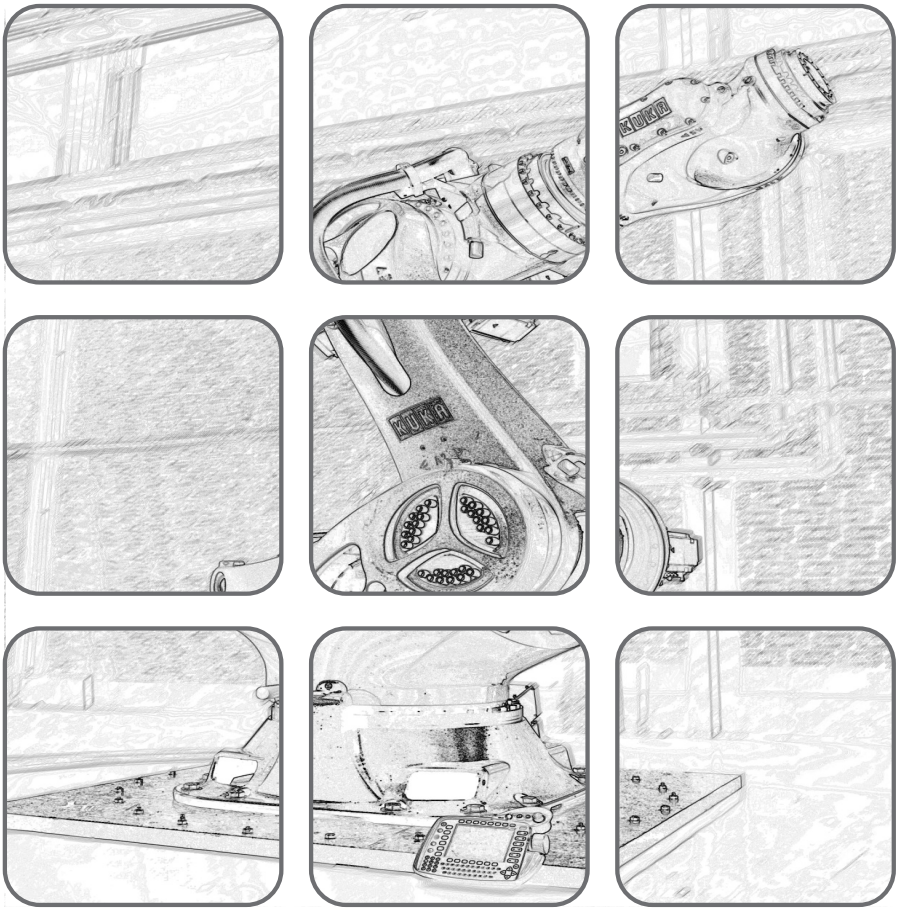
TECHNISCHE UNIVERSITÄT WIEN

AUTOMATION & CONTROL INSTITUTE
INSTITUT FÜR AUTOMATISIERUNGS- & REGELUNGSTECHNIK

Lecture
2024W

CONTROL SYSTEMS

Univ.-Prof. Dr.-Ing. Wolfgang Kemmetmüller
Univ.-Prof. Dr. techn. Andreas Kugi



Control Systems

Lecture
2024W

Univ.-Prof. Dr.-Ing. Wolfgang Kemmetmüller
Univ.-Prof. Dr. techn. Andreas Kugi

TU Wien
Automation and Control Institute
Complex Dynamical Systems Group

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

Contents

1 Identification Methods

This chapter deals with the process identification of linear dynamic systems. After a brief introduction to the concept of identification, some essential representatives of non-parametric and parametric identification methods will be discussed. It should be noted at this point that the terminology used is strongly based on the book by L. Ljung [1.1], as this book also forms the basis of the MATLAB Identification Toolbox, in which all the algorithms discussed in this chapter and many more are implemented.

