mathcal{X}$ is called *compact* or *relatively compact* if every sequence in $\mathcal{S}$ or $\bar{\mathcal{S}}$ contains a convergent subsequence with the limit in $\mathcal{S}$ or $\bar{\mathcal{S}}$.

The following theorems hold for subspaces of a Banach space:

Theorem 2.6. *In a Banach space, a subset is complete if and only if it is closed.*

Theorem 2.7. *In a normed linear vector space, every finite-dimensional subspace is complete.*

Next, consider an equation of the form $\mathbf{x} = T(\mathbf{x})$. A solution $\mathbf{x}^*$ of this equation is called a *fixed point* of the mapping T, since $\mathbf{x}^*$ is invariant under T. A classical approach to finding the fixed point is the so-called *successive approximation* using the recurrence equation $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$ with the initial value $\mathbf{x}_0$. The *contraction mapping theorem* provides sufficient conditions for the existence of a unique fixed point for the mapping T in a Banach space and for the convergence of the successive approximation sequence to this fixed point.

Theorem 2.8 (Contraction Theorem). *Let $\mathcal{S}$ be a non-empty closed subset of a Banach space $\mathcal{X}$ with the mapping $T : \mathcal{S} \to \mathcal{S}$. If for all $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}$ the inequality*

$$\|T(\mathbf{x}) - T(\mathbf{y})\| \leq \rho \|\mathbf{x} - \mathbf{y}\| \ , \quad 0 \leq \rho < 1 \ , \tag{2.73}$$

holds, then the equation

$$\mathbf{x} = T(\mathbf{x}) \tag{2.74}$$

has exactly one fixed point solution $\mathbf{x} = \mathbf{x}^*$*, and the sequence* $\mathbf{x}_{k+1} = T(\mathbf{x}_k)$ *converges for every initial value* $\mathbf{x}_0 \in \mathcal{S}$ *to* $\mathbf{x}^*$*. In this case,* T *is called a* contraction*.*

The following exercise demonstrates a simple application of the Contraction Theorem.

Exercise 2.15. Consider a linear system of equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.75}$$

with a real-valued $(n \times n)$ matrix $\mathbf{A}$. Suppose

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \ . \tag{2.76}$$

Show that the equation system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a unique solution, which can be computed using the recurrence equation

$$\mathbf{D}\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{A})\mathbf{x}_k + \mathbf{b} \ , \quad k \geq 0 \ , \quad \mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, \dots, a_{nn}) \tag{2.77}$$

for every $\mathbf{x}_0 \in \mathbb{R}^n$.

2.1.4 Hilbert Space

A so-called *pre-Hilbert space* is a linear vector space $\mathcal{X}$ equipped with an inner product.

Definition 2.12 (Pre-Hilbert Space)**.** Let $\mathcal{X}$ be a linear vector space over the scalar field K. A mapping $\langle \mathbf{x}, \mathbf{y} \rangle : \mathcal{X} \times \mathcal{X} \to K$, which assigns to each pair of elements $\mathbf{x}$, $\mathbf{y} \in \mathcal{X}$ a scalar, is called an *inner product* if it satisfies the following conditions:

$$\begin{aligned}
&(1)\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \quad \text{(Sesquilinear form)} \\
&(2)\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle^* \\
&(3)\langle a\mathbf{x}, \mathbf{y} \rangle = a\langle \mathbf{x}, \mathbf{y} \rangle \\
&(4)\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \quad \text{und} \quad \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = 0
\end{aligned} \tag{2.78}$$

where $\langle \mathbf{y}, \mathbf{x} \rangle^*$ denotes the complex conjugate of $\langle \mathbf{y}, \mathbf{x} \rangle$ and $a \in K$.

Examples of vector spaces with an inner product include vectors in $\mathbb{R}^n$ with

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^{\mathrm{T}}\mathbf{x} \tag{2.79}$$

or the vector space of continuous time functions on the interval $-1 \leq t \leq 1$ with the inner product

$$\langle x, y \rangle = \int_{-1}^{1} y(\tau)x(\tau)\, \mathrm{d}\tau \ . \tag{2.80}$$

As the examples show, the inner product also defines the specific norm

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \ . \tag{2.81}$$

To generalize this property, the following theorem is needed.

Theorem 2.9 (Cauchy-Schwarz Inequality)**.** *For all* $\mathbf{x}$, $\mathbf{y}$, *elements of a linear vector space* $\mathcal{X}$ *with scalar field* K *and an inner product, the following inequality holds:*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \ . \tag{2.82}$$

The equality in (2.82) is satisfied if and only if $\mathbf{x} = \lambda \mathbf{y}$ *or* $\mathbf{y} = \mathbf{0}$.

Proof. To prove this, consider the inequality valid for all $a \in K$:

$$\begin{aligned} 0 &\leq \langle \mathbf{x} - a\mathbf{y}, \mathbf{x} - a\mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle - \langle a\mathbf{y}, \mathbf{x} \rangle - \underbrace{\langle \mathbf{x}, a\mathbf{y} \rangle}_{=\langle a\mathbf{y}, \mathbf{x} \rangle^* = a^* \langle \mathbf{y}, \mathbf{x} \rangle^*} + |a|^2 \langle \mathbf{y}, \mathbf{y} \rangle \end{aligned} \tag{2.83}$$

with $\mathbf{y} \neq \mathbf{0}$. Choosing

$$a = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} \ , \tag{2.84}$$

it follows

$$0 \leq \langle \mathbf{x}, \mathbf{x} \rangle - \frac{\|\langle \mathbf{x}, \mathbf{y} \rangle\|^2}{\langle \mathbf{y}, \mathbf{y} \rangle} \tag{2.85}$$

or

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \ . \tag{2.86}$$

For $\mathbf{y} = \mathbf{0}$, nothing needs to be shown. □

Theorem 2.10 (Associated Norm in Pre-Hilbert Spaces)**.** *In a pre-Hilbert space* $\mathcal{X}$, *the inner product induces a function* $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ *that is a norm according to the definition in 2.4.*

In a pre-Hilbert space, there are other useful properties:

Theorem 2.11. *In a pre-Hilbert space* $\mathcal{X}$, *if* $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ *for all* $\mathbf{x} \in \mathcal{X}$, *then* $\mathbf{y} = \mathbf{0}$.

Exercise 2.16. Prove Theorem 2.11.

Theorem 2.12 (Parallelogram Equation)**.** *In a pre-Hilbert space* $\mathcal{X}$, *the following equation holds:*

$$\|\mathbf{x} + \mathbf{y}\|_2^2 + \|\mathbf{x} - \mathbf{y}\|_2^2 = 2\|\mathbf{x}\|_2^2 + 2\|\mathbf{y}\|_2^2 \ . \tag{2.87}$$

Exercise 2.17. Prove Theorem 2.12.

Definition 2.13 (Hilbert Space)**.** A complete pre-Hilbert space is called a *Hilbert space.*

Therefore, a Hilbert space is a Banach space equipped with an inner product that, according to Theorem 2.10, induces a norm. The spaces $(\mathbb{R}^n, \| \, \|_2))$ and $L_2[t_0, t_1]$ are Hilbert spaces with inner products

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^{\mathrm{T}} \mathbf{x} \tag{2.88}$$

for $\mathbf{x}^T = [x_1, \ldots, x_n]$ and $\mathbf{y}^T = [y_1, \ldots, y_n]$, and

$$\langle x, y \rangle_{L_2[t_0,t_1]} = \int_{t_0}^{t_1} x(t) y^*(t) \, \mathrm{d}t \tag{2.89}$$

for x, $y \in L_2[t_0, t_1]$. It is important to note that in this case, the Cauchy-Schwarz inequality (2.82) corresponds to Hölder's inequality (2.40) or (2.47) for $p = q = 2$.

2.1.5 Existence and Uniqueness

The solution of a differential equation does not have to be unique. To see this, consider the differential equation

$$\dot{x} = x^{1/3} \; , \quad x_0 = 0 \; . \tag{2.90}$$

It is easy to verify that

$$x(t) = 0 \; , \tag{2.91a}$$

$$x(t) = \left(\frac{2t}{3}\right)^{3/2} \tag{2.91b}$$

are solutions of (2.90). Although the right-hand side of the differential equation is continuous, the solution is not unique. In fact, continuity guarantees the *existence of a solution*, but further conditions are needed for *uniqueness*. In the following, the time-varying system

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}) \; , \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2.92}$$

is examined, as this also covers the non-autonomous case.

Theorem 2.13 (Local Existence and Uniqueness)**.** *Let* $\mathbf{f}(t, \mathbf{x})$ *be piecewise continuous in* t *and satisfy the estimate (*Lipschitz condition*)*

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\| \; , \quad 0 < L < \infty \tag{2.93}$$

for all $\mathbf{x}$, $\mathbf{y} \in B = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \le r\}$ *and all* $t \in [t_0, t_0 + \tau]$*. Then there exists a* $\delta > 0$ *such that*

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}) \; , \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2.94}$$

has exactly one *solution for* $t \in [t_0, t_0 + \delta]$*. In this case, the function* $\mathbf{f}(t, \mathbf{x})$ *is said to be* locally Lipschitz *on* $B \subset \mathbb{R}^n$*. If condition (2.93) holds in the entire* $\mathbb{R}^n$*, then the* *function* $\mathbf{f}(t, \mathbf{x})$ *is called* globally Lipschitz.

Proof. The proof of this theorem is based on the contraction theorem according to Theorem 2.8. In a first step, the Banach space $\mathcal{X} = \mathbf{C}^n[t_0, t_0 + \delta]$ of all vector-valued continuous time functions in the time interval $[t_0, t_0 + \delta]$ is defined with the norm $\|\mathbf{x}(t)\|_C = \sup_{t \in [t_0, t_0+\delta]} \|\mathbf{x}(t)\|$. For further explanation, see also (2.52). Furthermore, the differential equation (2.94) is transformed into an equivalent integral equation of the form

$$(P\mathbf{x})(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}(\tau))\, \mathrm{d}\tau \tag{2.95}$$

Within the proof, it is then shown that the mapping P on the closed subset $\mathcal{S} \subset \mathcal{X}$ with $\mathcal{S} = \{\mathbf{x} \in \mathbf{C}^n[t_0, t_0 + \delta] \mid \|\mathbf{x} - \mathbf{x}_0\|_C \leq r\}$ is a contraction and that P maps the subset $\mathcal{S}$ to itself. To do this, one calculates

$$(P\mathbf{x}_1)(t) - (P\mathbf{x}_2)(t) = \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_1(\tau))\, \mathrm{d}\tau - \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}_2(\tau))\, \mathrm{d}\tau \tag{2.96}$$

for $\mathbf{x}_1(t)$, $\mathbf{x}_2(t) \in \mathcal{S}$.
It now holds that

$$\begin{aligned}
\|(P\mathbf{x}_1)(t) - (P\mathbf{x}_2)(t)\|_C &= \left\| \int_{t_0}^{t} (\mathbf{f}(\tau, \mathbf{x}_1(\tau)) - \mathbf{f}(\tau, \mathbf{x}_2(\tau)))\, \mathrm{d}\tau \right\|_C \\
&\leq \int_{t_0}^{t} \|\mathbf{f}(\tau, \mathbf{x}_1(\tau)) - \mathbf{f}(\tau, \mathbf{x}_2(\tau))\|_C\, \mathrm{d}\tau \\
&\leq \int_{t_0}^{t} L\|\mathbf{x}_1(\tau) - \mathbf{x}_2(\tau)\|_C\, \mathrm{d}\tau \\
&\leq L\delta\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\|_C \ ,
\end{aligned} \tag{2.97}$$

and by choosing

$$\delta \leq \rho/L \ , \quad \rho < 1 \ , \tag{2.98}$$

and with (2.98), Theorem 2.8 shows that P is a contraction on $\mathcal{S}$. In the next step, it must be proven that the mapping P maps the subset $\mathcal{S} \subset \mathcal{X}$ to itself. Since $\mathbf{f}$ is piecewise continuous, it follows that $\mathbf{f}(t, \mathbf{x}_0)$ is bounded on the interval $[t_0, t_0 + \delta]$, hence

$$h = \max_{t \in [t_0, t_0+\delta]} \|\mathbf{f}(t, \mathbf{x}_0)\| \ . \tag{2.99}$$

This results in

$$\begin{aligned}
\|(P\mathbf{x})(t)-\mathbf{x}_0\|_C &\leq \int\limits_{t_0}^{t} \|\mathbf{f}(\tau,\mathbf{x}(\tau))\|_C \,\mathrm{d}\tau \\
&\leq \int\limits_{t_0}^{t} \|\mathbf{f}(\tau,\mathbf{x}(\tau))-\mathbf{f}(\tau,\mathbf{x}_0)+\mathbf{f}(\tau,\mathbf{x}_0)\|_C \,\mathrm{d}\tau \\
&\leq \int\limits_{t_0}^{t} \left(\|\mathbf{f}(\tau,\mathbf{x}(\tau))-\mathbf{f}(\tau,\mathbf{x}_0)\|_C+\|\mathbf{f}(\tau,\mathbf{x}_0)\|_C\right) \mathrm{d}\tau \\
&\leq \int\limits_{t_0}^{t} \left(L\|\mathbf{x}(\tau)-\mathbf{x}_0\|_C+h\right) \mathrm{d}\tau \\
&\leq \delta(Lr+h)\ .
\end{aligned} \tag{2.100}$$

Choosing

$$\delta \leq \frac{r}{Lr+h}\ , \tag{2.101}$$

ensures that $\mathcal{S}$ is mapped onto itself under P. Combining (2.98) and (2.101) and choosing δ to be less than or equal to the considered time interval τ from Theorem 2.13,

$$\delta = \min\left(\frac{\rho}{L}, \frac{r}{Lr+h}, \tau\right)\ , \quad \rho < 1\ , \tag{2.102}$$

the existence and uniqueness of the solution in $\mathcal{S}$ for $t \in [t_0, t_0+\delta]$ is thus demonstrated. □

Since the mapping P from (2.95) is a contraction, it follows from Theorem 2.8 that the sequence $\mathbf{x}_{k+1} = P\mathbf{x}_k$ with $\mathbf{x}_0 = \mathbf{x}(t_0)$ converges to the unique solution of the integral equation (2.95) or the equivalent differential equation (2.94). This method is also known as the *Picard iteration method*.

Exercise 2.18. Show that for linear, time-invariant systems of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}\ , \quad \mathbf{x}(t_0) = \mathbf{x}_0\ , \tag{2.103}$$

the Picard iteration method precisely iteratively calculates the transition matrix $\mathbf{\Phi}(t) = e^{\mathbf{A}t}$.

Exercise 2.19. Calculate, using the Picard iteration method, the transition matrix of a linear, time-varying system of the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}\ , \quad \mathbf{x}(t_0) = \mathbf{x}_0\ . \tag{2.104}$$

Tip: The transition matrix of (2.104) is calculated from the *Peano-Baker series* as

$$\mathbf{\Phi}(t) = \mathbf{I} + \int\limits_0^t \mathbf{A}(\tau)\,\mathrm{d}\tau + \int\limits_0^t \mathbf{A}(\tau)\int\limits_0^\tau \mathbf{A}(\tau_1)\,\mathrm{d}\tau_1\,\mathrm{d}\tau + \dots \tag{2.105}$$

For a scalar function $f(x) : \mathbb{R} \to \mathbb{R}$ that does not explicitly depend on time t, the Lipschitz condition (2.93) can be written very simply as

$$\frac{|f(y) - f(x)|}{|y - x|} \leq L \tag{2.106}$$

The condition (2.106) allows a very simple graphical interpretation, namely the function $f(x)$ must not have a slope greater than L. Therefore, functions $f(x)$ that have an infinite slope at a point (like the function $x^{1/3}$ from (2.90) at the point $x = 0$) are certainly not locally Lipschitz. This also implies that discontinuous functions $f(x)$ do not satisfy the Lipschitz condition (2.93) at the point of discontinuity. This connection between the Lipschitz condition and the boundedness of $\left|\frac{\partial}{\partial x} f(x)\right|$ is generalized in the following theorem without proof:

Theorem 2.14 (Lipschitz condition and continuity)**.** *If the functions* $\mathbf{f}(t, \mathbf{x})$ *from (2.92) and* $[\partial\mathbf{f}/\partial\mathbf{x}](t, \mathbf{x})$ *are continuous on the set* $[t_0, t_0 + \delta] \times B$ *with* $B \subset \mathbb{R}^n$*, then* $\mathbf{f}(t, \mathbf{x})$ *locally satisfies the Lipschitz condition of (2.93).*

To verify the *global existence and uniqueness* of a differential equation of type (2.92), the following theorem is provided:

Theorem 2.15 (Global Existence and Uniqueness)**.** *Assume that the function* $\mathbf{f}(t, \mathbf{x})$ *from (2.92) is piecewise continuous in* t *and globally Lipschitz for all* $t \in [t_0, t_0 + \tau]$ *according to Theorem 2.13. Then the differential equation (2.92) has a unique solution in the time interval* $t \in [t_0, t_0 + \tau]$*. If the function* $\mathbf{f}(t, \mathbf{x})$ *from (2.92) and* $[\partial\mathbf{f}/\partial\mathbf{x}](t, \mathbf{x})$ *are continuous on the set* $[t_0, t_0 + \tau] \times \mathbb{R}^n$*, then* $\mathbf{f}(t, \mathbf{x})$ *is globally Lipschitz if and only if* $[\partial\mathbf{f}/\partial\mathbf{x}](t, \mathbf{x})$ *on* $[t_0, t_0 + \tau] \times \mathbb{R}^n$ *is* uniformly bounded.

To explain, $[\partial\mathbf{f}/\partial\mathbf{x}](t, \mathbf{x})$ is *uniformly bounded* if, independently of $t_0 \geq 0$, for every positive, finite constant a, there exists a $\beta(a) > 0$ independent of t_0 such that

$$\left\|\frac{\partial\mathbf{f}}{\partial\mathbf{x}}(t_0, \mathbf{x}(t_0))\right\|_i \leq a \Rightarrow \left\|\frac{\partial\mathbf{f}}{\partial\mathbf{x}}(t, \mathbf{x}(t))\right\|_i \leq \beta(a) \tag{2.107}$$

with $\| \, \|_i$ denoting the induced norm according to (2.53) for all $t \in [t_0, t_0 + \tau]$ and all $\mathbf{x} \in \mathbb{R}^n$.

The proofs of the last two theorems can be found in the literature cited at the end of this chapter. As an example, consider the system

$$\underbrace{\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix}}_{\dot{\mathbf{x}}} = \underbrace{\begin{bmatrix} -x_1 + x_1 x_2 \\ x_2 - x_1 x_2 \end{bmatrix}}_{\mathbf{f}(\mathbf{x})} . \tag{2.108}$$

From Theorem 2.14, it can be immediately concluded that $\mathbf{f}(\mathbf{x})$ from (2.108) is locally Lipschitz on $\mathbb{R}^2$. However, the application of Theorem 2.15 shows that $\mathbf{f}(\mathbf{x})$ is not globally Lipschitz, since $\partial\mathbf{f}/\partial\mathbf{x}$ on $\mathbb{R}^2$ is not uniformly bounded.

In summary, it can be stated that the mathematical models of most physical systems in the form of (2.92) are locally Lipschitz, as this essentially corresponds to a requirement of continuous differentiability of the right-hand side, as stated in Theorem 2.14. In contrast, the global Lipschitz condition is very restrictive and is satisfied by only a few physical systems, as was already hinted at by the requirement for the uniform boundedness of $[\partial\mathbf{f}/\partial\mathbf{x}](t,\mathbf{x})$.

Exercise 2.20. Check for the following functions

$$(1) \quad f(x) = x^2 + |x| \tag{2.109}$$

$$(2) \quad f(x) = \sin(x)\,\mathrm{sgn}(x) \tag{2.110}$$

$$(3) \quad f(x) = \tan(x) \tag{2.111}$$

and

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} ax_1 + \tanh(bx_1) - \tanh(bx_2) \\ ax_2 + \tanh(bx_1) + \tanh(bx_2) \end{bmatrix} \tag{2.112}$$

and

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} -x_1 + a\|x_2\| \\ -(a+b)x_1 + bx_1^2 - x_1x_2 \end{bmatrix} , \tag{2.113}$$

whether they are (a) continuous, (b) continuously differentiable, (c) locally Lipschitz, and (d) globally Lipschitz.

Exercise 2.21. Show that the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -x_1 + \frac{2x_2}{1+x_2^2} \\ -x_2 + \frac{2x_1}{1+x_1^2} \end{bmatrix} , \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2.114}$$

has a unique solution for all $t \geq t_0$.

2.1.6 Influence of Parameters

Often one wants to investigate the influence of parameters on the solution of a differential equation of the form

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}) \ , \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{2.115}$$

with the parameter vector $\mathbf{p} \in \mathbb{R}^d$. Let $\mathbf{p}_0$ denote the nominal value of the parameter vector $\mathbf{p}$.

Theorem 2.16 (Influence of Parameters). *Assume that $\mathbf{f}(t,\mathbf{x},\mathbf{p})$ is continuous in $(t,\mathbf{x},\mathbf{p})$ and locally Lipschitz in $\mathbf{x}$ (Lipschitz condition (2.93)) on $[t_0,t_0+\tau]\times D\times\{\mathbf{p}| \, \|\mathbf{p}-\mathbf{p}_0\|\leq r\}$ with $D\subset\mathbb{R}^n$. Furthermore, let $\mathbf{y}(t,\mathbf{p}_0)$ be a solution of the differential equation $\dot{\mathbf{y}}=\mathbf{f}(t,\mathbf{y},\mathbf{p}_0)$ with the initial value $\mathbf{y}(t_0,\mathbf{p}_0)=\mathbf{y}_0\in D$, where the solution $\mathbf{y}(t,\mathbf{p}_0)$ remains in D for all times $t\in[t_0,t_0+\tau]$. Then, for a given $\varepsilon>0$, there exist δ_1, $\delta_2>0$ such that for*

$$\|\mathbf{z}_0-\mathbf{y}_0\|<\delta_1 \quad \textit{und} \quad \|\mathbf{p}-\mathbf{p}_0\|<\delta_2 \tag{2.116}$$

the differential equation $\dot{\mathbf{z}}=\mathbf{f}(t,\mathbf{z},\mathbf{p})$ with the initial value $\mathbf{z}(t_0,\mathbf{p})=\mathbf{z}_0$ has a unique solution $\mathbf{z}(t,\mathbf{p})$ for all times $t\in[t_0,t_0+\tau]$ and $\mathbf{z}(t,\mathbf{p})$ satisfies the condition

$$\|\mathbf{z}(t,\mathbf{p})-\mathbf{y}(t,\mathbf{p}_0)\|<\varepsilon \tag{2.117}$$

For the proof of this theorem, we refer to the literature cited at the end of this chapter. In essence, this theorem states that for all parameters $\mathbf{p}$ sufficiently close to the nominal value $\mathbf{p}_0$ ($\|\mathbf{p}-\mathbf{p}_0\|<\delta_2$), the differential equation (2.115) has a unique solution that is very close to the nominal solution of the differential equation $\dot{\mathbf{x}}=\mathbf{f}(t,\mathbf{x},\mathbf{p}_0)$, $\mathbf{x}(t_0)=\mathbf{x}_0$.

Assuming that $\mathbf{f}(t,\mathbf{x},\mathbf{p})$ satisfies the conditions of Theorem 2.16 and has continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{p}$ for all $(t,\mathbf{x},\mathbf{p})\in[t_0,t_0+\tau]\times\mathbb{R}^n\times\mathbb{R}^d$. The differential equation (2.115) can now be rewritten into an equivalent integral equation of the form

$$\mathbf{x}(t,\mathbf{p})=\mathbf{x}_0+\int_{t_0}^{t}\mathbf{f}(s,\mathbf{x}(s,\mathbf{p}),\mathbf{p})\,\mathrm{d}s \tag{2.118}$$

Due to the continuous differentiability of $\mathbf{f}(t,\mathbf{x},\mathbf{p})$ with respect to $\mathbf{x}$ and $\mathbf{p}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\mathbf{x}(t,\mathbf{p})=\underbrace{\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\mathbf{x}_0}_{=\mathbf{0}}+\int_{t_0}^{t}\frac{\partial}{\partial\mathbf{x}}\mathbf{f}(s,\mathbf{x}(s,\mathbf{p}),\mathbf{p})\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\mathbf{x}(s,\mathbf{p})+\frac{\partial}{\partial\mathbf{p}}\mathbf{f}(s,\mathbf{x}(s,\mathbf{p}),\mathbf{p})\,\mathrm{d}s\ . \tag{2.119}$$

Differentiating (2.119) with respect to t, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_{\mathbf{p}}(t,\mathbf{p})=\mathbf{A}(t,\mathbf{p})\mathbf{x}_{\mathbf{p}}(t,\mathbf{p})+\mathbf{B}(t,\mathbf{p})\ ,\quad \mathbf{x}_{\mathbf{p}}(t_0,\mathbf{p})=\mathbf{0} \tag{2.120}$$

and

$$\mathbf{x}_{\mathbf{p}}(t,\mathbf{p})=\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\mathbf{x}(t,\mathbf{p})\ , \tag{2.121a}$$

$$\mathbf{A}(t,\mathbf{p})=\frac{\partial}{\partial\mathbf{x}}\mathbf{f}(t,\mathbf{x},\mathbf{p})\bigg|_{\mathbf{x}=\mathbf{x}(t,\mathbf{p})}\ , \tag{2.121b}$$

$$\mathbf{B}(t,\mathbf{p})=\frac{\partial}{\partial\mathbf{p}}\mathbf{f}(t,\mathbf{x},\mathbf{p})\bigg|_{\mathbf{x}=\mathbf{x}(t,\mathbf{p})}\ . \tag{2.121c}$$

For parameters $\mathbf{p}$ sufficiently close to the nominal value $\mathbf{p}_0$, the matrices $\mathbf{A}(t,\mathbf{p})$ and $\mathbf{B}(t,\mathbf{p})$, and thus $\mathbf{x}_{\mathbf{p}}(t,\mathbf{p})$, are well-defined on the time interval $[t_0,t_0+\tau]$. Substituting

$\mathbf{p} = \mathbf{p}_0$ into $\mathbf{x}_\mathbf{p}(t, \mathbf{p})$ yields the so-called *sensitivity function*

$$\mathbf{S}(t) = \mathbf{x}_\mathbf{p}(t, \mathbf{p}_0) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\mathbf{x}(t, \mathbf{p})\Big|_{\mathbf{p}=\mathbf{p}_0} \tag{2.122}$$

which is the solution of the differential equation (compare with (2.120))

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}_0) \ , \tag{2.123a}$$

$$\mathbf{x}(t_0) = \mathbf{x}_0 \ , \tag{2.123b}$$

$$\dot{\mathbf{S}} = \left[\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(t, \mathbf{x}, \mathbf{p})\right]_{\mathbf{p}=\mathbf{p}_0} \mathbf{S} + \left[\frac{\partial}{\partial \mathbf{p}}\mathbf{f}(t, \mathbf{x}, \mathbf{p})\right]_{\mathbf{p}=\mathbf{p}_0} \ , \tag{2.123c}$$

$$\mathbf{S}(t_0) = \mathbf{0} \ . \tag{2.123d}$$

The matrix differential equation for $\mathbf{S}(t)$ is also referred to as the *sensitivity equation*. The sensitivity function can be interpreted as providing a first-order approximation for the effect of parameter variations on the solution. This allows for approximating the solution $\mathbf{x}(t, \mathbf{p})$ of (2.115) for small changes in the parameter vector $\mathbf{p}$ from the nominal value $\mathbf{p}_0$ in the form

$$\mathbf{x}(t, \mathbf{p}) \approx \mathbf{x}(t, \mathbf{p}_0) + \mathbf{S}(t)(\mathbf{p} - \mathbf{p}_0) \tag{2.124}$$

This approximation is, among other things, the basis for singular perturbation theory. While one could imagine determining the effect of parameter variations by simply varying the parameters in the differential equations, this approach has the disadvantage that small parameter variations often get lost in the round-off errors of the integration, thus not allowing for quantitative statements about the influence of parameters on the solution.

Exercise 2.22. The following differential equation system (Phase-Locked-Loop) is given

$$\dot{x}_1 = x_2 \tag{2.125}$$

$$\dot{x}_2 = -c\sin(x_1) - (a + b\cos(x_1))x_2 \tag{2.126}$$

with state $\mathbf{x}^\mathrm{T} = [x_1, x_2]$ and parameter vector $\mathbf{p}^\mathrm{T} = [a, b, c]$. The nominal values of the parameter vector $\mathbf{p}$ are $\mathbf{p}_0 = [1, 0, 1]$. The sensitivity function $\mathbf{S}(t)$ according to (2.122) is sought. Compare the solutions for the nominal parameter vector $\mathbf{p}_0$ and for the parameter vector $\mathbf{p}^\mathrm{T} = [1.2, -0.2, 0.8]$ for $\mathbf{x}_0^\mathrm{T} = [1, 1]$ by simulation in Matlab/Simulink.

Exercise 2.23. Calculate the sensitivity equation for the *Van der Pol oscillator*

$$\ddot{v} - \varepsilon\left(1 - v^2\right)\dot{v} + v = 0 \tag{2.127}$$

with state $\mathbf{x}^\mathrm{T} = [v, \dot{v}]$ and parameter $p = \varepsilon$. Compare the solutions for various small deviations from the nominal value $\varepsilon_0 = 0.01$ by simulation in Matlab/Simulink.

2.2 Literatur

[2.1] M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems and Linear Algebra.* San Diego: Academic Press, 1974.

[2.2] H. K. Khalil, *Nonlinear Systems (3rd Edition).* New Jersey: Prentice Hall, 2002.

[2.3] D. Luenberger, *Optimization by Vector Space Methods.* New York: John Wiley & Sons, 1969.

[2.4] D. Luenberger, *Introduction to Dynamic Systems.* New York: John Wiley & Sons, 1979.

[2.5] E. Slotine and W. Li, *Applied Nonlinear Control.* New Jersey: Prentice Hall, 1991.

[2.6] M. Vidyasagar, *Nonlinear Systems Analysis.* New Jersey: Prentice Hall, 1993.

3 Fundamentals of Lyapunov Theory

This chapter covers the theoretical foundations for investigating the stability of an equilibrium point for autonomous and non-autonomous nonlinear systems.

3.1 Autonomous Systems

In this section, we consider an autonomous system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{3.1}$$

with the smooth vector field $\mathbf{f}(\mathbf{x})$. Denoting the flow of (3.1) by $\mathbf{\Phi}_t(\mathbf{x})$, an equilibrium point $\mathbf{x}_R$ satisfies the relation

$$\mathbf{f}(\mathbf{x}_R) = \mathbf{0} \qquad \text{or} \qquad \mathbf{\Phi}_t(\mathbf{x}_R) = \mathbf{x}_R \ . \tag{3.2}$$

Without loss of generality, we can assume that the equilibrium point is $\mathbf{x}_R = \mathbf{0}$. If $\mathbf{x}_R \neq \mathbf{0}$, then by a simple coordinate transformation $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_R$, one can always achieve that in the new coordinates $\tilde{\mathbf{x}}_R = \mathbf{0}$. The concept of a vector field will now be briefly explained.

3.1.1 Vector Fields

An important concept in the study of (autonomous) systems of the form (3.1) is that of a *vector field*, where so-called *smooth vector fields* are of particular significance. The following definition applies:

Definition 3.1 (Smooth Function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *smooth* or C^∞ if f and all *partial derivatives* of any order l

$$\frac{\partial^l}{\prod_{i=1}^n \partial^{l_i} x_i} f(x_1, \dots, x_n), \qquad \sum_{i=1}^n l_i = l, \qquad l_i \geq 0 \tag{3.3}$$

are continuous.

This definition can now be easily extended to a mapping $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ by requiring that all components f_i, $i = 1, \dots, n$ of $\mathbf{f}$ are smooth.

Definition 3.2 (Vector Field)**.** A (smooth) *vector field* is a prescription that assigns to each point $\mathbf{x} \in \mathbb{R}^n$ the pair $(\mathbf{x}, \mathbf{f}(\mathbf{x})) \in \mathbb{R}^n \times \mathbb{R}^n$ through a (smooth) mapping $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$.

Note that a vector field is *not* a mapping of the form $\mathbb{R}^n \to \mathbb{R}^n$. A vector field assigns a linear vector space $\mathbb{R}^n$ to each point $\mathbf{x}$ in $\mathbb{R}^n$, where the specific coordinate system is the image set of the mapping $\mathbf{f}(\mathbf{x})$. Often, the explicit indication of the first argument in

$(\mathbf{x}, \mathbf{f}(\mathbf{x}))$ is suppressed and simply written as $\mathbf{f}(\mathbf{x})$. However, if we have two vector fields $\mathbf{f}_1 : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{f}_2 : \mathbb{R}^n \to \mathbb{R}^n$, then they can only be added $\mathbf{f}_1(\mathbf{x}_1) + \mathbf{f}_2(\mathbf{x}_2)$ if $\mathbf{x}_1 = \mathbf{x}_2$, as otherwise $\mathbf{f}_1$ and $\mathbf{f}_2$ would lie in different vector spaces.

As an example, consider the electrostatic field of two fixed point charges q_1 and q_2 in three-dimensional space. If q_1 is located at position $\mathbf{x}_{q_1}^{\mathrm{T}} = [x_{q_1,1},\, x_{q_1,2},\, x_{q_1,3}]$, then to each point $\mathbf{x}^{\mathrm{T}} = [x_1,\, x_2,\, x_3]$ the field strength $\mathbf{E}_1(\mathbf{x})$ is assigned in the form

$$\mathbf{E}_1(\mathbf{x}) = \frac{q_1}{4\pi\varepsilon_0} \frac{(\mathbf{x} - \mathbf{x}_{q_1})}{\left((x_{q_1,1} - x_1)^2 + (x_{q_1,2} - x_2)^2 + (x_{q_1,3} - x_3)^2\right)^{3/2}} \tag{3.4}$$

Analogously, charge q_2 generates the field $\mathbf{E}_2$. Both vector fields can be superimposed, and one obtains the force on a test charge q at position $\mathbf{x}$ as

$$\mathbf{F} = q\mathbf{E}_1(\mathbf{x}) + q\mathbf{E}_2(\mathbf{x}) \ . \tag{3.5}$$

Note that the sum $q\mathbf{E}_1(\mathbf{x}_1) + q\mathbf{E}_2(\mathbf{x}_2)$ is not a meaningful operation for $\mathbf{x}_1 \neq \mathbf{x}_2$. Figure 3.1 illustrates this fact.

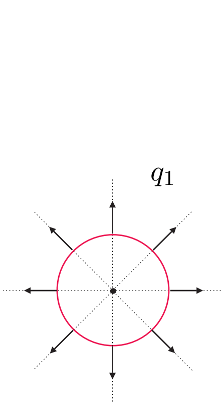


Figure 3.1: Illustration of the concept of a vector field using the example of the electric field of two point charges.

For second-order systems of the type (3.1), the solution trajectories can be easily obtained graphically by drawing the vector field $\mathbf{f}^{\mathrm{T}}(\mathbf{x}) = [f_1(x_1, x_2),\, f_2(x_1, x_2)]$. The reason for this is that for a solution curve of (3.1) passing through the point $\mathbf{x}^{\mathrm{T}} = [x_1,\, x_2]$, the vector field $\mathbf{f}(\mathbf{x})$ at point $\mathbf{x}$ is tangential to the solution curve.

Exercise 3.1. Draw the vector field for the system of differential equations

$$\dot{x}_1 = x_2 \tag{3.6a}$$
$$\dot{x}_2 = -\sin(x_1) - 1.5x_2 \ . \tag{3.6b}$$

Tip: Use MAPLE and the command `fieldplot` for this purpose.

3.1.2 Stability of the Equilibrium

These prerequisites allow us to define the stability of an equilibrium point in the sense of Lyapunov.

Definition 3.3 (Lyapunov Stability of Autonomous Systems)**.** The equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.1) is called *stable (in the sense of Lyapunov)* if for every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that

$$\|\mathbf{x}_0\| < \delta(\varepsilon) \quad \Rightarrow \quad \|\mathbf{\Phi}_t(\mathbf{x}_0)\| < \varepsilon \tag{3.7}$$

holds for all $t \geq 0$. Furthermore, the equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.1) is referred to as *attractive* if there exists a positive real number η such that

$$\|\mathbf{x}_0\| < \eta \quad \Rightarrow \quad \lim_{t \to \infty} \mathbf{\Phi}_t(\mathbf{x}_0) = \mathbf{0}\ . \tag{3.8}$$

If the equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.1) is *stable and attractive*, then it is also called *asymptotically stable.*

The choice of norms $\| \ \|$ in (3.7) and (3.8) is arbitrary, as shown in Section 2.1.1, where it is demonstrated that in a finite-dimensional vector space, norms are topologically equivalent. The distinction between stable and attractive in Definition 3.3 is important because an attractive equilibrium may not necessarily be stable. An example of this is given by the system

$$\dot{x}_1 = \frac{x_1^2(x_2 - x_1) + x_2^5}{(x_1^2 + x_2^2)\Big(1 + \big(x_1^2 + x_2^2\big)^2\Big)} \tag{3.9a}$$

$$\dot{x}_2 = \frac{x_2^2(x_2 - 2x_1)}{(x_1^2 + x_2^2)\Big(1 + \big(x_1^2 + x_2^2\big)^2\Big)} \tag{3.9b}$$

with the vector field shown in Figure 3.2.

3.1.3 Direct (Second) Method of Lyapunov

Before discussing the direct method of Lyapunov, the physical idea behind this method will be illustrated using the simple electrical system shown in Figure 3.3.

The network equations are

$$\frac{\mathrm{d}}{\mathrm{d}t} i_L = \frac{1}{L}(-u_C - R_1 i_L) \tag{3.10a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} u_C = \frac{1}{C}\left(i_L - \frac{u_C}{R_2}\right) \tag{3.10b}$$

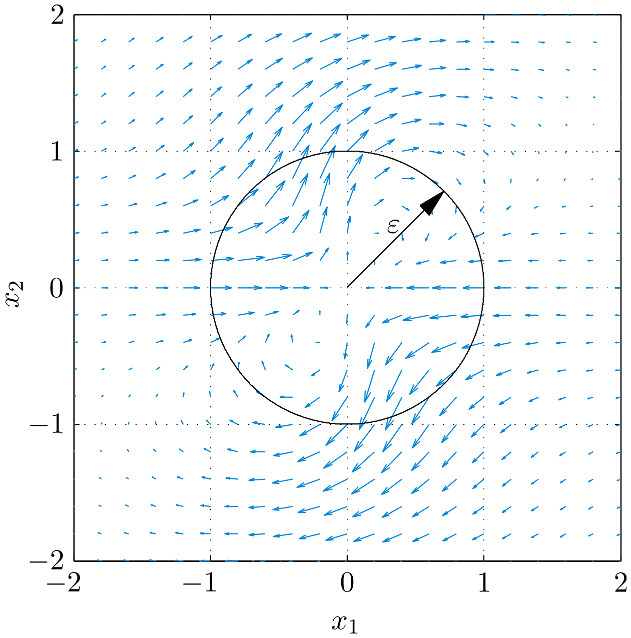


Figure 3.2: Vector field of an unstable but attractive point.

with the capacitor voltage u_C and the current through the inductance i_L. The energy stored in the capacitance C and inductance L

$$V = \frac{1}{2} L i_L^2 + \frac{1}{2} C u_C^2 \tag{3.11}$$

is positive for all $(u_C, i_L) \neq (0{,}0)$ and its time derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} V = -R_1 i_L^2 - \frac{1}{R_2} u_C^2 \tag{3.12}$$

is negative for all $(u_C, i_L) \neq (0{,}0)$. By introducing the norm

$$\left\| \begin{bmatrix} u_C \\ i_L \end{bmatrix} \right\| = \sqrt{C u_C^2 + L i_L^2} \tag{3.13}$$

it can be shown from Definition 3.3 for $\delta = \varepsilon$ that the equilibrium $u_C = i_L = 0$ is stable and attractive, hence asymptotically stable.

> *Exercise* 3.2. Show that (3.13) is a norm.

In the context of Lyapunov theory, for nonlinear systems of type (3.1), the energy function (3.11) is replaced by a function V with corresponding properties. For this purpose, the following definition is introduced:

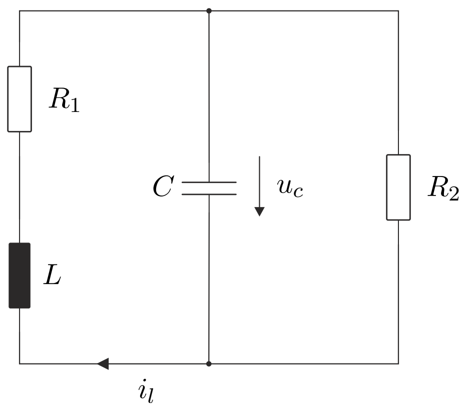


Figure 3.3: Simple electrical system.

Definition 3.4 (Positive/Negative (Semi-)Definiteness)**.** Let $\mathcal{D} \subseteq \mathbb{R}^n$ be an open neighborhood of $\mathbf{0}$. A function $V(\mathbf{x}) : \mathcal{D} \to \mathbb{R}$ is called *locally positive (negative) definite* if the following conditions are satisfied:

(1) $V(\mathbf{x})$ is continuously differentiable,

(2) $V(\mathbf{0}) = 0$, and

(3) $V(\mathbf{x}) > 0$, $(V(\mathbf{x}) < 0)$ for $\mathbf{x} \in \mathcal{D} - \{\mathbf{0}\}$.

If $\mathcal{D} = \mathbb{R}^n$ and there exists a constant $r > 0$ such that

$$\inf_{\|\mathbf{x}\|\geq r} V(\mathbf{x}) > 0 \quad \left(\sup_{\|\mathbf{x}\|\geq r} V(\mathbf{x}) < 0\right) , \tag{3.14}$$

then $V(\mathbf{x})$ is called *positive (negative) definite.*
If $V(\mathbf{x})$ in condition (3) satisfies only the following conditions:

(3) $V(\mathbf{x}) \geq 0$, $(V(\mathbf{x}) \leq 0)$ for $\mathbf{x} \in \mathcal{D} - \{\mathbf{0}\}$,

then $V(\mathbf{x})$ is called *(locally) positive (negative) semidefinite.*

Exercise 3.3. Which of the following functions are positive (negative) (semi)definite?

$$V(x_1,x_2,x_3) = x_1^2 + x_2^2 + 3x_3^4 \tag{3.15a}$$
$$V(x_1,x_2,x_3) = -x_1^2 - x_2^4 - ax_3^2 + x_3^4, \qquad a > 0 \tag{3.15b}$$
$$V(x_1,x_2,x_3) = (x_1 + x_2)^2 \tag{3.15c}$$
$$V(x_1,x_2,x_3) = x_1 - 2x_2 + x_3^2 \tag{3.15d}$$
$$V(x_1,x_2,x_3) = x_1^2 \exp\left(-x_1^2\right) + x_2^2 \tag{3.15e}$$

In analogy to the electrical example in Figure 3.3, one now tries to construct a positive definite function $V(\mathbf{x})$ (corresponding to the energy function), the so-called *Lyapunov function*, whose time derivative is negative definite. For the temporal change of $V(\mathbf{x})$ along a trajectory $\mathbf{\Phi}_t(\mathbf{x}_0)$ of (3.1), the following holds:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V(\mathbf{\Phi}_t(\mathbf{x}_0)) &= \frac{\partial}{\partial \mathbf{x}}V(\mathbf{\Phi}_t(\mathbf{x}_0))\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Phi}_t(\mathbf{x}_0) \\ &= \frac{\partial}{\partial \mathbf{x}}V(\mathbf{x})\mathbf{f}(\mathbf{x}) \ .\end{aligned} \tag{3.16}$$

Figure 3.4 illustrates this fact using the *level sets* $V(\mathbf{x}) = c$ for various positive constants c.

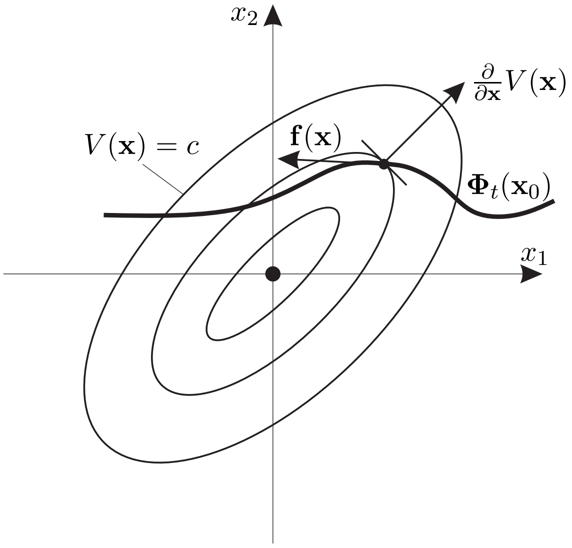

x_2
$\frac{\partial}{\partial \mathbf{x}}V(\mathbf{x})$
$V(\mathbf{x}) = c$
$\mathbf{f}(\mathbf{x})$
$\mathbf{\Phi}_t(\mathbf{x}_0)$
x_1

Figure 3.4: Constructing a Lyapunov function.

Exercise 3.4. Show that for second-order systems, the level sets near the equilibrium point are always ellipses. (This also justifies the choice of the schematic representation in Figure 3.4.)

Now we are able to formulate Lyapunov's direct method:

Theorem 3.1 (Lyapunov's Direct Method). *Let $\mathbf{x}_R = \mathbf{0}$ be an equilibrium point of (3.1) and $\mathcal{D} \subseteq \mathbb{R}^n$ be an open neighborhood of $\mathbf{0}$. If there exists a function $V(\mathbf{x}) : \mathcal{D} \to \mathbb{R}$ such that $V(\mathbf{x})$ is* positive definite *on $\mathcal{D}$ and $\dot{V}(\mathbf{x})$ is* negative semidefinite *on $\mathcal{D}$, then the equilibrium point $\mathbf{x}_R = \mathbf{0}$ is* stable. *If $\dot{V}(\mathbf{x})$ is even* negative definite*, then the equilibrium point $\mathbf{x}_R = \mathbf{0}$ is* asymptotically stable. *The function $V(\mathbf{x})$ is then called a* Lyapunov function.

The proof of this theorem is not provided here but can be found in the literature referenced at the end. It should be noted at this point that using the level sets of Figure 3.4 can help illustrate the statement of Theorem 3.1.

Exercise 3.5. Consider an RLC network described by the following system of differential equations:

$$\begin{bmatrix} \dot{\mathbf{x}}_C \\ \dot{\mathbf{x}}_L \end{bmatrix} = \begin{bmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{L} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_C \\ \mathbf{x}_L \end{bmatrix} \tag{3.17}$$

Here, $\mathbf{x}_C$ denotes the vector of capacitor voltages and $\mathbf{x}_L$ denotes the vector of inductance currents. The diagonal matrix $\mathbf{C}$ contains all capacitor values, and the positive definite matrix $\mathbf{L}$ consists of self and mutual inductances. The matrices $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are symmetric, and $\mathbf{R}_{12} = -\mathbf{R}_{21}^{\mathrm{T}}$. Show that for negative definite matrices $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$, the equilibrium point $\mathbf{x}_C = \mathbf{x}_L = \mathbf{0}$ is asymptotically stable.