1.1 General Aspects

Figure 1.1 shows the basic task of process identification. The process is affected by the

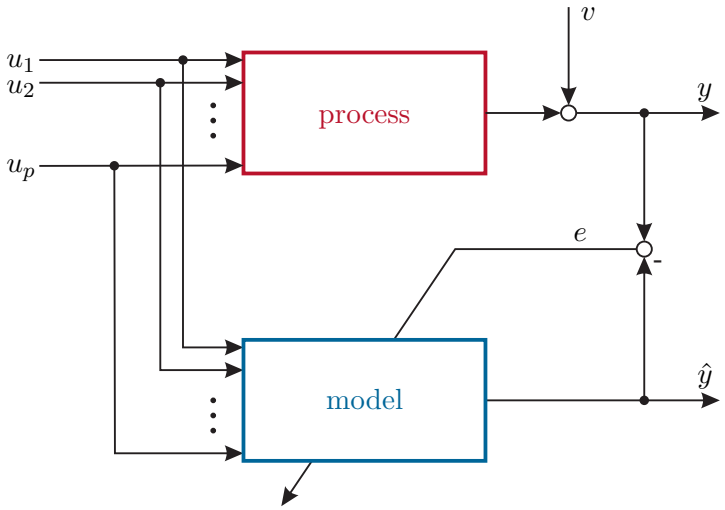


Figure 1.1: On the concept of process identification.

input variables $u_1, \ldots, u_p$, and it is assumed that only one output variable y is measured. This measured output variable y is disturbed by the noise signal v. The process model should now describe the process as accurately as possible, whereby as much a priori knowledge about the process as possible should be taken into account. The quality of the model is then evaluated, for example, using the error e between the disturbed measured output y and the model output $\hat{y}$. This error is then used to improve the model. The various steps to be carried out within the identification task are summarized in Figure 1.2 and must be iteratively repeated several times depending on the problem:

(A) **Choice of Input Variables:** In this step, it must be decided which input variables

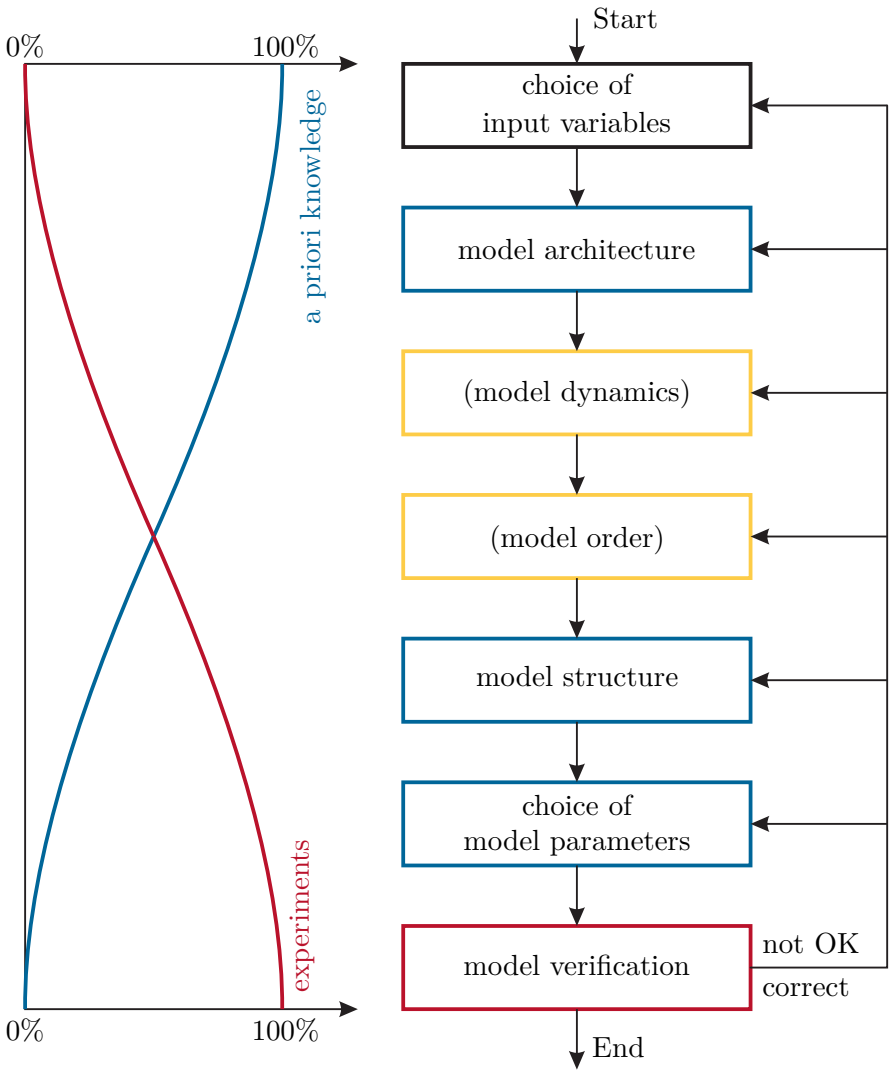


Figure 1.2: Identification procedure.

are used for process identification and which signals are suitable for the individual input variables.