Tip: Use as a Lyapunov function the total energy stored in the energy storage elements: $V(\mathbf{x}_C, \mathbf{x}_L) = \frac{1}{2}\mathbf{x}_C^{\mathrm{T}}\mathbf{C}\mathbf{x}_C + \frac{1}{2}\mathbf{x}_L^{\mathrm{T}}\mathbf{L}\mathbf{x}_L$.

Note that the failure of a candidate for $V(\mathbf{x})$ does *not* imply the instability of the equilibrium point. In such a case, a different function $V(\mathbf{x})$ must be chosen. However, the existence of a Lyapunov function is always guaranteed if the equilibrium point is stable in the Lyapunov sense, i.e., the main challenge is to find a suitable Lyapunov function $V(\mathbf{x})$. In most technical-physical applications, the Lyapunov function can be obtained from *physical considerations* by considering the stored energy in the system as a suitable candidate. If this is not possible, for example, if the physical structure is partially destroyed by control, then other methods must be used accordingly.

In the case of a scalar system of the form

$$\dot{x} = -f(x) \tag{3.18}$$

with continuous $f(x)$, $f(0) = 0$, and $xf(x) > 0$ for all $x \neq 0$ with $x \in (-a, a)$, one chooses candidates for the Lyapunov function as

$$V(x) = \int\limits_0^x f(z) dz \ . \tag{3.19}$$

Obviously, $V(\mathbf{x})$ is positive definite on the interval $(-a, a)$ and for the time derivative of

$V(\mathbf{x})$ we have

$$\dot{V}(x) = f(x)(-f(x)) = -f^2(x) < 0 \tag{3.20}$$

for all $x \neq 0$ with $x \in (-a, a)$. This proves the asymptotic stability of the equilibrium $x_R = 0$.

> *Exercise* 3.6. Show that a single-input system with an asymptotically stable equilibrium $x_R = 0$ can always be written in the form of (3.18) in a sufficiently small neighborhood $\mathcal{D} = \{x \in \mathbb{R} | -a < x < a\}$ around the equilibrium, with the condition $xf(x) > 0$ for all $x \in \mathcal{D} - \{0\}$.

3.1.4 Basin of Attraction

Although stability of an equilibrium can be assessed using the above methods, the allowed deviation $\mathbf{x}_0$ from the equilibrium $\mathbf{0}$ is only known to be sufficiently small. To quantitatively classify these possible deviations, the so-called basin of attraction is defined.

Definition 3.5 (Basin of Attraction)**.** Let $\mathbf{x}_R = \mathbf{0}$ be an asymptotically stable equilibrium of (3.1). Then the set

$$\mathcal{E} = \left\{ \mathbf{x}_0 \in \mathbb{R}^n | \lim_{t \to \infty} \mathbf{\Phi}_t(\mathbf{x}_0) = \mathbf{0} \right\} \tag{3.21}$$

is called the *basin of attraction* of $\mathbf{x}_R = \mathbf{0}$. If $\mathcal{E} = \mathbb{R}^n$, then the equilibrium $\mathbf{x}_R = \mathbf{0}$ is *globally asymptotically stable.*

If one can show that the Lyapunov function $V(\mathbf{x})$ is positive definite on a domain $\mathcal{X}$ and $\dot{V}(\mathbf{x})$ is negative definite on a domain $\mathcal{Y}$, where the domains $\mathcal{X}$ and $\mathcal{Y}$ include the equilibrium $\mathbf{x}_R = \mathbf{0}$, then a simple estimation of the basin of attraction is given by the largest *level set*

$$\mathcal{L}_c = \{\mathbf{x} \in \mathbb{R}^n | \ V(\mathbf{x}) \leq c\} \tag{3.22}$$

for which $\mathcal{L}_c \subset \mathcal{X} \cap \mathcal{Y}$.

> *Exercise* 3.7. Show that $\mathcal{L}_c \subset \mathcal{X} \cap \mathcal{Y}$ being a positively invariant set according to Definition 3.6. Provide a justification for why this is indeed a suitable estimation of the basin of attraction.

When proving global asymptotic stability, fundamental difficulties arise as for large c, the level sets (3.22) may no longer be *closed and bounded* (*compact*). If this property is lost, the level sets are no longer positively invariant sets and hence not suitable estimates for the basin of attraction. An example of this is given by the Lyapunov function

$$V(\mathbf{x}) = \frac{x_1^2}{(1 + x_1^2)} + x_2^2 \tag{3.23}$$

As can be seen from Figure 3.5, the level sets $\mathcal{L}_c$ are compact for small c, which directly follows from the fact that $V(\mathbf{x})$ is positive definite. In order for the level

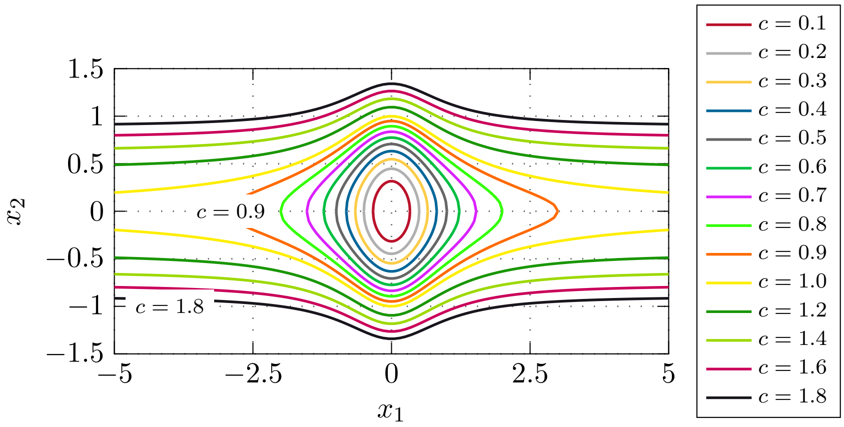


Figure 3.5: Regarding the compactness of level sets.

sets $\mathcal{L}_c$ to be completely contained in a region $\mathcal{B}_r = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x}\| < r\}$, the condition $c < \min_{\|\mathbf{x}\|=r} V(\mathbf{x}) < \infty$ must be satisfied, i.e., if

$$l = \lim_{r\to\infty} \min_{\|\mathbf{x}\|=r} V(\mathbf{x}) < \infty \ , \tag{3.24}$$

then the level sets $\mathcal{L}_c$ for $c < l$ are compact. For the Lyapunov function (3.23), it follows that

$$\begin{aligned} l &= \lim_{r\to\infty} \min_{\|\mathbf{x}\|=r} \left(\frac{x_1^2}{(1+x_1^2)} + x_2^2 \right) \\ &= \lim_{|x_1|\to\infty} \frac{x_1^2}{(1+x_1^2)} \\ &= 1 \ , \end{aligned} \tag{3.25}$$

which means that the level sets are compact only for $c < 1$. To ensure that the level sets $\mathcal{L}_c$ are compact for all $c > 0$, the additional requirement

$$\lim_{\|\mathbf{x}\|\to\infty} V(\mathbf{x}) = \infty \tag{3.26}$$

is established. A function that satisfies this condition is called *radially unbounded*. This leads to the following theorem.

Theorem 3.2 (Global asymptotic stability)**.** *Let $\mathbf{x}_R = \mathbf{0}$ be an equilibrium point of (3.1). If there exists a function $V(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ such that $V(\mathbf{x})$ is* positive definite*, $\dot{V}(\mathbf{x})$ is* negative definite*, and $V(\mathbf{x})$ is* radially unbounded*, then the equilibrium point $\mathbf{x}_R = \mathbf{0}$ is* globally asymptotically stable*.*

Again, for the detailed proof, one should refer to the literature.

Consider the dynamic system shown in Figure 3.6 with T_1, $T_2 > 0$, and the saturation characteristic

$$F(x_1) = \begin{cases} -1 & \text{for } x_1 \leq -1 \\ x_1 & \text{for } -1 < x_1 < 1 \\ 1 & \text{for } x_1 \geq 1 \end{cases} \tag{3.27}$$

or

$$\frac{x_1}{F(x_1)} = \begin{cases} -x_1 & \text{for } x_1 \leq -1 \\ 1 & \text{for } -1 < x_1 < 1 \\ x_1 & \text{for } x_1 \geq 1 \; . \end{cases} \tag{3.28}$$

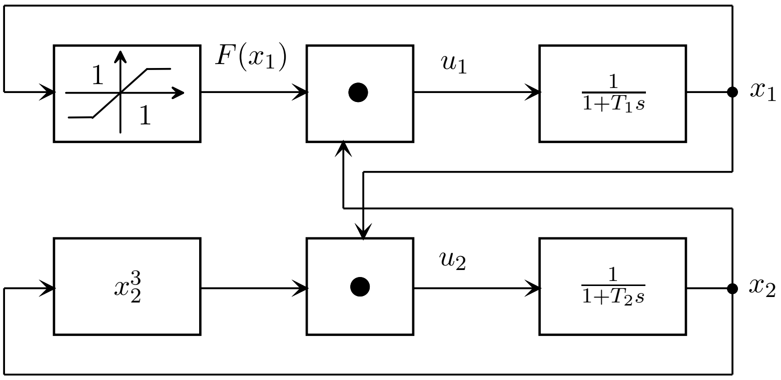


Figure 3.6: Block diagram of the analyzed dynamic system.

The corresponding mathematical model is

$$\dot{x}_1 = \frac{1}{T_1}(F(x_1)x_2 - x_1) \tag{3.29a}$$

$$\dot{x}_2 = \frac{1}{T_2}\Big(x_2^3 x_1 - x_2\Big) \; . \tag{3.29b}$$

Now, if we choose candidates for the Lyapunov function as

$$V(\mathbf{x}) = a^2 x_1^2 + b^2 x_2^2, \qquad a, b \neq 0 \; , \tag{3.30}$$

then we obtain the expression for $\dot{V}(\mathbf{x})$ as

$$\dot{V}(\mathbf{x}) = x_1^2 \frac{2a^2}{T_1}\left(\frac{F(x_1)}{x_1}x_2 - 1\right) + x_2^2 \frac{2b^2}{T_2}\Big(x_2^2 x_1 - 1\Big) \; . \tag{3.31}$$

Obviously, $\dot{V}(\mathbf{x})$ is negative definite for

$$x_2 < \frac{x_1}{F(x_1)} \quad \text{and} \quad x_1 < \frac{1}{x_2^2} \tag{3.32}$$

To estimate the basin of attraction, a level set $\mathcal{L}_c = \left\{\mathbf{x} \in \mathbb{R}^2 \middle| V(\mathbf{x}) \leq c\right\}$ is sought where $\dot{V}(\mathbf{x})$ is negative definite. For this purpose, we determine the ellipse $V(\mathbf{x}) = a^2 x_1^2 + b^2 x_2^2 = \left(\sqrt{c}\right)^2$, which touches the curves (3.32). The point of tangency between the ellipse

$$\frac{x_1^2}{\left(\sqrt{c}/a\right)^2} + \frac{x_2^2}{\left(\sqrt{c}/b\right)^2} = 1 \tag{3.33}$$

and the saturation characteristic $x_2 = \frac{x_1}{F(x_1)}$ immediately yields the relationship $\sqrt{c}/b = 1$. To determine the second point of tangency, we use the fact that at the point of tangency of the two curves

$$\frac{x_1^2}{\left(\sqrt{c}/a\right)^2} + x_2^2 = 1 \quad \text{and} \quad x_1 = \frac{1}{x_2^2} \tag{3.34}$$

the slopes

$$\frac{2x_1 \, \mathrm{d}x_1}{\left(\sqrt{c}/a\right)^2} + 2x_2 \, \mathrm{d}x_2 = 0 \quad \text{and} \quad \mathrm{d}x_1 = \frac{-2 \, \mathrm{d}x_2}{x_2^3} \tag{3.35}$$

and

$$\frac{\mathrm{d}x_2}{\mathrm{d}x_1} = \frac{-x_1}{x_2\left(\sqrt{c}/a\right)^2} \quad \text{and} \quad \frac{\mathrm{d}x_2}{\mathrm{d}x_1} = \frac{-x_2^3}{2} \tag{3.36}$$

must be equal. From (3.34) and (3.36) it follows that

$$\frac{-x_1}{\left(\sqrt{c}/a\right)^2} = \frac{-x_2^4}{2} \quad \text{and} \quad x_2^4 = \frac{1}{x_1^2} \tag{3.37}$$

and thus

$$x_1^3 = \frac{\left(\sqrt{c}/a\right)^2}{2} \; . \tag{3.38}$$

Substituting (3.38) into (3.34), we obtain

$$\sqrt{c}/a = \frac{3\sqrt{3}}{2} \; . \tag{3.39}$$

Thus, an estimation of the basin of attraction is calculated as the interior of the ellipse

$$\frac{x_1^2}{\frac{27}{4}} + x_2^2 = 1 \; . \tag{3.40}$$

Figure 3.7 shows the graphical representation of the situation.

Exercise 3.8. The following dynamic system is given

$$\dot{x}_1 = \frac{-6x_1}{u^2} + 2x_2, \qquad u = 1 + x_1^2 \tag{3.41a}$$

$$\dot{x}_2 = \frac{-2(x_1 + x_2)}{u^2} \; . \tag{3.41b}$$

(1) Calculate the equilibrium(s) of the system (3.41). Show that for all $\mathbf{x} \in \mathbb{R}^2$,

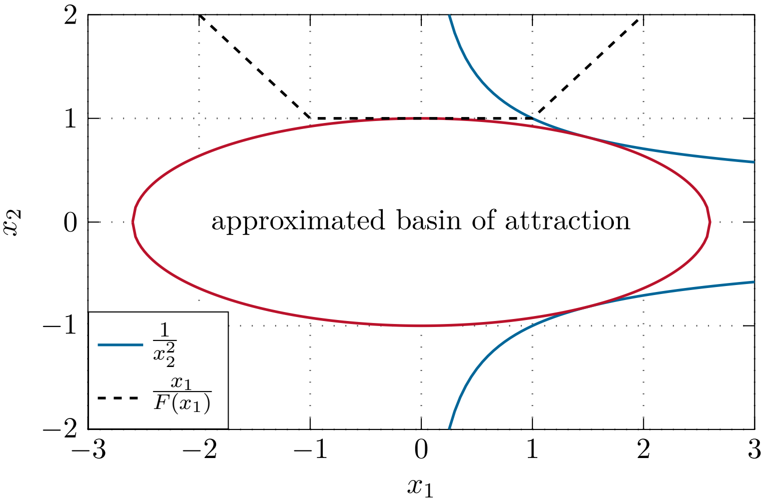


Figure 3.7: Calculation of the basin of attraction of Figure 3.6.

$V(\mathbf{x}) > 0$ and $\dot{V}(\mathbf{x}) < 0$ for

$$V(\mathbf{x}) = \frac{x_1^2}{1+x_1^2} + x_2^2 \ . \tag{3.42}$$

(2) Are the equilibrium(s) stable, asymptotically stable, globally stable, or globally asymptotically stable?

Exercise 3.9. The following dynamic system is given:

$$\dot{x}_1 = -x_1 + 2x_1^3 x_2 \tag{3.43a}$$

$$\dot{x}_2 = -x_2 \ . \tag{3.43b}$$

(1) Show that the equilibrium $\mathbf{x}_R = \mathbf{0}$ is asymptotically stable.

(2) Provide the largest possible estimate of the basin of attraction.

3.1.5 The Invariance Principle

Expanding on Theorem 3.1, there are systems where the equilibrium $\mathbf{x}_R = \mathbf{0}$ is asymptotically stable even though the time derivative of the Lyapunov function $\dot{V}(\mathbf{x})$ is only negative semidefinite. As an example, consider the simple spring-mass-damper system shown in Figure 3.8 with mass m, linear damping force $F_d = d\frac{\mathrm{d}}{\mathrm{d}t}z$, $d > 0$, and nonlinear spring force $F_c = \psi_F(z)$ satisfying $k_1 z^2 \leq \psi_F(z)z \leq k_2 z^2$ with $0 < k_1 < k_2$.

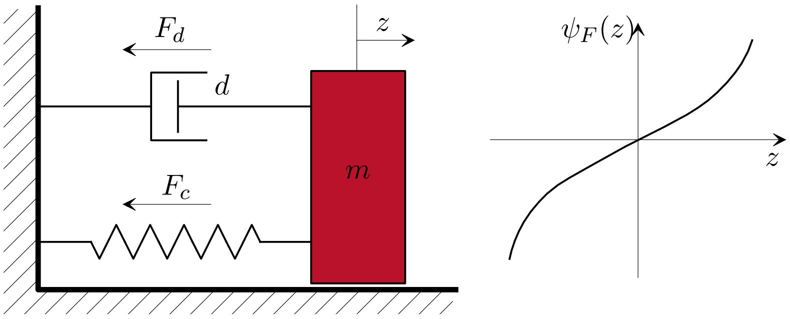


Figure 3.8: Simple mechanical system.

The equations of motion are

$$\frac{\mathrm{d}}{\mathrm{d}t} z = v \tag{3.44a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} v = -\frac{1}{m}(\psi_F(z) + dv) \tag{3.44b}$$

with the state $\mathbf{x}^{\mathrm{T}} = [z, v]$ and the only equilibrium $\mathbf{x}_R = \mathbf{0}$. The kinetic and potential energy stored in the system

$$V = \frac{1}{2}mv^2 + \int_0^z \psi_F(w)\,\mathrm{d}w \tag{3.45}$$

are naturally positive definite and serve as suitable candidates for a Lyapunov function. Clearly,

$$\frac{\mathrm{d}}{\mathrm{d}t} V = mv\Big(-\frac{1}{m}(\psi_F(z) + dv)\Big) + \psi_F(z)v = -dv^2 \tag{3.46}$$

is negative semidefinite, and according to Theorem 3.1, we can conclude that the equilibrium $\mathbf{x}_R = \mathbf{0}$ is stable in the sense of Lyapunov. That is, the energy V stored in the system always decreases, except when $v = 0$ where it remains constant. Substituting $v = 0$

into (3.44), we see that $z = \bar{z}$ and $\frac{\mathrm{d}}{\mathrm{d}t}v = -\frac{1}{m}\psi_F(\bar{z})$ for a constant $\bar{z}$. From the specific form of the characteristic curve $\psi_F(z)$ in Figure 3.8, it follows that $\frac{\mathrm{d}}{\mathrm{d}t}v$ only becomes zero for $\bar{z} = 0$. This demonstrates that the energy V stored in the system must decrease until the point $z = v = 0$ is reached, proving the asymptotic stability of the equilibrium.

The mathematical generalization of this procedure leads to the so-called Invariance Principle of Krassovskii-LaSalle. Before this is discussed in more detail, the concepts of limit points and limit sets should be explained. Without loss of generality, consider again the autonomous, smooth nth-order system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{3.47}$$

with the flow $\mathbf{\Phi}_t(\mathbf{x})$ according to (3.1).

Definition 3.6 (Positively Invariant Set)**.** A set $M \subset \mathbb{R}^n$ is called a *positively invariant set* of the system (3.47) if the image of set M under the flow $\mathbf{\Phi}_t$ is the set M itself, i.e., $\mathbf{\Phi}_t(M) \subseteq M$, for all $t > 0$.

Simple examples of a positively invariant set are the set $\{\mathbf{x}_R\}$ with $\mathbf{x}_R$ as an equilibrium point, the set of points of a limit cycle, etc. A set M is called a *negatively invariant set* of the system (3.47) if $\mathbf{\Phi}_{-t}(M)$ is positively invariant. Also of interest are points that are approached arbitrarily closely by a trajectory an infinite number of times. For this, the following definition is given:

Definition 3.7 (Limit Point and Limit Set)**.** A point $\mathbf{y} \in \mathbb{R}^n$ is called an *ω-limit point* of $\mathbf{x}$ of the system (3.47) if there exists a sequence (t_i) of real numbers from the interval $[0, \infty)$ with $t_i \to \infty$ such that

$$\lim_{i\to\infty} \|\mathbf{y} - \mathbf{\Phi}_{t_i}(\mathbf{x})\| = 0 \tag{3.48}$$

holds. The set of all *ω-limit points* of $\mathbf{x}$, the so-called *ω-limit set* of $\mathbf{x}$, is denoted by $L_\omega(\mathbf{x})$.

Equivalently to the above definition, limit points and limit sets can be considered for $t < 0$. In this case, the designations α-limit point and α-limit set $L_\alpha(\mathbf{x})$ are used.

Definition 3.8 (Limit Cycle)**.** A *limit cycle* of (3.47) is a *closed trajectory* γ that satisfies the conditions $\gamma \subset L_\omega(\mathbf{x})$ or $\gamma \subset L_\alpha(\mathbf{x})$ for certain $\mathbf{x} \in \mathbb{R}^n$. In the first case, the limit cycle is called an *ω-limit cycle*, and in the second case, an *α-limit cycle*.

In Figure 3.9, the concepts of limit set and limit cycle are illustrated based on a schematic representation of the trajectories of the Van der Pol oscillator. Here, γ describes the unique closed trajectory that, for every point $\mathbf{x} \in \mathbb{R}^2$ except for the point $\mathbf{x}_A$, forms the ω-limit set $L_\omega(\mathbf{x})$, i.e., γ describes an ω-limit cycle. Furthermore, the point $\mathbf{x}_A$ is the α-limit set $L_\alpha(\mathbf{x})$ for every point $\mathbf{x}$ inside γ. If $\mathbf{x}$ is outside γ, then $L_\alpha(\mathbf{x}) = \{\}$.

With these concepts, it is now possible to formulate the invariance principle of Krassovskii-LaSalle.

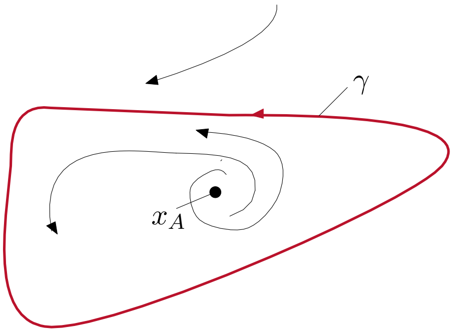


Figure 3.9: Limit points and limit sets.

Theorem 3.3 (Auxiliary lemma for the invariance theorem)**.** *If the solution* $\mathbf{x}(t) = \mathbf{\Phi}_t(\mathbf{x}_0)$ *of the system (3.1) is bounded for* $t \geq 0$*, then the* ω*-limit set* $L_\omega(\mathbf{x}_0)$ *of* $\mathbf{x}_0$ *according to Definition 3.7 is a nonempty, compact (bounded and closed), positively invariant set with the property*

$$\lim_{t\to\infty} \mathbf{\Phi}_t(\mathbf{x}_0) \in L_\omega(\mathbf{x}_0) \ . \tag{3.49}$$

The proof of this theorem can be found in the literature cited at the end.

Theorem 3.4 (Invariance principle of Krassovskii-LaSalle)**.** *Assume* $\mathcal{X}$ *is a compact, positively invariant set and* $V : \mathcal{X} \to \mathbb{R}$ *is a continuously differentiable function that satisfies* $\dot{V}(\mathbf{x}) \leq 0$ *on* $\mathcal{X}$*. Let* $\mathcal{Y}$ *be the subset of* $\mathcal{X}$ *for which* $\mathcal{Y} = \left\{\mathbf{x} \in \mathcal{X} | \dot{V}(\mathbf{x}) = 0\right\}$*. If* $\mathcal{M}$ *denotes the largest positively invariant set of* $\mathcal{Y}$*, then*

$$L_\omega(\mathcal{X}) \subseteq \mathcal{M} \ . \tag{3.50}$$

The proof of this theorem can also be found in the literature cited at the end. As seen from Theorem 3.4, $V(\mathbf{x})$ does not need to be positive definite. The difficulty here lies in finding the compact, positively invariant set $\mathcal{X}$. However, it is known from Section 3.1.4 that the level set of a positive definite function $V(\mathbf{x})$ is locally compact and positively invariant. If radial unboundedness can be proven, then this holds globally. Thus, it is possible to formulate the following theorem as a direct consequence of Theorem 3.4.

Theorem 3.5 (Application of the Invariance Theorem)**.** *Let* $\mathbf{x}_R = \mathbf{0}$ *be an equilibrium point of (3.1) and* $\mathcal{D} \subseteq \mathbb{R}^n$ *be an open neighborhood of* $\mathbf{0}$*. If there exists a function* $V(\mathbf{x}) : \mathcal{D} \to \mathbb{R}$ *such that* $V(\mathbf{x})$ *is* positive definite *on* $\mathcal{D}$ *and* $\dot{V}(\mathbf{x})$ *is* negative semidefinite *on* $\mathcal{D}$*, then the point* $\mathbf{x}_R = \mathbf{0}$ *is* asymptotically stable *if the largest positively invariant subset of* $\mathcal{Y} = \left\{\mathbf{x} \in \mathcal{D} | \dot{V}(\mathbf{x}) = 0\right\}$ *is the set* $\mathcal{M} = \{\mathbf{0}\}$*. Furthermore, if* $V(\mathbf{x})$ *is radially unbounded, then* $\mathbf{x}_R = \mathbf{0}$ *is* globally asymptotically stable*.*

Referring to the spring-mass-damper system in Figure 3.8, consider the example

$$\dot{x}_1 = x_2 \tag{3.51a}$$
$$\dot{x}_2 = -g(x_1) - h(x_2) \tag{3.51b}$$

with

$$g(0) = 0, \qquad x_1 g(x_1) > 0 \text{ for } x_1 \neq 0, \qquad x_1 \in (-a, a) \tag{3.52}$$
$$h(0) = 0, \qquad x_2 h(x_2) > 0 \text{ for } x_2 \neq 0, \qquad x_2 \in (-a, a) \tag{3.53}$$

being examined. It is assumed that $g(x_1)$ and $h(x_2)$ are continuous on the interval $(-a, a)$. It can be easily verified that $\mathbf{x}_R = \mathbf{0}$ in the set $\mathcal{D} = \left\{\mathbf{x} \in \mathbb{R}^2 \middle| -a < x_1 < a, -a < x_2 < a\right\}$ is the only equilibrium point. A candidate for a Lyapunov function is chosen as

$$V(\mathbf{x}) = \int_0^{x_1} g(x)\,\mathrm{d}x + \frac{x_2^2}{2} \tag{3.54}$$

Clearly, $V(\mathbf{x})$ is positive definite on $\mathcal{D}$ and for $\dot{V}$ we have

$$\dot{V}(\mathbf{x}) = g(x_1)\dot{x}_1 + x_2\dot{x}_2 = -x_2 h(x_2) \leq 0 \ . \tag{3.55}$$

In this example, the set $\mathcal{Y} = \left\{\mathbf{x} \in \mathcal{D} | \dot{V}(\mathbf{x}) = 0\right\}$ simplifies to $\mathcal{Y} = \{\mathbf{x} \in \mathcal{D} | x_1 \text{ arbitrary and } x_2 = 0\}$. Therefore, for the solution curves to remain in $\mathcal{Y}$ for all times $t \geq 0$, it follows immediately that $x_1 = 0$, meaning the largest positively invariant subset of $\mathcal{Y}$ is the set $\mathcal{M} = \{\mathbf{0}\}$. Hence, according to Theorem 3.5, the equilibrium point $\mathbf{x}_R = \mathbf{0}$ is asymptotically stable.

Exercise 3.10. Given is a first-order dynamic system

$$\dot{x}_1 = a x_1 + u \tag{3.56}$$

with an adaptive control law

$$\dot{x}_2 = \gamma x_1^2, \qquad \gamma > 0 \tag{3.57a}$$
$$u = -x_2 x_1 \ . \tag{3.57b}$$

Show using the invariance principle of Krassovskii-LaSalle that for the closed loop system, $\lim_{t\to\infty} x_1(t) = 0$ regardless of the plant parameter a. It is only known that the parameter a is bounded from above by $a < b$.

Tip: Choose as a candidate for the Lyapunov function

$$V(\mathbf{x}) = \frac{1}{2}x_1^2 + \frac{1}{2\gamma}(x_2 - b)^2, \qquad b > a \ . \tag{3.58}$$

3.1.6 Linear Systems

The stability analysis of linear systems

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{3.59}$$

can be carried out based on the eigenvalues of the matrix $\mathbf{A}$. By means of a regular state transformation $\mathbf{z} = \mathbf{T}\mathbf{x}$, the system can be transformed to *Jordan normal form*

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z} \tag{3.60}$$

with

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_n \end{bmatrix} \tag{3.61}$$

A Jordan block $\mathbf{J}_i$ has the form

$$\mathbf{J}_i = \begin{bmatrix} a_i & 1 & 0 & \cdots & 0 \\ 0 & a_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & a_i & 1 \\ 0 & \cdots & \cdots & 0 & a_i \end{bmatrix}_{m\times m} \tag{3.62}$$

for an m-fold real eigenvalue $\lambda_i = a_i$ of the matrix $\mathbf{A}$ or

$$\mathbf{J}_i = \begin{bmatrix} \mathbf{A}_i & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & & \ddots & \mathbf{A}_i & \mathbf{I} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{A}_i \end{bmatrix}_{2m\times 2m} , \qquad \mathbf{A}_i = \begin{bmatrix} a_i & -b_i \\ b_i & a_i \end{bmatrix} \tag{3.63}$$

for an m-fold complex conjugate eigenvalue $\lambda_i = a_i \pm jb_i$ of the matrix $\mathbf{A}$.

> *Exercise* 3.11. How should the transformation matrix $\mathbf{T}$ look like in order to obtain the Jordan form?
>
> **Tip:** Eigenvectors

Now, the following theorem holds for stability according to Lyapunov:

Theorem 3.6 (Stability of Linear Systems). *The equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.59) is stable in the sense of Lyapunov if and only if for each Jordan block $\mathbf{J}_i$ of (3.60), $a_i < 0$ or $a_i \leq 0$ and $m = 1$. If $a_i < 0$ holds for each Jordan block $\mathbf{J}_i$ of (3.60), then the equilibrium $\mathbf{x}_R = \mathbf{0}$ is asymptotically stable.*

Exercise 3.12. Prove Theorem 3.6.

Two more definitions are needed for the subsequent considerations.

Definition 3.9 (Hurwitz Matrix). An $(n \times n)$ matrix $\mathbf{A}$ is called a *Hurwitz matrix* if for all eigenvalues λ_i of $\mathbf{A}$, $\mathrm{Re}(\lambda_i) < 0$ for $i = 1, \ldots, n$.

Definition 3.10 (Positive Definite Matrix). A symmetric $(n \times n)$ matrix $\mathbf{P}$ is called *positive definite* if $\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{0}\}$. If $\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x} \geq 0$, then $\mathbf{P}$ is called *positive semidefinite.*

Exercise 3.13. Where are the eigenvalues of a positive (semi)definite matrix located? Prove your statements.

Now, if we choose candidates for a Lyapunov function of (3.59) as

$$V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x} \tag{3.64}$$

with a positive definite matrix $\mathbf{P}$, then for $\dot{V}$ we have

$$\begin{aligned} \dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}^{\mathrm{T}}\mathbf{P}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{P}\dot{\mathbf{x}} \\ &= \mathbf{x}^{\mathrm{T}}\Big(\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A}\Big)\mathbf{x} \\ &= -\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} \end{aligned} \tag{3.65}$$

with a square matrix $\mathbf{Q}$ that satisfies the relationship

$$\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{Q} = \mathbf{0} \tag{3.66}$$

(3.66) is also called the *Lyapunov equation.*

Exercise 3.14. Show that the Lyapunov equation (3.66) is a linear equation in the elements p_{ij} of $\mathbf{P}$.

If the matrix $\mathbf{Q}$ is positive definite, then from Theorem 3.1, it follows that the equilibrium $\mathbf{x}_R = \mathbf{0}$ is asymptotically stable and consequently $\mathbf{A}$ is a Hurwitz matrix. That is, for a given positive definite matrix $\mathbf{P}$, the matrix $\mathbf{Q}$ is computed for system (3.59) and checked for positive definiteness. For linear systems, this procedure can be reversed. A positive definite $\mathbf{Q}$ is specified, and $\mathbf{P}$ is computed accordingly. The following theorem states:

Theorem 3.7 (Lyapunov Equation)**.** *The matrix* $\mathbf{A}$ *is a* Hurwitz matrix *if and only if the* Lyapunov equation *(3.66) has a* positive definite *solution* $\mathbf{P}$ *for* every positive definite $\mathbf{Q}$. *In this case,* $\mathbf{P}$ *is uniquely determined.*

Proof. (⇐): Follows trivially from Theorem 3.1. (⇒): If $\mathbf{A}$ is a Hurwitz matrix, then the existence of the integral

$$\mathbf{P} = \int_0^\infty \mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t}\,\mathrm{d}t \tag{3.67}$$

is guaranteed. Furthermore, if $\mathbf{Q}$ is positive definite, then this must also hold for $\mathbf{P}$, because from

$$\mathbf{x}^\mathrm{T}\mathbf{P}\mathbf{x} = 0 \tag{3.68}$$

it follows

$$\int_0^\infty \underbrace{\mathbf{x}^\mathrm{T}\mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t}\mathbf{x}}_{>0}\,\mathrm{d}t = 0\ . \tag{3.69}$$

Since $\mathbf{Q}$ is positive definite, $\mathrm{e}^{\mathbf{A}t}\mathbf{x} = \mathbf{0}$ and due to the regularity of the transition matrix, $\mathbf{x} = \mathbf{0}$. The calculation

$$\begin{aligned}\mathbf{A}^\mathrm{T}\mathbf{P} + \mathbf{P}\mathbf{A} &= \int_0^\infty \mathbf{A}^\mathrm{T}\mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t}\,\mathrm{d}t + \int_0^\infty \mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t}\mathbf{A}\,\mathrm{d}t \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t}\right)\mathrm{d}t \\ &= \lim_{t\to\infty} \mathrm{e}^{\mathbf{A}^\mathrm{T} t}\mathbf{Q}\mathrm{e}^{\mathbf{A}t} - \mathbf{Q} \\ &= -\mathbf{Q}\end{aligned} \tag{3.70}$$

shows that $\mathbf{P}$ from (3.67) is indeed a solution of the Lyapunov equation (3.66). The uniqueness of the solution remains to be shown. Assuming $\mathbf{P}_0$ is another solution of the Lyapunov equation (3.66). For the time derivative of the expression

$$\mathbf{F}(\mathbf{X}) = \mathbf{X}^\mathrm{T}\mathbf{P}\mathbf{X} - \mathbf{X}^\mathrm{T}\mathbf{P}_0\mathbf{X} = \mathbf{X}^\mathrm{T}(\mathbf{P} - \mathbf{P}_0)\mathbf{X} \tag{3.71}$$

with $\mathbf{X}$ as a solution of the matrix differential equation

$$\dot{\mathbf{X}} = \mathbf{A}\mathbf{X} \tag{3.72}$$

we obtain

$$\dot{\mathbf{F}}(\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\left(\underbrace{\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A}}_{-\mathbf{Q}} - \underbrace{\left(\mathbf{A}^{\mathrm{T}}\mathbf{P}_0 + \mathbf{P}_0\mathbf{A}\right)}_{-\mathbf{Q}}\right)\mathbf{X} = \mathbf{0}\ . \tag{3.73}$$

Thus, $\mathbf{F}(\mathbf{X})$ is constant along a trajectory of (3.59). From

$$\mathbf{F}\left(\mathrm{e}^{\mathbf{A}t}\right) = \mathrm{e}^{\mathbf{A}^{\mathrm{T}}t}(\mathbf{P} - \mathbf{P}_0)\mathrm{e}^{\mathbf{A}t} \tag{3.74}$$

we then deduce, with

$$\begin{aligned}\lim_{t\to 0}\mathbf{F}\left(\mathrm{e}^{\mathbf{A}t}\right) &= \mathbf{F}(\mathbf{I})\\ &= (\mathbf{P} - \mathbf{P}_0)\\ &= \lim_{t\to+\infty}\mathbf{F}\left(\mathrm{e}^{\mathbf{A}t}\right)\\ &= \mathbf{0}\end{aligned} \tag{3.75}$$

the uniqueness of the solution of (3.66). □

Exercise 3.15. Given are two identical linear systems of the form

$$\dot{\mathbf{x}}_i = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\mathbf{x}_i + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u_i, \qquad i = 1, 2 \tag{3.76a}$$

$$y_i = \begin{bmatrix} 1 & 0 \end{bmatrix}\mathbf{x}_i\ . \tag{3.76b}$$

Check the stability of the equilibrium when the two systems are connected in series or in parallel. Provide a physical interpretation of the results when considering system (3.76) as an undamped mass-spring oscillator.

Exercise 3.16. Given is the linear autonomous time-invariant sampled system

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k, \qquad \mathbf{A} \in \mathbb{R}^{n\times n}\ . \tag{3.77}$$

Show that the existence of a positive definite solution $\mathbf{P} \in \mathbb{R}^{n\times n}$ of the inequality

$$\mathbf{A}^{\mathrm{T}}\mathbf{P}\mathbf{A} - \mathbf{P} < \mathbf{0} \tag{3.78}$$

is sufficient for $V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$ to be a Lyapunov function for (3.77).

Exercise 3.17. The linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{3.79a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{3.79b}$$

is completely observable. Show that $\mathbf{A}$ is a Hurwitz matrix if and only if the Lyapunov equation

$$\mathbf{P}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P} = -\mathbf{C}^{\mathrm{T}}\mathbf{C} \tag{3.80}$$

is satisfied for a positive definite $\mathbf{P}$. Show further that in this case, the solution for $\mathbf{P}$ is unique.

Tip: Use the invariance principle of Krassovskii-LaSalle and the fact that for the observable pair $(\mathbf{A}, \mathbf{C})$, $\mathbf{C}\mathrm{e}^{\mathbf{A}t}\mathbf{x} = \mathbf{0}$ for all $t \geq 0$ if and only if $\mathbf{x} = \mathbf{0}$ for all $t \geq 0$.

3.1.7 Indirect (First) Method of Lyapunov

In addition to the second method of Lyapunov discussed in Section 3.1.3, which is essentially based on the construction of a Lyapunov function, there is also the possibility to assess the stability of an equilibrium point based on the linearized system around this equilibrium point. Consider the nonlinear autonomous system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \tag{3.81}$$

with equilibrium point $\mathbf{x}_R = \mathbf{0}$. Assuming that $\mathbf{f}(\mathbf{x})$ is continuously differentiable on an open neighborhood $\mathcal{D}$ of $\mathbf{0}$, $\mathbf{f}(\mathbf{x})$ can be written in the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{0}) + \frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{0}} \mathbf{x} + \mathbf{r}(\mathbf{x}), \qquad \lim_{\|\mathbf{x}\| \to 0} \frac{\|\mathbf{r}(\mathbf{x})\|}{\|\mathbf{x}\|} = 0 \tag{3.82}$$

Then the following theorem holds:

Theorem 3.8 (Indirect (first) Method of Lyapunov)**.** *Let $\mathbf{x}_R = \mathbf{0}$ be an equilibrium point of (3.81) and $\mathbf{f}(\mathbf{x})$ be continuously differentiable on an open neighborhood $\mathcal{D} \subseteq \mathbb{R}^n$ of $\mathbf{0}$. With*

$$\mathbf{A} = \frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{0}} \tag{3.83}$$

the following holds:

*(1) If **all** eigenvalues λ_i of $\mathbf{A}$ have a real part less than zero, i.e., $\mathrm{Re}(\lambda_i) < 0$, then the equilibrium point is asymptotically stable.*

*(2) If **one** eigenvalue λ_i of $\mathbf{A}$ satisfies $\mathrm{Re}(\lambda_i) > 0$, then the origin is unstable.*

(3) For eigenvalues λ_i of $\mathbf{A}$ with $\mathrm{Re}(\lambda_i) = 0$, no statement can be made about the stability of the equilibrium point of the nonlinear system.

Proof. To prove the first part of this theorem, the function

$$V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x} \tag{3.84}$$

with positive definite $\mathbf{P}$ is considered as a candidate for a Lyapunov function. From (3.82), it follows for $\dot{V}$

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{f}(\mathbf{x}) + \mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathbf{P}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{P}(\mathbf{A}\mathbf{x} + \mathbf{r}(\mathbf{x})) + (\mathbf{A}\mathbf{x} + \mathbf{r}(\mathbf{x}))^{\mathrm{T}}\mathbf{P}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}\Big(\mathbf{P}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P}\Big)\mathbf{x} + 2\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{r}(\mathbf{x}) \ .\end{aligned} \tag{3.85}$$

Since $\mathbf{A}$ is a Hurwitz matrix, the Lyapunov equation

$$\mathbf{P}\mathbf{A} + \mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{Q} = \mathbf{0} \tag{3.86}$$

has a positive definite solution $\mathbf{P}$ for every positive definite $\mathbf{Q}$. It was also assumed that $\mathbf{f}(\mathbf{x})$ is continuously differentiable, and therefore for every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\|\mathbf{r}(\mathbf{x})\|_2 < \varepsilon\|\mathbf{x}\|_2, \qquad \|\mathbf{x}\|_2 < \delta \ . \tag{3.87}$$

For a positive definite matrix $\mathbf{P}$, the induced 2-norm satisfies the estimate (compare to (2.55))

$$\lambda_{\text{min}}(\mathbf{P}) \leq \|\mathbf{P}\|_{i,2} \leq \lambda_{\text{max}}(\mathbf{P}) \tag{3.88}$$

with $\lambda_{\text{min}}(\mathbf{P}) > 0$ or $\lambda_{\text{max}}(\mathbf{P}) > 0$ as the smallest or largest eigenvalue of $\mathbf{P}$. Thus, from the Cauchy-Schwarz inequality (2.82), (3.87), and (3.88), the estimate

$$\Big|\mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{r}(\mathbf{x})\Big| \leq \|\mathbf{P}\mathbf{r}(\mathbf{x})\|_2\|\mathbf{x}\|_2 \leq \|\mathbf{P}\|_{i,2}\underbrace{\|\mathbf{r}(\mathbf{x})\|_2}_{<\varepsilon\|\mathbf{x}\|_2}\|\mathbf{x}\|_2 \leq \varepsilon\lambda_{\text{max}}(\mathbf{P})\|\mathbf{x}\|_2^2 \tag{3.89}$$

or

$$\begin{aligned}\dot{V}(\mathbf{x}) &\leq -\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + 2\varepsilon\lambda_{\text{max}}(\mathbf{P})\|\mathbf{x}\|_2^2 \\ &\leq (-\lambda_{\text{min}}(\mathbf{Q}) + 2\varepsilon\lambda_{\text{max}}(\mathbf{P}))\|\mathbf{x}\|_2^2 \ ,\end{aligned} \tag{3.90}$$

is obtained, and $\dot{V}$ is definitely negative for

$$\varepsilon < \frac{\lambda_{\text{min}}(\mathbf{Q})}{2\lambda_{\text{max}}(\mathbf{P})} \tag{3.91}$$

This proves, according to Theorem 3.1, the asymptotic stability of the equilibrium $\mathbf{x}_R = \mathbf{0}$. The proof of the second part of Theorem 3.8 is not carried out here but can be found in the corresponding literature. □

Exercise 3.18. Search in the literature provided at the end for Lyapunov instability theorems and apply them to prove the second part of Theorem 3.8.

If the linearized system has eigenvalues λ_i with $\mathrm{Re}(\lambda_i) = 0$, then the indirect method does not allow any statement. Consider the nonlinear single-input system

$$\dot{x} = ax^3 \tag{3.92}$$

with the system linearized around the equilibrium $x_R = 0$

$$\dot{x} = 0 \ . \tag{3.93}$$

Choosing candidates for a Lyapunov function as

$$V(x) = x^4 \tag{3.94}$$

and obtaining $\dot{V}$ as

$$\dot{V}(x) = 4ax^6 \ . \tag{3.95}$$

It is easy to see that the origin is asymptotically stable for $a < 0$ but unstable for $a > 0$. For $a = 0$, the system is linear and has infinitely many equilibrium points.

> *Exercise* 3.19. Examine the stability of the equilibrium point(s) for systems (3.9), (3.29), (3.41), and (3.43) using the indirect method of Lyapunov.

3.2 Non-autonomous Systems

The following considerations are based on the non-autonomous nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}) \tag{3.96}$$

with $\mathbf{f} : [0, \infty) \times \mathcal{D} \to \mathbb{R}^n$ piecewise continuous in t and locally Lipschitz in $\mathbf{x}$ on $[0, \infty) \times \mathcal{D}$, $\mathcal{D} \subseteq \mathbb{R}^n$, (compare Theorem 2.13). The error systems that arise in trajectory tracking control of nonlinear systems typically have the structure of (3.96). One calls $\mathbf{x}_R \in \mathcal{D}$ an equilibrium of (3.96) for $t = t_0$, if for all times $t \geq t_0 \geq 0$ the relationship

$$\mathbf{f}(t, \mathbf{x}_R) = \mathbf{0} \tag{3.97}$$

is satisfied, where $\mathbf{x}_R$ must be independent of time t. Without loss of generality, one can assume that an equilibrium with $\mathbf{x}_R = \mathbf{0}$ for $t_0 = 0$ is given.

> *Exercise* 3.20. Show that for $\mathbf{x}_R \neq \mathbf{0}$, $t_0 \neq 0$, one can always achieve, through a simple coordinate and time transformation, that in the new coordinates the equilibrium $\tilde{\mathbf{x}}_R = \mathbf{0}$ for $\tilde{t} = 0$.

In the following, it will be briefly shown that the equilibrium of a non-autonomous system (3.96) can also be the transformed nontrivial solution of an autonomous system. This has the advantage that the stability analysis of a solution trajectory can be reduced to the stability of an equilibrium of a non-autonomous system. Consider the autonomous system

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{y} = \mathbf{g}(\mathbf{y}) \ , \tag{3.98}$$

where $\bar{\mathbf{y}}(\tau)$ denotes a solution of (3.98) for $\tau \geq \tau_0 \geq 0$. Now, performing a coordinate and time transformation of the form $\mathbf{x} = \mathbf{y} - \bar{\mathbf{y}}(\tau)$ and $t = \tau - \tau_0$, we obtain the transformed system

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{y}(t+\tau_0) - \frac{\mathrm{d}}{\mathrm{d}t}\bar{\mathbf{y}}(t+\tau_0)\\ &= \mathbf{g}(\mathbf{x} + \bar{\mathbf{y}}(t+\tau_0)) - \frac{\mathrm{d}}{\mathrm{d}t}\bar{\mathbf{y}}(t+\tau_0)\\ &:= \mathbf{f}(t,\mathbf{x})\ .\end{aligned} \tag{3.99}$$

Since $\bar{\mathbf{y}}(\tau)$ is a solution of (3.98) for $\tau \geq \tau_0 \geq 0$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\bar{\mathbf{y}}(\tau) = \mathbf{g}(\bar{\mathbf{y}}(\tau)), \qquad \tau \geq \tau_0 \geq 0 \tag{3.100}$$

or in the transformed time t

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\mathbf{y}}(t+\tau_0) = \mathbf{g}(\bar{\mathbf{y}}(t+\tau_0)), \qquad t \geq 0\ . \tag{3.101}$$

It is immediately clear from (3.99) and (3.101) that $\mathbf{x}_R = \mathbf{0}$ for $t_0 = 0$ is an equilibrium of the transformed system $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{f}(t,\mathbf{x})$.

The definition of Lyapunov stability according to Definition 3.3 can now also be applied to non-autonomous systems, but here the dependence of the system behavior on the initial time t_0 must be explicitly taken into account.

Definition 3.11 (Lyapunov Stability of Non-Autonomous Systems)**.** The equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.96) is called

- *stable (in the sense of Lyapunov)*, if for every $\varepsilon > 0$ there exists a $\delta(\varepsilon, t_0) > 0$ such that

$$\|\mathbf{x}(t_0)\| < \delta(\varepsilon, t_0) \quad \Rightarrow \quad \|\mathbf{x}(t)\| < \varepsilon \tag{3.102}$$

  holds for all $t \geq t_0 \geq 0$,

- *uniformly stable*, if for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ (independent of t_0) such that (3.102) is satisfied for all $t \geq t_0 \geq 0$,

- *asymptotically stable*, if it is stable and there exists a positive real number $\eta(t_0)$ such that from

$$\|\mathbf{x}(t_0)\| < \eta(t_0) \quad \Rightarrow \quad \lim_{t\to\infty} \mathbf{x}(t) = \mathbf{0}\ , \tag{3.103}$$

- *uniformly asymptotically stable*, if it is uniformly stable, there exists a positive real number η (independent of t_0) such that (3.103) is satisfied for all $t \geq t_0 \geq 0$, and for every $\mu > 0$ one can find a $T(\mu) > 0$ such that

$$\|\mathbf{x}(t_0)\| < \eta \quad \Rightarrow \quad \|\mathbf{x}(t)\| < \mu \quad \text{for all} \quad t \geq t_0 + T(\mu) \tag{3.104}$$

holds.

For non-autonomous systems of the form (3.96), a theorem analogous to Theorem 3.1 can now be given for checking uniform stability:

Theorem 3.9 (Uniform stability of non-autonomous systems)**.** *Let $\mathbf{x}_R = \mathbf{0}$ be an equilibrium of (3.96) for $t = 0$ and $\mathcal{D} \subseteq \mathbb{R}^n$ be an open neighborhood of $\mathbf{0}$. If there exists a continuously differentiable function $V(t, \mathbf{x}) : [0, \infty) \times \mathcal{D} \to \mathbb{R}$ and continuous positive definite functions $W_1(\mathbf{x})$ and $W_2(\mathbf{x})$ on $\mathcal{D}$ such that*

$$W_1(\mathbf{x}) \leq V(t, \mathbf{x}) \leq W_2(\mathbf{x}) \tag{3.105a}$$

$$\frac{\partial}{\partial t} V + \left(\frac{\partial}{\partial \mathbf{x}} V\right) \mathbf{f}(t, \mathbf{x}) \leq 0 \tag{3.105b}$$

holds for all $t \geq 0$ and all $\mathbf{x} \in \mathcal{D}$, then the equilibrium $\mathbf{x}_R = \mathbf{0}$ is uniformly stable*. If furthermore a continuous positive definite function $W_3(\mathbf{x})$ on $\mathcal{D}$ exists such that (3.105b) can be bounded as*

$$\frac{\partial}{\partial t} V + \left(\frac{\partial}{\partial \mathbf{x}} V\right) \mathbf{f}(t, \mathbf{x}) \leq -W_3(\mathbf{x}) < 0 \tag{3.106}$$

for all $t \geq 0$ and all $\mathbf{x} \in \mathcal{D}$, then the equilibrium $\mathbf{x}_R = \mathbf{0}$ is uniformly asymptotically stable*.*

The proof of this theorem can be found in the literature cited at the end.

Exercise 3.21. Show that the equilibrium $\mathbf{x} = \mathbf{0}$ of the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -x_1 - g(t)x_2 \\ x_1 - x_2 \end{bmatrix} \tag{3.107}$$

with the continuously differentiable time function $g(t)$, $0 \leq g(t) \leq k$ and $\frac{\mathrm{d}}{\mathrm{d}t} g(t) \leq g(t)$ for all $t \geq 0$ is uniformly asymptotically stable.

Exercise 3.22. Given is the following mathematical model (mathematical pendulum with time-varying damping)

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -\sin(x_1) - g(t)x_2 \end{bmatrix} \tag{3.108}$$

with the continuously differentiable time function $g(t)$, $0 < \alpha \leq g(t) \leq \beta < \infty$ and $\frac{d}{dt} g(t) \leq \gamma < 2$ for all $t \geq 0$. Show that the equilibrium $x_1 = x_2 = 0$ is uniformly asymptotically stable.

Besides uniform stability, exponential stability also plays a crucial role in the analysis of non-autonomous systems.

Definition 3.12 (Exponential Stability of Non-autonomous Systems)**.** The equilibrium $\mathbf{x}_R = \mathbf{0}$ of (3.96) is called *exponentially stable* if positive constants k_1, k_2, and k_3 exist such that

$$\|\mathbf{x}(t_0)\| < k_3 \quad \Rightarrow \quad \|\mathbf{x}(t)\| < k_1 \|\mathbf{x}(t_0)\| \mathrm{e}^{-k_2(t-t_0)} \ . \tag{3.109}$$

The verification of exponential stability can be done using the following theorem.

Theorem 3.10 (Exponential Stability of Non-autonomous Systems)**.** *Let $\mathbf{x}_R = \mathbf{0}$ be an equilibrium of (3.96) at $t = 0$ and $\mathcal{D} \subseteq \mathbb{R}^n$ be an open neighborhood of $\mathbf{0}$. If there exists a continuously differentiable function $V(t,\mathbf{x}) : [0,\infty) \times \mathcal{D} \rightarrow \mathbb{R}$ and positive constants α_j, $j = 1,\dots,4$, such that*

$$\alpha_1 \|\mathbf{x}(t)\|^{\alpha_4} \leq V(t,\mathbf{x}) \leq \alpha_2 \|\mathbf{x}(t)\|^{\alpha_4} \tag{3.110a}$$

$$\frac{\partial}{\partial t} V + \left(\frac{\partial}{\partial \mathbf{x}} V\right) \mathbf{f}(t,\mathbf{x}) \leq -\alpha_3 \|\mathbf{x}(t)\|^{\alpha_4} \tag{3.110b}$$

holds for all $t \geq 0$ and all $\mathbf{x} \in \mathcal{D}$, then the equilibrium $\mathbf{x}_R = \mathbf{0}$ is exponentially stable.

Proof. From the two inequalities (3.110), it can be seen that

$$\frac{\mathrm{d}}{\mathrm{d}t} V(t,\mathbf{x}) \leq -\alpha_3 \|\mathbf{x}(t)\|^{\alpha_4} \leq -\frac{\alpha_3}{\alpha_2} V(t,\mathbf{x}) \tag{3.111}$$

and thus

$$V(t,\mathbf{x}) \leq V(t_0,\mathbf{x}(t_0)) \mathrm{e}^{-\frac{\alpha_3}{\alpha_2}(t-t_0)} \ . \tag{3.112}$$

Furthermore, from (3.110a) it follows

$$V(t_0,\mathbf{x}(t_0)) \leq \alpha_2 \|\mathbf{x}(t_0)\|^{\alpha_4} \tag{3.113}$$

and

$$\|\mathbf{x}(t)\| \leq \left(\frac{V(t,\mathbf{x})}{\alpha_1}\right)^{\frac{1}{\alpha_4}} \ , \tag{3.114}$$

hence, with (3.112), the following estimation

$$\|\mathbf{x}(t)\| \leq \left(\frac{V(t,\mathbf{x})}{\alpha_1}\right)^{\frac{1}{\alpha_4}} \leq \left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{1}{\alpha_4}} \|\mathbf{x}(t_0)\| \mathrm{e}^{-\frac{\alpha_3}{\alpha_2 \alpha_4}(t-t_0)} \tag{3.115}$$

can be given. This directly shows the exponential stability according to Definition 3.12 for $k_1 = \left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{1}{\alpha_4}}$ and $k_2 = \frac{\alpha_3}{\alpha_2 \alpha_4}$. □

Exercise 3.23. Given is the following mathematical model

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}h(t)x_2 - g(t)x_1^3\\ -h(t)x_1 - g(t)x_2^3\end{bmatrix} \tag{3.116}$$

with the continuously differentiable and bounded time functions $h(t)$ and $g(t)$, $g(t) \geq k > 0$ for all $t \geq 0$. Is the equilibrium $x_1 = x_2 = 0$ uniformly asymptotically stable? Is the equilibrium $x_1 = x_2 = 0$ exponentially stable?

Exercise 3.24. Given is the following mathematical model

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-x_1 + x_2 + \left(x_1^2 + x_2^2\right)\sin(t)\\ -x_1 - x_2 + \left(x_1^2 + x_2^2\right)\cos(t)\end{bmatrix} . \tag{3.117}$$

Show that the equilibrium $x_1 = x_2 = 0$ is exponentially stable.

3.2.1 Linear Systems

The stability analysis of linear time-varying systems of the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \tag{3.118}$$

is significantly more challenging compared to the time-invariant case as in (3.59).

Example 3.1. Consider the system (3.118) with the dynamics matrix

$$\mathbf{A}(t) = \begin{bmatrix}-1 + 1.5(\cos(t))^2 & 1 - 1.5\sin(t)\cos(t)\\ -1 - 1.5\sin(t)\cos(t) & -1 + 1.5(\sin(t))^2\end{bmatrix} . \tag{3.119}$$

In this case, the eigenvalues $\lambda_{1,2} = -1/4 \pm I\sqrt{7}/4$ of $\mathbf{A}(t)$ are constant for all times t and have negative real parts, yet the equilibrium is unstable as shown by the calculation of the solution for $t_0 = 0$

$$\mathbf{x}(t) = \begin{bmatrix}\mathrm{e}^{t/2}\cos(t) & \mathrm{e}^{-t}\sin(t)\\ -\mathrm{e}^{t/2}\sin(t) & \mathrm{e}^{-t}\cos(t)\end{bmatrix}\mathbf{x}(0) \tag{3.120}$$

It is worth mentioning that linear time-varying systems arise naturally when linearizing nonlinear (autonomous) systems around a desired trajectory.

The stability analysis of the equilibrium can be carried out, for example, using Theorem 3.9. To do this, one chooses a suitable Lyapunov function of the form

$$V(t, \mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}(t)\mathbf{x}, \qquad 0 < \alpha_1\mathbf{I} \leq \mathbf{P}(t) \leq \alpha_2\mathbf{I} \tag{3.121}$$

with a continuously differentiable, bounded, and symmetric matrix $\mathbf{P}(t)$ and positive constants α_1 and α_2. The Lyapunov function satisfies the inequalities

$$\alpha_1\|\mathbf{x}\|_2^2 \leq V(t, \mathbf{x}) \leq \alpha_2\|\mathbf{x}\|_2^2 . \tag{3.122}$$

If $\mathbf{P}(t)$ satisfies the matrix differential equation

$$-\dot{\mathbf{P}}(t) = \mathbf{A}^{\mathrm{T}}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}(t) + \mathbf{Q}(t) \tag{3.123}$$

for a continuous, bounded, and symmetric matrix $\mathbf{Q}(t)$ such that

$$0 < \alpha_3 \mathbf{I} \leq \mathbf{Q}(t) \ , \tag{3.124}$$

then the change in $V(t, \mathbf{x})$ along a solution curve of (3.118) is given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} V(t, \mathbf{x}) &= \dot{\mathbf{x}}^{\mathrm{T}} \mathbf{P}(t)\mathbf{x} + \mathbf{x}^{\mathrm{T}} \dot{\mathbf{P}}(t)\mathbf{x} + \mathbf{x}^{\mathrm{T}} \mathbf{P}(t)\dot{\mathbf{x}} \\ &= \mathbf{x}^{\mathrm{T}} \Big(\mathbf{A}^{\mathrm{T}}(t)\mathbf{P}(t) + \dot{\mathbf{P}}(t) + \mathbf{P}(t)\mathbf{A}(t) \Big) \mathbf{x} \\ &= -\mathbf{x}^{\mathrm{T}} \mathbf{Q}(t)\mathbf{x} \\ &\leq -\alpha_3 \|\mathbf{x}\|_2^2 < 0 \ . \end{aligned} \tag{3.125}$$

From (3.122) and (3.125), it is immediately apparent that exponential stability for $\alpha_4 = 2$ is also demonstrated by Theorem 3.10. It is worth mentioning that for linear time-varying systems, uniform asymptotic stability and exponential stability are equivalent.

For the analysis of linear *periodically* time-varying systems of the form (3.118) with $\mathbf{A}(t) = \mathbf{A}(t + T)$, a comprehensive theory can be found in the literature under the term *Floquet theory*. Here, we refrain from further elaboration on this topic, but we provide a useful estimation for the trajectories of linear time-varying systems.

Theorem 3.11 (Ważewski's Inequality)**.** *A solution $\mathbf{x}(t)$ of the linear time-varying system (3.118) with the real-valued dynamics matrix $\mathbf{A}(t)$ satisfies the following inequality*

$$\|\mathbf{x}(t_0)\|_2 \exp\left(\int_{t_0}^{t} \lambda(\tau) \,\mathrm{d}\tau \right) \leq \|\mathbf{x}(t)\|_2 \leq \|\mathbf{x}(t_0)\|_2 \exp\left(\int_{t_0}^{t} \Lambda(\tau) \,\mathrm{d}\tau \right) \ , \tag{3.126}$$

where $\lambda(t)$ and $\Lambda(t)$ denote the smallest and largest eigenvalue of the symmetric part of the matrix $\mathbf{A}(t)$

$$\mathbf{A}_s(t) = \frac{1}{2} \Big(\mathbf{A}(t) + \mathbf{A}^{\mathrm{T}}(t) \Big) \tag{3.127}$$

Proof. For a fixed time t, according to (2.64), the relationship holds

$$\lambda(t)\|\mathbf{x}(t)\|_2^2 \le \mathbf{x}^{\mathrm{T}}(t)\mathbf{A}_s(t)\mathbf{x}(t) \le \Lambda(t)\|\mathbf{x}(t)\|_2^2 \tag{3.128}$$

and by substituting

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t)\|_2^2 &= \dot{\mathbf{x}}^{\mathrm{T}}(t)\mathbf{x}(t) + \mathbf{x}^{\mathrm{T}}(t)\dot{\mathbf{x}}(t)\\ &= \mathbf{x}^{\mathrm{T}}(t)\Big(\mathbf{A}(t) + \mathbf{A}^{\mathrm{T}}(t)\Big)\mathbf{x}(t)\\ &= 2\mathbf{x}^{\mathrm{T}}(t)\mathbf{A}_s(t)\mathbf{x}(t)\end{aligned} \tag{3.129}$$

we obtain

$$2\lambda(t)\|\mathbf{x}(t)\|_2^2 \le \frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t)\|_2^2 \le 2\Lambda(t)\|\mathbf{x}(t)\|_2^2\ . \tag{3.130}$$

Now, considering only the left part of the inequality (3.130) in the first step, the result immediately follows according to (3.126)

$$2\lambda(t)\|\mathbf{x}(t)\|_2^2 \le 2\|\mathbf{x}(t)\|_2\frac{\mathrm{d}(\|\mathbf{x}(t)\|_2)}{\mathrm{d}t} \tag{3.131a}$$

$$\lambda(t)\,\mathrm{d}t \le \frac{\mathrm{d}(\|\mathbf{x}(t)\|_2)}{\|\mathbf{x}(t)\|_2} \tag{3.131b}$$

$$\int_{t_0}^{t}\lambda(\tau)\,\mathrm{d}\tau \le \ln\left(\frac{\|\mathbf{x}(t)\|_2}{\|\mathbf{x}(t_0)\|_2}\right) \tag{3.131c}$$

$$\|\mathbf{x}(t_0)\|_2\exp\left(\int_{t_0}^{t}\lambda(\tau)\,\mathrm{d}\tau\right) \le \|\mathbf{x}(t)\|_2\ . \tag{3.131d}$$

□

Exercise 3.25. Show in the same way the right part of the inequality (3.130).

Taking again the system (3.118) with the dynamics matrix (3.119) as an example, the symmetric part of the dynamics matrix is calculated as

$$\begin{aligned}\mathbf{A}_s(t) &= \frac{1}{2}\Big(\mathbf{A}(t) + \mathbf{A}^{\mathrm{T}}(t)\Big)\\ &= \begin{bmatrix}-1 + 1.5(\cos(t))^2 & -1.5\sin(t)\cos(t)\\ -1.5\sin(t)\cos(t) & -1 + 1.5(\sin(t))^2\end{bmatrix}\end{aligned} \tag{3.132}$$

with the corresponding eigenvalues $\lambda_{s1} = 1/2$ and $\lambda_{s2} = -1$. According to Theorem 3.11, a solution $\mathbf{x}(t)$ satisfies the inequality

$$\|\mathbf{x}(t_0)\|_2\mathrm{e}^{-(t-t_0)} \le \|\mathbf{x}(t)\|_2 \le \|\mathbf{x}(t_0)\|_2\mathrm{e}^{\frac{1}{2}(t-t_0)}\ . \tag{3.133}$$

3.2.2 Lyapunov-like Theory: Barbalat's Lemma

In addition to the Lyapunov theory for non-autonomous nonlinear systems of the form (3.96) discussed in the previous section, one often finds a Lyapunov-like approach using what is called *Barbalat's Lemma*. It is based on the mathematical properties of the asymptotic behavior of functions and their derivatives. In the first step, let us review some asymptotic properties of functions and their temporal derivatives. For a function $f(t)$ differentiable with respect to time t, the following holds:

(1) From $\lim\limits_{t\to\infty} \dot{f}(t) = 0$, *it does not follow* $\lim\limits_{t\to\infty} f(t) = c$ with $|c| < \infty$.

As an example, consider the function $f(t) = \ln(t)$. While the derivative satisfies

$$\lim_{t\to\infty} \dot{f}(t) = \frac{1}{t} = 0\ , \tag{3.134}$$

the function itself goes to ∞ as $t \to \infty$.

(2) From $\lim\limits_{t\to\infty} f(t) = c$ with $|c| < \infty$, *it does not follow* $\lim\limits_{t\to\infty} \dot{f}(t) = 0$.

For example, consider the function $f(t) = \mathrm{e}^{-t}\sin(\mathrm{e}^{2t})$, for which $\lim\limits_{t\to\infty} f(t) = 0$, but

$$\lim_{t\to\infty} \dot{f}(t) = \lim_{t\to\infty} \Big(2\cos\Big(\mathrm{e}^{2t}\Big)\mathrm{e}^{t} - \mathrm{e}^{-t}\sin\Big(\mathrm{e}^{2t}\Big)\Big) \tag{3.135}$$

is not defined.

(3) If $f(t)$ is bounded from below and not increasing $\Big(\dot{f}(t) \leq 0\Big)$, then *it follows* $\lim\limits_{t\to\infty} f(t) = c$ with $|c| < \infty$.

Barbalat's Lemma now clarifies under which conditions the derivative $\dot{f}(t)$ of a bounded function converges to zero as $t \to \infty$.

Theorem 3.12 (Barbalat's Lemma)**.** *If the differentiable function $f(t)$ satisfies $\lim\limits_{t\to\infty} f(t) = c$ with $|c| < \infty$ and $\dot{f}(t)$ is uniformly continuous, then $\lim\limits_{t\to\infty} \dot{f}(t) = 0$.*

Before showing how this theorem is used for stability analysis, let us briefly revisit the concept of *uniform continuity* of a function $f(t)$.

Definition 3.13 ($\epsilon\delta$-Continuity)**.** A function $f(t)$ is said to be *continuous* at the point t_1 if for every $\epsilon > 0$ there exists $\delta = \delta(\epsilon, t_1) > 0$ such that

$$|t - t_1| < \delta \quad \Rightarrow \quad |f(t) - f(t_1)| < \epsilon \ . \tag{3.136}$$

A function $f(t)$ is called *uniformly continuous* if δ can always be found independently of t_1.

Consider the function $f(t) = t^2$ as an example. Let us choose an $\epsilon > 0$ and determine a δ such that

$$\left|t^2 - t_1^2\right| < \epsilon \quad \text{or} \quad |t - t_1||t + t_1| < \epsilon, \qquad |t - t_1| < \delta \ . \tag{3.137}$$

From (3.137), it can be seen that for $t > t_1 > 0$, for every ϵ, a δ can always be found such that

$$0 < t - t_1 < \delta \quad \Rightarrow \quad (t - t_1)(t + t_1) < \epsilon \ . \tag{3.138}$$

Replacing t in (3.138) with $t_n = t_1 + \delta - \frac{\delta}{n}$ and letting $n \to \infty$, we obtain

$$\delta(2t_1 + \delta) < \epsilon \tag{3.139}$$

or rather

$$\delta < \frac{\epsilon}{2t_1} \ . \tag{3.140}$$

It can be observed that as t_1 increases, keeping ϵ constant, the value of δ decreases, and thus there is no smallest δ that would be correct for all t_1. Therefore, the function $f(t) = t^2$ is continuous but not uniformly continuous. In contrast, for the function $f(t) = \sqrt{t}$ under the condition $t > t_1 > 0$,

$$\left|\sqrt{t} - \sqrt{t_1}\right| < \sqrt{|t - t_1|} < \epsilon \ , \tag{3.141}$$

and choosing $\delta = \epsilon^2$ immediately leads to *uniform continuity*, i.e.,

$$|t - t_1| < \delta \ , \tag{3.142a}$$

$$\sqrt{|t - t_1|} < \epsilon \ , \tag{3.142b}$$

$$\left|\sqrt{t} - \sqrt{t_1}\right| < \epsilon \ . \tag{3.142c}$$

Exercise 3.26. Prove the last implication in (3.142).

As can be seen, verifying uniform continuity in this manner is quite tedious. Therefore, a *sufficient criterion* of the following form is often used:

Theorem 3.13 (Sufficient condition for uniform continuity). *A differentiable function $f(t)$ is uniformly continuous if its derivative $\frac{\mathrm{d}}{\mathrm{d}t}f(t)$ is bounded.*

From Barbalat's Lemma, the following theorem for stability analysis of nonlinear, non-autonomous systems of the form (3.96) immediately follows.

Theorem 3.14 (Lyapunov-like method). *If a scalar function $V(t, \mathbf{x}) : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ satisfies the conditions*

(1) $V(t, \mathbf{x})$ is bounded from below,

(2) $\dot{V}(t, \mathbf{x}) \leq 0$, and

(3) $\dot{V}(t, \mathbf{x})$ is uniformly continuous in time t,

then $\lim\limits_{t\to\infty} \dot{V}(t, \mathbf{x}) = 0$.

As an application example, consider the following control engineering problem: We want to position a mass m sliding on a horizontal surface using the force F in the absence of friction. The corresponding system of differential equations is

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}x = F \ . \tag{3.143}$$

Suppose the desired position $r_\mathrm{d}(t)$ is specified by a person using a control stick, then a simple way to convert this external signal into a twice continuously differentiable reference signal $x_\mathrm{d}(t)$ is through a reference model of the form

$$\ddot{x}_\mathrm{d} + a_1\dot{x}_\mathrm{d} + a_0 x_\mathrm{d} = a_0 r_\mathrm{d}, \qquad G(s) = \frac{\hat{x}_\mathrm{d}}{\hat{r}_\mathrm{d}} = \frac{a_0}{s^2 + a_1 s + a_0} \tag{3.144}$$

for suitable parameters a_1 and a_0. The parameters a_1 and a_0 are chosen such that the reference model with transfer function $G(s)$ is stable and meets the performance requirements. Now, the simple control law

$$F(t) = m\Big(\ddot{x}_\mathrm{d} - 2\lambda\dot{e} - \lambda^2 e\Big), \qquad e = x - x_\mathrm{d} \tag{3.145}$$

for $\lambda > 0$ leads to an asymptotically stable closed loop with error dynamics

$$\ddot{e} + 2\lambda\dot{e} + \lambda^2 e = 0 \ . \tag{3.146}$$

Furthermore, assume that the mass m is constant but not precisely known, i.e., only the estimated value $\hat{m}$ is known. Substituting the estimated value $\hat{m}$ for m in the control law (3.145), we obtain for the closed loop

$$m\ddot{x} = \hat{m}\Big(\ddot{x}_{soll} - 2\lambda\dot{e} - \lambda^2 e\Big) \tag{3.147}$$

or

$$m\ddot{x} - m\Big(\ddot{x}_{soll} - 2\lambda\dot{e} - \lambda^2 e\Big) = \hat{m}\Big(\ddot{x}_{soll} - 2\lambda\dot{e} - \lambda^2 e\Big) - m\Big(\ddot{x}_{soll} - 2\lambda\dot{e} - \lambda^2 e\Big) \tag{3.148}$$

and by introducing a generalized control error $s = \dot{e} + \lambda e$, we get

$$m\frac{\mathrm{d}}{\mathrm{d}t}s + m\lambda s = e_m \underbrace{\left(\ddot{x}_{soll} - 2\lambda\dot{e} - \lambda^2 e\right)}_{w(t)} \tag{3.149}$$

with the parameter error $e_m = \hat{m} - m$.

The *adaptive control law*

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{m} = -\gamma w s, \qquad \gamma > 0 \tag{3.150}$$

guarantees that the generalized control error converges asymptotically to zero. To prove this, one considers the function bounded from below

$$V(s, e_m) = \frac{1}{2}\left(ms^2 + \frac{1}{\gamma}e_m^2\right) \tag{3.151}$$

and calculates its time derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V &= ms\left(-\lambda s + \frac{1}{m}e_m w\right) + \frac{1}{\gamma}e_m(-\gamma w s)\\ &= -\lambda m s^2 \leq 0\ .\end{aligned} \tag{3.152}$$

Since V is positive definite in s and e_m and $\dot{V}$ is negative semidefinite, the functions s and e_m are bounded. Taking another time derivative of $\dot{V}$, one obtains

$$\ddot{V} = -2\lambda m s\left(-\lambda s + \frac{1}{m}e_m w\right)\ , \tag{3.153}$$

and this function is also bounded due to the bounded quantities s and e_m and the assumption of bounded reference signals $r_\mathrm{d}(t)$ (hence $w(t)$ is also bounded). According to Theorem 3.13, $\dot{V}$ is uniformly continuous, the Barbalat's Lemma (Theorem 3.14) can be applied, leading to

$$\lim_{t\to\infty}\dot{V} = -\lim_{t\to\infty}\lambda m s^2 = 0 \tag{3.154}$$

thus

$$\lim_{t\to\infty} s = 0\ . \tag{3.155}$$

3.3 Literatur

[3.1] B. P. Demidovich, *Vorlesung zur Mathematischen Stabilitätstheorie*. Moskau: Verlag der Moskau Universität, 1998.

[3.2] O. Föllinger, *Nichtlineare Regelung I + II*. München: Oldenbourg, 1993.

[3.3] H. K. Khalil, *Nonlinear Systems (3rd Edition)*. New Jersey: Prentice Hall, 2002.

[3.4] E. Slotine and W. Li, *Applied Nonlinear Control*. New Jersey: Prentice Hall, 1991.

[3.5] M. Vidyasagar, *Nonlinear Systems Analysis*. New Jersey: Prentice Hall, 1993.

4 Lyapunov-based Controller Design

This chapter discusses some controller design methods based on Lyapunov's theory of stability. The basic idea of these methods is to find a *nonlinear state feedback* $\mathbf{u} = \boldsymbol{\alpha}(\mathbf{x})$ for a system of the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \ , \qquad \mathbf{f}(\mathbf{0}, \mathbf{0}) = \mathbf{0} \tag{4.1}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the control input $\mathbf{u} \in \mathbb{R}^p$, and $\boldsymbol{\alpha}(\mathbf{0}) = \mathbf{0}$, such that the equilibrium $\mathbf{x}_R = \mathbf{0}$ of the closed loop system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \boldsymbol{\alpha}(\mathbf{x})) \tag{4.2}$$

becomes stable or asymptotically stable in the sense of Lyapunov.

4.1 Integrator Backstepping

As a starting point and motivation for this nonlinear controller design method, consider the following nonlinear system

$$\dot{x}_1 = \cos(x_1) - x_1^3 + x_2 \tag{4.3a}$$
$$\dot{x}_2 = u \tag{4.3b}$$

with state $\mathbf{x}^{\mathrm{T}} = [x_1, x_2]$ and control input u. Now, a state feedback control $u = u(x_1, x_2)$ should be designed such that for every initial state $\mathbf{x}(0) = \mathbf{x}_0$, $\lim_{t\to\infty} x_1(t) = 0$ and $\lim_{t\to\infty} |x_2(t)| = c < \infty$. From (4.3), it can be seen that for $x_{1,R} = 0$, the only equilibrium with $\mathbf{x}_R^{\mathrm{T}} = [0, -1]$ is given. Considering the state x_2 as a *virtual control input* for the system (4.3a), then the state feedback

$$x_2 = \alpha(x_1) = -\cos(x_1) - c_1 x_1 \ , \qquad c_1 > 0 \tag{4.4}$$

would make the equilibrium $x_{1,R} = 0$ of the subsystem (4.3a), (4.4) asymptotically stable. To show this, let's choose the Lyapunov function

$$V(x_1) = \frac{1}{2} x_1^2 > 0 \ , \tag{4.5}$$

then the time derivative is calculated as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} V(x_1) &= x_1 \Big(-x_1^3 - c_1 x_1 \Big) \\ &= -x_1^4 - c_1 x_1^2 < 0 \ . \end{aligned} \tag{4.6}$$

Next, the deviation of the state x_2 from the "ideal" form (4.4)

$$z = x_2 - \alpha(x_1) = x_2 + \cos(x_1) + c_1 x_1 \tag{4.7}$$

is introduced as a new state variable, resulting in the differential equation (4.3) in the new state $[x_1, z]$

$$\begin{aligned} \dot{x}_1 &= \cos(x_1) - x_1^3 + \underbrace{(z - \cos(x_1) - c_1 x_1)}_{x_2} \\ &= -x_1^3 - c_1 x_1 + z \end{aligned} \tag{4.8a}$$

$$\begin{aligned} \dot{z} &= \dot{x}_2 - \frac{\mathrm{d}}{\mathrm{d}t}\alpha(x_1) \\ &= u - (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1 + z\Big) \ . \end{aligned} \tag{4.8b}$$

Now, assuming a Lyapunov function in the form

$$V_a(x_1, x_2) = V(x_1) + \frac{1}{2}z^2 = \frac{1}{2}x_1^2 + \frac{1}{2}(x_2 + \cos(x_1) + c_1 x_1)^2 \tag{4.9}$$

we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V_a(x_1, x_2) &= x_1\Big(-x_1^3 - c_1 x_1 + z\Big) + z\Big(u - (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1 + z\Big)\Big) \\ &= -c_1 x_1^2 - x_1^4 + z\underbrace{\Big\{x_1 + u - (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1 + z\Big)\Big\}}_{\chi} \ . \end{aligned} \tag{4.10}$$

The idea is now to determine the control input u in such a way that $\frac{\mathrm{d}}{\mathrm{d}t}V_a(x_1, x_2)$ becomes negative definite. This can be achieved, for example, by choosing

$$\chi = x_1 + u - (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1 + z\Big) = -c_2 z, \qquad c_2 > 0 \tag{4.11}$$

or

$$u = -x_1 + (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1 + z\Big) - c_2 z \ . \tag{4.12}$$

In conclusion, it can be easily verified that the state feedback (4.12) globally asymptotically stabilizes the equilibrium $x_{1,R} = z_R = 0$ or $x_{1,R} = 0$ and $x_{2,R} = -1$.

> *Exercise* 4.1. Show that $V_a(x_1, x_2)$ from (4.9) is radially unbounded.

The choice of u according to (4.11) is of course not unique, as on one hand, $\chi = -f(z)$ could be chosen with any arbitrary function $f(z)$ satisfying $f(z)z > 0$ for all $z \neq 0$, and on the other hand, it is not necessary to cancel all terms of χ. For example, the state feedback

$$u = -x_1 + (\sin(x_1) - c_1)\Big(-x_1^3 - c_1 x_1\Big) - c_2 z \tag{4.13}$$

would lead to a closed loop (4.8), (4.13) of the form

$$\dot{x}_1 = -x_1^3 - c_1 x_1 + z \tag{4.14a}$$
$$\dot{z} = -x_1 - c_2 z - (\sin(x_1) - c_1)z \tag{4.14b}$$

and for the choice of parameters $c_2 > c_1 + 1$, the Lyapunov function

$$V_a(x_1, z) = \frac{1}{2}x_1^2 + \frac{1}{2}z^2 \tag{4.15}$$

and its time derivative

$$\frac{\mathrm{d}}{\mathrm{d}t}V_a = -x_1^4 - c_1 x_1^2 - (c_2 - c_1 + \sin(x_1))z^2 \tag{4.16}$$

show the global asymptotic stability of the equilibrium $x_{1,R} = z_R = 0$ or $x_{1,R} = 0$ and $x_{2,R} = -1$.

Exercise 4.2. Show that for a suitable choice of parameters k_1 and k_2, even the simple state feedback

$$u = -k_1 z - k_2 x_1^2 z \tag{4.17}$$

leads to a closed loop with a globally asymptotically stable equilibrium.

> **Tip:** Choose the Lyapunov function as $V_a = \frac{1}{2}x_1^2 + \frac{1}{2}z^2$ and combine the terms of $\dot{V}_a$ appropriately.

These variations mentioned above demonstrate the design degrees of freedom of the method. The generalization of the example discussed above is now possible in the following form:

Theorem 4.1 (Integrator Backstepping)**.** *Consider the nonlinear system*

$$\dot{\mathbf{x}}_1 = \mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2 \tag{4.18a}$$
$$\dot{x}_2 = u \tag{4.18b}$$

with the state $\mathbf{x}^{\mathrm{T}} = \left[\mathbf{x}_1^{\mathrm{T}}, x_2\right] \in \mathbb{R}^{n+1}$*, the control input* $u \in \mathbb{R}$*, and* $\mathbf{x}_0 = \mathbf{x}(0)$*. Assume that a continuously differentiable function* $\alpha(\mathbf{x}_1)$ *with* $\alpha(\mathbf{0}) = 0$ *and a positive definite, radially unbounded function* $V(\mathbf{x}_1)$ *exist such that*

$$\frac{\partial}{\partial \mathbf{x}_1}V\{\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\alpha(\mathbf{x}_1)\} \leq W(\mathbf{x}_1) \leq 0 \tag{4.19}$$

and $\mathbf{f}(\mathbf{x}_1)$ *satisfies* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$*.*

(1) If $W(\mathbf{x}_1)$ *is negative definite, then there exists a state feedback* $u = \alpha_a(\mathbf{x}_1, x_2)$ *such that the equilibrium* $\mathbf{x}_{1,R} = 0$*,* $x_{2,R} = 0$ *of the closed loop system is globally*

asymptotically stable with the Lyapunov function

$$V_a(\mathbf{x}_1, x_2) = V(\mathbf{x}_1) + \frac{1}{2}(x_2 - \alpha(\mathbf{x}_1))^2 \ . \tag{4.20}$$

One possible state feedback is given by

$$\begin{aligned} u = -c(x_2 - \alpha(\mathbf{x}_1)) + \frac{\partial}{\partial \mathbf{x}_1}\alpha(\mathbf{x}_1)\{\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2\} \\ - \frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)\mathbf{g}(\mathbf{x}_1) \ , \qquad c > 0 \ . \end{aligned} \tag{4.21}$$

(2) If $W(\mathbf{x}_1)$ is only negative semidefinite, then there exists a state feedback $u = \alpha_a(\mathbf{x}_1, x_2)$ such that the state variables $\mathbf{x}_1(t)$ and $x_2(t)$ are bounded for all times $t \geq 0$, and the solution of the system converges for $t \to \infty$ to the largest positive invariant set $\mathcal{M}$ of the set

$$\mathcal{Y} = \left\{ \begin{bmatrix} \mathbf{x}_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^{n+1} \middle| W(\mathbf{x}_1) = 0 \quad \textit{und} \quad x_2 = \alpha(\mathbf{x}_1) \right\} \tag{4.22}$$

Proof. Introducing the new state variables $z = x_2 - \alpha(\mathbf{x}_1)$ transforms (4.18) to

$$\dot{\mathbf{x}}_1 = \mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\{z + \alpha(\mathbf{x}_1)\} \tag{4.23a}$$

$$\dot{z} = u - \frac{\partial}{\partial \mathbf{x}_1}\alpha(\mathbf{x}_1)\{\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\{z + \alpha(\mathbf{x}_1)\}\} \ . \tag{4.23b}$$

Substituting the state feedback (4.21) into (4.23), the time derivative of the positive definite, radially unbounded Lyapunov function $V_a(\mathbf{x}_1, x_2)$ from (4.20) satisfies

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V_a &= \frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\{z + \alpha(\mathbf{x}_1)\}) + z\left\{-cz - \frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)\mathbf{g}(\mathbf{x}_1)\right\} \\ &\leq W(\mathbf{x}_1) - cz^2 \ . \end{aligned} \tag{4.24}$$

For $W(\mathbf{x}_1) < 0$, the global asymptotic stability of the equilibrium $\mathbf{x}_{1,R} = 0$, $x_{2,R} = 0$ is thus proven. In the case when $W(\mathbf{x}_1) \leq 0$, according to the invariance principle of Krassovskii-LaSalle (see Theorem 3.4), it follows that

$$\lim_{t \to \infty} \mathbf{\Phi}_t(\mathbf{x}_0) \in \mathcal{M} \tag{4.25}$$

with $\mathcal{M}$ being the largest positive invariant subset of set $\mathcal{Y}$

$$\mathcal{Y} = \left\{ \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^{n+1} \middle| \frac{\mathrm{d}}{\mathrm{d}t}V_a = 0 \quad \text{bzw.} \quad W(\mathbf{x}_1) = 0 \quad \text{und} \quad x_2 = \alpha(\mathbf{x}_1) \right\}, \tag{4.26}$$

which concludes the proof of the theorem above. □

> *Exercise* 4.3. Design a nonlinear state feedback using the Integrator Backstepping method for the system
>
> $$\dot{x}_1 = x_1 x_2 \tag{4.27a}$$
> $$\dot{x}_2 = u \ . \tag{4.27b}$$

Satz 4.1 can now be extended to systems with a chain of integrators of the form

$$\begin{aligned}\dot{\mathbf{x}}_1 &= \mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2\\ \dot{x}_2 &= x_3\\ \dot{x}_3 &= x_4\\ &\vdots\\ \dot{x}_k &= u \ .\end{aligned} \tag{4.28}$$

Assuming that a continuously differentiable function $\alpha_1(\mathbf{x}_1)$ with $\alpha_1(\mathbf{0}) = 0$ and a positive definite, radially unbounded function $V(\mathbf{x}_1)$ exist such that condition (4.19) is satisfied, and $\mathbf{f}(\mathbf{x}_1)$ satisfies the relationship $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, the function

$$V_a(\mathbf{x}_1, x_2, \ldots, x_k) = V(\mathbf{x}_1) + \frac{1}{2}\sum_{j=2}^{k}(x_j - \alpha_{j-1}(\mathbf{x}_1, x_2, \ldots, x_{j-1}))^2 \tag{4.29}$$

can be assumed as the Lyapunov function of the closed loop. To explain the procedure in more detail, consider the case $k = 3$. The mathematical model (4.28) then reads

$$\dot{\mathbf{x}}_1 = \mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2 \tag{4.30a}$$
$$\dot{x}_2 = x_3 \tag{4.30b}$$
$$\dot{x}_3 = u \tag{4.30c}$$

and the Lyapunov function (4.29) results in

$$V_a(\mathbf{x}_1, x_2, x_3) = V(\mathbf{x}_1) + \frac{1}{2}(x_2 - \alpha_1(\mathbf{x}_1))^2 + \frac{1}{2}(x_3 - \alpha_2(\mathbf{x}_1, x_2))^2 \ . \tag{4.31}$$

In a first step, introduce the state variables

$$z_1 = x_2 - \alpha_1(\mathbf{x}_1) \tag{4.32a}$$
$$z_2 = x_3 - \alpha_2(\mathbf{x}_1, x_2) \tag{4.32b}$$

and calculate the time derivative of the Lyapunov function (4.31) along a solution of the system

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V_a = {}& \frac{\partial V(\mathbf{x}_1)}{\partial \mathbf{x}_1}(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\{z_1 + \alpha_1(\mathbf{x}_1)\})\\ &+ z_1\left(x_3 - \frac{\partial \alpha_1(\mathbf{x}_1)}{\partial \mathbf{x}_1}(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2)\right)\\ &+ z_2\left(u - \frac{\partial}{\partial \mathbf{x}_1}\alpha_2(\mathbf{x}_1, x_2)\{\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2\} - \frac{\partial}{\partial x_2}\alpha_2(\mathbf{x}_1, x_2)x_3\right) \ .\end{aligned} \tag{4.33}$$

Next, considering x_3 in the first row of (4.33) as the input and applying Theorem 4.1 for it, we obtain

$$\begin{aligned} x_3 &= \alpha_2(\mathbf{x}_1, x_2) \\ &= -c_1 z_1 + \frac{\partial}{\partial \mathbf{x}_1}\alpha_1(\mathbf{x}_1)(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2) - \frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)\mathbf{g}(\mathbf{x}_1) \end{aligned} \tag{4.34}$$

with $c_1 > 0$. By replacing $x_3 = z_2 + \alpha_2(\mathbf{x}_1, x_2)$ according to (4.32) in (4.33), we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V_a = {} & \underbrace{\frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)\alpha_1(\mathbf{x}_1))}_{\leq W(\mathbf{x}_1)} - c_1 z_1^2 + z_1 z_2 \\ & + z_2\left(u - \frac{\partial}{\partial \mathbf{x}_1}\alpha_2(\mathbf{x}_1, x_2)\{\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2\} - \frac{\partial}{\partial x_2}\alpha_2(\mathbf{x}_1, x_2)x_3\right) . \end{aligned} \tag{4.35}$$

Applying Theorem 4.1 again to (4.35) with the input u ultimately leads to the state feedback

$$u = -z_1 - c_2 z_2 + \frac{\partial}{\partial \mathbf{x}_1}\alpha_2(\mathbf{x}_1, x_2)(\mathbf{f}(\mathbf{x}_1) + \mathbf{g}(\mathbf{x}_1)x_2) + \frac{\partial}{\partial x_2}\alpha_2(\mathbf{x}_1, x_2)x_3 \tag{4.36}$$

with $c_2 > 0$ and $\alpha_2(\mathbf{x}_1, x_2)$ according to (4.34).

> *Exercise* 4.4. Prove that for a negatively definite $W(\mathbf{x}_1)$, the equilibrium $\mathbf{x}_1 = \mathbf{0}$, $x_2 = x_3 = 0$ is globally asymptotically stable. To which set do the solutions of the system converge if $W(\mathbf{x}_1)$ is only negatively semidefinite?

4.2 Generalized Backstepping

The method of Integrator Backstepping can now be extended to a class of nonlinear systems of the form

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) \tag{4.37a}$$

$$\dot{\mathbf{x}}_2 = \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{u} \tag{4.37b}$$

with the state $\mathbf{x}_1 \in \mathbb{R}^n$, $\mathbf{x}_2 \in \mathbb{R}^p$ and the control input $\mathbf{u} \in \mathbb{R}^p$. Without loss of generality, assume that $\mathbf{x}_{1,R} = \mathbf{0}$, $\mathbf{x}_{2,R} = \mathbf{0}$ is an equilibrium of the free system, i.e., for $\mathbf{u} = \mathbf{0}$. If this is not the case, then a state transformation $\tilde{\mathbf{x}}_1 = \mathbf{x}_1 - \mathbf{x}_{1,R}$ and $\tilde{\mathbf{x}}_2 = \mathbf{x}_2 - \mathbf{x}_{2,R}$ and a control input transformation $\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}_R$ can always be found such that this holds in the new variables.

Theorem 4.2. *Assume there exists a Lyapunov function $V(\mathbf{x}_1)$ and a state feedback $\mathbf{x}_2 = \boldsymbol{\alpha}(\mathbf{x}_1)$ with $\boldsymbol{\alpha}(\mathbf{0}) = \mathbf{0}$ such that the equilibrium $\mathbf{x}_{1,R} = \mathbf{0}$ of the system*

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1)) \tag{4.38}$$

is globally (locally) asymptotically stable. Then, a state feedback $\mathbf{u} = \mathbf{u}(\mathbf{x}_1, \mathbf{x}_2)$ with $\mathbf{u}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$ can always be specified such that the equilibrium $\mathbf{x}_{1,R} = \mathbf{0}$, $\mathbf{x}_{2,R} = \mathbf{0}$ of the closed loop system (4.37) is globally (locally) asymptotically stable.

Proof. The following proof is constructive and thus provides a computational procedure to obtain the state feedback law.