(B) **Model Architecture:** Here it should be determined whether it is a static or dynamic model, for what purpose the model is needed (simulation, controller design, error detection, etc.), the resulting dynamic range, whether the identification should be on- or off-line, etc.

(C,D) **Model Dynamics** and **Model Order:** In these steps, a description of the system dynamics is carried out, for example, in the form of a transfer function or a state-space model, and a system order should be determined, which, however, if not known from a priori knowledge, can only be sensibly estimated within the framework of targeted experiments.

(E) **Model Structure:** In this step, the model structure underlying the identification task is determined. Reference is made here to Section 1.3.1 for the different models of parametric identification methods for linear dynamic systems (ARMA, ARX, ARMAX, etc.).

(F) **Choice of Model Parameters:** Often, the model parameters are already fixed by the previous steps, in particular the choice of model structure and system order. However, there is still the possibility to specifically select parameters adapted to the identification task.

(G) **Model Verification:** Depending on the purpose of the model (step (B)), it must be checked whether the model has the corresponding quality. It is particularly important to ensure in this step that not the same data set is used for model verification as for solving the identification task.

In the literature, a distinction is also made between *white-box*, *black-box*, and *grey-box* models:

- In white-box models, all equations and parameters can be derived on the basis of physical considerations. White-box models are also referred to as such when the model is completely derived from physical laws and some so-called constitutive parameters (friction parameters, leakage parameters, stray inductances, etc.) are determined from experiments. The advantages of these models lie in the very good extrapolatability of the model beyond the data obtained through experiments, high reliability, good insight into the model, and the scalability of the model, which makes it applicable even for systems that have not yet been realized (prototyping). As a disadvantage of white-box models, it can be stated that the creation is generally relatively time-consuming and requires precise knowledge of the system.

- Black-box models are based solely on experimental results and have no (or very little) a priori knowledge of the system. Of course, it should be borne in mind that the model thus obtained is only valid in the data set covered by the identification. The main advantage is that relatively little knowledge about the system is required. All advantages of white-box models can be listed here as disadvantages.

- Grey-box models combine model building based on physical considerations with principles of process identification.

1.2 Non-parametric Methods

Linear, time-invariant systems can be characterized by their transfer function (s-transfer function in continuous time and z-transfer function in discrete time) or their impulse response. The goal of *non-parametric identification methods* is now to determine the frequency response or impulse response directly from measurements of the input and output variables, without resorting to a specific model structure underlying the identification task. Since these methods do not rely on a finite number of parameters to be identified, they are called non-parametric. The following considerations are based on a system of the form shown in Figure 1.3 with the deterministic input sequence (u_k), the stochastic disturbance (noise) (v_k), the undisturbed output sequence $(\bar{y}_k)$, the output sequence (y_k), and the BIBO-stable z-transfer function $G(z)$ with the impulse sequence $(g_k) = \mathcal{Z}^{-1}\{G(z)\}$.

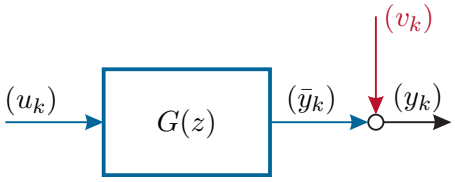


Figure 1.3: System considered for non-parametric identification.

For the output sequence, the following applies:

$$y_k = \sum_{i=0}^{k} g_{k-i} u_i + v_k \tag{1.1}$$

or in the z-domain

$$y_z(z) = G(z) u_z(z) + v_z(z) \tag{1.2}$$

with $y_z(z)$, $v_z(z)$, and $u_z(z)$ as the z-transforms of the sequences (y_k), (v_k), and (u_k).

1.2.1 Impulse Response Analysis

If the input variable (u_k) is chosen as the impulse sequence

$$u_k = \delta_k = \begin{cases} \alpha & \text{for } k = 0 \\ 0 & \text{for } k > 0, \end{cases} \tag{1.3}$$

then the output sequence is

$$y_k = \alpha g_k + v_k \ . \tag{1.4}$$

If the noise is small compared to the impulse response, i.e., $|v_k| \ll |\alpha g_k|$, then the impulse response can be determined from the measured sequence values y_k in the form

$$\hat{g}_k = \frac{y_k}{\alpha} \tag{1.5}$$

The problem with this method is that many technical systems do not allow impulse-like inputs.