(1) For the Lyapunov function $V(\mathbf{x}_1)$, due to the asymptotic stability of system (4.38), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\mathbf{x}_1) = \frac{\partial}{\partial \mathbf{x}_1}V(\mathbf{x}_1)\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1)) < 0 \; . \tag{4.39}$$

(2) Now, introduce an auxiliary quantity $\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)$ in the form

$$\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2) = \int_0^1 \frac{\partial}{\partial \mathbf{v}}\mathbf{f}_1(\mathbf{x}_1, \mathbf{v})\bigg|_{\mathbf{v}=\boldsymbol{\alpha}(\mathbf{x}_1)+\lambda\mathbf{x}_2} \mathrm{d}\lambda \tag{4.40}$$

such that $\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1) + \mathbf{x}_2)$ can be expressed as follows

$$\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1) + \mathbf{x}_2) = \mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1)) + \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2 \tag{4.41}$$

To show this, multiply (4.40) from the right by $\mathbf{x}_2$ and replace the integrand with the left-hand side of the subsequent expression

$$\begin{aligned}\frac{\partial}{\partial \lambda}\mathbf{f}_1\Bigg(\mathbf{x}_1, \underbrace{\boldsymbol{\alpha}(\mathbf{x}_1) + \lambda\mathbf{x}_2}_{\mathbf{v}}\Bigg) &= \begin{bmatrix} \frac{\partial f_{1,1}(\mathbf{x}_1,\mathbf{v})}{\partial v_1}x_{2,1} + \cdots + \frac{\partial f_{1,1}(\mathbf{x}_1,\mathbf{v})}{\partial v_p}x_{2,p} \\ \vdots \\ \frac{\partial f_{1,n}(\mathbf{x}_1,\mathbf{v})}{\partial v_1}x_{2,1} + \cdots + \frac{\partial f_{1,n}(\mathbf{x}_1,\mathbf{v})}{\partial v_p}x_{2,p} \end{bmatrix} \\ &= \frac{\partial}{\partial \mathbf{v}}\mathbf{f}_1(\mathbf{x}_1, \mathbf{v})\bigg|_{\mathbf{v}=\boldsymbol{\alpha}(\mathbf{x}_1)+\lambda\mathbf{x}_2} \mathbf{x}_2 \; ,\end{aligned} \tag{4.42}$$

which yields

$$\begin{aligned}\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2 &= \int_0^1 \frac{\partial}{\partial \mathbf{v}}\mathbf{f}_1(\mathbf{x}_1, \mathbf{v})\bigg|_{\mathbf{v}=\boldsymbol{\alpha}(\mathbf{x}_1)+\lambda\mathbf{x}_2} \mathbf{x}_2 \, \mathrm{d}\lambda \\ &= \int_0^1 \frac{\partial}{\partial \lambda}\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1) + \lambda\mathbf{x}_2) \, \mathrm{d}\lambda\end{aligned} \tag{4.43}$$

and consequently (4.41)

$$\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2 = \mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1) + \mathbf{x}_2) - \mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1)) \; . \tag{4.44}$$

(3) The state feedback law

$$\begin{aligned}\mathbf{u}(\mathbf{x}_1, \mathbf{x}_2) = &-\mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) + \frac{\partial \boldsymbol{\alpha}(\mathbf{x}_1)}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) \\ &- \left[\frac{\partial V(\mathbf{x}_1)}{\partial \mathbf{x}_1}\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))\right]^{\mathrm{T}} \\ &- c(\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)), \qquad c > 0\end{aligned} \tag{4.45}$$

guarantees the asymptotic stability of the equilibrium of the closed loop system. The candidate for the Lyapunov function of the closed loop system is the positive definite function

$$V_a(\mathbf{x}_1, \mathbf{x}_2) = V(\mathbf{x}_1) + \frac{1}{2}\|\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)\|_2^2 \tag{4.46}$$

The time derivative of V_a along a solution of the system is

$$\frac{\mathrm{d}}{\mathrm{d}t}V_a(\mathbf{x}_1, \mathbf{x}_2) = \begin{bmatrix} \frac{\partial V_a}{\partial \mathbf{x}_1} & \frac{\partial V_a}{\partial \mathbf{x}_2} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) \\ \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{u} \end{bmatrix} \tag{4.47}$$

Substituting $\mathbf{u}(\mathbf{x}_1, \mathbf{x}_2)$ and $V_a(\mathbf{x}_1, \mathbf{x}_2)$ from (4.45) and (4.46) into the equations, we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V_a = & \frac{\partial V}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) + (\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))^{\mathrm{T}}\Bigg\{-\frac{\partial \boldsymbol{\alpha}(\mathbf{x}_1)}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) \\ & - \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) + \frac{\partial \boldsymbol{\alpha}(\mathbf{x}_1)}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) \\ & - \left[\frac{\partial V(\mathbf{x}_1)}{\partial \mathbf{x}_1}\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))\right]^{\mathrm{T}} - c(\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))\Bigg\} \\ = & \frac{\partial V}{\partial \mathbf{x}_1}\{\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) - \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))(\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))\} \\ & - c\|\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)\|_2^2 \; . \end{aligned} \tag{4.48}$$

Replacing $\mathbf{x}_2$ with $\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)$ in (4.44), we get

$$\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1))(\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)) = \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) - \mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1)) \tag{4.49}$$

Hence, for (4.48) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}V_a = \underbrace{\frac{\partial V}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \boldsymbol{\alpha}(\mathbf{x}_1))}_{=\frac{\mathrm{d}}{\mathrm{d}t}V(\mathbf{x}_1)<0} - c\|\mathbf{x}_2 - \boldsymbol{\alpha}(\mathbf{x}_1)\|_2^2 < 0 \; . \tag{4.50}$$

Thus, Theorem 4.2 is proven.

□

As an application example, consider the *active damping system* of a vehicle shown in Figure 4.1.

A hydraulic actuator is mounted in parallel to a spring-damper system with the spring constant k_s and the damping constant d_s between the vehicle chassis and the suspension. The inflow q of oil into the hydraulic actuator can be adjusted via a current-controlled servo valve. The dynamics of the servo valve are approximated by a first-order time delay

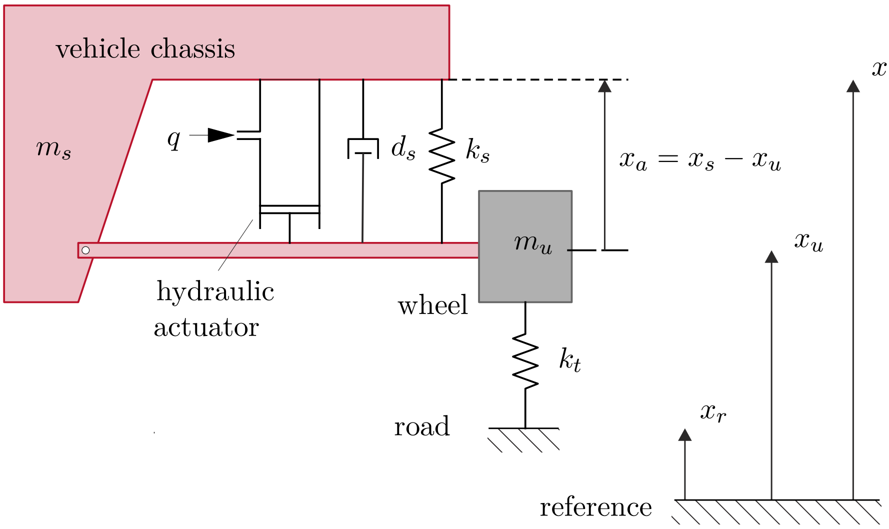


Figure 4.1: Active vehicle damping system.

element in the form

$$\dot{x}_v = -c_v x_v + k_v i_v, \qquad c_v,\, k_v > 0 \tag{4.51}$$

describing the spool position x_v and the servo current as input i_v. The oil flow q then results from the relationship (compare to (1.49))

$$q = \begin{cases} K_{v,1}\sqrt{p_S - p}x_v & \text{for} \quad x_v \geq 0 \\ K_{v,2}\sqrt{p - p_T}x_v & \text{for} \quad x_v \leq 0 \end{cases} \tag{4.52}$$

with the tank pressure p_T, the supply pressure p_S, the pressure in the cylinder p, and the valve coefficients $K_{v,1}$ and $K_{v,2}$. For simplicity, assuming the oil is incompressible, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}p = 0$, and neglecting the leakage oil flows, (4.51) and (4.52) can be written as follows

$$\frac{\dot{q}}{K_{v,1}\sqrt{p_S - p}} = -c_v \frac{q}{K_{v,1}\sqrt{p_S - p}} + k_v i_v, \qquad x_v \geq 0 \tag{4.53a}$$

$$\frac{\dot{q}}{K_{v,2}\sqrt{p - p_T}} = -c_v \frac{q}{K_{v,2}\sqrt{p - p_T}} + k_v i_v, \qquad x_v \leq 0 \tag{4.53b}$$

The state feedback, also called *servo compensation*,

$$i_v = \begin{cases} \dfrac{i_v^*}{K_{v,1}\sqrt{p_S - p}} & \text{for} \quad x_v \geq 0 \\ \dfrac{i_v^*}{K_{v,2}\sqrt{p - p_T}} & \text{for} \quad x_v \leq 0 \end{cases} \tag{4.54}$$

with the new input i_v^* then leads to the differential equation for the oil flow

$$\dot{q} = -c_v q + k_v i_v^* \ . \tag{4.55}$$

Furthermore, due to the assumption of oil incompressibility, the relation

$$\dot{x}_a = \frac{q}{A} \tag{4.56}$$

holds with the piston area A. Now, a damping behavior of the form

$$q = \alpha(x_a) = -A\Big(d_1 x_a + d_2 x_a^3\Big), \qquad d_1,\, d_2 > 0 \ , \tag{4.57}$$

is desired, where for small displacements ($x_a \ll$) a linear behavior is assumed (x_a^3 is negligible compared to x_a), and for larger displacements, damping proportional to the third power of x_a is considered. This allows the application of the backstepping method from Theorem 4.2 with $n = p = 1$, $\mathbf{x}_1 = x_a$, $\mathbf{x}_2 = q$, $\mathbf{u} = k_v i_v^*$, $\mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{q}{A}$, and $\mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2) = -c_v q$:

(1) The equilibrium $x_a = 0$ of the system (4.56) with the fictitious state feedback (4.57) is asymptotically stable, which can be directly shown with the Lyapunov function

$$V(x_a) = \frac{1}{2} x_a^2 \tag{4.58}$$

and its time derivative along a solution of the system

$$\frac{\mathrm{d}}{\mathrm{d}t} V(x_a) = -\Big(d_1 x_a^2 + d_2 x_a^4\Big) < 0 \tag{4.59}$$

(2) In this case, the auxiliary quantity (4.40) reads

$$G(x_a, q) = \int_0^1 \frac{\partial}{\partial q}\left(\frac{q}{A}\right)\bigg|_{q = \boldsymbol{\alpha}(x_a) + \lambda q} \mathrm{d}\lambda = \frac{1}{A} \ . \tag{4.60}$$

(3) The state feedback according to (4.45) is given by

$$k_v i_v^* = c_v q + \frac{\partial \alpha(x_a)}{\partial x_a} \frac{q}{A} - \frac{\partial V(x_a)}{\partial x_a} \frac{1}{A} - c(q - \alpha(x_a)), \qquad c > 0 \tag{4.61}$$

or with the choice $c = c_v$ we obtain

$$i_v^* = \frac{1}{k_v}\left(-c_v A\Big(d_1 x_a + d_2 x_a^3\Big) - \Big(d_1 + 3 d_2 x_a^2\Big) q - x_a \frac{1}{A}\right) \ . \tag{4.62}$$

As one can easily verify,

$$V_a(x_a, q) = \underbrace{\frac{1}{2} x_a^2}_{V(x_a)} + \frac{1}{2}\left(q + \underbrace{A\Big(d_1 x_a + d_2 x_a^3\Big)}_{-\alpha(x_a)}\right)^2 \tag{4.63}$$

is the corresponding Lyapunov function of the closed loop system given by (4.46).

Therefore, the state feedback for the servo current command of the servo valve consists of (4.54) and (4.62).

> *Exercise* 4.5. Given is the mathematical model (1.15) of the rotational motion of a satellite as shown in Figure 1.1
>
> $$\Theta_{11}\dot{\omega}_1 = -(\Theta_{33} - \Theta_{22})\omega_2\omega_3 + M_1 \tag{4.64a}$$
> $$\Theta_{22}\dot{\omega}_2 = -(\Theta_{11} - \Theta_{33})\omega_1\omega_3 + M_2 \tag{4.64b}$$
> $$\Theta_{33}\dot{\omega}_3 = -(\Theta_{22} - \Theta_{11})\omega_1\omega_2 + M_3 \tag{4.64c}$$
>
> with the angular velocities ω_1, ω_2, ω_3, the moments of inertia Θ_{11}, Θ_{22}, Θ_{33}, and the moments M_1, M_2, and M_3 around the principal axes of inertia.
>
> (1) In a first step, design a controller using the Computed-Torque method according to Section 4.5 so that the equilibrium $\omega_{1,R} = \omega_{2,R} = \omega_{3,R} = 0$ is asymptotically stabilized.
>
> (2) Now assume that the cold gas thrusters in the x_3 axis have failed, i.e., $M_3 = 0$. Design a state feedback controller according to Theorem 4.2 in such a way that for this case, the equilibrium of the closed loop system $\omega_{1,R} = \omega_{2,R} = \omega_{3,R} = 0$ remains globally asymptotically stable. Why can the Computed-Torque method no longer be applied here?

4.3 Adaptive Control

In this section, some basic concepts of Lyapunov-based adaptive control are discussed using simple examples. To illustrate the idea, consider the simple nonlinear system

$$\dot{x} = u + \theta\varphi(x) \tag{4.65}$$

with the state $x \in \mathbb{R}$, the control input $u \in \mathbb{R}$, and the unknown but constant parameter $\theta \in \mathbb{R}$. Assuming in a first step that the parameter θ is known, the equilibrium $x = 0$ is asymptotically stabilized by the state feedback

$$u = -\theta\varphi(x) - c_1 x, \qquad \text{with} \qquad c_1 > 0\ . \tag{4.66}$$

A possible Lyapunov function is given by

$$V(x) = \frac{1}{2}x^2 > 0, \qquad \dot{V}(x) = -c_1 x^2 < 0\ . \tag{4.67}$$

Substituting an estimated value $\hat{\theta}$ for the unknown parameter θ in the state feedback (4.66), the change of $V(x) = \frac{1}{2}x^2$ along a solution curve of the closed loop system is given by

$$\dot{x} = -c_1 x - \hat{\theta}\varphi(x) + \theta\varphi(x) = -c_1 x - \underbrace{\left(\hat{\theta} - \theta\right)}_{=\tilde{\theta}}\varphi(x)\ . \tag{4.68}$$

The expression for the change of $V(x) = \frac{1}{2}x^2$ along a solution curve of the closed loop system is

$$\dot{V}(x) = -c_1 x^2 - \tilde{\theta}\varphi(x)x \ . \tag{4.69}$$

To eliminate the indefinite term in the *estimation error* $\tilde{\theta}$, the Lyapunov function is extended by an additional quadratic term

$$V_e\Big(x, \tilde{\theta}\Big) = V(x) + \frac{1}{2\gamma}\tilde{\theta}^2 = \frac{1}{2}x^2 + \frac{1}{2\gamma}\tilde{\theta}^2 > 0, \qquad \gamma > 0 \tag{4.70}$$

and the change of $V_e\Big(x, \tilde{\theta}\Big)$ along a solution curve of (4.68) is calculated as

$$\dot{V}_e\Big(x, \tilde{\theta}\Big) = -c_1 x^2 + \tilde{\theta}\Big(-\varphi(x)x + \frac{1}{\gamma}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\theta}\Big) \ . \tag{4.71}$$

The differential equation of the estimated value $\hat{\theta}$ is then determined such that the bracketed expression in (4.71) vanishes, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\theta} = \frac{\mathrm{d}}{\mathrm{d}t}\Big(\hat{\theta} - \theta\Big) = \frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta} = \gamma\varphi(x)x \ , \tag{4.72}$$

resulting in $\dot{V}_e\Big(x, \tilde{\theta}\Big)$ as

$$\dot{V}_e\Big(x, \tilde{\theta}\Big) = -c_1 x^2 \leq 0 \tag{4.73}$$

From Theorem 3.4, it is immediately clear that $\lim_{t\to\infty} x(t) = 0$.

The assumption that the (nonlinear) state feedback stabilizes the system for known parameters θ is also referred to in the literature as the *certainty equivalence property*, which is essential for a variety of adaptive controller design methods. Furthermore, it is easy to see that the unknown parameter θ affects the system (4.65) in the same way as the control input u, and thus the effect of the term $\theta\varphi(x)$ can be easily compensated for known θ through the control input. This structural property is also known in the literature as the *matching condition*. In the next part of this section, it will be shown that the design of the parameter estimator still analogous even when the matching condition is violated to the extent that the control input u affects the system with the unknown θ only after one integrator. In this context, it is also referred to as the *extended matching condition*. Hence, the associated system with the extended matching condition for the parameter θ takes the form

$$\dot{x}_1 = x_2 + \theta\varphi(x_1) \tag{4.74a}$$

$$\dot{x}_2 = u \ . \tag{4.74b}$$

In the first step, design a state feedback using the simple integrator backstepping method assuming that the parameter θ is known (certainty equivalence property). For the fictitious control input

$$x_2 = -\theta\varphi(x_1) - c_1 x_1, \qquad c_1 > 0 \tag{4.75}$$

the asymptotic stability of the equilibrium $x_1 = 0$ of the first subsystem immediately follows with the Lyapunov function

$$V(x_1) = \frac{1}{2}x_1^2 > 0, \qquad \dot{V}(x_1) = -c_1 x_1^2 < 0\ . \tag{4.76}$$

Setting the Lyapunov function of the overall system as

$$V_a(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{1}{2}(x_2 + \theta\varphi(x_1) + c_1 x_1)^2 \tag{4.77}$$

and calculating the control input u from

$$\begin{aligned}\dot{V}_a(x_1, x_2) = {} & \underbrace{x_1(x_2 + \theta\varphi(x_1))}_{=-c_1x_1^2+(x_2+\theta\varphi(x_1)+c_1x_1)x_1} + (x_2 + \theta\varphi(x_1) + c_1 x_1)\\ & \times\left(u + \left(\theta\frac{\partial}{\partial x_1}\varphi(x_1) + c_1\right)(x_2 + \theta\varphi(x_1))\right)\\ = {} & -c_1 x_1^2 + (x_2 + \theta\varphi(x_1) + c_1 x_1)\\ & \times \underbrace{\left(u + \left(\theta\frac{\partial}{\partial x_1}\varphi(x_1) + c_1\right)(x_2 + \theta\varphi(x_1)) + x_1\right)}_{=-c_2(x_2+\theta\varphi(x_1)+c_1x_1), \quad c_2>0}\end{aligned} \tag{4.78}$$

yields

$$u = -\left(\theta\frac{\partial}{\partial x_1}\varphi(x_1) + c_1\right)(x_2 + \theta\varphi(x_1)) - x_1 - c_2(x_2 + \theta\varphi(x_1) + c_1 x_1)\ . \tag{4.79}$$

To calculate the state feedback and the parameter estimator for a constant but unknown parameter θ, the following Lyapunov function

$$V_e\left(x_1, x_2, \tilde{\theta}\right) = \frac{1}{2}x_1^2 + \frac{1}{2}\left(x_2 + \hat{\theta}\varphi(x_1) + c_1 x_1\right)^2 + \frac{1}{2\gamma}\tilde{\theta}^2, \qquad \gamma > 0 \tag{4.80}$$

with the parameter estimation error $\tilde{\theta} = \hat{\theta} - \theta$ is used. The time derivative of $V_a\left(x_1, x_2, \tilde{\theta}\right)$ is given by

$$
\begin{aligned}
\dot{V}_e = {} & \underbrace{x_1(x_2+\theta\varphi(x_1))}_{=-c_1x_1^2+\big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\big)x_1-\tilde{\theta}\varphi(x_1)x_1} + \Big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\Big) \\
& \times \left(u+\left(\hat{\theta}\frac{\partial}{\partial x_1}\varphi(x_1)+c_1\right)(x_2+\theta\varphi(x_1))+\varphi(x_1)\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}\right)+\frac{1}{\gamma}\tilde{\theta}\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta} \\
= {} & -c_1x_1^2+\Big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\Big) \\
& \underbrace{\times\left(u+\left(\hat{\theta}\frac{\partial}{\partial x_1}\varphi(x_1)+c_1\right)\Big(x_2+\hat{\theta}\varphi(x_1)\Big)+x_1+\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}\varphi(x_1)\right)}_{=-c_2\left(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\right),\quad c_2>0} \\
& +\tilde{\theta}\underbrace{\left(-\varphi(x_1)x_1+\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}\frac{1}{\gamma}-\Big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\Big)\left(\hat{\theta}\frac{\partial}{\partial x_1}\varphi(x_1)+c_1\right)\varphi(x_1)\right)}_{=0}\;.
\end{aligned}
\tag{4.81}
$$

The state feedback and the parameter estimator then follow as

$$
u=-\left(\hat{\theta}\frac{\partial}{\partial x_1}\varphi(x_1)+c_1\right)\Big(x_2+\hat{\theta}\varphi(x_1)\Big)-x_1-\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}\varphi(x_1)-c_2\Big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\Big)
\tag{4.82}
$$

and

$$
\frac{\mathrm{d}}{\mathrm{d}t}\hat{\theta}=\gamma\varphi(x_1)\left(x_1+\Big(x_2+\hat{\theta}\varphi(x_1)+c_1x_1\Big)\left(\hat{\theta}\frac{\partial}{\partial x_1}\varphi(x_1)+c_1\right)\right)\;.
\tag{4.83}
$$

As an application example, consider the mathematical model of a simplified biochemical process of the form

$$
\dot{x}_1=[\varphi_0(x_2)+\theta_1\varphi_1(x_2)+\theta_2\varphi_2(x_2)]x_1-Dx_1 \tag{4.84a}
$$

$$
\dot{x}_2=-k[\varphi_0(x_2)+\theta_1\varphi_1(x_2)+\theta_2\varphi_2(x_2)]x_1-Dx_2+u \tag{4.84b}
$$

with x_1 as the concentration of the bacterial population, x_2 as the concentration of the substrate, the specific growth rate $\mu(x_2)=[\varphi_0(x_2)+\theta_1\varphi_1(x_2)+\theta_2\varphi_2(x_2)]$ with the unknown but constant parameters θ_1 and θ_2, the substrate feed rate u as the input, and the system parameters D and k. Note that both the state variables x_1 and x_2 as well as the specific growth rate $\mu(x_2)$ are always non-negative. The task of control is now to regulate the concentration of the bacterial population x_1 to a predetermined reference value $x_{1,d}$.

In the first step, one performs a regular state transformation of the form

$$
z_1=\ln(x_1)-\ln(x_{1,d}) \quad \text{bzw.} \quad x_1=x_{1,d}\exp(z_1) \tag{4.85a}
$$

$$
z_2=x_2 \quad \text{bzw.} \quad x_2=z_2 \tag{4.85b}
$$

and the system (4.84) in the new state $\mathbf{z}^{\mathrm{T}} = [z_1, z_2]$ reads

$$\begin{aligned}\dot{z}_1 &= [\varphi_0(z_2) + \theta_1\varphi_1(z_2) + \theta_2\varphi_2(z_2)] - D && \text{(4.86a)}\\ \dot{z}_2 &= -k[\varphi_0(z_2) + \theta_1\varphi_1(z_2) + \theta_2\varphi_2(z_2)]x_{1,d}\exp(z_1) - Dz_2 + u\ . && \text{(4.86b)}\end{aligned}$$

If one interprets $\varphi_0(z_2)$ as a fictitious input in the first differential equation of (4.86), it can be easily verified that the control law

$$\varphi_0(z_2) = -\theta_1\varphi_1(z_2) - \theta_2\varphi_2(z_2) + D - c_1 z_1, \qquad c_1 > 0 \tag{4.87}$$

asymptotically stabilizes the desired equilibrium $z_{1,d} = 0$ ($x_1 = x_{1,d}$). In this context, one chooses the Lyapunov function as

$$V(z_1) = \frac{1}{2}z_1^2 > 0, \qquad \dot{V}(z_1) = -c_1 z_1^2 < 0\ . \tag{4.88}$$

To derive the state feedback and the parameter estimator for $\boldsymbol{\theta}^{\mathrm{T}} = [\theta_1, \theta_2]$, one chooses a similar Lyapunov function as shown before, i.e.,

$$V_e\Big(\mathbf{z}, \tilde{\boldsymbol{\theta}}\Big) = \frac{1}{2}z_1^2 + \frac{1}{2}\Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)^2 + \frac{1}{2}\tilde{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\tilde{\boldsymbol{\theta}} \tag{4.89a}$$

with

$$\hat{\boldsymbol{\theta}}^{\mathrm{T}} = \Big[\hat{\theta}_1, \hat{\theta}_2\Big], \qquad \boldsymbol{\varphi}_{12}(z_2) = \begin{bmatrix}\varphi_1(z_2)\\ \varphi_2(z_2)\end{bmatrix}, \qquad \tilde{\boldsymbol{\theta}} = \begin{bmatrix}\tilde{\theta}_1\\ \tilde{\theta}_2\end{bmatrix} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \tag{4.89b}$$

and the positive definite matrix $\boldsymbol{\Gamma}$. The change of the Lyapunov function $V_e(\mathbf{z}, \tilde{\boldsymbol{\theta}})$ along a solution of the system (4.86) is calculated as

$$
\begin{aligned}
\dot{V}_e\Big(\mathbf{z},\tilde{\boldsymbol{\theta}}\Big) &= z_1\Big(\varphi_0(z_2) + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D\Big) + \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big) \\
&\quad \times \bigg(\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg)\dot{z}_2 + c_1\dot{z}_1 + \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2)\bigg) + \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}} \\
&= z_1\bigg(\Big[\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big] - c_1 z_1 - \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2)\bigg) \\
&\quad + \bigg(\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg)\dot{z}_2 + c_1\dot{z}_1 + \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2)\bigg) \\
&\quad \times \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big) + \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}} \\
&= -c_1 z_1^2 + \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)\bigg(\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg)\dot{z}_2 \\
&\quad + c_1\dot{z}_1 + \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) + z_1\bigg) + \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\bigg(-z_1\boldsymbol{\varphi}_{12}(z_2) + \boldsymbol{\Gamma}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}}\bigg) \\
&= -c_1 z_1^2 + \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)\bigg\{\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg) \\
&\quad \times \Bigg(-k\Bigg[\varphi_0(z_2) + \underbrace{\boldsymbol{\theta}^{\mathrm{T}}}_{=\hat{\boldsymbol{\theta}}^{\mathrm{T}}-\tilde{\boldsymbol{\theta}}^{\mathrm{T}}}\boldsymbol{\varphi}_{12}(z_2)\Bigg]x_{1,d}\exp(z_1) - Dz_2 + u\Bigg) \\
&\quad + c_1\Bigg(\Bigg[\varphi_0(z_2) + \underbrace{\boldsymbol{\theta}^{\mathrm{T}}}_{=\hat{\boldsymbol{\theta}}^{\mathrm{T}}-\tilde{\boldsymbol{\theta}}^{\mathrm{T}}}\boldsymbol{\varphi}_{12}(z_2)\Bigg] - D\Bigg) + \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) + z_1\Bigg\} \\
&\quad + \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\bigg(-z_1\boldsymbol{\varphi}_{12}(z_2) + \boldsymbol{\Gamma}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}}\bigg) \\
&= -c_1 z_1^2 + \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)\underline{\bigg\{\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg)} \\
&\quad \times \underline{\bigg(-k\Big[\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2)\Big]x_{1,d}\exp(z_1) - Dz_2 + u\bigg)} \\
&\quad \underline{+ c_1\bigg(\Big[\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2)\Big] - D\bigg) + \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) + z_1\bigg\}} \\
&\quad + \tilde{\boldsymbol{\theta}}^{\mathrm{T}}\underline{\underline{\bigg\{-z_1\boldsymbol{\varphi}_{12}(z_2) + \boldsymbol{\Gamma}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}} + \Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)}} \\
&\quad \underline{\underline{\times\bigg[\bigg(\frac{\partial}{\partial z_2}\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\frac{\partial}{\partial z_2}\boldsymbol{\varphi}_{12}(z_2)\bigg)k\boldsymbol{\varphi}_{12}(z_2)x_{1,d}\exp(z_1) - c_1\boldsymbol{\varphi}_{12}(z_2)\bigg]\bigg\}}}\;.
\end{aligned}
\tag{4.90}
$$

Exercise 4.6. Calculate the relation (4.90).

> **Tip:** Take your time for this task.

The state feedback is obtained by setting the simply underlined expression in (4.90) equal to $-c_2\Big(\varphi_0(z_2) + \hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\varphi}_{12}(z_2) - D + c_1 z_1\Big)$, where $c_2 > 0$, and the parameter estimator follows directly by setting to zero the double underlined expression in (4.90) and the fact that $\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\boldsymbol{\theta}} = \frac{\mathrm{d}}{\mathrm{d}t}\hat{\boldsymbol{\theta}}$.

4.4 PD control law for rigid body systems

If $\mathbf{q}^{\mathrm{T}} = [q_1, q_2, \ldots, q_n]$ denotes the generalized coordinates of a mechanical rigid body system, then the equations of motion are obtained from the so-called Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{q}_k}L\right) - \frac{\partial}{\partial q_k}L = \tau_k \ , \qquad k = 1, \ldots, n \tag{4.91}$$

with the generalized velocities $\dot{\mathbf{q}} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}$, the generalized forces or moments $\boldsymbol{\tau}^{\mathrm{T}} = [\tau_1, \tau_2, \ldots, \tau_n]$, and the Lagrangian L. For rigid body systems, the Lagrangian always results from the difference between kinetic and potential energy, that is, $L = T - V$. Under the assumption that

(1) the kinetic energy T can be expressed as a quadratic function of the generalized velocities $\dot{\mathbf{q}}$ in the form

$$T = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}(\mathbf{q})\dot{q}_i\dot{q}_j = \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\mathbf{D}(\mathbf{q})\dot{\mathbf{q}} \tag{4.92}$$

with the symmetric, positive definite generalized mass matrix $\mathbf{D}(\mathbf{q})$, and

(2) the potential energy $V(\mathbf{q})$ is independent of $\dot{\mathbf{q}}$,

the equations of motion (4.91) can be written in the form

$$\mathbf{D}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \tag{4.93}$$

To show this, substitute T from (4.92) and $V(\mathbf{q})$ into the Euler-Lagrange equations (4.91) and with

$$\frac{\partial}{\partial \dot{q}_k}L = \sum_{j=1}^{n} d_{kj}(\mathbf{q})\dot{q}_j \ , \tag{4.94a}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial}{\partial \dot{q}_k}L\right) &= \sum_{j=1}^{n} d_{kj}(\mathbf{q})\ddot{q}_j + \sum_{j=1}^{n}\frac{\mathrm{d}}{\mathrm{d}t}d_{kj}(\mathbf{q})\dot{q}_j \\ &= \sum_{j=1}^{n} d_{kj}(\mathbf{q})\ddot{q}_j + \sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial}{\partial q_i}d_{kj}(\mathbf{q})\dot{q}_i\dot{q}_j \ ,\end{aligned} \tag{4.94b}$$

$$\frac{\partial}{\partial q_k}L = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial}{\partial q_k}d_{ij}(\mathbf{q})\dot{q}_i\dot{q}_j - \frac{\partial}{\partial q_k}V \tag{4.94c}$$

(4.91) finally simplifies to

$$\sum_{j=1}^{n} d_{kj}(\mathbf{q})\ddot{q}_j + \underbrace{\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial q_i}d_{kj}(\mathbf{q}) - \frac{1}{2}\frac{\partial}{\partial q_k}d_{ij}(\mathbf{q})\right)\dot{q}_i\dot{q}_j}_{B} + \frac{\partial}{\partial q_k}V = \tau_k \ . \tag{4.95}$$

Now, writing for

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial}{\partial q_i}d_{kj}(\mathbf{q})\dot{q}_i\dot{q}_j = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial q_i}d_{kj}(\mathbf{q}) + \frac{\partial}{\partial q_j}d_{ki}(\mathbf{q})\right)\dot{q}_i\dot{q}_j \ , \tag{4.96}$$

the term B from (4.95) follows as

$$B = \sum_{j=1}^{n}\sum_{i=1}^{n}\underbrace{\frac{1}{2}\left(\frac{\partial}{\partial q_i}d_{kj}(\mathbf{q}) + \frac{\partial}{\partial q_j}d_{ki}(\mathbf{q}) - \frac{\partial}{\partial q_k}d_{ij}(\mathbf{q})\right)}_{c_{ijk}(\mathbf{q})}\dot{q}_i\dot{q}_j \ , \tag{4.97}$$

where the terms $c_{ijk}(\mathbf{q})$ are referred to as *Christoffel symbols of the first kind*. Furthermore, if we set $\frac{\partial V}{\partial q_k}(\mathbf{q}) = g_k(\mathbf{q})$, then from (4.95) and (4.97) we immediately obtain the equations of motion in the form

$$\sum_{j=1}^{n} d_{kj}(\mathbf{q})\ddot{q}_j + \sum_{j=1}^{n}\sum_{i=1}^{n} c_{ijk}(\mathbf{q})\dot{q}_i\dot{q}_j + g_k(\mathbf{q}) = \tau_k \ . \tag{4.98}$$

As can be seen, the equations of motion (4.98) contain three different terms - those involving the second derivative of the generalized coordinates (*acceleration terms*), those where the product $\dot{q}_i\dot{q}_j$ appears (*centrifugal terms* for $i = j$ and *Coriolis terms* for $i \neq j$), and those that depend solely on $\mathbf{q}$ (*potential forces*). As stated above, the equations of motion can thus be written in matrix form

$$\mathbf{D}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \tag{4.99}$$

with the (k, j)-th element of the matrix $\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})$ given by

$$\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})[k,j] = \sum_{i=1}^{n} c_{ijk}(\mathbf{q})\dot{q}_i \tag{4.100}$$

Exercise 4.7. Transform the mathematical models from Exercise 1.6 and 1.7 into the structure of (4.99).

For stability considerations, the following essential theorem now applies:

Theorem 4.3. *The matrix*

$$\mathbf{N}(\mathbf{q},\dot{\mathbf{q}}) = \dot{\mathbf{D}}(\mathbf{q}) - 2\mathbf{C}(\mathbf{q},\dot{\mathbf{q}}) \tag{4.101}$$

is skew-symmetric, i.e.,

$$n_{jk}(\mathbf{q},\dot{\mathbf{q}}) = -n_{kj}(\mathbf{q},\dot{\mathbf{q}}) \ . \tag{4.102}$$

Proof. To prove this, consider the (j,k)-th component of the matrix $\mathbf{N}(\mathbf{q},\dot{\mathbf{q}})$ in the form

$$\begin{aligned} n_{jk} &= \sum_{i=1}^{n}\left(\frac{\partial}{\partial q_i}d_{jk}(\mathbf{q}) - 2c_{ikj}(\mathbf{q})\right)\dot{q}_i \\ &= \sum_{i=1}^{n}\left(\frac{\partial}{\partial q_i}d_{jk}(\mathbf{q}) - \frac{\partial}{\partial q_i}d_{jk}(\mathbf{q}) - \frac{\partial}{\partial q_k}d_{ji}(\mathbf{q}) + \frac{\partial}{\partial q_j}d_{ik}(\mathbf{q})\right)\dot{q}_i \end{aligned} \tag{4.103}$$

then it follows

$$n_{jk} = \sum_{i=1}^{n}\left(-\frac{\partial}{\partial q_k}d_{ji}(\mathbf{q}) + \frac{\partial}{\partial q_j}d_{ik}(\mathbf{q})\right)\dot{q}_i \tag{4.104}$$

or by interchanging the indices j and k

$$n_{kj} = \sum_{i=1}^{n}\left(-\frac{\partial}{\partial q_j}d_{ki}(\mathbf{q}) + \frac{\partial}{\partial q_k}d_{ij}(\mathbf{q})\right)\dot{q}_i \tag{4.105}$$

and taking into account the symmetry of the mass matrix $\mathbf{D}(\mathbf{q})$, i.e., $d_{ki}(\mathbf{q}) = d_{ik}(\mathbf{q})$, we immediately obtain the result $n_{jk} = -n_{kj}$. □

In the next step, we will show how a *PD control law* can asymptotically stabilize a constant desired position of the generalized coordinates $\mathbf{q}_d$. For this purpose, a control law of the form

$$\boldsymbol{\tau} = \mathbf{K}_P \underbrace{(\mathbf{q}_d - \mathbf{q})}_{\mathbf{e}_q} - \mathbf{K}_D\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) \tag{4.106}$$

is used with the positive definite matrices $\mathbf{K}_P$ and $\mathbf{K}_D$, where the compensation of the potential forces $\mathbf{g}(\mathbf{q})$ guarantees that $\mathbf{q} = \mathbf{q}_d$ is an equilibrium of the closed loop. With the positive definite function

$$V(\mathbf{q},\dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\mathbf{D}(\mathbf{q})\dot{\mathbf{q}} + \frac{1}{2}\mathbf{e}_q^{\mathrm{T}}\mathbf{K}_P\mathbf{e}_q \tag{4.107}$$

as the Lyapunov function and its time derivative along the solution of the closed loop (4.99) and (4.106)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}V(\mathbf{q},\dot{\mathbf{q}}) &= \dot{\mathbf{q}}^{\mathrm{T}}\mathbf{D}(\mathbf{q})\ddot{\mathbf{q}} + \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\dot{\mathbf{D}}(\mathbf{q})\dot{\mathbf{q}} + \mathbf{e}_q^{\mathrm{T}}\mathbf{K}_P\dot{\mathbf{e}}_q \\ &= \dot{\mathbf{q}}^{\mathrm{T}}(-\mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{K}_P(\mathbf{q}_d - \mathbf{q}) - \mathbf{K}_D\dot{\mathbf{q}}) + \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\dot{\mathbf{D}}(\mathbf{q})\dot{\mathbf{q}} + \mathbf{e}_q^{\mathrm{T}}\mathbf{K}_P\underbrace{\dot{\mathbf{e}}_q}_{-\dot{\mathbf{q}}} \\ &= \underbrace{\dot{\mathbf{q}}^{\mathrm{T}}\left(\frac{1}{2}\dot{\mathbf{D}}(\mathbf{q}) - \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\right)\dot{\mathbf{q}}}_{=0} + \underbrace{\dot{\mathbf{q}}^{\mathrm{T}}\mathbf{K}_P(\mathbf{q}_d - \mathbf{q}) - \mathbf{e}_q^{\mathrm{T}}\mathbf{K}_P\dot{\mathbf{q}}}_{=0} - \dot{\mathbf{q}}^{\mathrm{T}}\mathbf{K}_D\dot{\mathbf{q}} \\ &\leq 0 \end{aligned} \tag{4.108}$$

the asymptotic stability of the desired position $\mathbf{q}_d$ follows directly from the invariance principle of Krassovskii-LaSalle (see Theorem 3.4). It should be noted at this point that this PD control law (4.106) also leads to very good results for slowly varying desired trajectories $\mathbf{q}_d(t)$ (i.e., where $\dot{\mathbf{q}}_d(t)$ is very small).

> *Exercise* 4.8. Design a PD controller for the mechanical systems in Exercise 1.6 and 1.7 according to (4.106). Choose suitable parameters and perform simulations of the closed-loop systems in Matlab/Simulink.

> *Exercise* 4.9. Figure 4.2 shows a robot with three degrees of freedom with rod masses m_i, rod lengths l_i, distances from the rod base to the center of mass l_{ci}, and moments of inertia I_{xxi}, I_{yyi}, I_{zzi} (all cross-moments are assumed to be zero) in the body-fixed coordinate system (x_i, y_i, z_i) for $i = 1, 2, 3$. A mass m_L is attached at the end of the third rod. The three degrees of freedom of the robot are the rotation around the z_1 axis of rod 1, the rotation around the x_2 axis of rod 2, and the rotation around the x_3 axis of rod 3. The action of the actuators is idealized as torque τ_i in the connecting joints.
> Design a PD controller to stabilize a given desired position and simulate the control loop in Matlab/Simulink. Use the following numerical values: $m_1, m_2, m_3, m_\mathrm{L} = 1$ kg, $l_{c1}, l_{c2}, l_{c3} = 1/2$ m, $l_1, l_2, l_3 = 1$ m, $I_{xx1} = I_{yy1} = I_{xx2} = I_{zz2} = I_{xx3} = I_{zz3} = 0.1$ m^4, and $I_{zz1} = I_{yy2} = I_{yy3} = 0.02$ m^4.

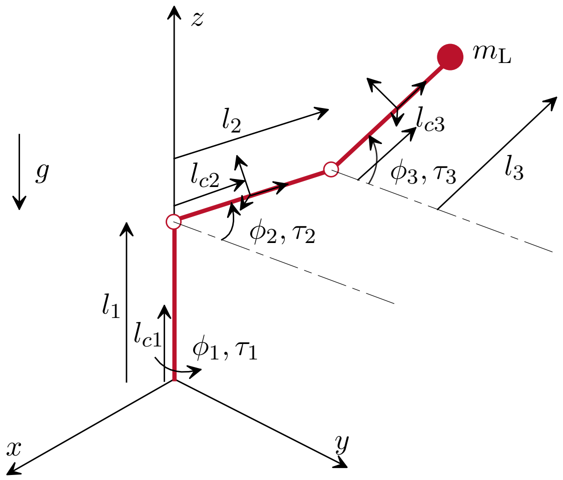


Figure 4.2: Robot with three degrees of freedom.

4.5 Inverse Dynamics (Computed-Torque)

Since the inertia matrix $\mathbf{D}(\mathbf{q})$ in (4.99) is positive definite, it can also be inverted, and thus the *control law of inverse dynamics (Computed-Torque)*

$$\boldsymbol{\tau} = \mathbf{D}(\mathbf{q})\mathbf{v} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) \tag{4.109}$$

leads to a closed loop of the form

$$\ddot{\mathbf{q}} = \mathbf{v} \tag{4.110}$$

with the new input $\mathbf{v}$. One can now specify a controller for $\mathbf{v}$ such that the error system converges globally asymptotically to a trajectory $\mathbf{q}_d(t)$ that is twice continuously differentiable. For this purpose, $\mathbf{v}$ is given in the form

$$\mathbf{v} = \ddot{\mathbf{q}}_d - \mathbf{K}_0 \underbrace{(\mathbf{q} - \mathbf{q}_d)}_{\mathbf{e}_q} - \mathbf{K}_1 \underbrace{(\dot{\mathbf{q}} - \dot{\mathbf{q}}_d)}_{\dot{\mathbf{e}}_q} \tag{4.111}$$

with suitable positive definite diagonal matrices $\mathbf{K}_0$ and $\mathbf{K}_1$, and the error dynamics then reads

$$\ddot{\mathbf{e}}_q + \mathbf{K}_1\dot{\mathbf{e}}_q + \mathbf{K}_0\mathbf{e}_q = \mathbf{0} \ . \tag{4.112}$$

Hence, the error dynamics can be freely adjusted by choosing the matrices $\mathbf{K}_0$ and $\mathbf{K}_1$.

> *Exercise* 4.10. Design a controller for the mechanical systems of exercises 1.6 and 1.7 using the Computed-Torque method according to (4.109) and (4.111). Choose suitable parameters and perform simulations of the closed control loops in MATLAB/SIMULINK. Compare the results with those of exercise 4.8.

It is well known that system parameters such as masses, moments of inertia, etc., are generally not precisely known and therefore cannot be ideally compensated for, as shown in (4.109). However, the rigid body systems of the form (4.99) have the property that a parameter vector $\mathbf{p} \in \mathbb{R}^m$ can always be found in such a way that it appears *linearly* in the equations of motion, i.e.,

$$\mathbf{D}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \mathbf{Y}_0(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}}) + \mathbf{Y}_1(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})\mathbf{p} = \boldsymbol{\tau} \tag{4.113}$$

with an (n,m)-matrix $\mathbf{Y}_1(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})$ and a vector $\mathbf{Y}_0(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})$ consisting of known functions. It should be noted that the entries of the parameter vector $\mathbf{p}$ might themselves depend nonlinearly on the system's masses, lengths, etc. Now, if an estimated value $\hat{\mathbf{p}}$ of the parameter vector $\mathbf{p}$ is substituted into the control law (4.109), then the control law (4.109) and (4.111) becomes

$$\boldsymbol{\tau} = \hat{\mathbf{D}}(\mathbf{q})(\ddot{\mathbf{q}}_d - \mathbf{K}_0\mathbf{e}_q - \mathbf{K}_1\dot{\mathbf{e}}_q) + \hat{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \hat{\mathbf{g}}(\mathbf{q}) \tag{4.114}$$

and the error system (4.112) results in

$$\begin{aligned}\hat{\mathbf{D}}(\mathbf{q})(\ddot{\mathbf{e}}_q + \mathbf{K}_0\mathbf{e}_q + \mathbf{K}_1\dot{\mathbf{e}}_q) = {} & \underbrace{\hat{\mathbf{D}}(\mathbf{q})\ddot{\mathbf{q}} + \hat{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \hat{\mathbf{g}}(\mathbf{q})}_{\mathbf{Y}_0(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})+\mathbf{Y}_1(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})\hat{\mathbf{p}}} \\ & - \left(\underbrace{\mathbf{D}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q})}_{\mathbf{Y}_0(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})+\mathbf{Y}_1(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})\mathbf{p}}\right) .\end{aligned} \tag{4.115}$$

It should be mentioned at this point that the quantities $\mathbf{D}$ and $\hat{\mathbf{D}}$, $\mathbf{C}$ and $\hat{\mathbf{C}}$, as well as $\mathbf{g}$ and $\hat{\mathbf{g}}$ differ only in that the parameter vector $\mathbf{p}$ is replaced by $\hat{\mathbf{p}}$, but their entries remain functionally the same. Assuming the invertibility of $\hat{\mathbf{D}}(\mathbf{q})$, one can ultimately rewrite (4.115) in the form

$$\ddot{\mathbf{e}}_q + \mathbf{K}_0\mathbf{e}_q + \mathbf{K}_1\dot{\mathbf{e}}_q = \hat{\mathbf{D}}(\mathbf{q})^{-1}\mathbf{Y}_1(\mathbf{q},\dot{\mathbf{q}},\ddot{\mathbf{q}})\tilde{\mathbf{p}} = \boldsymbol{\Phi}\tilde{\mathbf{p}} \tag{4.116}$$

or as a first-order differential equation system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{0}_{n,n} & \mathbf{E}_{n,n}\\ -\mathbf{K}_0 & -\mathbf{K}_1\end{bmatrix}}_{\mathbf{A}}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} + \underbrace{\begin{bmatrix}\mathbf{0}_{n,n}\\ \mathbf{E}_{n,n}\end{bmatrix}}_{\mathbf{B}}\boldsymbol{\Phi}\tilde{\mathbf{p}} \tag{4.117}$$

with $\tilde{\mathbf{p}} = \hat{\mathbf{p}} - \mathbf{p}$ and the identity matrix $\mathbf{E}$. Since the matrices $\mathbf{K}_0$ and $\mathbf{K}_1$ were chosen in such a way that the error system is asymptotically stable, the matrix $\mathbf{A}$ is a Hurwitz matrix, and according to Theorem 3.7, for every positive definite matrix $\bar{\mathbf{Q}}$, there exists a unique positive definite solution $\mathbf{P}$ of the Lyapunov equation

$$\mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A} + \bar{\mathbf{Q}} = \mathbf{0}\ . \tag{4.118}$$

To develop an adaptation law for the estimated value $\hat{\mathbf{p}}$ of the parameter $\mathbf{p}$, a Lyapunov function of the form

$$V(\mathbf{e}_q, \dot{\mathbf{e}}_q, \tilde{\mathbf{p}}) = \begin{bmatrix}\mathbf{e}_q^{\mathrm{T}} & \dot{\mathbf{e}}_q^{\mathrm{T}}\end{bmatrix}\mathbf{P}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} + \tilde{\mathbf{p}}^{\mathrm{T}}\boldsymbol{\Gamma}\tilde{\mathbf{p}} \tag{4.119}$$

is assumed with a symmetric, positive definite matrix $\boldsymbol{\Gamma}$, and its time derivative along a solution is calculated

$$\frac{\mathrm{d}}{\mathrm{d}t}V = -\begin{bmatrix}\mathbf{e}_q^{\mathrm{T}} & \dot{\mathbf{e}}_q^{\mathrm{T}}\end{bmatrix}\bar{\mathbf{Q}}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} + 2\tilde{\mathbf{p}}^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{P}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} + \boldsymbol{\Gamma}\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mathbf{p}}\right)\ . \tag{4.120}$$

Assuming that the parameter vector $\mathbf{p}$ is constant (or changes sufficiently slowly compared to the system dynamics in practice) yields the adaptation law

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mathbf{p}} = \frac{\mathrm{d}}{\mathrm{d}t}\hat{\mathbf{p}} = -\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{P}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix}\ , \tag{4.121}$$

which results in (4.120) becoming

$$\frac{\mathrm{d}}{\mathrm{d}t}V = -\begin{bmatrix}\mathbf{e}_q^{\mathrm{T}} & \dot{\mathbf{e}}_q^{\mathrm{T}}\end{bmatrix}\bar{\mathbf{Q}}\begin{bmatrix}\mathbf{e}_q\\ \dot{\mathbf{e}}_q\end{bmatrix} \leq 0\ . \tag{4.122}$$

This immediately demonstrates the stability of the equilibrium of the error system $\mathbf{e}_{q,R} = \dot{\mathbf{e}}_{q,R} = \mathbf{0}$.

To prove asymptotic stability, Barbalat's Lemma is used (see Theorem 3.14). From the fact that $V(\mathbf{e}_q, \dot{\mathbf{e}}_q, \tilde{\mathbf{p}})$ from (4.119) is positive definite and $\frac{\mathrm{d}}{\mathrm{d}t}V$ from (4.122) is negative semidefinite, the boundedness of $\mathbf{e}_q$, $\dot{\mathbf{e}}_q$, and $\tilde{\mathbf{p}}$ directly follows. Assuming that the matrix $\hat{\mathbf{D}}(\mathbf{q})$ remains positive definite and invertible through parameter estimation guarantees that the entries of $\boldsymbol{\Phi}$ in (4.116) are also bounded. From (4.116) and (4.121), it can then be immediately seen that $\ddot{\mathbf{e}}_q$ and $\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\mathbf{p}}$ are bounded. This implies that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}V$ is bounded, and consequently, according to Theorem 3.13, $\frac{\mathrm{d}}{\mathrm{d}t}V$ is uniformly continuous. This allows the application of Barbalat's Lemma, resulting in

$$\lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t}V = 0 \tag{4.123a}$$

or

$$\lim_{t\to\infty} \mathbf{e}_q = \lim_{t\to\infty} \dot{\mathbf{e}}_q = \mathbf{0} \ . \tag{4.123b}$$

One disadvantage of this method is that to calculate $\mathbf{Y}$ from (4.113) or $\boldsymbol{\Phi}$ from (4.116), either the acceleration $\ddot{\mathbf{q}}$ must be measured or approximated by differentiating the velocity $\dot{\mathbf{q}}$. In practice, $\ddot{\mathbf{q}}$ is often simply replaced by $\ddot{\mathbf{q}}_d$.

Exercise 4.11. Design a controller using the Computed-Torque method with parameter adaptation according to (4.114) and (4.121) for the mechanical systems in exercises 1.6 and 1.7. Choose a deviation of +15% from the nominal parameters and simulate the closed-loop systems in MATLAB/SIMULINK. Compare the results with those from exercise 4.10 where the actual parameters deviate by +15% from the nominal values.

Exercise 4.12. Design a trajectory tracking controller using the Computed-Torque method for the three-degree-of-freedom robot shown in Figure 4.2 and perform an adaptation for the end mass m_{Last} according to (4.121). Simulate the closed-loop system in MATLAB/SIMULINK for an end mass $m_{\text{Last}} = 20$ kg. Note that for the nominal value of the end mass, $\hat{m}_{\text{Last}} = 1$ kg.

Exercise 4.13. Show that the *controller according to Slotine and Li*

$$\boldsymbol{\tau} = \mathbf{D}(\mathbf{q})\dot{\mathbf{v}} + \mathbf{C}(\mathbf{q},\dot{\mathbf{q}})\mathbf{v} + \mathbf{g}(\mathbf{q}) - \mathbf{K}_D(\dot{\mathbf{q}} - \mathbf{v}), \quad \mathbf{v} = \dot{\mathbf{q}}_d - \boldsymbol{\Lambda}(\mathbf{q} - \mathbf{q}_d) \tag{4.124}$$

leads to an asymptotically stable error system for $\mathbf{e}_q = \mathbf{q} - \mathbf{q}_d$ with a positive definite diagonal matrix $\boldsymbol{\Lambda}$.

Tip: Introduce the generalized control error

$$\mathbf{s} = \dot{\mathbf{e}}_q + \boldsymbol{\Lambda}\mathbf{e}_q \tag{4.125}$$

as an auxiliary quantity and consider the Lyapunov function

$$V = \frac{1}{2}\mathbf{s}^{\mathrm{T}}\mathbf{D}(\mathbf{q})\mathbf{s} \tag{4.126}$$

4.6 Literatur

[4.1] H. K. Khalil, *Nonlinear Systems (3rd Edition)*. New Jersey: Prentice Hall, 2002.

[4.2] M. Krstić, I. Kanellakopoulos, and P. Kokotović, *Nonlinear and Adaptive Control Design*. New York: John Wiley & Sons, 1995.

[4.3] E. Slotine and W. Li, *Applied Nonlinear Control*. New Jersey: Prentice Hall, 1991.

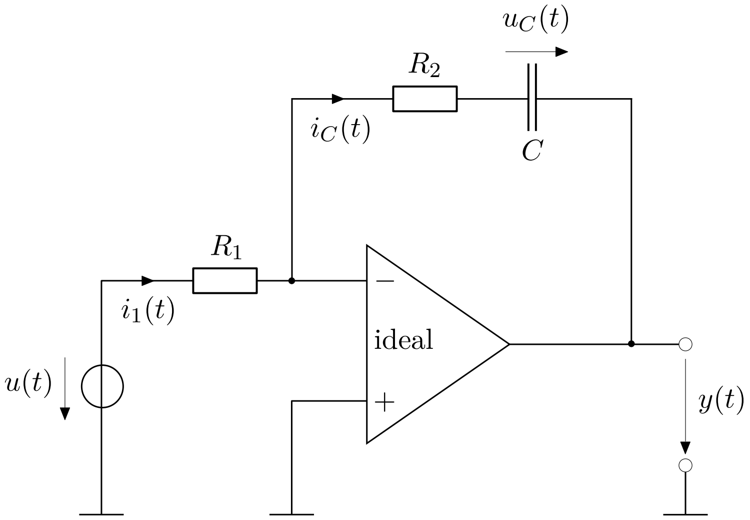
[4.4] E. D. Sontag, *Mathematical Control Theory (2nd Edition)*. New York: Springer, 1998.

[4.5] M. W. Spong, *Robot Dynamics and Control*. New York: John Wiley & Sons, 1989.

[4.6] M. Vidyasagar, *Nonlinear Systems Analysis*. New Jersey: Prentice Hall, 1993.

5 Linear State Space Control

Having discussed concrete strategies for applying Lyapunov theory to nonlinear control systems in Chapter 4, the questions is how such strategies can be applied to real-world systems such as the ones presented in Chapter 1.
For linear time-invariant (LTI) systems in continuous time, the simplest and most traditional approach is to design an analog electronic circuit which directly represents the transfer function $R(s)$ of the desired controller, designed, e.g. using the loop-shaping technique covered in the course *Quantum Technology and Devices: Experimental Techniques and Platforms.* A proportional-integral (PI) controller, as depicted in Figure 5.1, can be designed as follows:

> *Example* 5.1. Due to the ideal operational amplifier, one can directly conclude that in the Laplace domain, $i_1(s) = i_C(s) = u(s)/R_1$. Similarly, as the voltage over the feedback approximately keeps the noninverting input of the operational amplifier at ground level, we have $y(s) = -R_2 i_C(s) - \frac{i_C(s)}{Cs}$. Therefore, the complete circuit has the transfer function
>
> $$R(s) = \frac{y(s)}{u(s)} = V\frac{1+sT}{s} = -\frac{1}{R_1 C}\frac{1+sR_2C}{s} \; . \tag{5.1}$$
>
> 
>
>
> Figure 5.1: Electronic circuit of a PI controller.

The problem now is that this methodology cannot be extended to nonlinear control as the Laplace domain does not lend itself to nonlinear systems. If linearization around a stationary point is not feasible, A different approach has to be taken.

Digital hardware such as programmable logic controllers (PLCs), microcontrollers, or desktop computers could simply evaluate (linear or nonlinear) control laws $\mathbf{g}$ in state space of the form

$$\mathbf{u}(t) = \mathbf{f}_{\mathbf{u}}(\mathbf{x}(t), t) \ , \tag{5.2}$$

provided the system state $\mathbf{x}(t)$ is known. The problem here, however, is that digital hardware is fundamentally constrained by a sample time which is given by the hardware architecture. Thus, (5.2) can only be computed at discrete points in time.
Therefore, the remainder of these lecture notes, unless stated otherwise, will focus on discrete-time systems, which emerge naturally from the sampling of continuous-time systems. There are, however, also examples of dynamical systems that are fundamentally discrete-time systems, for instance in economics [5.1].

5.1 Sampling of Control Systems

Figure 5.2 shows the basic structure of a digital control loop. In contrast to continuous-time control loops, the continuous-time measured variable $\bar{y}(t)$ recorded by the sensor must be discretized in time using an *A/D(Analog/Digital) converter* so that it can be further processed in the digital hardware. In the following, an ideal sensor is usually assumed, i. e. $n(t) = 0$, so that $\bar{y}(t) = y(t)$ applies. An ideal A/D converter, also called a *sample element* or Sampler, then generates a *sequence of samples* $(y_k) = (y_0, y_1, y_2, \ldots)$ $= (y(t_0), y(t_1), y(t_2), \ldots)$ at the times $t_0, t_1, t_2, \ldots$ from a continuous-time signal $y(t)$ (see Figure 5.3). In the following, it is assumed that the sampling is *equidistant*, i. e. the sampling times t_k are integer multiples of the so-called *sampling time* T_a and $t_k = kT_a$, $k = 0, 1, 2, \ldots$. For continuous and right-continuous functions $y(t)$, the sample element thus satisfies the relationship

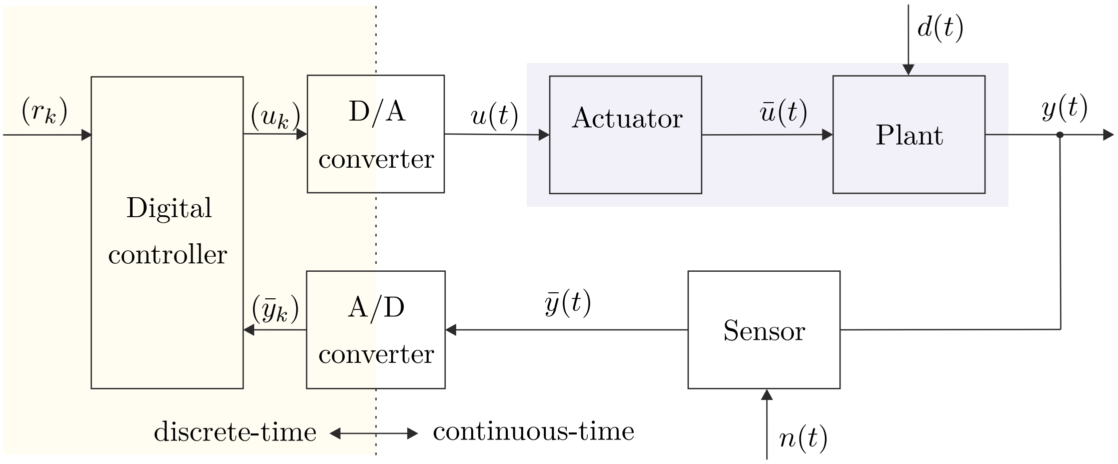


Figure 5.2: Block diagram of a digital control loop.

$$y_k = y(kT_a) \quad \text{for} \quad k \in \mathbb{Z} \ , \tag{5.3}$$

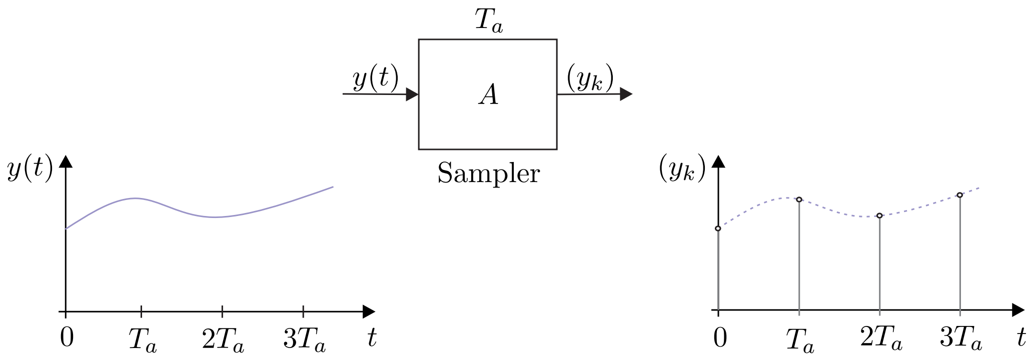


Figure 5.3: Functionality of the sample element.

and for left-continuous functions $y(t)$, we have

$$y_k = \lim_{t \to +kT_a} y(t) \quad \text{for} \quad k \in \mathbb{Z} \ . \tag{5.4}$$

In the digital hardware, the manipulated variable sequence (u_k) is then calculated from the sequence values (y_k) of the sampled sensor signal $y(t)$ and from the sequence values (r_k) of the setpoint or reference signal at discrete times $t_k = kT_a$. To generate a continuous-time signal $u(t)$ from a sequence of samples (u_k), a *D/A(Digital/Analog) converter* is required. In digital control loops, the D/A converter is usually designed in such a way that it holds the sample value at the output constant until the next sampling period (see Figure 5.4). This is also referred to as a *zero-order hold* in the case of ideal behavior. The output variable $u(t)$ of the zero-order hold is thus related to the input

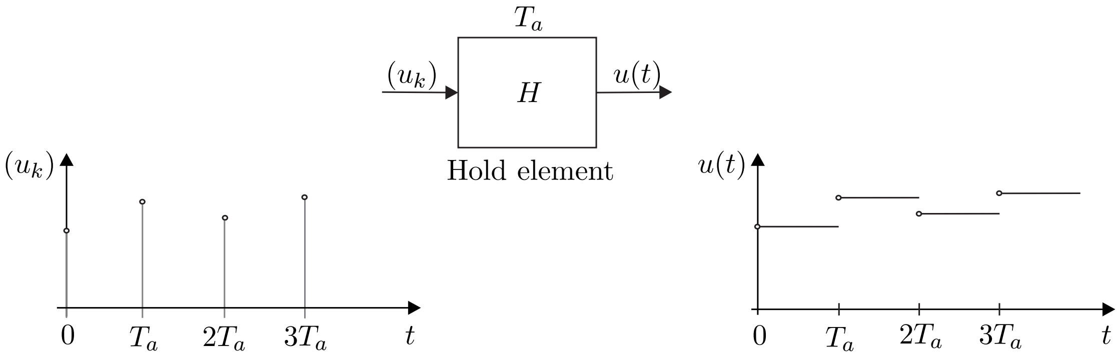


Figure 5.4: Functionality of the zero-order hold.

variable (u_k) as follows,

$$u(t) = u_k \quad \text{for} \quad kT_a \leq t < (k+1)T_a \tag{5.5}$$

and, more generally,

$$u(t) = \sum_{k=0}^{\infty} u_k \Big(\sigma(t - kT_a) - \sigma(t - (k+1)T_a) \Big) \ . \tag{5.6}$$

Transforming (5.6) into the Laplace domain yields

$$\hat{u}(s) = \sum_{k=0}^{\infty} u_k \frac{1}{s}\left(\mathrm{e}^{-kT_a s} - \mathrm{e}^{-(k+1)T_a s}\right) = \frac{1}{s}\left(1 - \mathrm{e}^{-T_a s}\right) \underbrace{\sum_{k=0}^{\infty} u_k \mathrm{e}^{-kT_a s}}_{\hat{v}(s)} , \tag{5.7}$$

from which the transfer function of the hold element immediately results in

$$G(s) = \frac{\hat{u}(s)}{\hat{v}(s)} = \frac{1}{s}\left(1 - \mathrm{e}^{-T_a s}\right) \tag{5.8}$$

with the input variable

$$(u_k) = v(t) = \sum_{k=0}^{\infty} u(kT_a)\delta(t - kT_a) \tag{5.9}$$

results.

> *Exercise* 5.1. Show that the series connection of a precisely synchronized hold and sample element with the sampling time T_a and input sequence (u_k) corresponds to a direct connection.

5.1.1 Sampled-Data Systems

For further considerations, let the discrete-time system be according to Figure 5.5 with ideally synchronized sample-and-hold with sampling time T_a as well as the input sequence (u_k) and the output sequence (y_k).

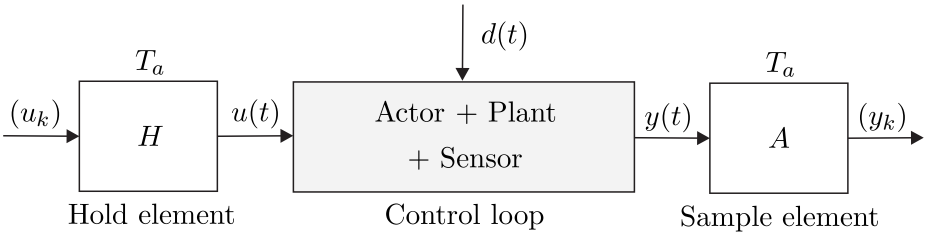


Figure 5.5: Sampled-data System.

The Nonlinear Case

Assume that the plant in Figure 5.5 can be written in the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) , \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{5.10a}$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \tag{5.10b}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input $\mathbf{u} \in \mathbb{R}^p$ and the output $\mathbf{y} \in \mathbb{R}^q$. The relationship between $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ of the associated sampled-data system then results as the solution of the differential equation

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{5.11}$$

for the initial value $\mathbf{x}_k = \mathbf{x}(kT_a)$ and the input $\mathbf{u}(t) = \mathbf{u}_k = \mathbf{u}(kT_a)$ for $kT_a \leq t < (k+1)T_a$, terminating at time $t = (k+1)T_a$. Doing this recursively gives the mathematical model of the sampled-data system as a *difference equation system*

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k)\,, \qquad \mathbf{x}_0 = \mathbf{x}(t_0) \tag{5.12a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k)\ . \tag{5.12b}$$

Example 5.2. As an example, consider the nonlinear first-order plant

$$\dot{x}(t) = -\sin(t)x(t) + x(t)u(t) \tag{5.13a}$$

$$y(t) = tx(t)^2 + u(t)\ . \tag{5.13b}$$

The solution of the differential equation for the initial value $x(t_0) = x_0$ and the constant input $u(t) = u_C$ is

$$x(t) = \frac{x_0 \exp(u_C t + \cos(t))}{\cosh(\cos(t_0) + u_C t_0) + \sinh(\cos(t_0) + u_C t_0)}\ , \tag{5.14}$$

from which the associated sampled-data system immediately follows for $t_0 = kT_a$, $x_0 = x_k$, $t = (k+1)T_a$, $x(t) = x_{k+1}$ and $u_C = u_k$ to

$$x_{k+1} = \frac{x_k \exp(u_k(k+1)T_a + \cos((k+1)T_a))}{\cosh(\cos(kT_a) + u_k kT_a) + \sinh(\cos(kT_a) + u_k kT_a)} \tag{5.15}$$

results. The output $y(t)$ at time $t = kT_a$ then follows in the form

$$y_k = kT_a x_k^2 + u_k\ . \tag{5.16}$$

One thus recognizes that it is *generally impossible* to calculate the corresponding sampled-data system *exactly* for nonlinear continuous-time systems since knowledge of the analytic solution of the nonlinear system of differential equations is required. The system (5.10) can be approximately converted into a difference equation system by means of various numerical integration methods. For example, the difference equation system for (5.10) according to the explicit Euler method is

$$\mathbf{x}_{k+1} = \mathbf{x}_k + T_a\mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, kT_a)\ . \tag{5.17}$$

For the stability of the integration method and the sampling time T_a (step size) to be chosen, please refer to the numerical literature.

The Linear Time-Invariant Case

For linear, time-invariant (LTI) plants of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{5.18a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}\ , \tag{5.18b}$$

the general solution of this system of differential equations is given by the following theorem:

Theorem 5.1. *The general solution of the system* (5.18) *is*

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{\Phi}(t-t_0)\mathbf{x}_0 + \int_{t_0}^{t} \mathbf{\Phi}(t-\tau)\mathbf{B}\mathbf{u}(\tau)\,\mathrm{d}\tau\,, \qquad \mathbf{x}(t_0)=\mathbf{x}_0 \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)\ ,\end{aligned} \tag{5.19}$$

with the continuous-time transition matrix

$$\mathbf{\Phi}(t) = \exp(\mathbf{A}t)\ . \tag{5.20}$$

This transition matrix (5.20) corresponds to the Peano-Baker series from (2.105) for the case of $\mathbf{A}(t)$ not depending on time.
Substituting $t_0 = kT_a$, $t = (k+1)T_a$ and $\mathbf{u}(\tau) = \mathbf{u}_k$ for $kT_a \le \tau < (k+1)T_a$ into (5.19), one obtains the sampled-data system associated with (5.18)

$$\mathbf{x}_{k+1} = \mathbf{\Phi}(T_a)\mathbf{x}_k + \int_{kT_a}^{(k+1)T_a} \mathbf{\Phi}\big((k+1)T_a - \tau\big)\,\mathrm{d}\tau\mathbf{B}\mathbf{u}_k \tag{5.21a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \tag{5.21b}$$

or, in short,

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \tag{5.22a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k\ , \tag{5.22b}$$

with the new dynamics and input matrices

$$\mathbf{\Phi} = \exp(\mathbf{A}T_a) \tag{5.22c}$$

$$\mathbf{\Gamma} = \int_0^{T_a} \exp(\mathbf{A}\tau)\,\mathrm{d}\tau\mathbf{B}\ . \tag{5.22d}$$

Exercise 5.2. Show, using the variable transformation $\bar{\tau} = (k+1)T_a - \tau$, that the following equation

$$\int_{kT_a}^{(k+1)T_a} \mathbf{\Phi}\big((k+1)T_a - \tau\big)\,\mathrm{d}\tau = \int_0^{T_a} \mathbf{\Phi}(\bar{\tau})\,\mathrm{d}\bar{\tau} \tag{5.23}$$

holds.

Example 5.3. As an example, calculate the associated sampled-data system for the state model of a double integrator

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}u = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\mathbf{x} + \begin{bmatrix}0\\ 1\end{bmatrix}u \tag{5.24a}$$

$$y = \mathbf{c}^{\mathrm{T}}\mathbf{x} = \begin{bmatrix}1 & 0\end{bmatrix}\mathbf{x} \tag{5.24b}$$

for the sampling time T_a. The transition matrix is calculated in the form

$$\mathbf{\Phi}(t) = \exp(\mathbf{A}t) = \mathbf{E} + \mathbf{A}t + \mathbf{A}^2\frac{t^2}{2} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \tag{5.25}$$

and thus the dynamic matrix $\mathbf{\Phi}$ and the input vector $\mathbf{\Gamma}$ of the sampled-data system are obtained according to (5.22) as

$$\mathbf{\Phi} = \mathbf{\Phi}(T_a) = \begin{bmatrix} 1 & T_a \\ 0 & 1 \end{bmatrix} \tag{5.26}$$

and

$$\mathbf{\Gamma} = \left(\int_0^{T_a} \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \mathrm{d}\tau\right) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} T_a & \frac{T_a^2}{2} \\ 0 & T_a \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{T_a^2}{2} \\ T_a \end{bmatrix} . \tag{5.27}$$

As a result, one obtains the sampled-data system

$$\begin{aligned} \mathbf{x}_{k+1} &= \begin{bmatrix} 1 & T_a \\ 0 & 1 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} \frac{T_a^2}{2} \\ T_a \end{bmatrix} u_k && \text{(5.28a)} \\ y_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k \ . && \text{(5.28b)} \end{aligned}$$

A discrete-time system is linear if and only if it can be written in the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{\Phi}_k \mathbf{x}_k + \mathbf{\Gamma}_k \mathbf{u}_k && \text{(5.29a)} \\ \mathbf{y}_k &= \mathbf{C}_k \mathbf{x}_k + \mathbf{D}_k \mathbf{u}_k \ , && \text{(5.29b)} \end{aligned}$$

Which is a form often obtained by linearizing (5.12) around a desired trajectory. A discrete-time system is linear and time-invariant if and only if it can be written in the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{\Phi} \mathbf{x}_k + \mathbf{\Gamma} \mathbf{u}_k && \text{(5.29c)} \\ \mathbf{y}_k &= \mathbf{C} \mathbf{x}_k + \mathbf{D} \mathbf{u}_k \ , && \text{(5.29d)} \end{aligned}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input $\mathbf{u} \in \mathbb{R}^p$, the output $\mathbf{y} \in \mathbb{R}^q$ and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n \times p}$, $\mathbf{C} \in \mathbb{R}^{q \times n}$ and $\mathbf{D} \in \mathbb{R}^{q \times p}$. As a reminder, it should be noted at this point that the subscript k of a quantity ξ_k refers to the value of ξ at time kT_a, i. e. $\xi_k = \xi(kT_a)$.

Exercise 5.3. Show that the matrices $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ of the sampled-data system (5.22) can be calculated in the form

$$\begin{bmatrix} \mathbf{\Phi} & \mathbf{\Gamma} \\ \mathbf{0} & \mathbf{E} \end{bmatrix} = \exp\left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} T_a\right) ,$$

with $\mathbf{E}$ as the identity matrix.

Hint: The following relationships hold:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}T_a}\mathbf{\Phi} &= \mathbf{A}\mathbf{\Phi}(T_a)\\ \frac{\mathrm{d}}{\mathrm{d}T_a}\mathbf{\Gamma} &= \mathbf{\Phi}(T_a)\mathbf{B}\ .\end{aligned}$$

Exercise 5.4. For the system

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} &= \begin{bmatrix}-1 & 0\\ 1 & 0\end{bmatrix}\mathbf{x} + \begin{bmatrix}1\\0\end{bmatrix}u\\ y &= \begin{bmatrix}0 & 1\end{bmatrix}\mathbf{x}\end{aligned}$$

calculate the associated sampled-data system for the sampling time T_a. Check the result in MATLAB using the command `c2d` for the sampling time $T_a = 0.1\,\mathrm{s}$.

Exercise 5.5. Assume that the manipulated variable $u(t)$ is applied to the process with a known delay $\Delta T \leq T_a$, i. e., the mathematical model (5.18) is modified in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t-\Delta T) \tag{5.30a}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)\ . \tag{5.30b}$$

What is the associated sampled-data system for the sampling time T_a?

5.1.2 The Transition Matrix

In the following, consider the linear, time-invariant, autonomous sampled-data system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k\ , \qquad \mathbf{x}(0) = \mathbf{x}_0\ . \tag{5.31}$$

As can easily be verified, the solution of the difference equation (5.31) at time kT_a for a given initial value $\mathbf{x}_0$ is

$$\mathbf{x}_k = \mathbf{\Phi}^k\mathbf{x}_0 = \mathbf{\Psi}(k)\mathbf{x}_0\ . \tag{5.32}$$

Thus, the $(n \times n)$ matrix $\mathbf{\Psi}(k)$ defines the transition matrix of the discrete-time system (5.31). Analogous to the transition matrix (5.20) in continuous time, $\mathbf{\Psi}(k)$ satisfies the following discrete-time relationships:

$$\boldsymbol{\Psi}(0) = \mathbf{E} \tag{5.33a}$$
$$\boldsymbol{\Psi}(k+l) = \boldsymbol{\Psi}(k)\boldsymbol{\Psi}(l) \tag{5.33b}$$
$$\boldsymbol{\Psi}^{-1}(k) = \boldsymbol{\Psi}(-k) \tag{5.33c}$$
$$\boldsymbol{\Psi}(k+1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(k) \ , \tag{5.33d}$$

where it is assumed for relation (5.33c) that the dynamic matrix $\boldsymbol{\Phi}$ of the discrete-time system is non-singular. Note that the transition matrix of continuous-time systems for all times t is non-singular regardless of the dynamic matrix $\mathbf{A}$, whereas the transition matrix $\boldsymbol{\Psi}$ of discrete-time systems is non-singular if and only if the dynamic matrix $\boldsymbol{\Phi}$ of the associated discrete-time system is non-singular.

> *Exercise* 5.6. Prove the properties (5.33) of the transition matrix $\boldsymbol{\Psi}$.

The general solution of a linear, time-invariant, sampled-data system of the form (5.22) is easily obtained by successively applying the iteration rule

$$\begin{aligned}
\mathbf{x}_1 &= \boldsymbol{\Phi}\mathbf{x}_0 + \boldsymbol{\Gamma}\mathbf{u}_0 \\
\mathbf{x}_2 &= \boldsymbol{\Phi}\underbrace{(\boldsymbol{\Phi}\mathbf{x}_0 + \boldsymbol{\Gamma}\mathbf{u}_0)}_{\mathbf{x}_1} + \boldsymbol{\Gamma}\mathbf{u}_1 = \boldsymbol{\Phi}^2\mathbf{x}_0 + \boldsymbol{\Phi}\boldsymbol{\Gamma}\mathbf{u}_0 + \boldsymbol{\Gamma}\mathbf{u}_1 \\
&\ \ \vdots
\end{aligned} \tag{5.34}$$

in the form

$$\mathbf{x}_k = \boldsymbol{\Phi}^k\mathbf{x}_0 + \sum_{j=0}^{k-1}\boldsymbol{\Phi}^{k-j-1}\boldsymbol{\Gamma}\mathbf{u}_j \tag{5.35a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \ . \tag{5.35b}$$

It can be seen from (5.35) that the convolution integral of the general solution of continuous-time systems from Theorem 5.1 is replaced in the discrete-time domain by the convolution sum.

The following theorem for the stability of a linear, time-invariant, autonomous, sampled-data system of the form

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}\mathbf{x}_k \tag{5.36}$$

can be stated:

Theorem 5.2 (Global Asymptotic Stability for Sampled-Data LTI Systems). *For all initial values* $\mathbf{x}_0 \in \mathbb{R}^n$ *of the system* (5.36),

$$\lim_{k\to\infty}\mathbf{x}_k = \lim_{k\to\infty}\boldsymbol{\Phi}^k\mathbf{x}_0 = \mathbf{0} \tag{5.37}$$

holds if and only if all eigenvalues *of the dynamic matrix* $\boldsymbol{\Phi}$ *are* less than 1 in absolute value. *One then also says that the equilibrium point* $\mathbf{x}_R = \mathbf{0}$*, which satisfies the relation* $\mathbf{x}_R = \boldsymbol{\Phi}\mathbf{x}_R$*, is* globally asymptotically stable.

In other words: While global asymptotic stability of continuous-time LTI systems is equivalent to all eigenvalues of $\mathbf{A}$ having negative real part, the discrete-time counterpart is all eigenvalues of $\mathbf{\Phi}$ being within the unit circle in the plane of complex numbers.

5.1.3 Input-Output Behavior

Similar to how linear differential equations representing a continuous-time LTI system can be converted to a transfer function $G(s)$, an analogous route can be taken for systems of the form (5.22) to yield a discrete-time transfer function $G(z)$. We will keep the treatment here very brief as this lecture focuses on state-space methods. Details can be found, e.g., in the lecture notes of *Automatisierung* [5.2]. The main points are as follows:

- Domain transform: Instead of the Laplace transform, one may use the so-called *z-transform*

$$\mathcal{Z} : (f_k) = (f_0, f_1, \dots, f_n, \dots) \mapsto \mathcal{Z}((f_k)) = f_z(z) = \sum_{k=0}^{\infty} f_k z^{-k} \tag{5.38}$$

- Transformed variable: While in the Laplace domain, s and $1/s$ correspond to derivatives and integrals in time, respectively, one can see with the z-transform above that z and $1/z$ correspond to the time shifts $f_k \mapsto f_{k+1}$ and $f_k \mapsto f_{k-1}$, respectively.

- Transfer matrix: The transfer matrix (matrix of transfer functions) of (5.22) is given by

$$\mathbf{G}(z) = \frac{\mathcal{Z}\{(\mathbf{y}_k)\}}{\mathcal{Z}\{(\mathbf{u}_k)\}} = \frac{\mathbf{y}_z(z)}{\mathbf{u}_z(z)} = \mathbf{C}(z\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma} + \mathbf{D} \ . \tag{5.39}$$

> *Exercise* 5.7. Verify the validity of these statements based on (5.38).

Using the so-called *Tustin transform*

$$z = \frac{1 + T_a q/2}{1 - T_a q/2} \ ,$$

one can establish a frequency domain for discrete-time systems and, for instance, apply loop-shaping techniques for controller design based in the discrete-time frequency response $q = I\Omega$. We will not further pursue this approach in the following.

5.1.4 Choice of the sampling time

For the choice of the sampling time, one is met with a trade-off:

- Long sampling times might lead to insufficient control over the dynamics of a sampled-data system since the input sequence (u_k) must be held constant for long stretches of time. Furthermore, if the system cannot be discretized exactly, such as in the case of a nonlinear system without an analytic solution, one must employ a numerical integration scheme, for example (5.17). These schemes become less accurate the longer T_a is chosen. The primary upside of a long T_a is that long sampling times lead to simple and efficient implementations on the control hardware as well as fast simulations.

- Short sampling times, on the contrary, are more difficult to implement on the control hardware and lead to more simulation effort, but the accuracy of the numerical integration in the simulation of the sampled-data system is improved.

For the minimum admissible value of T_a for a given control system, one may orient oneself based on the following two heuristics:

1. Information-theoretic considerations: Suppose one knows, at least in terms of the maximal desired frequency ω_{max}, the spectral properties of the planned input signal $u(t)$ beforehand. In that case, the maximum-frequency component of that sampled signal is given by

$$(u_k) = (A_0 \sin(\omega_{max} T_a k + \phi_0)) \ . \tag{5.40}$$

The output sequence (y_k) of an LTI system will then, as soon as transient processes have decayed, oscillate at the same maximal frequency ω_{max} with some phase delay. Based on Shannon's sampling theorem [5.3], we then know that a unique assignment between the harmonic underlying oscillation $y(t)$ and the sequence (y_k) is possible only if the frequency ω_{max} and the sampling time T_a satisfy the inequality

$$T_a < \frac{\pi}{\omega_{max}} \ . \tag{5.41}$$

2. Measured rise times: As a rule of thumb, one can apply step functions to all inputs $u_i(t), 1 \leq i \leq p$ of the system (5.10) or (5.18). Given the resulting step responses $h_i(t)$, it makes sense to choose T_a such that the shortest of the corresponding rise times $t_{r,i}$ is still sufficiently resolved in time. We thus obtain

$$T_a \leq \frac{\min_i\{t_{r,i}\}}{\beta} \ , \ 4 \leq \beta \leq 10 \tag{5.42}$$

as another rule for the sampling time.

Example 5.4. A damped harmonic oscillator with gain V, damping ratio ξ and time constant T can be represented in state space as

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1/T^2 & -2\xi/T \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ V/T^2 \end{bmatrix} u \tag{5.43}$$

$$y = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x} \tag{5.44}$$

or, equivalently, by the transfer function $G(s) = \frac{V}{1+2\xi(sT)+(sT)^2}$. The rise time for this system can be found by first computing the step response

$$h(t) = V\left(1 - \frac{1}{\sqrt{1-\xi^2}}\left(\xi \sin\left(\sqrt{1-\xi^2}\frac{t}{T}\right) + \sqrt{1-\xi^2}\cos\left(\sqrt{1-\xi^2}\frac{t}{T}\right)\right)e^{-\xi\frac{t}{T}}\right)\sigma(t)\ . \tag{5.45}$$

the inflection point t^* is then found by setting the second derivative of (5.45) to zero. The solution for t_r is then the point where the linear interpolation reaches the stationary value of $h(t)$, i.e., $\dot{h}(t^*)t = V$, which results in

$$t_r = T\exp\left(\frac{\tan(\arccos(\xi))}{\arccos(\xi)}\right)\ . \tag{5.46}$$

Therefore, a reasonable choice for the sampling time is to choose $T_a = t_r/5$ in accordance with (5.42), with t_r from (5.46).

5.2 Reachability & Observability

As soon as one is presented with a state-space representation of a dynamical system, the question is whether the inputs $\mathbf{u}_k$ and outputs $\mathbf{y}_k$ of this representation, together with the state $\mathbf{x}_k$, are suitable for designing an effective control system. In the following, we will focus on the case of discrete-time LTI systems

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k\ ,\ \mathbf{x}(0) = \mathbf{x}_0 \tag{5.47a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k\ , \tag{5.47b}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input $\mathbf{u} \in \mathbb{R}^p$, the output $\mathbf{y} \in \mathbb{R}^q$ and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$ and $\mathbf{D} \in \mathbb{R}^{q\times p}$.

Example 5.5. If the state-space dynamics in (5.47) are, for example, given by

$$\mathbf{\Phi} = \begin{bmatrix} 0 & 2 & 0 \\ -0.5 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}\ ,\ \mathbf{\Gamma} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\ ,\ \mathbf{C} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}\ ,\ \mathbf{D} = 0\ , \tag{5.48}$$

then one can see that the actuators and sensors of the control loop exclusively act on the subspace spanned by the third component of $\mathbf{x}_k$. This subspace, however, is only mapped onto itself by $\mathbf{\Phi}$. Therefore, we can conclude that

- the actuator represented by $\mathbf{\Gamma}$ will not be able to attain arbitrary states in the full state space $\mathbb{R}^n$ and that
- the sensor represented by $\mathbf{C}$ will not provide sufficient information about the

> whole system state if only the input and output sequences (u_k) and (y_k) are known.

This example can be generalized to yield the concepts of reachability and observability, as will be demonstrated in the following.

5.2.1 Reachability (discrete-time case)

Reachability analysis assesses whether arbitrary states in $\mathbb{R}^n$ can be reached with a finite amount of control actions, in detail, it is defined as follows:

Definition 5.1 (Reachability in discrete time)**.** The system (5.47) is called *completely reachable* if, starting from the initial state $\mathbf{x}_0 = \mathbf{0}$, *any arbitrary state* $\mathbf{x}_N$ can be reached with a finite control sequence $(\mathbf{u}_k) = (\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{N-1}, \mathbf{0}, \dots)$.

The following theorem now gives a more practical criterion for the reachability of the system (5.47) based on the system matrices $(\mathbf{\Phi}, \mathbf{\Gamma})$:

Theorem 5.3 (Reachability via the reachability matrix)**.** *The system* (5.47) *is* completely reachable *if and only if the so-called* reachability matrix

$$\mathcal{R}(\mathbf{\Phi}, \mathbf{\Gamma}) = \left[\mathbf{\Gamma}, \mathbf{\Phi}\mathbf{\Gamma}, \mathbf{\Phi}^2\mathbf{\Gamma}, \dots, \mathbf{\Phi}^{n-1}\mathbf{\Gamma}\right] \tag{5.49}$$

has rank n.

Proof. One has to look at the analytic solution for the system state, which is given by (5.35). Since the definition of reachability assumes that $\mathbf{x}_0 = \mathbf{0}$, the solution for the state $\mathbf{x}_N$ reads as

$$\mathbf{x}_N = \sum_{j=0}^{k-1} \mathbf{\Phi}^{k-j-1}\mathbf{\Gamma}\mathbf{u}_j \ . \tag{5.50}$$

Another way to write (5.50) is

$$\mathbf{x}_N = \left[\mathbf{\Gamma} \ \mathbf{\Phi}\mathbf{\Gamma} \ \dots \ \mathbf{\Phi}^{N-1}\mathbf{\Gamma}\right] \begin{bmatrix} \mathbf{u}_{N-1} \\ \mathbf{u}_{N-2} \\ \dots \\ \mathbf{u}_0 \end{bmatrix} . \tag{5.51}$$

For any arbitrary $\mathbf{x}_N \in \mathbb{R}^n$, this has at least one solution in the input sequence if and only if the matrix in (5.51) has full row rank n. Since the matrix is exactly the reachability matrix (5.49), the theorem follows. □

> *Exercise* 5.8. Show that the system from Example 5.5 is not completely reachable.

The analogous theorem to Theorem 5.3 for the continuous-time case (5.18) goes completely analogously, except with a piecewise-continuous input function $\mathbf{u}(t)$ instead of the input sequence $(\mathbf{u}_k)$ and the matrices $(\mathbf{A}, \mathbf{B})$ instead of $(\mathbf{\Phi}, \mathbf{\Gamma})$.

5.2.2 Observability (discrete-time case)

Suppose again a discrete-time LTI system given by (5.47). As a next step, one can address the question when the sensor represented by $\mathbf{C}$ will provide sufficient information about the whole system state if only the input and output sequences (u_k) and (y_k) can be measured rather than the state itself. to this end, we define the concept of observability as follows:

Definition 5.2 (Observability in discrete time)**.** The discrete-time system (5.47) is called *completely observable* if the knowledge of the input and output sequences $(\mathbf{u}_k) = (\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{N-1}, \mathbf{0}, \dots)$ and $(\mathbf{y}_k) = (\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1}, \mathbf{0}, \dots)$ with finite N as well as the system matrices $\mathbf{\Phi}$, $\mathbf{\Gamma}$, $\mathbf{C}$ and $\mathbf{D}$ allows the calculation of the initial state $\mathbf{x}_0$.

In analogy to Theorem 5.3 on reachability, the following theorem on observability applies:

Theorem 5.4 (Observability via the observability matrix)**.** *The discrete-time system* (5.47) *is* completely observable *if and only if the so-called* observability matrix

$$\mathcal{O}(\mathbf{C}, \mathbf{\Phi}) = \begin{bmatrix} \mathbf{C} \\ \mathbf{C\Phi} \\ \mathbf{C\Phi}^2 \\ \vdots \\ \mathbf{C\Phi}^{n-1} \end{bmatrix} \tag{5.52}$$

has rank n.

Exercise 5.9. Prove Theorem 5.4, keeping in mind what is measured and what is to be calculated.

Exercise 5.10. Show that the system from Example 5.5 is not completely observable.

The combination of reachability and observability considerations can be based on the following exercise:

Example 5.6. The motion of a hot air balloon according to Figure 5.6 is approximately described by a mathematical model of the form

$$\Delta\dot{T} = -\frac{1}{\tau_1}\Delta T + u \tag{5.53a}$$

$$\dot{v} = -\frac{1}{\tau_2}(v - w) + \sigma\Delta T \tag{5.53b}$$

$$\dot{h} = v \ , \tag{5.53c}$$

with the temperature difference ΔT to the equilibrium temperature, the altitude h of the balloon, the vertical velocity v of the balloon, the vertical wind speed w (disturbance variable) and the control variable u proportional to the heat supplied.

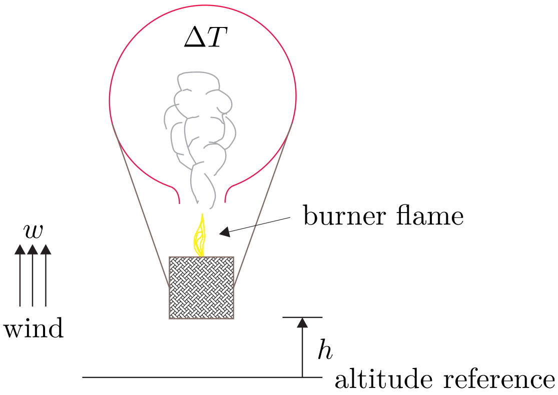


Figure 5.6: Schematic representation of a hot air balloon.

1. Is it possible to observe the temperature change ΔT and a constant wind speed w solely from the measurement of the altitude h?
   Result: One must extend the mathematical model for the constant wind speed w by the differential equation $\dot{w} = 0$. The observability matrix $\mathcal{O}$ then is

$$
\begin{aligned}
\mathcal{O} &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ \sigma & -\frac{1}{\tau_2} & 0 & \frac{1}{\tau_2} \\ -\frac{\sigma}{\tau_1} - \frac{\sigma}{\tau_2} & \frac{1}{\tau_2^2} & 0 & -\frac{1}{\tau_2^2} \end{bmatrix} \\
\det(\mathcal{O}) &= \frac{-\sigma}{\tau_1 \tau_2}
\end{aligned}
$$

   and thus, the extended system is completely observable for $\sigma \neq 0$ with the measurement $y = h$.

2. From now on assume $w \equiv 0$, making the state space three-dimensional again. Is the system (5.53) completely reachable with the input u?
   Result: Yes, for $\sigma \neq 0$.

3. Is the system (5.53) completely reachable with the input w?
   Result: No.

5.3 Linear State Space Controllers & Observers

Now that we have a discrete-time description of dynamical systems and can assess their reachability and observability (globally for LTI systems and locally around a reference point for nonlinear or time-varying systems), this information can be leveraged to derive state feedback controllers and state observers (to be defined in this section) for these systems. Due to the discrete-time and linear nature of the methods, they are straightworward to implement on digital hardware and thus are popular methods for practitioners.

5.3.1 State Space Controllers via Pole Placement

A state feedback control law is a *memoryless, functional dependence* of the control input $\mathbf{u}$ on the state variables $\mathbf{x}$ and possibly other external input variables (e.g., reference variables) $\mathbf{r}$ in the general form

$$\mathbf{u}_k = \mathbf{f}_{\mathbf{u},k}(\mathbf{x}_k, \mathbf{r}_k) \tag{5.54}$$

or for linear, time-invariant systems

$$\mathbf{u}_k = \mathbf{K}\mathbf{x}_k + \mathbf{G}\mathbf{r}_k \ . \tag{5.55}$$

Without loss of generality, the following considerations are based on the linear, time-invariant, discrete-time, single-input system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k \,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.56a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k + du_k \tag{5.56b}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input $u \in \mathbb{R}$, the output $y \in \mathbb{R}$ and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma}, \mathbf{c} \in \mathbb{R}^n$ and $d \in \mathbb{R}$. If a state feedback control law of the form

$$u_k = \mathbf{k}^{\mathrm{T}}\mathbf{x}_k + gr_k \tag{5.57}$$

is inserted into (5.56), the closed loop system is given by

$$\mathbf{x}_{k+1} = \underbrace{\left(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)}_{\mathbf{\Phi}_g}\mathbf{x}_k + \mathbf{\Gamma}gr_k \,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.58a}$$

$$y_k = \left(\mathbf{c}^{\mathrm{T}} + d\mathbf{k}^{\mathrm{T}}\right)\mathbf{x}_k + dgr_k \tag{5.58b}$$

with the closed-loop system matrix $\mathbf{\Phi}_g$ and the input variable r. Obviously, the variables $\mathbf{k} \in \mathbb{R}^n$ and $g \in \mathbb{R}$ have to be determined within the state controller design in such a way that the output sequence (y_k) as a response of the closed loop (5.58) to special input sequences, such as the step sequence $(r_k) = r_0\left(1^k\right) = (r_0, r_0, r_0, \ldots)$, satisfies certain conditions.

In the first step, the input sequence (r_k) is disregarded, i.e. $(r_k) = \left(0^k\right)$, and the variable $\mathbf{k}$ should be designed in such a way that the eigenvalues of the closed-loop system matrix $\mathbf{\Phi}_g$ are located at arbitrarily specified desired locations. Therefore, this design is also called *pole placement in state space*. Before that, however, an essential theorem for the following, namely the Cayley-Hamilton theorem, is formulated and proven:

Theorem 5.5 (Cayley-Hamilton Theorem). *Let*

$$p(z) = a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n \tag{5.59}$$

be the characteristic polynomial of the matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$*, then* $\mathbf{\Phi}$ *satisfies the relation*

$$p(\mathbf{\Phi}) = a_0 \mathbf{E} + a_1 \mathbf{\Phi} + \cdots + a_{n-1} \mathbf{\Phi}^{n-1} + \mathbf{\Phi}^n = \mathbf{0} \ . \tag{5.60}$$

Proof. For the inverse of the matrix $(z\mathbf{E} - \mathbf{\Phi}) \in \mathbb{R}^{n \times n}$ holds

$$(z\mathbf{E} - \mathbf{\Phi})^{-1} = \frac{\operatorname{adj}(z\mathbf{E} - \mathbf{\Phi})}{\det(z\mathbf{E} - \mathbf{\Phi})} \ , \tag{5.61}$$

where the adjugate matrix $\operatorname{adj}(z\mathbf{E} - \mathbf{\Phi})$ only has polynomials of $n-1$-order and can therefore be written in the form

$$\operatorname{adj}(z\mathbf{E} - \mathbf{\Phi}) = \mathbf{R}_0 + \mathbf{R}_1 z + \cdots + \mathbf{R}_{n-2} z^{n-2} + \mathbf{R}_{n-1} z^{n-1} \tag{5.62}$$

From (5.61) and (5.62) one finally obtains

$$\det(z\mathbf{E} - \mathbf{\Phi})\mathbf{E} = (z\mathbf{E} - \mathbf{\Phi})\Big(\mathbf{R}_0 + \mathbf{R}_1 z + \cdots + \mathbf{R}_{n-2} z^{n-2} + \mathbf{R}_{n-1} z^{n-1}\Big) \tag{5.63}$$

or

$$\begin{aligned}&\Big(a_0 + a_1 z + \cdots + a_{n-1} z^{n-1} + z^n\Big)\mathbf{E} \\ &\qquad = -\mathbf{\Phi}\mathbf{R}_0 + \cdots + (\mathbf{R}_{n-2} - \mathbf{\Phi}\mathbf{R}_{n-1}) z^{n-1} + \mathbf{R}_{n-1} z^n \ .\end{aligned} \tag{5.64}$$

Comparing the coefficients of the powers of z in (5.64) gives the following system of equations

$$\begin{aligned} a_0 \mathbf{E} &= -\mathbf{\Phi}\mathbf{R}_0 \\ a_1 \mathbf{E} &= \mathbf{R}_0 - \mathbf{\Phi}\mathbf{R}_1 \\ &\vdots \\ a_{n-2}\mathbf{E} &= \mathbf{R}_{n-3} - \mathbf{\Phi}\mathbf{R}_{n-2} \\ a_{n-1}\mathbf{E} &= \mathbf{R}_{n-2} - \mathbf{\Phi}\mathbf{R}_{n-1} \\ \mathbf{E} &= \mathbf{R}_{n-1} \ . \end{aligned} \tag{5.65}$$

If one now multiplies the j-th line of (5.65) by $\mathbf{\Phi}^j$, $j = 0, \ldots, n$, and adds them up, then (5.60) follows, which proves Theorem 5.5. □

If the system (5.56) is in *1st standard form* (*reachable canonical form*) $\{\mathbf{\Phi}_R, \mathbf{\Gamma}_R, \mathbf{c}_R, d_R\}$, then one immediately obtains a rule how to determine $\mathbf{k}^{\mathrm{T}}$ in the state feedback control law (5.57) so that the eigenvalues of the system matrix $\mathbf{\Phi}_g = \Big(\mathbf{\Phi}_R + \mathbf{\Gamma}_R \mathbf{k}^{\mathrm{T}}\Big)$ of

(5.58) are located at arbitrarily specified desired locations. In this case, we have for $\mathbf{\Phi}_g$

$$\mathbf{\Phi}_g = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}}_{\mathbf{\Phi}_R} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{\mathbf{\Gamma}_R} \underbrace{\begin{bmatrix} k_0 & k_1 & \cdots & k_{n-1} \end{bmatrix}}_{\mathbf{k}^{\mathrm{T}}} \tag{5.66}$$

or

$$\mathbf{\Phi}_g = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ k_0 - a_0 & k_1 - a_1 & \cdots & k_{n-2} - a_{n-2} & k_{n-1} - a_{n-1} \end{bmatrix} \tag{5.67}$$

with the corresponding characteristic polynomial of $\mathbf{\Phi}_g$

$$p_g(z) = (a_0 - k_0) + (a_1 - k_1)z + \cdots + (a_{n-1} - k_{n-1})z^{n-1} + z^n \ . \tag{5.68}$$

The procedure for pole placement in state space for a system in reachable canonical form $\{\mathbf{\Phi}_R, \mathbf{\Gamma}_R, \mathbf{c}_R, d_R\}$ is therefore as follows: One specifies the n desired eigenvalues λ_j, $j = 1, \dots, n$, of the closed loop and determines from this a desired characteristic polynomial for the closed loop

$$p_{g,soll}(z) = \prod_{j=1}^{n} (z - \lambda_j) = p_0 + p_1 z + p_2 z^2 + \cdots + p_{n-1} z^{n-1} + z^n \ . \tag{5.69}$$

By comparing coefficients of (5.68) and (5.69) one directly obtains the state feedback coefficients

$$k_j = a_j - p_j, \qquad j = 0, \dots, n-1 \ . \tag{5.70}$$

If now the system (5.56) is not in reachable canonical form, then the pole placement in state space can be performed using *Ackermann's formula.*

Theorem 5.6 (Ackermann's Formula)**.** *The eigenvalues of the closed-loop system matrix* $\mathbf{\Phi}_g$ *of* (5.58) *can be* arbitrarily placed *by a state feedback of the form* (5.57) *if and only if the system* (5.56) *is* completely reachable. *The feedback vector* $\mathbf{k}^{\mathrm{T}}$ *is calculated according to the relation*

$$\underbrace{\begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}}_{\mathbf{e}_n^{\mathrm{T}} = \mathbf{\Gamma}_R^{\mathrm{T}}} = \mathbf{v}_1^{\mathrm{T}} \underbrace{\begin{bmatrix} \mathbf{\Gamma} & \mathbf{\Phi\Gamma} & \mathbf{\Phi}^2\mathbf{\Gamma} & \dots & \mathbf{\Phi}^{n-1}\mathbf{\Gamma} \end{bmatrix}}_{\mathcal{R}(\mathbf{\Phi}, \mathbf{\Gamma})} \tag{5.71a}$$

$$\mathbf{k}^{\mathrm{T}} = -p_0 \mathbf{v}_1^{\mathrm{T}} - p_1 \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi} - \cdots - p_{n-1} \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi}^{n-1} - \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi}^n = -\mathbf{v_1}^{\mathrm{T}} p_{g,soll}(\mathbf{\Phi}) \tag{5.71b}$$

with $p_{g,soll}(z) = p_0 + p_1 z + p_2 z^2 + \cdots + p_{n-1} z^{n-1} + z^n$ as the desired characteristic polynomial of the closed loop system.

Proof. At the beginning of the proof it should be noted that the property of complete reachability of the system (5.56) by a regular state transformation of the form $\mathbf{x}_k = \mathbf{V}\mathbf{z}_k$ with a regular $(n \times n)$ matrix $\mathbf{V}$ neither can be lost nor gained. This is immediately obvious because the reachability matrix of the equivalent transformed system

$$\mathbf{z}_{k+1} = \underbrace{\mathbf{V}^{-1}\boldsymbol{\Phi}\mathbf{V}}_{\tilde{\boldsymbol{\Phi}}}\mathbf{z}_k + \underbrace{\mathbf{V}^{-1}\boldsymbol{\Gamma}}_{\tilde{\boldsymbol{\Gamma}}} u_k\,, \qquad \mathbf{z}(0) = \mathbf{V}^{-1}\mathbf{x}_0 \tag{5.72a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k + d u_k \tag{5.72b}$$

associated with (5.56) is

$$\mathcal{R}\Big(\tilde{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\Gamma}}\Big) = \Big[\tilde{\boldsymbol{\Gamma}} \quad \tilde{\boldsymbol{\Phi}}\tilde{\Gamma} \quad \tilde{\boldsymbol{\Phi}}^2\tilde{\boldsymbol{\Gamma}} \quad \dots \quad \tilde{\boldsymbol{\Phi}}^{n-1}\tilde{\boldsymbol{\Gamma}}\Big] =$$
$$\Big[\mathbf{V}^{-1}\boldsymbol{\Gamma} \quad \mathbf{V}^{-1}\boldsymbol{\Phi}\mathbf{V}\mathbf{V}^{-1}\boldsymbol{\Gamma} \quad \mathbf{V}^{-1}\boldsymbol{\Phi}\mathbf{V}\mathbf{V}^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma} \quad \dots \quad \mathbf{V}^{-1}\boldsymbol{\Phi}^{n-1}\boldsymbol{\Gamma}\Big] = \mathbf{V}^{-1}\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{\Gamma})\,. \tag{5.73}$$

From (5.73) one can see that due to the regularity of $\mathbf{V}$, the regularity of the reachability matrix $\mathcal{R}\Big(\tilde{\boldsymbol{\Phi}}, \tilde{\boldsymbol{\Gamma}}\Big)$ of the transformed system (5.72) follows immediately from the regularity of the reachability matrix $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{\Gamma})$ of the original system (5.56) and vice versa. Furthermore, it can be that a system which is in reachable canonical form is always completely reachable. Combining these two findings, it follows that a completely reachable system can always be transformed to reachable canonical form.

The idea for determining the relationship (5.71) is now to transform the system (5.56) into reachable canonical form in a first step and then perform the pole placement in state space, as shown in (5.70). Thus, we are looking for a regular state transformation of the form

$$\mathbf{z} = \mathbf{V}\mathbf{x} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix} \mathbf{x} \tag{5.74}$$

so that the system (5.56) in the new state $\mathbf{z}$ is in reachable canonical form

$$\mathbf{z}_{k+1} = \underbrace{\mathbf{V}\boldsymbol{\Phi}\mathbf{V}^{-1}}_{\boldsymbol{\Phi}_R}\mathbf{z}_k + \underbrace{\mathbf{V}\boldsymbol{\Gamma}}_{\boldsymbol{\Gamma}_R = \mathbf{e}_n} u_k\,. \tag{5.75}$$

From the equation

$$\boldsymbol{\Phi}_R = \mathbf{V}\boldsymbol{\Phi}\mathbf{V}^{-1} \tag{5.76a}$$

or

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix} \mathbf{\Phi} \tag{5.76b}$$

one gets

$$\mathbf{v}_{j+1}^{\mathrm{T}} = \mathbf{v}_j^{\mathrm{T}} \mathbf{\Phi}\,, \qquad j = 1, \dots, n-1 \tag{5.77a}$$

$$-a_0 \mathbf{v}_1^{\mathrm{T}} - a_1 \mathbf{v}_2^{\mathrm{T}} - \cdots - a_{n-1} \mathbf{v}_n^{\mathrm{T}} = \mathbf{v}_n^{\mathrm{T}} \mathbf{\Phi}\ . \tag{5.77b}$$

By inserting the relations for $\mathbf{v}_j^{\mathrm{T}}$, $j = 2, \dots, n$, into the last equation of (5.77)

$$\mathbf{v}_1^{\mathrm{T}} \Big(a_0 + a_1 \mathbf{\Phi} + \ldots + a_{n-1} \mathbf{\Phi}^{n-1} + \mathbf{\Phi}^n \Big) = \mathbf{0}^{\mathrm{T}} \tag{5.78}$$

and applying Theorem 5.5 one realizes that (5.78) is trivially fulfilled. The missing equation for determining $\mathbf{v}_1^{\mathrm{T}}$ is obtained from

$$\mathbf{\Gamma}_R = \mathbf{e}_n = \mathbf{V}\mathbf{\Gamma} \tag{5.79a}$$

or

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix} \mathbf{\Gamma} \tag{5.79b}$$

and with (5.77) follows

$$\mathbf{e}_n^{\mathrm{T}} = \mathbf{v}_1^{\mathrm{T}} \underbrace{\begin{bmatrix} \mathbf{\Gamma} & \mathbf{\Phi}\mathbf{\Gamma} & \mathbf{\Phi}^2\mathbf{\Gamma} & \dots & \mathbf{\Phi}^{n-1}\mathbf{\Gamma} \end{bmatrix}}_{\mathcal{R}(\mathbf{\Phi},\mathbf{\Gamma})}\ . \tag{5.80}$$

Provided that the system (5.56) is completely reachable, the state transformation reads

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi} \\ \vdots \\ \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi}^{n-1} \end{bmatrix} \tag{5.81}$$

with

$$\mathbf{v}_1^{\mathrm{T}} = \mathbf{e}_n^{\mathrm{T}} \mathcal{R}(\mathbf{\Phi}, \mathbf{\Gamma})^{-1} \; . \tag{5.82}$$

For the system in reachable canonical form (5.75), by choosing

$$u_k = \mathbf{k}_R^{\mathrm{T}} \mathbf{z}_k = \begin{bmatrix} a_0 - p_0 & a_1 - p_1 & \dots & a_{n-1} - p_{n-1} \end{bmatrix} \mathbf{z}_k \tag{5.83}$$

one directly obtains the desired characteristic polynomial

$$p_{g,soll}(z) = p_0 + p_1 z + p_2 z^2 + \dots + p_{n-1} z^{n-1} + z^n \tag{5.84}$$

and thus the eigenvalues of the closed loop can be assigned (compare also (5.66) - (5.70)).

Since one does not always want to carry out these two steps (transformation to reachable canonical form followed by pole placement) separately, one transforms the closed loop (5.75) and (5.83) back to the original state $\mathbf{x}$, i.e.

$$\mathbf{x}_{k+1} = \underbrace{\mathbf{V}^{-1} \mathbf{\Phi}_R \mathbf{V}}_{\mathbf{\Phi}} \mathbf{x}_k + \underbrace{\mathbf{V}^{-1} \mathbf{\Gamma}_R}_{\mathbf{\Gamma}} \underbrace{\mathbf{k}_R^{\mathrm{T}} \mathbf{V}}_{\mathbf{k}^{\mathrm{T}}} \mathbf{x}_k \; . \tag{5.85}$$

The feedback vector $\mathbf{k}^{\mathrm{T}}$ in the original system then reads as follows

$$\mathbf{k}^{\mathrm{T}} = \mathbf{k}_R^{\mathrm{T}} \mathbf{V} = \begin{bmatrix} a_0 - p_0 & a_1 - p_1 & \dots & a_{n-1} - p_{n-1} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi} \\ \vdots \\ \mathbf{v}_1^{\mathrm{T}} \mathbf{\Phi}^{n-1} \end{bmatrix} \tag{5.86}$$

$$= \mathbf{v}_1^{\mathrm{T}} \underbrace{\left(a_0 + a_1 \mathbf{\Phi} + \dots + a_{n-1} \mathbf{\Phi}^{n-1} \right)}_{= -\mathbf{\Phi}^n \text{ according to Theorem 5.5}} - \mathbf{v}_1^{\mathrm{T}} \left(p_0 + p_1 \mathbf{\Phi} + \dots + p_{n-1} \mathbf{\Phi}^{n-1} \right) \tag{5.87}$$

$$= -\mathbf{v}_1^{\mathrm{T}} \left(p_0 + p_1 \mathbf{\Phi} + \dots + p_{n-1} \mathbf{\Phi}^{n-1} + \mathbf{\Phi}^n \right) = -\mathbf{v}_1^{\mathrm{T}} p_{g,soll}(\mathbf{\Phi}) \tag{5.88}$$

with the desired characteristic polynomial $p_{g,soll}$ of (5.84). It has thus been proved that if the system (5.56) is completely reachable, then the poles of the closed loop can be arbitrarily placed with the state feedback $u_k = \mathbf{k}^{\mathrm{T}} \mathbf{x}_k$ and $\mathbf{k}^{\mathrm{T}}$ according to (5.71).

Exercise 5.11. Prove the converse that the fact that (5.56) is not completely reachable implies that not all poles of the closed loop can be arbitrarily assigned.

Hint: Use the fact that every not completely reachable system can be transformed to the form

$$\begin{bmatrix}\mathbf{x}_{1,k+1}\\ \mathbf{x}_{2,k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi}_{11} & \mathbf{\Phi}_{12}\\ \mathbf{0} & \mathbf{\Phi}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{1,k}\\ \mathbf{x}_{2,k}\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}_1\\ \mathbf{0}\end{bmatrix}u_k$$

$$y_k = \begin{bmatrix}\mathbf{c}_1^{\mathrm{T}} & \mathbf{c}_2^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{1,k}\\ \mathbf{x}_{2,k}\end{bmatrix} + du_k\ .$$

With the solution of Exercise 5.11, however, Theorem 5.6 is proved. □

As we have seen, the input sequence (r_k) of (5.58) plays no role for the pole placement. Now, the state feedback control law

$$u_k = \mathbf{k}^{\mathrm{T}}\mathbf{x}_k + g r_k \tag{5.89}$$

of (5.57) still contains the parameter g, with the help of which one can achieve, for example, that for the closed loop

$$\mathbf{x}_{k+1} = \underbrace{\left(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)}_{\mathbf{\Phi}_g}\mathbf{x}_k + \mathbf{\Gamma} g r_k\,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.90a}$$

$$y_k = \left(\mathbf{c}^{\mathrm{T}} + d\mathbf{k}^{\mathrm{T}}\right)\mathbf{x}_k + d g r_k \tag{5.90b}$$

holds

$$\lim_{k\to\infty} y_k = r_0 \tag{5.91}$$

with the step sequence $(r_k) = r_0\left(1^k\right) = (r_0, r_0, r_0, \ldots)$ as input. If we now calculate the z-transform of (y_k), we obtain

$$y_z(z) = \left(\mathbf{c}^{\mathrm{T}} + d\mathbf{k}^{\mathrm{T}}\right)\left(z\mathbf{E} - \mathbf{\Phi} - \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)^{-1}\left(\mathbf{\Gamma} g\, \underbrace{r_0\frac{z}{z-1}}_{r_z(z)} + z\mathbf{x}_0\right) + dg\, \underbrace{r_0\frac{z}{z-1}}_{r_z(z)} \tag{5.92}$$

or by applying the final value theorem it follows

$$\lim_{k\to+\infty} y_k = \lim_{z\to 1}(z-1)y_z(z) = \left(\mathbf{c}^{\mathrm{T}} + d\mathbf{k}^{\mathrm{T}}\right)\left(\mathbf{E} - \mathbf{\Phi} - \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)^{-1}\mathbf{\Gamma} g r_0 + d g r_0 = r_0\ , \tag{5.93}$$

since all zeros of $\det\left(z\mathbf{E} - \mathbf{\Phi} - \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)$ lie inside the unit circle of the complex z-plane. Thus g is calculated from (5.93) to

$$g = \frac{1}{(\mathbf{c}^{\mathrm{T}} + d\mathbf{k}^{\mathrm{T}})(\mathbf{E} - \mathbf{\Phi} - \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}})^{-1}\mathbf{\Gamma} + d}\ . \tag{5.94}$$

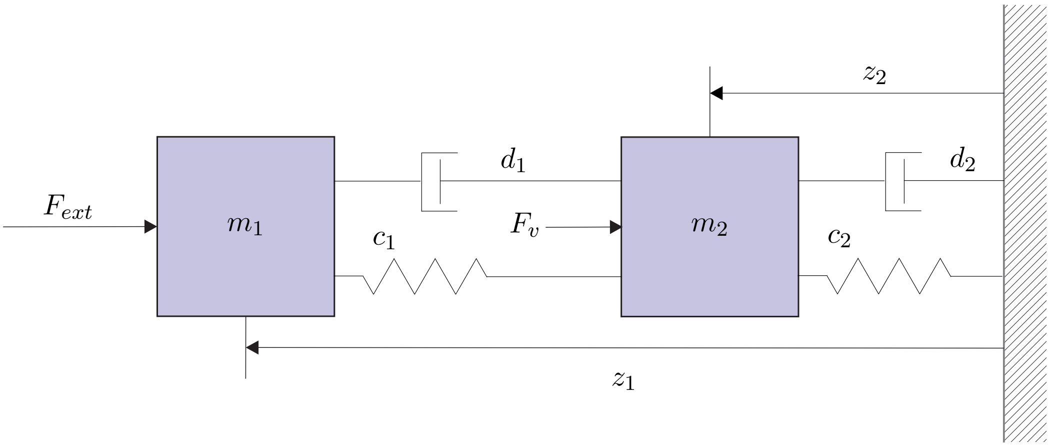


Figure 5.7: Two-mass oscillator.

Example 5.7 (Simulation example). As an example, consider the two-mass oscillator shown in Figure 5.7, consisting of two masses m_1 and m_2, two linear springs with spring constants c_1 and c_2, and two viscous dampers with damping constants d_1 and d_2.

The external force F_{ext}, which is also the control input $u = F_{ext}$ of the system, acts on the first mass m_1, and the force $F_v = v$ acting on the second mass m_2 is to be considered as an unknown disturbance. The mathematical model can be calculated directly by applying Newton's second law of motion to the two masses m_1 and m_2 in the form

$$m_1\ddot{z}_1 = -c_1(z_1 - z_2) - d_1(\dot{z}_1 - \dot{z}_2) - F_{ext} \tag{5.95a}$$

$$m_2\ddot{z}_2 = c_1(z_1 - z_2) + d_1(\dot{z}_1 - \dot{z}_2) - c_2 z_2 - d_2\dot{z}_2 - F_v \ , \tag{5.95b}$$

where z_1 and z_2 describe the displacements of the masses m_1 and m_2 from the relaxed position of the springs. With the state variables $\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} z_1 & v_1 = \dot{z}_1 & z_2 & v_2 = \dot{z}_2 \end{bmatrix}$, the control input $u = F_{ext}$, the disturbance $v = F_v$ and the output $y = z_2$, the state space representation of (5.95) is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}u + \mathbf{b}_v v \,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.96a}$$

$$y = \mathbf{c}^{\mathrm{T}}\mathbf{x} \tag{5.96b}$$

with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{c_1}{m_1} & -\frac{d_1}{m_1} & \frac{c_1}{m_1} & \frac{d_1}{m_1} \\ 0 & 0 & 0 & 1 \\ \frac{c_1}{m_2} & \frac{d_1}{m_2} & -\frac{c_1+c_2}{m_2} & -\frac{d_1+d_2}{m_2} \end{bmatrix}, \tag{5.97a}$$

$$\mathbf{b} = \begin{bmatrix} 0 \\ -\frac{1}{m_1} \\ 0 \\ 0 \end{bmatrix}, \tag{5.97b}$$

$$\mathbf{b}_v = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\frac{1}{m_2} \end{bmatrix}, \tag{5.97c}$$

$$\mathbf{c}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}. \tag{5.97d}$$

If one chooses the values $m_1 = 1$, $m_2 = 10$, $c_1 = c_2 = 1$ and $d_1 = d_2 = 1$ for the parameters, then the mathematical model is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} z_1 \\ v_1 \\ z_2 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0.1 & 0.1 & -0.2 & -0.2 \end{bmatrix}\begin{bmatrix} z_1 \\ v_1 \\ z_2 \\ v_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -0.1 \end{bmatrix} v \tag{5.98a}$$

$$y = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} z_1 \\ v_1 \\ z_2 \\ v_2 \end{bmatrix}. \tag{5.98b}$$

For the controller design, one first calculates the corresponding discrete-time system with input u and output y for the sampling time $T_a = 2$ (MATLAB command `c2d`). Subsequently, a state feedback controller should be designed for the discrete-time system using pole placement in such a way that the poles of the closed loop are located at $\exp(\lambda_j T_a)$, $j = 1, \dots, 4$, with $\lambda_{1,2} = -0.5 \pm 0.5I$, $\lambda_3 = -1$ and $\lambda_4 = -2$ (MATLAB command `place` or `acker`). Note that in MATLAB, in contrast to the script, see (5.89), the state feedback controller is defined with a negative sign in the form $u_k = -\mathbf{k}^{\mathrm{T}}\mathbf{x}_k$! Furthermore, the pre-factor g in (5.89) should be determined in such a way that for a step sequence $(r_k) = r_0\left(1^k\right)$ as reference signal holds $\lim_{k\to+\infty} y_k = r_0$.

With these specifications, the state feedback controller (5.89) for this example is

$$u_k = \begin{bmatrix} 0.0189 & 0.4627 & 0.5245 & 3.8538 \end{bmatrix} \mathbf{x}_k - 1.5624 r_k \ . \tag{5.99}$$

A simulation example in the form of Matlab/Simulink files for the state controller of the two-mass oscillator of Figure 5.7 is available on our homepage.

Dead-Beat Controller

If one now wants to design a state feedback control law of the form (5.89) with $g = 0$ in such a way that any initial deviation $\mathbf{x}_0$ of the system (5.56) is driven to $\mathbf{0}$ as quickly as possible, one arrives at the so-called *dead-beat controller*. The following theorem holds:

Theorem 5.7 (Dead-Beat Controller)**.** *If for a completely reachable system (5.56), according to Theorem 5.6, all eigenvalues of the closed-loop system matrix are set to zero, i.e. the desired characteristic polynomial is $p_{g,soll}(z) = z^n$, then every initial state $\mathbf{x}_0$ is transferred to $\mathbf{0}$ in at most n steps.*

Proof. The closed-loop system matrix $\mathbf{\Phi}_g$ for $p_{g,soll}(z) = z^n$ is

$$\mathbf{\Phi}_g = \mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} . \tag{5.100}$$

This matrix is nilpotent of order n, i.e. it holds $\mathbf{N}^k = \mathbf{0}$ for $k \geq n$. With this, however, it can be shown for any initial state $\mathbf{x}_0$ for the closed loop $\mathbf{x}_{k+1} = \mathbf{\Phi}_g \mathbf{x}_k$, that because of

$$\mathbf{x}_k = \mathbf{\Phi}_g^k \mathbf{x}_0 \tag{5.101}$$

holds $\mathbf{x}_k = \mathbf{0}$ for $k \geq n$. □

Exercise 5.12. Design a dead-beat controller for the linear, time-invariant, discrete-time system

$$\mathbf{x}_{k+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k$$

> and determine the region $\mathcal{D}$ in the $(x_{1,0},\, x_{2,0})$-plane, where the initial states $\mathbf{x}_0$ are allowed to lie, so that the magnitude of the control input is always less than or equal to 1, i.e., $|u_k| \leq 1$ for $k = 0, 1, 2, \ldots$.

The PI State Controller

In (5.94), the pre-factor g of the state feedback controller (5.57) was calculated in such a way that the steady-state control error

$$\lim_{k\to\infty} e_k = \lim_{k\to\infty} (r_k - y_k) \tag{5.102}$$

due to the input step sequence $(r_k) = r_0\left(1^k\right) = (r_0, r_0, r_0, \ldots)$ becomes zero. This is of course no longer fulfilled if the plant parameters deviate from the nominal value or disturbances act on the plant. Consider the system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k + \mathbf{\Gamma}_v v_k, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.103a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k \tag{5.103b}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input u, the output y, the disturbance v as well as the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$ and $\mathbf{\Gamma}, \mathbf{\Gamma}_v, \mathbf{c} \in \mathbb{R}^n$. If one inserts the state feedback controller (5.57) with $(r_k) = \left(0^k\right)$ for u_k, then one obtains

$$\mathbf{x}_{k+1} = \underbrace{\left(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)}_{\mathbf{\Phi}_g}\mathbf{x}_k + \mathbf{\Gamma}_v v_k\,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.104a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k \tag{5.104b}$$

or for a constant disturbance sequence $v_k = v_0\left(1^k\right)$ the steady-state control error is calculated to be

$$\lim_{k\to+\infty} (r_k - y_k) = -\lim_{z\to1}(z-1)y_z(z) = -\mathbf{c}^{\mathrm{T}}\left(\mathbf{E} - \mathbf{\Phi} - \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}\right)^{-1}\mathbf{\Gamma}_v v_0 \neq 0\,. \tag{5.105}$$

For this reason, as already done with the frequency response method, an integral part must be incorporated in the controller in order to be able to suppress at least constant disturbances and parameter variations in steady state. For this purpose, a so-called *PI state controller* of the form

$$x_{I,k+1} = x_{I,k} + \left(r_k - \underbrace{y_k}_{\mathbf{c}^{\mathrm{T}}\mathbf{x}_k}\right) \tag{5.106a}$$

$$u_k = \begin{bmatrix}\mathbf{k}_x^{\mathrm{T}} & k_I\end{bmatrix}\begin{bmatrix}\mathbf{x}_k \\ x_{I,k}\end{bmatrix} + k_P\left(r_k - \underbrace{y_k}_{\mathbf{c}^{\mathrm{T}}\mathbf{x}_k}\right) \tag{5.106b}$$

is used. The controller parameters $\mathbf{k}_x^{\mathrm{T}}$, k_I and k_P are now designed in two steps:

Step 1: In the first step, for the system (5.103) augmented by an integrator

$$\begin{bmatrix}\mathbf{x}_{k+1}\\ x_{I,k+1}\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{\Phi} & \mathbf{0}\\ -\mathbf{c}^{\mathrm{T}} & 1\end{bmatrix}}_{\mathbf{\Phi}_I}\begin{bmatrix}\mathbf{x}_k\\ x_{I,k}\end{bmatrix} + \underbrace{\begin{bmatrix}\mathbf{\Gamma}\\ 0\end{bmatrix}}_{\mathbf{\Gamma}_I}u_k + \underbrace{\begin{bmatrix}\mathbf{0}\\ 1\end{bmatrix}}_{\mathbf{\Gamma}_{r,I}}r_k + \underbrace{\begin{bmatrix}\mathbf{\Gamma}_v\\ 0\end{bmatrix}}_{\mathbf{\Gamma}_{v,I}}v_k \tag{5.107a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k \tag{5.107b}$$

a state feedback controller

$$u_k = \begin{bmatrix}\mathbf{k}_1^{\mathrm{T}} & k_2\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ x_{I,k}\end{bmatrix} \tag{5.108}$$

is designed according to Theorem 5.6. A comparison of (5.106b) with (5.108) shows that

$$\mathbf{k}_x^{\mathrm{T}} - k_P\mathbf{c}^{\mathrm{T}} = \mathbf{k}_1^{\mathrm{T}} \tag{5.109a}$$

and

$$k_I = k_2 \ . \tag{5.109b}$$

Note that this is always possible under certain conditions, because the following theorem holds:

Theorem 5.8. *If the system matrix $\mathbf{\Phi}$ of the system (5.107) has no eigenvalue at 1 and the transfer function from u_k to y_k has no zero at 1, then the complete reachability of $(\mathbf{\Phi}, \mathbf{\Gamma})$ implies the complete reachability of $(\mathbf{\Phi}_I, \mathbf{\Gamma}_I)$.*

Proof: The proof is based on the PBH eigenvector test and will be omitted here. See, e.g., the lecture notes of *Automatisierung* for details [5.2].

Step 2: In the second step, according to (5.109), the parameters $\mathbf{k}_x^{\mathrm{T}}$ and k_P must be determined. Since this problem is underdetermined, k_P is usually fixed and $\mathbf{k}_x^{\mathrm{T}}$ is calculated from (5.109). Assuming that $\mathbf{x}_0 = \mathbf{0}$ and $x_{I,0} = 0$ at time $t = 0$, then it follows from (5.106b)

$$u_0 = k_P r_0 \ . \tag{5.110}$$

If the system matrix $\mathbf{\Phi}$ is stable, i.e., all eigenvalues are inside the unit circle, then the output in steady state is

$$\lim_{k\to\infty} y_k = y_\infty = \mathbf{c}^{\mathrm{T}}(\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}u_\infty \ . \tag{5.111}$$

In this case, it is now appropriate to choose the proportional gain k_P in such a way that at time zero the control input u_0 assumes the same value that is also required for $k \to \infty$ to satisfy the condition $y_\infty = r_0$, i.e.

$$u_0 = k_P r_0 = \frac{r_0}{\mathbf{c}^{\mathrm{T}}(\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}} = u_\infty \tag{5.112}$$

or

$$k_P = \frac{1}{\mathbf{c}^{\mathrm{T}}(\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}} \ . \tag{5.113}$$

Figure 5.8 shows the block diagram of a discrete-time, linear, time-invariant system with PI state controller.

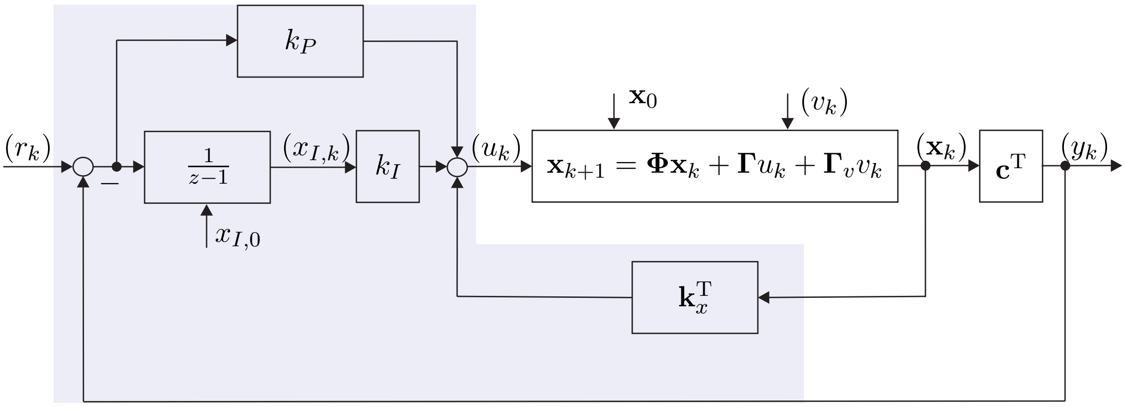
Example 5.8 (Simulation Example). For the two-mass oscillator in Figure 5.7 with the corresponding state space model (5.98), a discrete-time PI state controller according to (5.106) for a sampling time $T_a = 2$ should be designed in such a way that the poles of the closed-loop system are at $\exp(\lambda_j T_a)$, $j = 1, \dots, 5$, with $\lambda_{1,2} = -0.5 \pm 0.5I$, $\lambda_3 = -1$, $\lambda_4 = -2$ and $\lambda_5 = -3$ (Matlab command `place` or `acker`).

With these specifications, the PI state controller (5.106) for this example is given by

$$x_{I,k+1} = x_{I,k} + (r_k - y_k) \tag{5.114a}$$

$$u_k = \begin{bmatrix} 0.2163 & 0.7201 & 2.4323 & 6.7892 & -1.5585 \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ x_{I,k} \end{bmatrix} + (-1)(r_k - y_k) \, . \tag{5.114b}$$

A simulation example in the form of Matlab/Simulink files for the PI state controller of the two-mass oscillator of Figure 5.7 is available on our homepage.



k_P
$\mathbf{x}_0$
(v_k)
(r_k)
$-$
$\frac{1}{z-1}$
$(x_{I,k})$
k_I
(u_k)
$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k + \mathbf{\Gamma}_v v_k$
$(\mathbf{x}_k)$
$\mathbf{c}^{\mathrm{T}}$
(y_k)
$x_{I,0}$
$\mathbf{k}_x^{\mathrm{T}}$

Figure 5.8: Block diagram of the PI state controller for the discrete-time case.

5.3.2 State Observers

The disadvantage of the state feedback controller is obviously that its implementation requires the measurement of the entire state $\mathbf{x}$. In many cases, of course, this is not possible, which is why one asks the question whether it is possible to reconstruct the state $\mathbf{x}$ solely from the knowledge of the output y and the control input u. To answer this question, consider the linear, time-invariant, discrete-time single-input system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k \, , \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.115a}$$

$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k + du_k \tag{5.115b}$$

with the state $\mathbf{x} \in \mathbb{R}^n$, the input u, the output y as well as the matrices $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$, $\mathbf{\Gamma}, \mathbf{c} \in \mathbb{R}^n$ and $d \in \mathbb{R}$. A device that estimates the state $\mathbf{x}$ at time k from the knowledge of the inputs $(u_0, u_1, \ldots, u_k)$ and the outputs $(y_0, y_1, \ldots, y_k)$ is also called an *observer*. It will be shown in the following that such an observer for (5.115) can be constructed if and only if the system (5.115) is completely observable. All considerations can be directly transferred to the continuous-time case without additional effort.

Trivial Observer (Simulator)

The simplest way to estimate the state $\mathbf{x}$ is to simulate the mathematical model of the plant according to (5.115)

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma} u_k \,, \qquad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \tag{5.116a}$$

$$\hat{y}_k = \mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_k + d u_k \tag{5.116b}$$

in the computer with the *estimated state* $\hat{\mathbf{x}}$. The deviation of the estimated state $\hat{\mathbf{x}}$ from the actual state $\mathbf{x}$, the so-called *estimation error* $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$, then satisfies the following difference equation

$$\mathbf{e}_{k+1} = \mathbf{\Phi}\mathbf{e}_k \,, \qquad \mathbf{e}(0) = \mathbf{e}_0 = \hat{\mathbf{x}}_0 - \mathbf{x}_0 \;. \tag{5.117}$$

An observer of the form (5.116), which simply represents a copy of the plant model in the computer, is also called a *trivial observer* or *simulator* and it has the following disadvantages:

(1) The error dynamics (5.117) are obviously only stable if the plant is stable, i.e., all eigenvalues of $\mathbf{\Phi}$ lie inside the unit circle and

(2) the decay of estimation errors $\mathbf{e}_0$ for stable plants cannot be influenced but is determined by the plant dynamics.

The trivial observer (5.116) does not yet make use of the fact that a measurement, namely the measurement of y, is available for the system (5.115). This consideration leads finally to the so-called *full-order Luenberger observer*, which is discussed in the following.

> *Example* 5.9 (Simulation Example). For the two-mass oscillator in Figure 5.7 with the corresponding mathematical model in state space representation (5.98), the

discrete-time trivial observer according to (5.116) for a sampling time $T_a = 2$ is

$$\begin{bmatrix} \hat{z}_{1,k+1} \\ \hat{v}_{1,k+1} \\ \hat{z}_{2,k+1} \\ \hat{v}_{2,k+1} \end{bmatrix} = \begin{bmatrix} 0.2036 & 0.5061 & 0.6941 & 1.3366 \\ -0.3724 & -0.1688 & 0.2387 & 0.9328 \\ 0.0694 & 0.1337 & 0.7589 & 1.5754 \\ 0.0239 & 0.0933 & -0.1814 & 0.5774 \end{bmatrix} \begin{bmatrix} \hat{z}_{1,k} \\ \hat{v}_{1,k} \\ \hat{z}_{2,k} \\ \hat{v}_{2,k} \end{bmatrix} + \begin{bmatrix} -0.8987 \\ -0.5061 \\ -0.1023 \\ -0.1337 \end{bmatrix} u_k \tag{5.118a}$$

$$\hat{y}_k = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{z}_{1,k} \\ \hat{v}_{1,k} \\ \hat{z}_{2,k} \\ \hat{v}_{2,k} \end{bmatrix} . \tag{5.118b}$$

A simulation example in the form of Matlab/Simulink files for the trivial observer of the two-mass oscillator of Figure 5.7 is available on our homepage.

Full-Order Luenberger Observer

If one adds an additional term $\hat{\mathbf{k}}(\hat{y}_k - y_k)$, $\hat{\mathbf{k}} \in \mathbb{R}^n$, to the trivial observer of (5.116), then one obtains the so-called *full-order Luenberger observer*

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}u_k + \hat{\mathbf{k}}(\hat{y}_k - y_k)\,, \qquad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \tag{5.119a}$$

$$\hat{y}_k = \mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_k + du_k\ . \tag{5.119b}$$

The corresponding error dynamics for $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ with $\mathbf{x}$ from (5.115) and $\hat{\mathbf{x}}$ from (5.119) is given by

$$\mathbf{e}_{k+1} = \underbrace{\left(\mathbf{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\right)}_{\mathbf{\Phi}_e}\mathbf{e}_k\,, \qquad \mathbf{e}(0) = \mathbf{e}_0 = \hat{\mathbf{x}}_0 - \mathbf{x}_0\ . \tag{5.120}$$

Now the question arises under which conditions $\hat{\mathbf{k}}$ can be designed such that the eigenvalues of the error dynamics matrix $\mathbf{\Phi}_e$ of (5.120) are at prescribed desired locations. This problem is very similar to the problem of state feedback design, discussed in detail in Section 5.3.1. Indeed, a theorem completely analogous to Theorem 5.6 can be stated here:

Theorem 5.9 (Ackermann's formula for state observer design)**.** *The eigenvalues of the error dynamics matrix $\mathbf{\Phi}_e$ of* (5.120) *of the full-order observer* (5.119) *for the system* (5.115) *can be* arbitrarily placed *by $\hat{\mathbf{k}}$ if and only if the system* (5.115) *is* completely observable. *The vector* $\hat{\mathbf{k}}$ *is calculated according to the relation*

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}}_{\mathbf{e}_n} = \underbrace{\begin{bmatrix} \mathbf{c}^{\mathrm{T}} \\ \mathbf{c}^{\mathrm{T}}\mathbf{\Phi} \\ \vdots \\ \mathbf{c}^{\mathrm{T}}\mathbf{\Phi}^{n-1} \end{bmatrix}}_{\mathcal{O}(\mathbf{c}^{\mathrm{T}},\mathbf{\Phi})} \hat{\mathbf{v}}_1 \tag{5.121a}$$

$$\hat{\mathbf{k}} = -\hat{p}_0\hat{\mathbf{v}}_1 - \hat{p}_1\mathbf{\Phi}\hat{\mathbf{v}}_1 - \cdots - \hat{p}_{n-1}\mathbf{\Phi}^{n-1}\hat{\mathbf{v}}_1 - \mathbf{\Phi}^n\hat{\mathbf{v}}_1 = -\hat{p}_{g,soll}(\mathbf{\Phi})\hat{\mathbf{v}}_1 \tag{5.121b}$$

with $\hat{p}_{g,soll}(z) = \hat{p}_0 + \hat{p}_1 z + \hat{p}_2 z^2 + \cdots + \hat{p}_{n-1}z^{n-1} + z^n$ *as the desired characteristic polynomial of the error dynamics matrix* $\mathbf{\Phi}_e$.

Proof. Due to the fact that the characteristic polynomial of a matrix is equal to the characteristic polynomial of the transpose of this matrix, i.e.

$$\det(z\mathbf{E} - \mathbf{\Phi}_e) = \det\left(z\mathbf{E} - \mathbf{\Phi} - \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\right) = \det\left(z\mathbf{E} - \mathbf{\Phi}^{\mathrm{T}} - \mathbf{c}\hat{\mathbf{k}}^{\mathrm{T}}\right) = \det\left(z\mathbf{E} - \mathbf{\Phi}_e^{\mathrm{T}}\right) , \tag{5.122}$$

the design of $\hat{\mathbf{k}}$ for placing the eigenvalues of $\mathbf{\Phi}_e = \mathbf{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}$ can be performed based on $\mathbf{\Phi}_e^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}} + \mathbf{c}\hat{\mathbf{k}}^{\mathrm{T}}$. Comparing $\mathbf{\Phi}_e^{\mathrm{T}}$ with $\mathbf{\Phi}_g = \mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}}$ of (5.58), one can see that the *observer design* according to the so-called *duality principle* can be mapped to the state feedback design by simply replacing

$$\begin{array}{ccc} \mathbf{\Phi}^{\mathrm{T}} & \text{by} & \mathbf{\Phi} \\ \mathbf{c} & \text{by} & \mathbf{\Gamma} \\ \hat{\mathbf{k}} & \text{by} & \mathbf{k}\,. \end{array} \tag{5.123}$$

Using this duality principle, Theorem 5.9 can be directly derived from Theorem 5.6. □

Example 5.10 (Simulation Example). For the two-mass oscillator in Figure 5.7 with the corresponding mathematical model in state space representation (5.98), a discrete-time full-order Luenberger observer according to (5.119) for a sampling time $T_a = 2$ should be designed in such a way that the poles of the error dynamics matrix are at $\exp(\lambda_j T_a)$, $j = 1, \ldots, 4$, with $\lambda_{1,2} = -3 \pm 3I$ and $\lambda_{3,4} = -1 \pm I$ (Matlab command `place` or `acker`).

With these specifications, the discrete-time full-order Luenberger observer (5.119) for this example is given by

$$\hat{\mathbf{x}}_{k+1} = \underbrace{\left(\mathbf{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\right)}_{\mathbf{\Phi}_e} \hat{\mathbf{x}}_k + \mathbf{\Gamma} u_k - \hat{\mathbf{k}} y_k \tag{5.124a}$$

$$\hat{y}_k = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \hat{\mathbf{x}}_k \tag{5.124b}$$

with

$$\mathbf{\Phi}_e = \begin{bmatrix} 0.2036 & 0.5061 & -0.6543 & 1.3366 \\ -0.3724 & -0.1688 & 0.1383 & 0.9328 \\ 0.0694 & 0.1337 & -0.7201 & 1.5754 \\ 0.0239 & 0.0933 & -0.2390 & 0.5774 \end{bmatrix} \tag{5.125a}$$

$$\mathbf{\Gamma} = \begin{bmatrix} -0.8987 \\ -0.5061 \\ -0.1023 \\ -0.1337 \end{bmatrix} \tag{5.125b}$$

$$\hat{\mathbf{k}} = \begin{bmatrix} -1.3483 \\ -0.1004 \\ -1.4790 \\ -0.0576 \end{bmatrix} . \tag{5.125c}$$

A simulation example in the form of Matlab/Simulink files for the full-order Luenberger observer of the two-mass oscillator of Figure 5.7 is available on our homepage.

Exercise 5.13. Show that if all eigenvalues of the error dynamics matrix $\mathbf{\Phi}_e = \mathbf{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}$ of (5.119) are set to zero - i.e., the desired characteristic polynomial is $\hat{p}_{g,soll}(z) = z^n$ - then any initial error $\mathbf{e}_0 = \hat{\mathbf{x}}_0 - \mathbf{x}_0$ is driven to $\mathbf{0}$ in at most n steps. In analogy to the controller of the same name, such an observer is called a *dead-beat observer*.

Exercise 5.14. Design a dead-beat observer for the system

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k$$

$$y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} .$$

Furthermore, determine the region $\mathcal{D}$ of admissible initial errors $\mathbf{e}_0$ in the $(e_{0,1}, e_{0,2})$-plane such that

$$\|\mathbf{e}_j\|_2^2 < 1\,, \qquad j = 0, 1, \dots\ .$$

Hint: Use the MATLAB command `place` or `acker`.

5.3.3 Separation Principle

If one cannot measure the entire state $\mathbf{x}$ but still wants to use a state feedback controller, it is natural to combine the state feedback controller with a state observer. For this purpose, for the linear, time-invariant, discrete-time, single-input system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k\,, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{5.126a}$$
$$y_k = \mathbf{c}^{\mathrm{T}}\mathbf{x}_k \tag{5.126b}$$

a state observer according to (5.119)

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}u_k + \hat{\mathbf{k}}(\hat{y}_k - y_k)\,, \qquad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \tag{5.127a}$$
$$\hat{y}_k = \mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_k \tag{5.127b}$$

is designed and in the state feedback control law (5.57), instead of the actual state $\mathbf{x}$, the observed state $\hat{\mathbf{x}}$ is used in the form

$$u_k = \mathbf{k}^{\mathrm{T}}\hat{\mathbf{x}}_k + g r_k\ . \tag{5.128}$$

Figure 5.9 illustrates this so-called *state feedback controller/state observer* configuration.

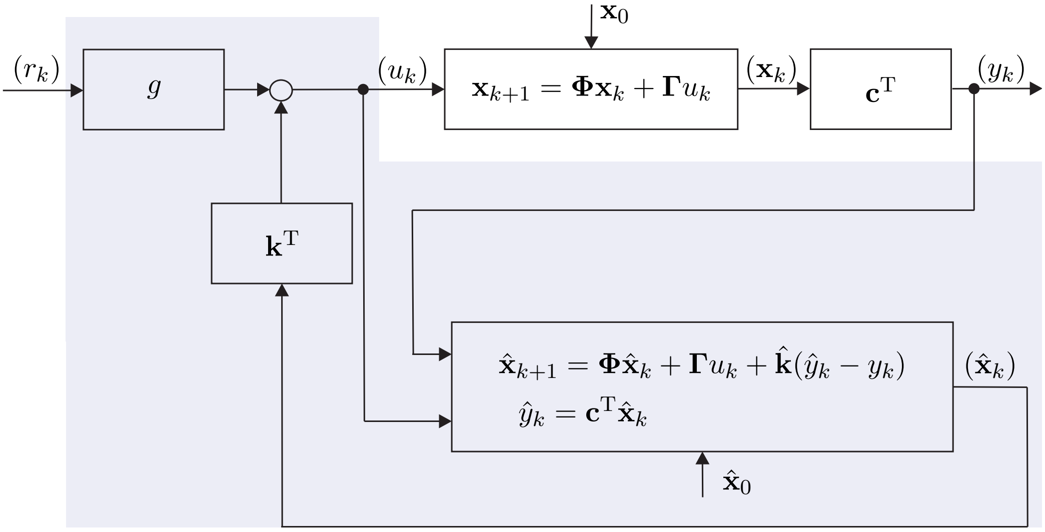

$\mathbf{x}_0$
(r_k)
g
(u_k)
$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}u_k$
$(\mathbf{x}_k)$
$\mathbf{c}^{\mathrm{T}}$
(y_k)
$\mathbf{k}^{\mathrm{T}}$
$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}u_k + \hat{\mathbf{k}}(\hat{y}_k - y_k)$
$\hat{y}_k = \mathbf{c}^{\mathrm{T}}\hat{\mathbf{x}}_k$
$(\hat{\mathbf{x}}_k)$
$\hat{\mathbf{x}}_0$

Figure 5.9: State feedback controller/state observer configuration.

If the state feedback controller and the state observer are designed separately according to Theorems 5.6 and 5.9, i.e., the eigenvalues are assigned separately, the question arises where the eigenvalues of the closed loop according to Figure 5.9 are located? The answer to this question is given by the so-called *separation principle*:

Theorem 5.10 (Separation Principle)**.** *If the system* (5.126) *is completely reachable and completely observable, then the characteristic polynomial of the closed loop of Figure 5.9 according to equations* (5.126)–(5.128) *is*

$$\begin{aligned} p_{ges}(z) &= \det\Big(z\mathbf{E}_{n\times n} - \big(\boldsymbol{\Phi} + \boldsymbol{\Gamma}\mathbf{k}^{\mathrm{T}}\big)\Big)\det\Big(z\mathbf{E}_{n\times n} - \big(\boldsymbol{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\big)\Big) \\ &= p_{g,soll}(z)\hat{p}_{g,soll}(z) \end{aligned} \tag{5.129}$$

with the desired characteristic polynomials $p_{g,soll}(z)$ for the state feedback design according to Theorem 5.6 and $\hat{p}_{g,soll}(z)$ for the state observer design according to Theorem 5.9.

Proof. To prove this central theorem, one writes the closed loop (5.126)-(5.128) as a difference equation system in the state $\mathbf{x}_{ges}^{\mathrm{T}} = \begin{bmatrix}\mathbf{x}^{\mathrm{T}} & \mathbf{e}^{\mathrm{T}}\end{bmatrix}$ with the estimation error $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ in the form

$$\underbrace{\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{e}_{k+1}\end{bmatrix}}_{\mathbf{x}_{ges,k+1}} = \underbrace{\begin{bmatrix}\boldsymbol{\Phi}+\boldsymbol{\Gamma}\mathbf{k}^{\mathrm{T}} & \boldsymbol{\Gamma}\mathbf{k}^{\mathrm{T}}\\ \mathbf{0} & \boldsymbol{\Phi}+\hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\end{bmatrix}}_{\boldsymbol{\Phi}_{ges}} \underbrace{\begin{bmatrix}\mathbf{x}_{k}\\ \mathbf{e}_{k}\end{bmatrix}}_{\mathbf{x}_{ges,k}} + \underbrace{\begin{bmatrix}\boldsymbol{\Gamma}g\\ \mathbf{0}\end{bmatrix}}_{\boldsymbol{\Gamma}_{ges}} r_k \tag{5.130a}$$

$$y_k = \underbrace{\begin{bmatrix}\mathbf{c}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}}\end{bmatrix}}_{\mathbf{c}_{ges}^{\mathrm{T}}} \underbrace{\begin{bmatrix}\mathbf{x}_{k}\\ \mathbf{e}_{k}\end{bmatrix}}_{\mathbf{x}_{ges,k}} \tag{5.130b}$$

One can immediately see that due to the block diagonal structure of the closed-loop system matrix $\boldsymbol{\Phi}_{ges}$, the characteristic polynomial can be calculated in the form

$$\det(z\mathbf{E}_{2n\times 2n} - \boldsymbol{\Phi}_{ges}) = \det\Big(z\mathbf{E}_{n\times n} - \big(\boldsymbol{\Phi} + \boldsymbol{\Gamma}\mathbf{k}^{\mathrm{T}}\big)\Big)\det\Big(z\mathbf{E}_{n\times n} - \big(\boldsymbol{\Phi} + \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}}\big)\Big)\ . \tag{5.131}$$

Exercise 5.15. Show the validity of relation (5.131).

□

Example 5.11 (Simulation Example). For the two-mass oscillator in Figure 5.7 with the corresponding state space model (5.98), the PI state controller (5.114) and the full-order Luenberger observer (5.124) and (5.125) should be combined.

A simulation example in the form of Matlab/Simulink files for the PI state controller with full-order Luenberger observer of the two-mass oscillator of Figure 5.7 is available on our homepage.

Exercise 5.16. Calculate the transfer function $R(z) = y_z(z)/r_z(z)$ of the *dynamic controller* resulting from the interconnection of the full-order observer with the state feedback controller. What conclusion can you draw from the order of the transfer function $R(z)$?

Exercise 5.17. In how many sampling steps can an initial deviation $\mathbf{x}_0$ of the system (5.126) be driven to $\mathbf{0}$ as fast as possible using the state feedback controller/state observer configuration? Prove your answer.

Exercise 5.18. Design a state feedback controller according to Theorem 5.6 for the system

$$\begin{aligned}
\mathbf{x}_{k+1} &= \frac{1}{8}\begin{bmatrix} -1 & 0 & -4 & 1 \\ 4 & -2 & 4 & 0 \\ -5 & 0 & -2 & -1 \\ 5 & 0 & 4 & 3 \end{bmatrix}\mathbf{x}_k + \begin{bmatrix} 2 \\ 5 \\ 1 \\ 4 \end{bmatrix} u_k \\
y_k &= \begin{bmatrix} 1 & 1 & -2 & 0 \end{bmatrix}\mathbf{x}_k
\end{aligned}$$

such that all eigenvalues of the closed-loop system are located at $\lambda = 1/5$. Furthermore, calculate a full-order observer according to Theorem 5.9 for the desired eigenvalues of the error dynamics matrix at $\lambda_j = 1/20$, $j = 1, \ldots, 4$. Combine the state feedback controller and the full-order state observer according to Figure 5.9 and simulate the closed loop in Matlab/Simulink. Compare the result when replacing the full-order observer with the trivial observer of (5.116).

5.4 Literatur

[5.1] N. L. Stokey, R. E. L. Jr., and E. C. Prescott, *Recursive Methods in Economic Dynamics*. Cambridge, MA: Harvard University Press, 1989.

[5.2] A. Kugi, *Skriptum zur VU Automatisierung (WS 2024/2025)*, Institut für Automatisierungs- und Regelungstechnik, TU Wien, 2024. [Online]. Available: https://www.acin.tuwien.ac.at/file/teaching/bachelor/automatisierung/AutomatisierungVO.pdf.

[5.3] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 447–457, 1949.

[5.4] J. Ackermann, *Abtastregelung*, 3rd ed. Berlin Heidelberg: Springer, 1988.

[5.5] K. J. Åström and B. Wittenmark, *Computer-Controlled Systems*, 3rd ed. New Jersey: Prentice Hall, 1997.

[5.6] F. Gausch, A. Hofer, and K. Schlacher, *Digitale Regelkreise*. München: Oldenbourg, 1991.

[5.7] J. Lunze, *Regelungstechnik 2*, 3rd ed. Berlin Heidelberg New York: Springer, 2005.

[5.8] A. Weinmann, *Regelungen: Analyse und technischer Entwurf*, 3rd ed. Wien New York: Springer, 1998, vol. 1 und 2.

[5.9] G. Ludyk, *Theoretische Regelungstechnik 1*. Berlin Heidelberg: Springer, 1995.

[5.10] T. Kailath, *Linear Systems*. New York: Prentice Hall, 1980.

6 Optimal Parameter & State Estimation and State Feedback

The discrete-time description of dynamical systems, as introduced in the previous chapter, facilitates the optimization-based parameter and state estimation as well as optimal feedback control. Optimization-based methods represent a control-theoretic design philosophy which has a long tradition and continued relevance due to its importance to state-of-the-art control concepts such as moving horizon estimation, model predictive control, and reinforcement leaning. So far, the lectures on *Quantum Technology and Devices: Experimental Techniques and Platforms* and previous chapters of this course, broadly speaking, have covered control-theoretic concepts based on

- the frequency domain and loop-shaping,
- Lyapunov Theory, and
- pole placement.

Optimization-based methods pose an alternative set of concepts that revolve around the minimization of some cost function $\mathbf{C}(\mathbf{x})$ or, in case stochastic signals are involved, the minimization of the expected cost $\mathrm{E}(\mathbf{C}(\mathbf{x}))$, where E is the expected value of a random variable with respect to its underlying probability distribution. This domain is vast, thus only selected results will be presented, focusing on

- the minimum-variance estimation of a dynamical system's parameters based on time series data,
- the recursive minimum-variance estimation of the state variable, leading to the Kalman filter as an optimal state observer, and
- the design of an optimal feedback controller gain with respect to a quadratic cost function of the state and input variables, yielding the linear-quadratic regulator (LQR).

6.1 Minimum-variance parameter estimation

First, the estimation of system parameters based on complete batches of data will be discussed. Afterwards, the theory is amended to make the estimator recursive, i.e., adapting the estimated parameters from one data sample to another.

6.1.1 Batch-based Minimum-Variance Estimation

Similar to the parameter-based variations in continuous time introduced towards the end of Chapter 2, the dynamics of an LTI system or variations in a linearized nonlinear system can be written as

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \ , \tag{6.1}$$

where the stochastic disturbance (or model error) is $\mathbf{v}$, the known $(m \times n)$ data matrix is $\mathbf{S} \in \mathbb{R}^{m\times n}$, the n-dimensional random vector to be determined is $\mathbf{p} \in \mathbb{R}^n$, as well as the m-dimensional output vector is $\mathbf{y} \in \mathbb{R}^m$, with the following holding:

$$\begin{aligned} &\mathrm{E}(\mathbf{v}) = \mathbf{0}, \qquad \mathrm{cov}(\mathbf{v}) = \mathrm{E}\left(\mathbf{v}\mathbf{v}^{\mathrm{T}}\right) = \mathbf{Q} \qquad \text{with} \quad \mathbf{Q} \geq 0 \\ &\mathrm{E}(\mathbf{p}) = \mathbf{0}, \qquad \mathrm{cov}(\mathbf{p}) = \mathrm{E}\left(\mathbf{p}\mathbf{p}^{\mathrm{T}}\right) = \mathbf{R} \qquad \text{with} \quad \mathbf{R} \geq 0 \\ &\mathrm{E}\left(\mathbf{p}\mathbf{v}^{\mathrm{T}}\right) = \mathbf{N} \ . \end{aligned} \tag{6.2}$$

A structure of the form (6.1) could be used, among other use cases, to represent the following example:

Example 6.1. The inputs (u_k) of an autoregressive model with exogenous input (ARX model), used, for instance, in economic research, may be related to the outputs (y_k) and the disturbances (v_k) via

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + a_3 y_{k-3} + b_0 u_k + b_1 u_{k-1} + v_k \ . \tag{6.3}$$

Assuming that sequential data of (u_k) and (y_k) are known, one may then write different time samples of (6.3) in individual rows and thus obtain

$$\mathbf{y} := \begin{bmatrix} y_3 \\ y_4 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} y_2 & y_1 & y_0 & u_3 & u_2 \\ y_3 & y_2 & y_1 & u_4 & u_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{N-1} & y_{N-2} & y_{N-3} & u_N & u_{N-1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} v_3 \\ v_4 \\ \vdots \\ v_N \end{bmatrix} =: \mathbf{S}\mathbf{p} + \mathbf{v} \ ,$$

where the row dimension m of $\mathbf{S}$ can be (much) larger than the column dimension n. This is exactly of the structure given by (6.1).

Exercise 6.1. Based on (6.3), derive a state-space model of the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\nu_k \\ y_k &= \mathbf{c}^{\mathrm{T}}\mathbf{x}_k + d\nu_k \ . \end{aligned}$$

> *Hint:* define the state as
>
> $$\mathbf{x}_k := \begin{bmatrix} y_k \\ y_{k-1} \\ y_{k-2} \\ u_{k-1} \end{bmatrix}$$
>
> and the input variable as $\nu_k := u_{k-1}$.

In the following, it is assumed that the matrix $\left(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)$ is regular (i.e., invertible). Now, the goal is to find a linear estimator

$$\hat{\mathbf{p}} = \mathbf{Ky} \tag{6.4}$$

with a constant $(n \times m)$-matrix $\mathbf{K} \in \mathbb{R}^{n\times m}$, such that the following minimization problem

$$\min_{\mathbf{K}} \mathrm{E}\left(\|\mathbf{e}\|_2^2\right) = \min_{\mathbf{K}} \mathrm{E}\left([\mathbf{p} - \mathbf{Ky}]^{\mathrm{T}}[\mathbf{p} - \mathbf{Ky}]\right) \tag{6.5}$$

is solved. By expanding (6.5) and using the relation $\mathrm{trace}(\mathbf{KSR}) = \mathrm{trace}\left(\mathbf{R}(\mathbf{KS})^{\mathrm{T}}\right)$, one obtains

$$\begin{aligned}
&\min_{\mathbf{K}} \mathrm{E}\left([\mathbf{p} - \mathbf{Ky}]^{\mathrm{T}}[\mathbf{p} - \mathbf{Ky}]\right) = \\
&\min_{\mathbf{K}} \left\{ \underbrace{\mathrm{E}\left([\mathbf{p} - \mathbf{KSp}]^{\mathrm{T}}[\mathbf{p} - \mathbf{KSp}]\right)}_{\mathrm{trace}\left(\mathrm{E}\left([\mathbf{E}-\mathbf{KS}]\mathbf{pp}^{\mathrm{T}}[\mathbf{E}-\mathbf{KS}]^{\mathrm{T}}\right)\right)} - 2 \underbrace{\mathrm{E}\left([\mathbf{p} - \mathbf{KSp}]^{\mathrm{T}}\mathbf{Kv}\right)}_{\mathrm{trace}(\mathrm{E}([\mathbf{E}-\mathbf{KS}]\mathbf{pv}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}))} + \underbrace{\mathrm{E}\left(\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\mathbf{Kv}\right)}_{\mathrm{trace}(\mathrm{E}(\mathbf{Kvv}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}))} \right\} = \\
&\min_{\mathbf{K}} \left\{ \mathrm{trace}\left([\mathbf{E} - \mathbf{KS}]\mathbf{R}[\mathbf{E} - \mathbf{KS}]^{\mathrm{T}} - 2\mathbf{KN}^{\mathrm{T}} - \mathbf{K}\left[\mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right]\mathbf{K}^{\mathrm{T}} + \mathbf{KQK}^{\mathrm{T}}\right) \right\} = \\
&\min_{\mathbf{K}} \left\{ \mathrm{trace}\left(\mathbf{K}\left(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)\mathbf{K}^{\mathrm{T}} - 2\mathbf{K}\left(\mathbf{SR} + \mathbf{N}^{\mathrm{T}}\right)\right) \right\} .
\end{aligned} \tag{6.6}$$

The optimal estimator $\mathbf{K}$ can then be found by taking the derivative of (6.6) and equating to the zero matrix, i.e.,

$$\frac{\partial}{\partial \mathbf{K}} \mathrm{trace}\left(\mathbf{K}\left(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)\mathbf{K}^{\mathrm{T}} - 2\mathbf{K}\left(\mathbf{SR} + \mathbf{N}^{\mathrm{T}}\right)\right) = \mathbf{0} , \tag{6.7}$$

which yields the **minimum-variance estimator**

$$\hat{\mathbf{p}} = \mathbf{Ky} = \left(\mathbf{RS}^{\mathrm{T}} + \mathbf{N}\right)\left(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)^{-1}\mathbf{y} . \tag{6.8}$$

Note that we made the necessary requirement for the regularity of the matrix $\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}$ above.

> *Exercise* 6.2. Verify the step from (6.7) to (6.8), using index notation and Einstein's summation convention, for instance. Keep in mind that covariance matrices are symmetric.

At this point, (6.8) can be related to other, simpler estimation methods which are special cases of the minimum-variance estimator:

- If there is no a priori information available on the parameter vector $\mathbf{p}$, i.e. $\mathbf{R}^{-1} = \mathbf{0}$, then one obtains the **Gauss-Markov estimator**, which is also known as the BLUE (best linear unbiased estimate). It is given by
$$\hat{\mathbf{p}} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} \; . \tag{6.9}$$
- If there is also no information available on the disturbance vector $\mathbf{v}$, then $\mathbf{Q}^{-1} = \mathbf{0}$, which yields the **least squares estimator**
$$\hat{\mathbf{p}} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y} \; . \tag{6.10}$$

The minimum-variance estimator (6.8) can be given a geometric interpretation which will lead to recursive minimum-variance estimation in the next section. For this, we first re-state the estimator, using the identities

$$\mathrm{E}\left(\mathbf{p}\mathbf{y}^{\mathrm{T}}\right) = \mathrm{E}\left(\mathbf{p}\mathbf{p}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} + \mathbf{p}\mathbf{v}^{\mathrm{T}}\right) = \left(\mathbf{R}\mathbf{S}^{\mathrm{T}} + \mathbf{N}\right) \tag{6.11}$$

and

$$\mathrm{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right) = \mathrm{E}\left([\mathbf{S}\mathbf{p}+\mathbf{v}][\mathbf{S}\mathbf{p}+\mathbf{v}]^{\mathrm{T}}\right) = \left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}} + \mathbf{Q} + \mathbf{S}\mathbf{N} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right) \; . \tag{6.12}$$

This allows us to formulate the following theorem:

Theorem 6.1 (Minimum-Variance Estimator)**.** *For the system of equations* (6.1)

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{6.13}$$

with the stochastic quantities $\mathbf{p}$, $\mathbf{v}$, *and* $\mathbf{y}$, *it is assumed that* $\mathrm{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right)$ *is invertible. The optimal linear estimate* $\hat{\mathbf{p}}$ *of* $\mathbf{p}$, *which minimizes the expected value of the quadratic error* $\mathrm{E}\left([\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}[\mathbf{p}-\hat{\mathbf{p}}]\right)$, *is given by*

$$\hat{\mathbf{p}} = \mathrm{E}\left(\mathbf{p}\mathbf{y}^{\mathrm{T}}\right)\left[\mathrm{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right)\right]^{-1}\mathbf{y} \; , \tag{6.14}$$

with the corresponding error covariance matrix

$$\begin{aligned} \mathrm{cov}(\mathbf{e}) &= \mathrm{E}\left([\mathbf{p}-\hat{\mathbf{p}}][\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\right) = \mathrm{E}\left(\mathbf{p}\mathbf{p}^{\mathrm{T}}\right) - \mathrm{E}\left(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}}\right) \\ &= \mathrm{E}\left(\mathbf{p}\mathbf{p}^{\mathrm{T}}\right) - \mathrm{E}\left(\mathbf{p}\mathbf{y}^{\mathrm{T}}\right)\left[\mathrm{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right)\right]^{-1}\mathrm{E}\left(\mathbf{y}\mathbf{p}^{\mathrm{T}}\right) \; . \end{aligned} \tag{6.15}$$

> *Exercise* 6.3. Show the validity of the relationship
>
> $$\mathrm{E}\left([\mathbf{p}-\hat{\mathbf{p}}][\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\right) = \mathrm{E}\left(\mathbf{p}[\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\right) \quad \text{or} \quad \mathrm{E}\left(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}}\right) = \mathrm{E}\left(\mathbf{p}\hat{\mathbf{p}}^{\mathrm{T}}\right) \; . \tag{6.16}$$

> *Exercise* 6.4. Assume that the expected values $\mathrm{E}(\mathbf{y})$ and $\mathrm{E}(\mathbf{p})$ are not zero as in (6.2), but rather $\mathrm{E}(\mathbf{y}) = \mathbf{y}_0 \neq \mathbf{0}$ and $\mathrm{E}(\mathbf{p}) = \mathbf{p}_0 \neq \mathbf{0}$. Show that the minimum-variance estimate of the form
>
> $$\hat{\mathbf{p}} = \mathbf{K}\mathbf{y} + \mathbf{b}$$
>
> with the constant vector $\mathbf{b}$ is then given by
>
> $$\hat{\mathbf{p}} = \mathbf{p}_0 + \mathrm{E}\Big([\mathbf{p} - \mathbf{p}_0][\mathbf{y} - \mathbf{y}_0]^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big([\mathbf{y} - \mathbf{y}_0][\mathbf{y} - \mathbf{y}_0]^{\mathrm{T}}\Big)\Big]^{-1}(\mathbf{y} - \mathbf{y}_0) \ .$$

By the *minimum-variance estimation of a linear function*

$$\mathbf{z} = \mathbf{C}\mathbf{p} \tag{6.17}$$

with the optimal estimator

$$\hat{\mathbf{z}} = \mathbf{K}_{\mathbf{z}}\mathbf{y} \tag{6.18}$$

based on the measurements

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \ , \tag{6.19}$$

one understands the solution to the minimization problem

$$\min_{\mathbf{K}_{\mathbf{z}}} \mathrm{E}\Big([\mathbf{z} - \hat{\mathbf{z}}]^{\mathrm{T}}[\mathbf{z} - \hat{\mathbf{z}}]\Big) \ . \tag{6.20}$$

The following theorem now holds:

Theorem 6.2 (Minimum-Variance Estimator of a Linear Function)**.** *The linear minimum-variance estimate* (6.18) *of a linear function* $\mathbf{C}\mathbf{p}$ *based on the measurements* (6.19) *is equivalent to the linear function of the minimum-variance estimate* $\hat{\mathbf{p}}$ *itself, i.e., it holds that the best estimate of* $\mathbf{C}\mathbf{p}$ *is* $\mathbf{C}\hat{\mathbf{p}}$*.*

6.1.2 Recursive Minimum-Variance Estimation

In the next step, we shall investigate how the optimal estimate $\hat{\mathbf{p}}$ from (6.14) improves through the addition of new measurements. This is particularly essential for on-line applications. The method is based on the properties of the projection theorem in a Hilbert space. If $\mathcal{U}_1$ and $\mathcal{U}_2$ denote two subspaces of a Hilbert space, then the projection of a vector $\mathbf{p}$ onto the subspace $\mathcal{U}_1 + \mathcal{U}_2$ is identical to the projection of $\mathbf{p}$ onto $\mathcal{U}_1$ plus the projection onto $\mathcal{U}_2^*$, where $\mathcal{U}_2^*$ is orthogonal to $\mathcal{U}_1$ and satisfies the relationship $\mathcal{U}_1 \oplus \mathcal{U}_2^* = \mathcal{U}_1 + \mathcal{U}_2$. When $\mathcal{U}_2$ is spanned by a finite number of vectors, then the differences of these vectors with their projections onto $\mathcal{U}_1$ span the subspace $\mathcal{U}_2^*$. Figure 6.1 illustrates this situation.

Thus one can state the following theorem:

Theorem 6.3 (Recursive Minimum-Variance Estimation)**.** *Let* $\mathbf{p}$ *be a random vector of a Hilbert space* $\mathcal{H}$ *of random variables and let* $\hat{\mathbf{p}}_1$ *denote the orthogonal projection of* $\mathbf{p}$ *onto a closed subspace* $\mathcal{U}_1$ *of* $\mathcal{H}$*. According to the projection theorem,* $\hat{\mathbf{p}}_1$ *is thus the best estimate of* $\mathbf{p}$ *in* $\mathcal{U}_1$*. Furthermore, let* $\mathbf{y}_2$ *describe all those random vectors that span the subspace* $\mathcal{U}_2$ *of* $\mathcal{H}$*, and let* $\hat{\mathbf{y}}_2$ *be the orthogonal projection of* $\mathbf{y}_2$ *onto* $\mathcal{U}_1$*.*

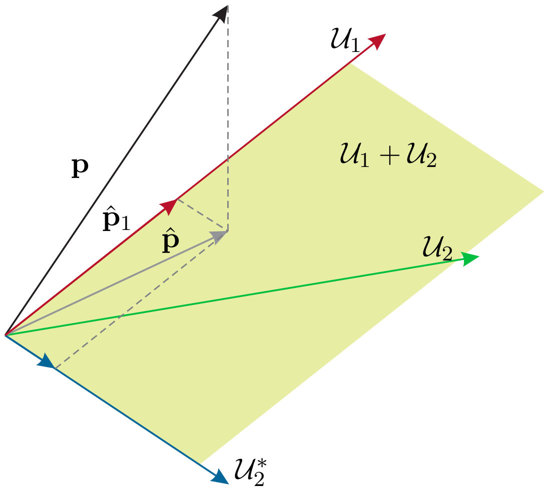


Figure 6.1: On projection onto the sum of orthogonal subspaces.

According to the projection theorem, $\hat{\mathbf{y}}_2$ is thus the best estimate of $\mathbf{y}_2$ in $\mathcal{U}_1$. With $\tilde{\mathbf{y}}_2 = \mathbf{y}_2 - \hat{\mathbf{y}}_2$, the projection $\hat{\mathbf{p}}$ of $\mathbf{p}$ onto $\mathcal{U}_1 + \mathcal{U}_2$ reads

$$\hat{\mathbf{p}} = \hat{\mathbf{p}}_1 + \mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_2 \ . \tag{6.21}$$

Thus the best estimate $\hat{\mathbf{p}}$ on $\mathcal{U}_1 + \mathcal{U}_2$ is composed of the sum of the best estimate of $\mathbf{p}$ on $\mathcal{U}_1$ ($\hat{\mathbf{p}}_1$) and the best estimate of $\mathbf{p}$ on $\mathcal{U}_2^*$ (that subspace generated by $\tilde{\mathbf{y}}_2$).

Proof of Theorem 6.3. One easily convinces oneself that $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{U}_1 \oplus \mathcal{U}_2^*$ holds and that $\mathcal{U}_2^*$ is orthogonal to $\mathcal{U}_1$. The relationship (6.21) then follows from the fact that the projection onto a sum of subspaces equals the sum of the projections onto the individual subspaces, provided these are orthogonal. □

The result of Theorem 6.3 can now also be interpreted in the following form: If $\hat{\mathbf{p}}_1$ gives the optimal estimate based on measurements that span the subspace $\mathcal{U}_1$, then when receiving new measurements that span the subspace $\mathcal{U}_2$, one needs to consider only that part that is not yet described by the measurements in $\mathcal{U}_1$, i.e., that part of the new data that is orthogonal to the old data and thus lies in the subspace $\mathcal{U}_2^*$.

As an application example, consider a system of equations of the form of (6.1)

$$\mathbf{y}_1 = \mathbf{S}_1\mathbf{p} + \mathbf{v}_1 \ . \tag{6.22}$$

Furthermore, let $\hat{\mathbf{p}}_1 = \mathrm{E}\Big(\mathbf{p}\mathbf{y}_1^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\mathbf{y}_1\mathbf{y}_1^{\mathrm{T}}\Big)\Big]^{-1}\mathbf{y}_1$ denote the optimal minimum-variance estimation of $\mathbf{p}$ according to (6.10) or (6.14) based on $\dim(\mathbf{y}_1)$ measurements with the error covariance matrix

$$\mathrm{cov}(\mathbf{p} - \hat{\mathbf{p}}_1) = \mathrm{E}\Big([\mathbf{p} - \hat{\mathbf{p}}_1][\mathbf{p} - \hat{\mathbf{p}}_1]^{\mathrm{T}}\Big) = \mathbf{P}_1 \ . \tag{6.23}$$

The question now arises of how one can improve the estimate of $\mathbf{p}$ by adding new measurements

$$\mathbf{y}_2 = \mathbf{S}_2\mathbf{p} + \mathbf{v}_2 \ . \tag{6.24}$$

For the stochastic disturbance $\mathbf{v}_2$ and the random parameter vector $\mathbf{p}$, let

$$\begin{aligned} \mathrm{E}(\mathbf{v}_2) &= \mathbf{0}, \qquad \mathrm{cov}(\mathbf{v}_2) = \mathrm{E}\Big(\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\Big) = \mathbf{Q}_2 \quad \text{with} \quad \mathbf{Q}_2 \geq 0 \\ \mathrm{E}(\mathbf{p}) &= \mathbf{0}, \qquad \mathrm{E}\Big(\mathbf{p}\mathbf{v}_2^{\mathrm{T}}\Big) = \mathbf{N}_2 \ . \end{aligned} \tag{6.25}$$

Furthermore, it is naturally sensible to assume that the disturbance $\mathbf{v}_2$ is not correlated with past measured quantities $\mathbf{y}_1$ and thus

$$\mathrm{E}\Big(\mathbf{v}_2\mathbf{y}_1^{\mathrm{T}}\Big) = \mathbf{0} \quad \text{or} \quad \mathrm{E}\Big(\mathbf{v}_2\hat{\mathbf{p}}_1^{\mathrm{T}}\Big) = \mathbf{0} \ . \tag{6.26}$$

The best estimate $\hat{\mathbf{y}}_2$ of $\mathbf{y}_2$ based on the past measured values $\mathbf{y}_1$ reads

$$\hat{\mathbf{y}}_2 = \mathbf{S}_2\hat{\mathbf{p}}_1 \ . \tag{6.27}$$

Thus according to Theorem 6.3 with $\tilde{\mathbf{y}}_2 = \mathbf{y}_2 - \hat{\mathbf{y}}_2$, one obtains the improved estimate $\hat{\mathbf{p}}_2$ as

$$\hat{\mathbf{p}}_2 = \hat{\mathbf{p}}_1 + \mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_2 \tag{6.28}$$

with

$$\mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{p}(\mathbf{p} - \hat{\mathbf{p}}_1)^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}} + \mathbf{p}\mathbf{v}_2^{\mathrm{T}}\Big) = \mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{N}_2 \tag{6.29}$$

and

$$\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big) = \mathrm{E}\Big([\mathbf{S}_2(\mathbf{p} - \hat{\mathbf{p}}_1) + \mathbf{v}_2][\mathbf{S}_2(\mathbf{p} - \hat{\mathbf{p}}_1) + \mathbf{v}_2]^{\mathrm{T}}\Big) = \mathbf{S}_2\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{Q}_2 + \mathbf{S}_2\mathbf{N}_2 + \mathbf{N}_2^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}} \ . \tag{6.30}$$

> *Exercise* 6.5. Show that the error covariance matrix can be calculated in the form
>
> $$\begin{aligned} \mathbf{P}_2 &= \mathrm{cov}(\mathbf{p} - \hat{\mathbf{p}}_2) \\ &= \mathbf{P}_1 - \Big(\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{N}_2\Big)\Big(\mathbf{S}_2\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{Q}_2 + \mathbf{S}_2\mathbf{N}_2 + \mathbf{N}_2^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}}\Big)^{-1}\Big(\mathbf{S}_2\mathbf{P}_1 + \mathbf{N}_2^{\mathrm{T}}\Big) \end{aligned} \tag{6.31}$$

Therefore, repeating this procedure recursively, the *recursive minimum-variance estimator* results as

$$\hat{\mathbf{p}}_k = \hat{\mathbf{p}}_{k-1} + \Big(\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{N}_k\Big)\Big(\mathbf{S}_k\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{Q}_k + \mathbf{S}_k\mathbf{N}_k + \mathbf{N}_k^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\Big)^{-1}(\mathbf{y}_k - \mathbf{S}_k\hat{\mathbf{p}}_{k-1}) \tag{6.32}$$

with

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \Big(\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{N}_k\Big)\Big(\mathbf{S}_k\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{Q}_k + \mathbf{S}_k\mathbf{N}_k + \mathbf{N}_k^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\Big)^{-1}\Big(\mathbf{S}_k\mathbf{P}_{k-1} + \mathbf{N}_k^{\mathrm{T}}\Big) \tag{6.33}$$

and the initial values $\mathbf{P}_{-1}$ as well as $\hat{\mathbf{p}}_{-1}$.

If one now assumes that exactly one new measurement is added in each iteration step, i.e., the quantities y_k and v_k are scalars, then by substituting $\mathbf{S}_k = \mathbf{s}_k^{\mathrm{T}}$, $\mathbf{Q}_k = \mathrm{E}(v_k^2) = q_k$ and $\mathbf{N}_k = \mathrm{E}(\mathbf{p}v_k) = \mathbf{n}_k$ into (6.32), (6.33), the recursive minimum-variance estimator becomes

$$\mathbf{k}_k = \frac{\mathbf{P}_{k-1}\mathbf{s}_k + \mathbf{n}_k}{\left(q_k + 2\mathbf{s}_k^{\mathrm{T}}\mathbf{n}_k + \mathbf{s}_k^{\mathrm{T}}\mathbf{P}_{k-1}\mathbf{s}_k\right)} \tag{6.34a}$$

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{k}_k\Big(\mathbf{s}_k^{\mathrm{T}}\mathbf{P}_{k-1} + \mathbf{n}_k^{\mathrm{T}}\Big) \tag{6.34b}$$

$$\hat{\mathbf{p}}_k = \hat{\mathbf{p}}_{k-1} + \mathbf{k}_k\Big(y_k - \mathbf{s}_k^{\mathrm{T}}\hat{\mathbf{p}}_{k-1}\Big)\ . \tag{6.34c}$$

6.2 Optimal State Observers and the Kalman Filter

Building on the previous considerations, particularly the recursive minimum variance estimation, the next step is to derive the Kalman filter, an *optimal state observer* in the sense of control theory. For the fundamentals of state observer theory, please refer to the previous chapter.

6.2.1 From parameter estimation to state estimation

Countless versions of the Kalman filter exist in the literature. In the context of this lecture, we will initially consider a linear, time-invariant, discrete-time system of the form

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}\mathbf{x}_k + \boldsymbol{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6.35a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{6.35b}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional deterministic input $\mathbf{u} \in \mathbb{R}^p$, the q-dimensional output $\mathbf{y} \in \mathbb{R}^q$, the r-dimensional disturbance $\mathbf{w} \in \mathbb{R}^r$, the measurement noise $\mathbf{v}$, and the matrices $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times p}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$ and $\mathbf{D} \in \mathbb{R}^{q\times p}$. It should be noted at this point that the Kalman filter can also be designed for linear time-varying and continuous-time systems. The following assumptions apply:

(1) For the disturbance $\mathbf{w}$ and the measurement noise $\mathbf{v}$, it is assumed that

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\Big(\mathbf{v}_k\mathbf{v}_j^{\mathrm{T}}\Big) = \mathbf{R}\delta_{kj} \tag{6.36a}$$

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\Big(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}\Big) = \mathbf{Q}\delta_{kj} \tag{6.36b}$$

$$\mathrm{E}\Big(\mathbf{w}_k\mathbf{v}_j^{\mathrm{T}}\Big) = \mathbf{0} \tag{6.36c}$$

with $\mathbf{Q} \geq 0$ and $\mathbf{R} > 0$ and the Kronecker symbol δ_{kj}.

(2) The expected value of the initial value and the covariance matrix of the initial error are given by

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\Big([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\Big) = \mathbf{P}_0 \geq 0 \tag{6.37}$$

with the estimate $\hat{\mathbf{x}}_0$ of the initial value $\mathbf{x}_0$.

(3) The disturbance $\mathbf{w}_k$, $k \geq 0$, and the measurement noise $\mathbf{v}_l$, $l \geq 0$, are not correlated

with the initial value $\mathbf{x}_0$, i. e., it holds

$$\mathrm{E}\left(\mathbf{w}_k \mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0} \tag{6.38a}$$

$$\mathrm{E}\left(\mathbf{v}_l \mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0} \ . \tag{6.38b}$$

However, this implies due to

$$\mathbf{x}_j = \mathbf{\Phi}^j \mathbf{x}_0 + \sum_{l=0}^{j-1} \mathbf{\Phi}^l (\mathbf{\Gamma} \mathbf{u}_{j-1-l} + \mathbf{G} \mathbf{w}_{j-1-l}) \tag{6.39}$$

and (6.36) also the relationship

$$\mathrm{E}\left(\mathbf{w}_k \mathbf{x}_j^{\mathrm{T}}\right) = \mathbf{0} \quad \forall k \geq j \tag{6.40a}$$

$$\mathrm{E}\left(\mathbf{v}_l \mathbf{x}_j^{\mathrm{T}}\right) = \mathbf{0} \quad \forall l, j \ . \tag{6.40b}$$

For further considerations, the following notation is introduced:

Definition 6.1. The optimal estimate of $\mathbf{x}_k$ considering $0, \dots, j$ measurements is abbreviated as $\hat{\mathbf{x}}(k|j)$.

Theorem 6.4 (Kalman Filter)**.** *The optimal estimate $\hat{\mathbf{x}}(k+1|k)$ of the state $\mathbf{x}_{k+1}$ of system (6.35) considering $l = 0, \dots, k$ measurements is calculated according to the iteration rule*

$$\begin{aligned}\hat{\mathbf{x}}(k+1|k) &= \mathbf{\Phi}\hat{\mathbf{x}}(k|k-1) + \mathbf{\Gamma}\mathbf{u}_k + \\ &\mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\right)^{-1}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k)\end{aligned} \tag{6.41}$$

with the covariance matrix of the estimation error

$$\begin{aligned}\mathbf{P}(k+1|k) &= \mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} \\ &- \mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\left(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\right)^{-1}\mathbf{C}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}}\end{aligned} \tag{6.42}$$

and the initial values $\hat{\mathbf{x}}(0|-1) = \mathbf{m}_0$ and $\mathbf{P}(0|-1) = \mathbf{P}_0$.

Proof of Theorem 6.4. Assume that the measurements $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{k-1}$ were used for the optimal estimate $\hat{\mathbf{x}}(k|k-1)$ with the error covariance matrix

$$\mathbf{P}(k|k-1) = \mathrm{E}\left([\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)]^{\mathrm{T}}\right) \tag{6.43}$$

At time k, the measurement

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{6.44}$$

is now used to improve the estimate of $\mathbf{x}_k$. According to Theorem 6.3, the estimate

$\hat{\mathbf{x}}(k|k)$ of $\mathbf{x}_k$ is

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathrm{E}\Big(\mathbf{x}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_k \tag{6.45a}$$

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k \tag{6.45b}$$

or with (6.16) in

$$\mathrm{E}\Big(\mathbf{x}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{x}_k(\mathbf{C}\mathbf{x}_k + \mathbf{v}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1))^{\mathrm{T}}\Big) = \mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} \tag{6.46}$$

and

$$\mathrm{E}\Big(\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big) = \mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R} \tag{6.47}$$

follows

$$\begin{aligned}&\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1)+\\&\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k)\ .\end{aligned} \tag{6.48}$$

Thus, the error covariance matrix can be written in the form (compare to (6.33))

$$\begin{aligned}&\mathbf{P}(k|k) = \mathrm{E}\Big([\mathbf{x}_k - \hat{\mathbf{x}}(k|k)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k)]^{\mathrm{T}}\Big) =\\&\mathbf{P}(k|k-1) - \mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}(k|k-1)\end{aligned}\ . \tag{6.49}$$

According to Theorem 6.2, the optimal estimate of $\mathbf{\Phi}\mathbf{x}_k$ equals the optimal estimate $\hat{\mathbf{x}}_k$ of $\mathbf{x}_k$ multiplied by $\mathbf{\Phi}$, thus

$$\hat{\mathbf{x}}(k+1|k) = \mathbf{\Phi}\hat{\mathbf{x}}(k|k) + \mathbf{\Gamma}\mathbf{u}_k\ . \tag{6.50}$$

For the covariance matrix of the estimation error, one obtains

$$\begin{aligned}\mathbf{P}(k+1|k) &= \mathrm{E}\Big([\mathbf{x}_{k+1} - \hat{\mathbf{x}}(k+1|k)][\mathbf{x}_{k+1} - \hat{\mathbf{x}}(k+1|k)]^{\mathrm{T}}\Big)\\&= \mathrm{E}\Big([\mathbf{\Phi}(\mathbf{x}_k - \hat{\mathbf{x}}(k|k)) + \mathbf{G}\mathbf{w}_k][\mathbf{\Phi}(\mathbf{x}_k - \hat{\mathbf{x}}(k|k)) + \mathbf{G}\mathbf{w}_k]^{\mathrm{T}}\Big)\ .\\&= \mathbf{\Phi}\mathbf{P}(k|k)\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}\end{aligned} \tag{6.51}$$

By combining (6.48)–(6.51), the result of Theorem 6.4 follows immediately. □

The composition of the covariance matrix of the estimation error (6.42) should be interpreted at this point: The term $\mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}}$ describes the change of the covariance matrix due to the system dynamics, $\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}$ indicates the increase in error variance due to the disturbance $\mathbf{w}$ and the remaining expression with negative sign describes how the error variance decreases through the inclusion of information from new measurements.

6.2.2 The Kalman Filter as Optimal Observer

If we introduce the abbreviations $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}(k+1|k)$, $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}(k|k-1)$, $\mathbf{P}_{k+1} = \mathbf{P}(k+1|k)$ and $\mathbf{P}_k = \mathbf{P}(k|k-1)$, then (6.41) and (6.42) can also be represented in the compact form

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}_k(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k) \tag{6.52}$$

with

$$\hat{\mathbf{K}}_k = \mathbf{\Phi}\mathbf{P}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1} \tag{6.53}$$

and

$$\mathbf{P}_{k+1} = \mathbf{\Phi}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} - \mathbf{\Phi}\mathbf{P}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} \ . \tag{6.54}$$

Equation (6.54) is also called the *discrete Riccati equation.* Comparing (6.52) with a complete Luenberger observer, one recognizes that the Kalman filter is a complete observer with a *time-varying observer gain matrix* $\hat{\mathbf{K}}_k$. The expected value of the observation error $\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k$ satisfies the iteration rule

$$\begin{aligned} \mathrm{E}(\mathbf{e}_{k+1}) &= \mathrm{E}\Big(\mathbf{\Phi}\mathbf{x}_k + \mathbf{G}\mathbf{w}_k - \mathbf{\Phi}\hat{\mathbf{x}}_k - \hat{\mathbf{K}}_k(\mathbf{C}\mathbf{x}_k + \mathbf{v}_k - \mathbf{C}\hat{\mathbf{x}}_k)\Big) \\ &= \mathrm{E}\Big(\Big(\mathbf{\Phi} - \hat{\mathbf{K}}_k\mathbf{C}\Big)(\mathbf{x}_k - \hat{\mathbf{x}}_k) + \mathbf{G}\mathbf{w}_k - \hat{\mathbf{K}}_k\mathbf{v}_k\Big) \\ &= \Big(\mathbf{\Phi} - \hat{\mathbf{K}}_k\mathbf{C}\Big)\,\mathrm{E}(\mathbf{e}_k) \ . \end{aligned} \tag{6.55}$$

Therefore, if $\hat{\mathbf{x}}_0 = \mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0$ is set, then $\mathrm{E}(\mathbf{e}_k) = \mathbf{0}$ for all $k \geq 0$. Furthermore, one can see from (6.53) and (6.54) that starting from the initial value $\mathbf{P}_0$, the error covariance matrix $\mathbf{P}_k$ and thus also $\hat{\mathbf{K}}_k$ can be pre-calculated and stored in the computer for all $k \geq 0$ without knowledge of the measurements $\mathbf{y}_k$. When no previous measurements of the process are available, one typically sets $\hat{\mathbf{x}}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \alpha\mathbf{E}$ for $\alpha \gg 1$. When the observer runs for a very long time, the problem can be treated mathematically as if it were running for infinite time. It turns out that for infinite time, the covariance matrix of the estimation error converges to a stationary value $\mathbf{P}_\infty$. In this case, the observer gain matrix $\hat{\mathbf{K}}_\infty$ is also constant and is calculated as

$$\hat{\mathbf{K}}_\infty = \mathbf{\Phi}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1} \tag{6.56}$$

with $\mathbf{P}_\infty$ as the solution of the so-called *discrete algebraic Riccati equation*

$$\mathbf{P}_\infty = \mathbf{\Phi}\mathbf{P}_\infty\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} - \mathbf{\Phi}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}_\infty\mathbf{\Phi}^{\mathrm{T}} \ . \tag{6.57}$$

Equation (6.57) has a unique symmetric solution $\mathbf{P}_\infty$ with the property that all eigenvalues of $\Big(\mathbf{\Phi} - \hat{\mathbf{K}}_\infty\mathbf{C}\Big)$ lie in the open interior of the unit circle, if the following conditions are satisfied:

(1) The pair $(\mathbf{C}, \mathbf{\Phi})$ is *detectable*, i. e. all eigenvalues outside the unit circle are observable.

(2) The pair $\left(\mathbf{\Phi}, \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}\right)$ is *stabilizable*, i. e. all eigenvalues outside the unit circle are controllable via the input $\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}$.

(3) The matrix $\mathbf{R}$ is positive definite.

Such a solution of the discrete algebraic Riccati equation (6.57) is also called a *stabilizing solution*. Since for this stabilizing solution, all eigenvalues of $\left(\mathbf{\Phi} - \hat{\mathbf{K}}_{\infty}\mathbf{C}\right)$ lie in the open interior of the unit circle, according to (6.55), the expected value of the observation error decreases, and $\lim_{k\to\infty} \mathrm{E}(\mathbf{e}_k) = \mathbf{0}$ holds. The solution $\mathbf{P}_{\infty}$ of the discrete algebraic Riccati equation (6.57) can be simply obtained by iterating the discrete Riccati equation (6.54) starting from the initial value $\mathbf{P}_0$ until $\mathbf{P}_k$ changes only sufficiently little in the sense of a norm. Although the iteration rule generally converges very quickly to a stationary value, in practice, including in Matlab, the algebraic Riccati equation (6.57) is solved numerically more efficiently via an eigenvector decomposition, see the Matlab commands `care` or `dare`.

> *Exercise* 6.6. The motion of a satellite around an axis is modeled in the form
>
> $$I\frac{\mathrm{d}^2}{\mathrm{d}t^2}\varphi = M_c - M_d$$
>
> with the moment of inertia I, the control torque M_c as control variable, the torque M_d as disturbance and the angle φ. Determine for the sampling time $T_a = 1\mathrm{s}$ the corresponding discrete-time system of the form (6.35) for $I = 1$ and the output variable φ. Assume that the measurement of the angle φ is superimposed with measurement noise v and that the disturbance M_d corresponds to the w acting on the process. Let
>
> $$\begin{aligned} \mathrm{E}(w) &= 0 & \mathrm{E}\left(w^2\right) &= q \\ \mathrm{E}(v) &= 0 & \mathrm{E}\left(v^2\right) &= 0.1\ . \end{aligned}$$
>
> Present the elements of $\mathbf{P}_k$ of the Kalman filter according to the iteration rule (6.54) for the initial value $\mathbf{P}_0 = \mathbf{E}$ and $q = \{0.1, 0.01, 0.001\}$. Implement the Kalman filter in Matlab/Simulink.

> **Tip:** The relationship (6.56) with the constant observer gain matrix $\hat{\mathbf{K}}_{\infty}$ provides an *optimal complete observer* that is *usable for both single-input and multi-input systems*. In contrast to observer design using the pole placement method and Ackermann's formula, no poles of the error system need to be chosen for the Kalman filter, which can be difficult especially in the multi-input case. Instead, the behavior of the error system is influenced by specifying the covariance matrices $\mathbf{Q}$ of the disturbance $\mathbf{w}$ and $\mathbf{R}$ of the measurement noise $\mathbf{v}$.
>
> For the choice of the covariance matrix $\mathbf{R}$ of the measurement noise $\mathbf{v}$, very often an interpretable approach based on the (noise) characteristics of the sensor can be found. Furthermore, the weighting of the entries of the covariance matrix $\mathbf{R}$ allows

discrimination between reliable and less reliable measurements. If a *measurement is less reliable*, the *corresponding entry* in the main diagonal *of the covariance matrix is chosen very large.* This assumed large variance of the measurement causes the observer to weight this measurement less compared to other measurements in state estimation. In practical application of the Kalman filter, it is even common to switch the covariance matrix during operation when one or more sensors provide implausible measurements or when entering an operating range where it is known a priori that certain sensors no longer provide reliable information.

For the process disturbance $\mathbf{w}$ and thus the covariance matrix $\mathbf{Q}$, these assumptions generally do not apply. The assumption that $\mathbf{w}$ is white noise is usually not valid. One might now think of choosing the generally unknown matrix $\mathbf{Q}$ very small or even zero. However, the choice $\mathbf{Q} = \mathbf{0}$ corresponds to the scenario of a system without disturbance $\mathbf{w}$. As can be seen from condition (2) for the solution of the discrete algebraic Riccati equation, the choice $\mathbf{Q} = \mathbf{0}$ (or generally also for $\mathbf{Q} \ll 1$) does not lead to a stabilizing solution.

The choice of $\mathbf{Q}$ (and also $\mathbf{R}$) in practical application is often done according to the "trial and error" method.

Beyond the linear, time-invariant systems considered so far, one can also design the Kalman Filter for nonlinear systems by means of (local) linearization. This approach will be presented in the following.

6.2.3 The Extended Kalman Filter

Before the Extended Kalman Filter is discussed as an observer for nonlinear systems, the Kalman filter from Section 6.2.2 shall be written for linear time-variant sampled systems of the form

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}_k \mathbf{x}_k + \boldsymbol{\Gamma}_k \mathbf{u}_k + \mathbf{G}_k \mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6.58a}$$

$$\mathbf{y}_k = \mathbf{C}_k \mathbf{x}_k + \mathbf{D}_k \mathbf{u}_k + \mathbf{v}_k \tag{6.58b}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional deterministic input $\mathbf{u} \in \mathbb{R}^p$, the q-dimensional output $\mathbf{y} \in \mathbb{R}^q$, the r-dimensional disturbance $\mathbf{w} \in \mathbb{R}^r$, the measurement noise $\mathbf{v}$ as well as the time-variant matrices $\boldsymbol{\Phi}_k \in \mathbb{R}^{n\times n}$, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{n\times p}$, $\mathbf{G}_k \in \mathbb{R}^{n\times r}$, $\mathbf{C}_k \in \mathbb{R}^{q\times n}$ and $\mathbf{D}_k \in \mathbb{R}^{q\times p}$. Analogous to Section 6.2.2, the following assumptions are made:

(1) For the disturbance $\mathbf{w}$ and the measurement noise $\mathbf{v}$, it is assumed that

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{v}_k \mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{R}_k \delta_{kj} \tag{6.59a}$$

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k \mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}_k \delta_{kj} \tag{6.59b}$$

$$\mathrm{E}\left(\mathbf{w}_k \mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{0} \tag{6.59c}$$

with $\mathbf{Q}_k \geq 0$ as well as $\mathbf{R}_k > 0$ and the Kronecker symbol $\delta_{kj} = 1$ for $k = j$ and $\delta_{kj} = 0$ otherwise.

(2) The expected value of the initial value and the covariance matrix of the initial error are given by

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\Big([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\Big) = \mathbf{P}_0 \geq 0 \tag{6.60}$$

with the estimate $\hat{\mathbf{x}}_0$ of the initial value $\mathbf{x}_0$.

(3) The disturbance $\mathbf{w}_k$, $k \geq 0$, and the measurement noise $\mathbf{v}_l$, $l \geq 0$, are not correlated with the initial value $\mathbf{x}_0$, i.e., it holds that

$$\mathrm{E}\Big(\mathbf{w}_k \mathbf{x}_0^{\mathrm{T}}\Big) = \mathbf{0} \tag{6.61a}$$

$$\mathrm{E}\Big(\mathbf{v}_l \mathbf{x}_0^{\mathrm{T}}\Big) = \mathbf{0}\ . \tag{6.61b}$$

The derivation of the Kalman filter for system (6.58) follows in a completely analogous manner as in Section 6.2.2 and reads for $k \geq 0$, compare to (6.52)–(6.54),

$$\hat{\mathbf{K}}_k = \mathbf{\Phi}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} \Big(\mathbf{C}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \tag{6.62a}$$

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}_k \hat{\mathbf{x}}_k + \mathbf{\Gamma}_k \mathbf{u}_k + \hat{\mathbf{K}}_k (\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_k - \mathbf{D}_k \mathbf{u}_k) \tag{6.62b}$$

$$\mathbf{P}_{k+1} = \mathbf{\Phi}_k \mathbf{P}_k \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^{\mathrm{T}} - \mathbf{\Phi}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} \Big(\mathbf{C}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \mathbf{C}_k \mathbf{P}_k \mathbf{\Phi}_k^{\mathrm{T}}\ . \tag{6.62c}$$

If the initial value $\mathbf{x}_0$ is known, then one sets $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ and $\mathbf{P}_0 = \mathbf{0}$, and for the case where no information about the initial value is available, typically $\hat{\mathbf{x}}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \alpha \mathbf{E}$ with $\alpha \gg 1$ is chosen.

In the literature, the Kalman filter is commonly descried in a somewhat different form. Here, the optimal estimation of the state $\mathbf{x}_k$ and the error covariance matrix $\mathbf{P}_k$ taking into account $0, \ldots, k-1$ measurements (compare Definition 6.1),

$$\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}(k|k-1) \tag{6.63a}$$

$$\mathbf{P}_k^- = \mathbf{P}(k|k-1) = \mathrm{E}\bigg(\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^-\Big]\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^-\Big]^{\mathrm{T}}\bigg) \tag{6.63b}$$

is called the *a priori estimate* and the optimal estimation of $\mathbf{x}_k$ and $\mathbf{P}_k$ taking into account $0, \ldots, k$ measurements

$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}(k|k) \tag{6.64a}$$

$$\mathbf{P}_k^+ = \mathbf{P}(k|k) = \mathrm{E}\bigg(\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^+\Big]\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^+\Big]^{\mathrm{T}}\bigg) \tag{6.64b}$$

is called the *a posteriori estimate*. (6.62) can thus be written in the equivalent form

Kalman gain matrix: $$\hat{\mathbf{L}}_k = \mathbf{P}_k^- \mathbf{C}_k^{\mathrm{T}} \left(\mathbf{C}_k \mathbf{P}_k^- \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k \right)^{-1} \tag{6.65a}$$

State estimate update: $$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k \left(\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_k^- - \mathbf{D}_k \mathbf{u}_k \right) \tag{6.65b}$$

Error covariance update: $$\mathbf{P}_k^+ = \left(\mathbf{E} - \hat{\mathbf{L}}_k \mathbf{C}_k \right) \mathbf{P}_k^- \tag{6.65c}$$

State extrapolation (6.50): $$\hat{\mathbf{x}}_{k+1}^- = \mathbf{\Phi}_k \hat{\mathbf{x}}_k^+ + \mathbf{\Gamma}_k \mathbf{u}_k \tag{6.65d}$$

Error covariance extrapolation (6.51): $$\mathbf{P}_{k+1}^- = \mathbf{\Phi}_k \mathbf{P}_k^+ \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^{\mathrm{T}} \tag{6.65e}$$

for $k \geq 0$ and the initial values $\hat{\mathbf{x}}_0^- = \hat{\mathbf{x}}_0$ and $\mathbf{P}_0^- = \mathbf{P}_0$.

> *Exercise* 6.7. Show the equivalence of relationships (6.62) and (6.65). To do this, perform the following substitutions in (6.65): $\hat{\mathbf{x}}_{k+1}^- = \hat{\mathbf{x}}_{k+1}$, $\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}_k$, $\mathbf{P}_{k+1}^- = \mathbf{P}_{k+1}$, $\mathbf{P}_k^- = \mathbf{P}_k$ and $\mathbf{\Phi}_k \hat{\mathbf{L}}_k = \hat{\mathbf{K}}_k$.

The Extended Kalman Filter (EKF) design is generally based on a nonlinear, time-variant, sampled-data multi-variable system of the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k) \qquad \mathbf{x}(0) = \mathbf{x}_0 && (6.66a)\\ \mathbf{y}_k &= \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k) \ . && (6.66b) \end{aligned}$$

The *idea of the Extended Kalman Filter* is based on performing a Taylor series expansion for the right-hand side of (6.66a) around the point $\mathbf{x}_k = \hat{\mathbf{x}}_k^+$, $\mathbf{u}_k = \mathbf{u}_k$ and $\mathbf{w}_k = \mathbf{0}$ and truncating after the linear term, i.e.,

$$\mathbf{x}_{k+1} \approx \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right) + \frac{\partial}{\partial \mathbf{x}_k} \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right)\left(\mathbf{x}_k - \hat{\mathbf{x}}_k^+\right) + \frac{\partial}{\partial \mathbf{w}_k} \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right)\mathbf{w}_k \ . \tag{6.67}$$

Analogously, the right side of the output equation (6.66b) is expanded in a Taylor series around the point $\mathbf{x}_k = \hat{\mathbf{x}}_k^-$, $\mathbf{u}_k = \mathbf{u}_k$ and $\mathbf{v}_k = \mathbf{0}$ and truncated after the linear term, i.e.,

$$\mathbf{y}_k \approx \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right) + \frac{\partial}{\partial \mathbf{x}_k} \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right)\left(\mathbf{x}_k - \hat{\mathbf{x}}_k^-\right) + \frac{\partial}{\partial \mathbf{v}_k} \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right)\mathbf{v}_k \ . \tag{6.68}$$

Note that throughout this, the following simplified notation

$$\frac{\partial}{\partial \mathbf{x}_k} \mathbf{F}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right) = \frac{\partial}{\partial \mathbf{x}_k} \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k)\Big|_{\mathbf{x}_k = \hat{\mathbf{x}}_k^-, \mathbf{u}_k = \mathbf{u}_k, \mathbf{w}_k = \mathbf{0}} \tag{6.69}$$

was used. The relationships (6.67) and (6.68) can be written more compactly for further considerations in the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{\Phi}_k \mathbf{x}_k + \bar{\mathbf{u}}_k + \mathbf{G}_k \mathbf{w}_k && (6.70a)\\ \mathbf{y}_k &= \mathbf{C}_k \mathbf{x}_k + \breve{\mathbf{u}}_k + \breve{\mathbf{v}}_k && (6.70b) \end{aligned}$$

with

$$
\begin{aligned}
\boldsymbol{\Phi}_k &= \frac{\partial}{\partial \mathbf{x}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) & \bar{\mathbf{u}}_k &= \mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) - \boldsymbol{\Phi}_k\hat{\mathbf{x}}_k^+ \\
\mathbf{G}_k &= \frac{\partial}{\partial \mathbf{w}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) & \mathbf{C}_k &= \frac{\partial}{\partial \mathbf{x}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) \\
\breve{\mathbf{u}}_k &= \mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) - \mathbf{C}_k\hat{\mathbf{x}}_k^- & \breve{\mathbf{v}}_k &= \frac{\partial}{\partial \mathbf{v}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big)\mathbf{v}_k
\end{aligned}
\tag{6.71}
$$

It is now obvious that the structure of system (6.70) directly enables the application of the Kalman filter according to (6.65). The computational steps required for implementation are summarized again in the following.

(1) In the case of a nonlinear, time-variant, continuous-time multi-variable system, first calculate a sampled-data system of the form (6.66).

(2) The estimated state and the covariance matrix of the estimation error must be initialized for the initial time point with $\hat{\mathbf{x}}_0^-$ and $\mathbf{P}_0^-$.

(3) For the disturbance $\mathbf{w}_k$ and the measurement noise $\breve{\mathbf{v}}_k$ in (6.70), it is again assumed that

$$
\mathrm{E}(\breve{\mathbf{v}}_k) = \mathbf{0} \qquad \mathrm{E}\Big(\breve{\mathbf{v}}_k\breve{\mathbf{v}}_j^{\mathrm{T}}\Big) = \mathbf{R}_k\delta_{kj} \tag{6.72a}
$$

$$
\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\Big(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}\Big) = \mathbf{Q}_k\delta_{kj} \tag{6.72b}
$$

$$
\mathrm{E}\Big(\mathbf{w}_k\breve{\mathbf{v}}_j^{\mathrm{T}}\Big) = \mathbf{0} \tag{6.72c}
$$

with $\mathbf{Q}_k \geq 0$ as well as $\mathbf{R}_k > 0$ and the Kronecker symbol $\delta_{kj} = 1$ for $k = j$ and $\delta_{kj} = 0$ otherwise.

(4) The iteration equations of the Extended Kalman Filter are then for $k \geq 0$

$$
\mathbf{C}_k = \frac{\partial}{\partial \mathbf{x}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) \tag{6.73a}
$$

$$
\hat{\mathbf{L}}_k = \mathbf{P}_k^-\mathbf{C}_k^{\mathrm{T}}\Big(\mathbf{C}_k\mathbf{P}_k^-\mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \tag{6.73b}
$$

$$
\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k\Big(\mathbf{y}_k - \mathbf{C}_k\hat{\mathbf{x}}_k^- - \breve{\mathbf{u}}_k\Big) = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k\Big(\mathbf{y}_k - \mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big)\Big) \tag{6.73c}
$$

$$
\mathbf{P}_k^+ = \Big(\mathbf{E} - \hat{\mathbf{L}}_k\mathbf{C}_k\Big)\mathbf{P}_k^- \tag{6.73d}
$$

$$
\boldsymbol{\Phi}_k = \frac{\partial}{\partial \mathbf{x}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{6.73e}
$$

$$
\mathbf{G}_k = \frac{\partial}{\partial \mathbf{w}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{6.73f}
$$

$$
\hat{\mathbf{x}}_{k+1}^- = \boldsymbol{\Phi}_k\hat{\mathbf{x}}_k^+ + \underbrace{\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) - \boldsymbol{\Phi}_k\hat{\mathbf{x}}_k^+}_{\bar{\mathbf{u}}_k} = \mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{6.73g}
$$

$$
\mathbf{P}_{k+1}^- = \boldsymbol{\Phi}_k\mathbf{P}_k^+\boldsymbol{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k\mathbf{Q}_k\mathbf{G}_k^{\mathrm{T}}\ . \tag{6.73h}
$$

Tip: In Extended Kalman Filter design, it is assumed that the linearized transformation of mean and covariance corresponds with good accuracy to the mean and covariance of the nonlinear transformation. This assumption generally does not hold, which can be accounted for with the so-called *Unscented Kalman Filter* which is often used to improve observer design for nonlinear systems. We will not consider the Unscented Kalman Filter in this lecture.

Exercise 6.8. The mathematical model

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_1 &= x_2 + w_1\\ \frac{\mathrm{d}}{\mathrm{d}t}x_2 &= \frac{1}{2}\rho_0 \exp\left(-\frac{x_1}{k}\right) C_w \frac{A}{m} x_2^2 - g + w_2\end{aligned}$$

describes the free fall of a body with mass m and cross-sectional area A in the Earth's atmosphere with altitude x_1 and velocity x_2. The term $\rho_0 \exp(-x_1/k)$ corresponds to the altitude-dependent density in the atmosphere (ρ_0 density at sea level), whereby the term with x_2^2 describes the deceleration due to air resistance with the drag coefficient C_w. Furthermore, g represents the gravitational acceleration.

The process noise is given by the stochastic quantities w_1 and w_2. The altitude x_1 can be measured via the output equation

$$y = x_1 + v$$

with the measurement noise v.

Design an Extended Kalman Filter for the parameters $\rho_0 = 1.2\,\mathrm{kg/m^3}$, $g = 9.81\,\mathrm{m/s^2}$, $k = 9100\,\mathrm{m}$, $A = 0.5\,\mathrm{m^2}$, $m = 100\,\mathrm{kg}$, and $C_w = 0.5$ that estimates not only the height x_1 and velocity x_2 but also the constant drag coefficient C_w. To do this, extend the differential equation system by the state $x_3 = C_w$ with

$$\frac{\mathrm{d}}{\mathrm{d}t}x_3 = 0 + w_3$$

and the process disturbance component w_3. Use the Euler method to determine the sampled-data system. Assume that the nominal values or initial conditions of the quantities C_w, x_1 and x_2 are normally distributed. The corresponding values of the means and variances can be taken from the following Table 6.1. For the simulation, use the following values: $C_w = 0.6$, $x_1(0) = 39\,500\,\mathrm{m}$ and $x_2(0) = -10\,\mathrm{m/s}$.

Variable	Mean	Variance
C_w	0.5	1
$x_1(0)$	$39 \cdot 10^3\,\mathrm{m}$	$1 \cdot 10^4\,\mathrm{m^2}$
$x_2(0)$	$0\,\mathrm{m/s}$	$1\,\mathrm{m^2/s^2}$

Table 6.1: Means and variances of the parameters or initial conditions.

Exercise 6.9. For position determination of a vehicle in a two-dimensional space (o-axis: East coordinate, n-axis: North coordinate), an Extended Kalman Filter shall be used. Several measurement stations with coordinates (O_i, N_i), $i = 1, \dots, M$ measure the distance to the vehicle. The acceleration of the vehicle in North and East directions is modeled by white noise. The difference equation system

$$\underbrace{\begin{bmatrix} o_{k+1} \\ n_{k+1} \\ o_{v,k+1} \\ n_{v,k+1} \end{bmatrix}}_{\mathbf{x}_{k+1}} = \underbrace{\begin{bmatrix} 1 & 0 & T_a & 0 \\ 0 & 1 & 0 & T_a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{\Phi}} \underbrace{\begin{bmatrix} o_k \\ n_k \\ o_{v,k} \\ n_{v,k} \end{bmatrix}}_{\mathbf{x}_k} + \underbrace{\begin{bmatrix} w_{1,k} \\ w_{2,k} \\ w_{3,k} \\ w_{4,k} \end{bmatrix}}_{\mathbf{w}_k}$$

describes the vehicle behavior, where o_k and n_k or $o_{v,k}$ and $n_{v,k}$ denote the coordinates or velocities of the vehicle with respect to the origin of a fixed coordinate system in East and North directions at time kT_a with sampling time T_a, and $w_{j,k}$, $j = 1, \dots, 4$ are the components of the process noise. Furthermore, the distance measurements of the vehicle from the stations are given by

$$y_{i,k} = \sqrt{(n_k - N_i)^2 + (o_k - O_i)^2} + v_{i,k}, \quad i = 1, \dots, M$$

with the measurement noise $v_{i,k}$, $i = 1, \dots, M$. Assume that all stochastic quantities $(w_{j,k})$ and $(v_{i,k})$ are normally distributed, uncorrelated and zero-mean. Let the sampling time be given as $T_a = 0.1\,\mathrm{s}$. For the covariance matrix of the process noise, let

$$\mathrm{E}\left(\mathbf{w}_k \mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}\delta_{kj} \quad \text{with} \quad \mathbf{Q} = \mathrm{diag}(0, 0, 4, 4)$$

and let the covariance of the measurement noise be

$$\mathrm{E}(v_{i,k} v_{i,j}) = R_i \delta_{kj} \quad \text{with} \quad R_i = 1, \quad i = 1, \dots, M \ .$$

The initial state $\mathbf{x}_0^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & 50 & 50 \end{bmatrix}$ is exactly known. Simulate the system for 60 seconds and design an Extended Kalman Filter for state estimation. Vary the number and position of the measurement stations.

6.3 Optimal State Controllers

The goal of this section is the development of an optimal state controller for linear, time-invariant systems and the combination of this state controller with the optimal state observer from the previous section. The starting point of the considerations is the linear, time-invariant, discrete-time system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6.74a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \tag{6.74b}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional deterministic input $\mathbf{u} \in \mathbb{R}^p$, the q-dimensional output $\mathbf{y} \in \mathbb{R}^q$ as well as the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$ and $\mathbf{D} \in \mathbb{R}^{q\times p}$. We now seek a control sequence $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{N-1}$ that minimizes the cost functional

$$\begin{aligned} J(\mathbf{x}_0) &= \sum_{k=0}^{N-1}\Big(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k\Big) + \mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N \\ &= \sum_{k=0}^{N-1}\begin{bmatrix}\mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}}\end{bmatrix}\underbrace{\begin{bmatrix}\mathbf{Q} & \mathbf{N}^{\mathrm{T}}\\ \mathbf{N} & \mathbf{R}\end{bmatrix}}_{\mathbf{J}}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{u}_k\end{bmatrix} + \mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N \end{aligned} \tag{6.75}$$

for suitable *weighting matrices* $\mathbf{Q} \in \mathbb{R}^{n\times n}$, $\mathbf{R} \in \mathbb{R}^{p\times p}$, $\mathbf{N} \in \mathbb{R}^{p\times n}$ and $\mathbf{S} \in \mathbb{R}^{n\times n}$. Due to the quadratic cost criterion (6.75), this controller design can also be found in the literature under the name *LQR (Linear Quadratic Regulator) problem.* To solve this task, the method of *dynamic programming according to Bellman* is employed.

6.3.1 Dynamic Programming according to Bellman

The foundation of dynamic programming is formed by the *optimality principle*:

Theorem 6.5 (Optimality Principle). *An optimal solution has the property that starting from any point of this solution, the remaining solution is optimal in the sense of the problem to be solved with the chosen point as initial condition.*

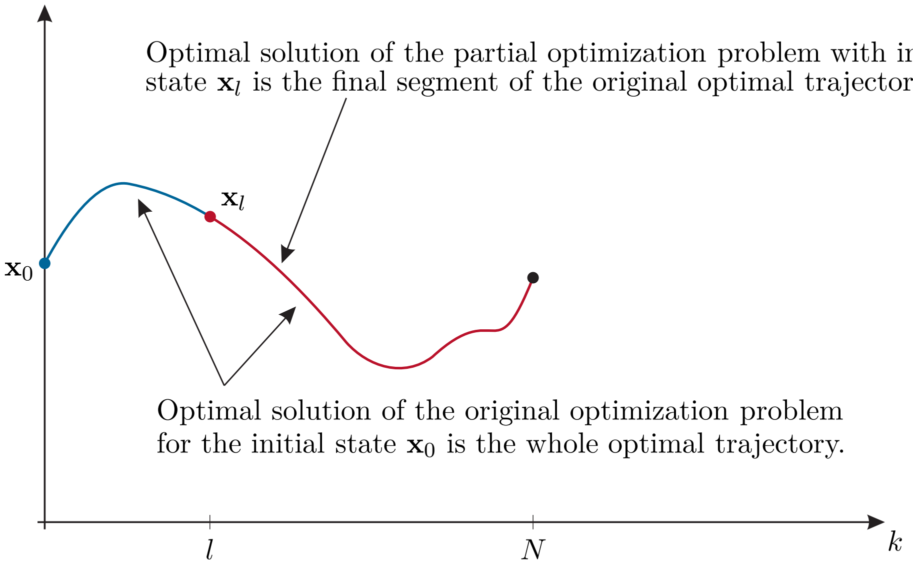

Optimal solution of the partial optimization problem with initial state $\mathbf{x}_l$ is the final segment of the original optimal trajectory.
$\mathbf{x}_l$
$\mathbf{x}_0$
Optimal solution of the original optimization problem for the initial state $\mathbf{x}_0$ is the whole optimal trajectory.
l
N
k

Figure 6.2: On the optimality principle.

Figure 6.2 illustrates Theorem 6.5. This idea is now used in the sense of *dynamic programming according to Bellman* such that the optimization problem (6.75) is solved backwards starting from the final time point N. Here, the value of the optimal control for time point N, i.e., $\mathbf{u}_{N-1}$, can be solved independently of the reached state $\mathbf{x}_{N-1}$. In the next step, starting from the optimal solution $\mathbf{u}_{N-1}$, the optimal $\mathbf{u}_{N-2}$ is calculated. If this procedure is repeated until $k = 0$, then the optimal control strategy is found.

Since the linearity of the system is not necessary for dynamic programming, we first examine the optimization problem

$$\min_{(\mathbf{u}_0,\dots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) \quad \text{with} \quad J(\mathbf{x}_0) = \sum_{k=0}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k) + s(\mathbf{x}_N) \tag{6.76}$$

subject to the constraint

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \ . \tag{6.77}$$

As already mentioned, the optimization problem (6.76) is solved backwards starting from the final time point $k = N$. Since $J(\mathbf{x}_N)$ is independent of the control input $\mathbf{u}$, it follows trivially that

$$J^*(\mathbf{x}_N) = s(\mathbf{x}_N), \tag{6.78}$$

where $J^*(\mathbf{x}_N)$ describes the optimal value of $J(\mathbf{x}_N)$. According to the optimality principle, for the optimal control sequence $\mathbf{u}_0^*, \mathbf{u}_1^*, \dots, \mathbf{u}_{N-1}^*$ with

$$J^*(\mathbf{x}_0) = \sum_{k=0}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + s(\mathbf{x}_N) \tag{6.79}$$

the relationship

$$J^*(\mathbf{x}_0) = \sum_{k=0}^{l} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + \underbrace{\sum_{k=l+1}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + s(\mathbf{x}_N)}_{J^*(\mathbf{x}_{l+1})} \tag{6.80}$$

holds with

$$J^*(\mathbf{x}_{l+1}) = \min_{(\mathbf{u}_{l+1},\dots,\mathbf{u}_{N-1})} J(\mathbf{x}_{l+1}) \tag{6.81}$$

as well as

$$J(\mathbf{x}_{l+1}) = \sum_{k=l+1}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k) + s(\mathbf{x}_N) \tag{6.82}$$

and subject to the constraint (6.77). Note that (6.81) with (6.82) is solved for the initial value $\mathbf{x}_{l+1}$. If one now wants to go back one step based on (6.81) and determine the optimal value of the value function $J^*(\mathbf{x}_l)$, then the substitute problem follows from the optimality principle, yielding

$$J^*(\mathbf{x}_l) = \min_{\mathbf{u}_l} \left(j_l(\mathbf{x}_l, \mathbf{u}_l) + J^*\Big(\underbrace{\mathbf{x}_{l+1}}_{\mathbf{f}(\mathbf{x}_l,\mathbf{u}_l)} \Big) \right) . \tag{6.83}$$

This equation is known as the Bellman equation and serves as the basis of many optimal control approaches.

Example 6.2. As a non-control-theoretic application, consider a simple allocation problem. Given an investment sum A that is to be divided among N projects. It is further assumed that when allocating a sum u_k to project k, the project yields a profit $g_k(u_k)$. The optimization problem to be solved therefore reads

$$\max_{(u_0,\dots,u_{N-1})} J(x_0) \quad \text{with} \quad J(x_0) = \sum_{k=0}^{N-1} g_k(u_k) \quad \text{subject to} \quad \sum_{k=0}^{N-1} u_k = A \ . \tag{6.84}$$

This problem can now be reformulated into an equivalent control problem of the form

$$\max_{(u_0,\dots,u_{N-1})} J(x_0) \quad \text{with} \quad J(x_0) = \sum_{k=0}^{N-1} g_k(u_k) \tag{6.85}$$

subject to the constraint

$$x_{k+1} = x_k - u_k \quad \text{with} \quad x_0 = A \quad \text{and} \quad x_N = 0 \tag{6.86}$$

If one chooses, for example, $g_k(u_k) = \sqrt{u_k}$, then using dynamic programming one obtains

$$\begin{aligned}
J^*(x_N) &= 0 \\
J^*(x_{N-1}) &= \max_{u_{N-1}}\{\sqrt{u_{N-1}}\} \overset{\text{subject to } x_{N-1}-u_{N-1}=0}{=} \sqrt{x_{N-1}} && u^*_{N-1} = x_{N-1} \\
J^*(x_{N-2}) &= \max_{u_{N-2}}\{\sqrt{u_{N-2}} + \sqrt{x_{N-2}-u_{N-2}}\} = \sqrt{2x_{N-2}} && u^*_{N-2} = x_{N-2}/2 \\
J^*(x_{N-3}) &= \max_{u_{N-3}}\left\{\sqrt{u_{N-3}} + \sqrt{2(x_{N-3}-u_{N-3})}\right\} = \sqrt{3x_{N-3}} && u^*_{N-3} = x_{N-3}/3 \\
&\vdots && \vdots \\
J^*(x_0) &= \sqrt{Nx_0} && u^*_0 = x_0/N \ .
\end{aligned} \tag{6.87}$$

Exercise 6.10. Interpret the result (6.87).

6.3.2 The LQR Problem

When applying the principle of dynamic programming to the task (6.75) with the constraint (6.74), one obtains for $k = N$

$$J^*(\mathbf{x}_N) = \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{6.88}$$

and for $k = N - 1$

$$J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \left\{ \left(\mathbf{x}_{N-1}^{\mathrm{T}} \mathbf{Q} \mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{R} \mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{N} \mathbf{x}_{N-1} \right) + J^*\Big(\underbrace{\mathbf{x}_N}_{\mathbf{\Phi}\mathbf{x}_{N-1}+\mathbf{\Gamma}\mathbf{u}_{N-1}} \Big) \right\} \tag{6.89}$$

or

$$\begin{aligned} J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \Big\{ &\left(\mathbf{x}_{N-1}^{\mathrm{T}} \mathbf{Q} \mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{R} \mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{N} \mathbf{x}_{N-1} \right) + \\ &(\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1})^{\mathrm{T}} \mathbf{S} (\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1}) \Big\} \ . \end{aligned} \tag{6.90}$$

Minimizing (6.90) with respect to $\mathbf{u}_{N-1}$ yields the optimal solution $\mathbf{u}_{N-1}^*$ of $\mathbf{u}_{N-1}$ as

$$\mathbf{u}_{N-1}^* = -\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} \ . \tag{6.91}$$

By substituting (6.91) into (6.90) it follows that

$$\begin{aligned} &J^*(\mathbf{x}_{N-1}) \\ &= \mathbf{x}_{N-1}^{\mathrm{T}} \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} + (\mathbf{u}_{N-1}^*)^{\mathrm{T}} \left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right) \mathbf{u}_{N-1}^* + 2(\mathbf{u}_{N-1}^*)^{\mathrm{T}} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} \\ &= \mathbf{x}_{N-1}^{\mathrm{T}} \left\{ \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) - \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right)^{\mathrm{T}} \left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \right\} \mathbf{x}_{N-1} \end{aligned} \tag{6.92}$$

Thus, the following theorem can be stated immediately:

Theorem 6.6 (Linear Quadratic Regulator). *The unique solution of the optimization problem* (6.75) *for the linear, time-invariant, discrete-time system* (6.74) *with the symmetric positive semi-definite matrix* $\mathbf{S} = \mathbf{P}_N$, *the symmetric positive semi-definite matrix*

$$\mathbf{J} = \begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix} \tag{6.93}$$

and the positive definite matrix $\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right)$ *is given by the control law*

$$\mathbf{u}_k^* = \mathbf{K}_k \mathbf{x}_k \tag{6.94}$$

with

$$\mathbf{K}_k = -\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Phi} \right) \tag{6.95}$$

and

$$\mathbf{P}_k = \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Phi} \right) - \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Phi} \right)^{\mathrm{T}} \left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_{k+1} \mathbf{\Phi} \right) \tag{6.96}$$

The minimum value of the cost functional (6.75) *is calculated as*

$$\min_{(\mathbf{u}_0,\dots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) = J^*(\mathbf{x}_0) = \mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0 \tag{6.97}$$

and it holds that $\mathbf{P}_k \geq 0$ *for all* $k = 0, 1, \dots, N$.

Proof of Theorem 6.6. The control law (6.94), (6.95) as well as the iteration rule (6.96) and also the relation (6.97) are obtained directly by repeated application of the iteration rule of dynamic programming from equations (6.91) and (6.92). It remains to show that $\mathbf{P}_k$ is positive semi-definite for all $k = 0, 1, \dots, N$. For this, substitute $\mathbf{u}_{N-1}^* = \mathbf{K}_{N-1}\mathbf{x}_{N-1}$ into (6.92), which gives

$$\begin{aligned} J^*(\mathbf{x}_{N-1}) = \mathbf{x}_{N-1}^{\mathrm{T}}\Big(\big(\mathbf{Q} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}_N\boldsymbol{\Phi}\big) + \mathbf{K}_{N-1}^{\mathrm{T}}\big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_N\boldsymbol{\Gamma}\big)\mathbf{K}_{N-1} \\ + 2\mathbf{K}_{N-1}^{\mathrm{T}}\big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_N\boldsymbol{\Phi}\big)\Big)\mathbf{x}_{N-1} = \mathbf{x}_{N-1}^{\mathrm{T}}\mathbf{P}_{N-1}\mathbf{x}_{N-1} \end{aligned} \tag{6.98}$$

and thus for $\mathbf{P}_k$ from (6.96)

$$\mathbf{P}_k = (\boldsymbol{\Phi} + \boldsymbol{\Gamma}\mathbf{K}_k)^{\mathrm{T}}\mathbf{P}_{k+1}(\boldsymbol{\Phi} + \boldsymbol{\Gamma}\mathbf{K}_k) + \begin{bmatrix}\mathbf{E} & \mathbf{K}_k^{\mathrm{T}}\end{bmatrix}\underbrace{\begin{bmatrix}\mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R}\end{bmatrix}}_{\mathbf{J}}\begin{bmatrix}\mathbf{E} \\ \mathbf{K}_k\end{bmatrix}. \tag{6.99}$$

Since now the matrices $\mathbf{P}_N = \mathbf{S}$ and $\mathbf{J}$ are positive semi-definite, the positive semi-definiteness of $\mathbf{P}_k$ for all $k = 0, 1, \dots, N$ is directly shown. □

As already with the Kalman filter as an observer (see (6.54)), equation (6.96) is also a *discrete Riccati equation*, which is why the *time-varying state controller* (6.94), (6.95) is also called a *Riccati controller*. Note, however, that the discrete Riccati equation (6.96), in contrast to the Kalman filter, *runs backward*! For a real-time implementation of the controller (6.94), (6.95), the end time N must therefore be known and the matrices $\mathbf{P}_k$ and $\mathbf{K}_k$ must be pre-calculated.

When the end time $N \to \infty$, one can, as already with the Kalman filter (compare (6.56), (6.57)), calculate a stationary solution $\mathbf{P}_s$ and $\mathbf{K}_s$ from (6.94)–(6.96). One could now determine the stationary solution $\mathbf{P}_s$ of the discrete Riccati equation (6.96) by starting from the initial value $\mathbf{P}_\infty = \alpha\mathbf{E}$ for $\alpha \gg 1$ and iterating until $\mathbf{P}_k$ changes only sufficiently little in the sense of a norm. Another solution possibility consists of solving the associated *discrete algebraic Riccati equation* for $\mathbf{P}_{k+1} = \mathbf{P}_k = \mathbf{P}_s$ in (6.96)

$$\mathbf{P}_s = \big(\mathbf{Q} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Phi}\big) - \big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Phi}\big)^{\mathrm{T}}\big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Gamma}\big)^{-1}\big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Phi}\big) \tag{6.100}$$

Thus, however, the *stationary Riccati controller*

$$\mathbf{u}_k^* = \mathbf{K}_s\mathbf{x}_k \tag{6.101a}$$

$$\mathbf{K}_s = -\big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Gamma}\big)^{-1}\big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_s\boldsymbol{\Phi}\big) \tag{6.101b}$$

has the structure of a classical state controller similar to one obtained from pole placement.

The discrete algebraic Riccati equation (6.100) has a unique symmetric positive semi-definite solution $\mathbf{P}_s$ with the property that all eigenvalues of $(\mathbf{\Phi} + \mathbf{\Gamma K}_s)$ lie in the open interior of the unit circle (asymptotic stability of the closed-loop system) if the following conditions are satisfied:

(1) The pair $(\mathbf{\Phi}, \mathbf{\Gamma})$ is *stabilizable*, i.e., all eigenvalues outside the unit circle are reachable, and

(2) the pair $(\mathbf{C_J}, \mathbf{\Phi})$ with

$$0 \leq \mathbf{J} = \begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\mathbf{J}}^{\mathrm{T}} \\ \mathbf{D}_{\mathbf{J}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{C_J} & \mathbf{D_J} \end{bmatrix} \tag{6.102}$$

is *detectable*, i.e., all eigenvalues outside the unit circle are observable through the output $\mathbf{C_J}$.

If one now wants to achieve that all poles of the closed loop with the dynamics matrix $(\mathbf{\Phi} + \mathbf{\Gamma K}_s)$ lie not only inside the unit circle, but inside a circle with radius $r < 1$ for robustness considerations, then the controller design must be carried out for the substitute system

$$\mathbf{x}_{k+1} = \tilde{\mathbf{\Phi}} \mathbf{x}_k + \tilde{\mathbf{\Gamma}} \mathbf{u}_k \tag{6.103}$$

with

$$\tilde{\mathbf{\Phi}} = \frac{1}{r} \mathbf{\Phi} \quad \text{and} \quad \tilde{\mathbf{\Gamma}} = \frac{1}{r} \mathbf{\Gamma} \tag{6.104}$$

Since then the eigenvalues of the matrix $\left(\tilde{\mathbf{\Phi}} + \tilde{\mathbf{\Gamma}} \mathbf{K}_s\right)$ lie inside the unit circle, it follows from (6.103) that the eigenvalues of $(\mathbf{\Phi} + \mathbf{\Gamma K}_s)$ come to lie inside a circle with radius r.

If in the cost functional (6.75) only the p-dimensional input $\mathbf{u}$ and the q-dimensional output $\mathbf{y}$ should be weighted, i.e.,

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \left(\mathbf{y}_k^{\mathrm{T}} \mathbf{Q_y} \mathbf{y}_k + \mathbf{u}_k^{\mathrm{T}} \mathbf{R} \mathbf{u}_k + 2 \mathbf{u}_k^{\mathrm{T}} \mathbf{N_y} \mathbf{y}_k \right) + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{6.105a}$$

$$= \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{y}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q_y} & \mathbf{N}_{\mathbf{y}}^{\mathrm{T}} \\ \mathbf{N_y} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{y}_k \\ \mathbf{u}_k \end{bmatrix} + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N, \tag{6.105b}$$

then (6.105) can be transformed to the form of (6.75) via the relation $\mathbf{y}_k = \mathbf{C x}_k + \mathbf{D u}_k$ as well as

$$\begin{aligned} &\sum_{k=0}^{N-1} \Bigg[\mathbf{x}_k^{\mathrm{T}} \underbrace{\mathbf{C}^{\mathrm{T}} \mathbf{Q_y} \mathbf{C}}_{\tilde{\mathbf{Q}}} \mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}} \underbrace{\left(\mathbf{R} + \mathbf{D}^{\mathrm{T}} \mathbf{Q_y} \mathbf{D} + \mathbf{N_y} \mathbf{D} + \mathbf{D}^{\mathrm{T}} \mathbf{N}_{\mathbf{y}}^{\mathrm{T}} \right)}_{\tilde{\mathbf{R}}} \mathbf{u}_k \\ &\quad + 2 \mathbf{u}_k^{\mathrm{T}} \underbrace{\left(\mathbf{N_y} + \mathbf{D}^{\mathrm{T}} \mathbf{Q_y} \right) \mathbf{C}}_{\tilde{\mathbf{N}}} \mathbf{x}_k \Bigg] \\ &= \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{N}}^{\mathrm{T}} \\ \tilde{\mathbf{N}} & \tilde{\mathbf{R}} \end{bmatrix}}_{\tilde{\mathbf{J}}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \end{aligned} \tag{6.106}$$

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{N}}^{\mathrm{T}} \\ \tilde{\mathbf{N}} & \tilde{\mathbf{R}} \end{bmatrix}}_{\tilde{\mathbf{J}}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{6.107}$$

and solved using Theorem 6.6.

> *Exercise* 6.11. Compare the MATLAB commands `lqrd`, `dlqr` and `dlqry`. What do these commands accomplish? Look at the respective `help` text and then establish the connection to the theory shown so far.

6.3.3 The LQG Problem

In the context of optimal estimation and control, one often speaks about the so-called *Linear-Quadratic-Gaussian (LQG) Problem.* In essence, this comprises a combination of

1. **Linear** state dynamics

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{6.108a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \ , \tag{6.108b}$$

2. **Quadratic** cost functions, leading to the value functional

$$\begin{aligned} J(\mathbf{x}_0) &= \mathrm{E}\left(\sum_{k=0}^{N-1}\left(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k\right)\right) + \mathrm{E}\left(\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N\right) \\ &= \mathrm{E}\left(\sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}\right) + \mathrm{E}\left(\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N\right) \ , \end{aligned} \tag{6.109}$$

and

3. **Gaussian** random variables $\mathbf{x}_0$, $\mathbf{w}$, and $\mathbf{v}$, satisfying the assumptions

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\left([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\right) = \hat{\mathbf{P}}_0 \geq 0 \tag{6.110}$$

as well as

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}\right) = \hat{\mathbf{Q}}\delta_{kj} \tag{6.111a}$$

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{v}_k\mathbf{v}_j^{\mathrm{T}}\right) = \hat{\mathbf{R}}\delta_{kj} \tag{6.111b}$$

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{0} \tag{6.111c}$$

and

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0} \tag{6.112a}$$

$$\mathrm{E}\left(\mathbf{v}_l\mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0} \ , \tag{6.112b}$$

where only the output $\mathbf{y}_k$ of the system (6.108) can be measured. Without proof, we will claim here that the additional stochastic terms in (6.108) do not change the fact that the LQR from Theorem 6.6 minimizes the value functional (6.109). Furthermore, if only the output $\mathbf{y}_k$ can be measured rather than the state $\mathbf{x}_k$, then the control law

$$\mathbf{u}_k = \mathbf{K}_k \hat{\mathbf{x}}_k \ , \tag{6.113}$$

with $\hat{\mathbf{x}}_k$ provided by the Kalman Filter from Theorem 6.4, provides the optimal solution for the resulting control problem. Therefore, we can conclude that the LQG problem is solved by a combination of

- a Kalman Filter (Theorem 6.4) designed with the system matrices from (6.108) and the covariance matrices $\hat{\mathbf{Q}}$, $\hat{\mathbf{R}}$, and $\hat{\mathbf{P}}_0$, and
- a Linear-Quadratic Regulator (LQR, Theorem 6.6) designed with the system matrices from (6.108) and the cost matrices $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{N}$, and $\mathbf{S}$.

6.4 Literatur

[6.1] L. Ljung, *System Identification.* New Jersey, USA: Prentice Hall, 1999.

[6.2] D. Luenberger, *Optimization by Vector Space Methods.* New York, USA: John Wiley & Sons, 1969.

[6.3] O. Nelles, *Nonlinear System Identification.* Berlin, Deutschland: Springer, 2001.

[6.4] R. Isermann, *Identifikation dynamischer Systeme 1 und 2*, 2nd ed. Berlin, Deutschland: Springer, 1992.

[6.5] A. Gelb, *Applied Optimal Estimation.* Cambridge, USA: MIT Pre, 74.

[6.6] D. Simon, *Optimal State Estimation.* New Jersey, USA: John Wiley & Sons, 2006.

[6.7] A. S. Deshpande, "Bridging the gap in applied kalman filtering - estimating outputs when measurements are correlated with the process noise," *IEEE Control Systems Magazine*, pp. 87–93, 2017.

[6.8] G. Franklin, J. Powell, and M. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, USA: Addison–Weseley, 1998.

[6.9] K. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design.* New York, USA: Prentice Hall, 1997.

[6.10] A. Bryson and Y. Ho, *Applied Optimal Control.* Washington, USA: He, 1975.

[6.11] P. Dorato, C. Abdallah, and V. Cerone, *Linear Quadratic Control: An Introduction.* Florida, USA: Krieger Publishing Company, 2000.