1.2.2 Frequency Response Analysis - Excitation with a Harmonic Function

For the system in Figure 1.3, in the steady state, the response to the harmonic input sequence

$$(u_k) = (U \sin(\omega_0 k T_a)) \tag{1.6}$$

with the sampling time T_a is the harmonic output sequence (y_k)

$$y_k = Y \sin(\omega_0 k T_a + \varphi) + v_k \tag{1.7}$$

with

$$Y = U\left|G\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right| \quad \text{and} \quad \varphi = \arg\left(G\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right) \tag{1.8}$$

or

$$y_k = Y_c \cos(\omega_0 k T_a) + Y_s \sin(\omega_0 k T_a) + v_k \tag{1.9}$$

with

$$Y_c = Y \sin(\varphi) \quad \text{and} \quad Y_s = Y \cos(\varphi)\ . \tag{1.10}$$

In view of (1.7)-(1.10), it is reasonable to choose an approach of the form

$$\hat{y}_k = \hat{Y} \sin(\omega_0 k T_a + \hat{\varphi}) = \hat{Y}_c \cos(\omega_0 k T_a) + \hat{Y}_s \sin(\omega_0 k T_a) \tag{1.11}$$

with the parameters to be estimated $\hat{Y}_c = \hat{Y} \sin(\hat{\varphi})$ and $\hat{Y}_s = \hat{Y} \cos(\hat{\varphi})$ for the estimate $\hat{y}_k$ of the measured output variable y_k. Obviously, the parameters $\hat{Y}$ and $\hat{\varphi}$ can then be easily determined using the relationships

$$\hat{Y} = \sqrt{\hat{Y}_c^2 + \hat{Y}_s^2} \quad \text{and} \quad \hat{\varphi} = \arctan\left(\frac{\hat{Y}_c}{\hat{Y}_s}\right) \tag{1.12}$$

It must now be clarified by which choice of $\hat{Y}_c$ and $\hat{Y}_s$ the quadratic error

$$L = \frac{1}{N} \sum_{k=0}^{N-1} (y_k - \hat{y}_k)^2 \tag{1.13}$$

for N measurements of y_k is minimized. By substituting (1.11) into (1.13), differentiating with respect to $\hat{Y}_c$ and setting to zero, we obtain

$$\frac{\partial L}{\partial \hat{Y}_c} = -\frac{2}{N} \sum_{k=0}^{N-1} \Big(y_k - \hat{Y}_c \cos(\omega_0 k T_a) - \hat{Y}_s \sin(\omega_0 k T_a)\Big) \cos(\omega_0 k T_a) = 0 \tag{1.14}$$

or with $(\cos(\alpha))^2 = \frac{1}{2}(1+\cos(2\alpha))$ and $\cos(\alpha)\sin(\alpha) = \frac{1}{2}\sin(2\alpha)$

$$\hat{Y}_c - \frac{2}{N}\sum_{k=0}^{N-1} y_k \cos(\omega_0 k T_a) + \frac{\hat{Y}_c}{N}\sum_{k=0}^{N-1}\cos(2\omega_0 k T_a) + \frac{\hat{Y}_s}{N}\sum_{k=0}^{N-1}\sin(2\omega_0 k T_a) = 0\ . \tag{1.15}$$

Using the Euler formulas $\cos(\alpha) = \frac{1}{2}\left(\mathrm{e}^{\mathrm{I}\alpha} + \mathrm{e}^{-\mathrm{I}\alpha}\right)$ and $\sin(\alpha) = \frac{1}{2\mathrm{I}}\left(\mathrm{e}^{\mathrm{I}\alpha} - \mathrm{e}^{-\mathrm{I}\alpha}\right)$, the second and third sums of (1.15) can be rewritten as follows

$$\sum_{k=0}^{N-1}\cos(2\omega_0 k T_a) = \frac{1}{2}\sum_{k=0}^{N-1}\left(\mathrm{e}^{\mathrm{I}2\omega_0 k T_a} + \mathrm{e}^{-\mathrm{I}2\omega_0 k T_a}\right) \tag{1.16a}$$

$$\sum_{k=0}^{N-1}\sin(2\omega_0 k T_a) = \frac{1}{2\mathrm{I}}\sum_{k=0}^{N-1}\left(\mathrm{e}^{\mathrm{I}2\omega_0 k T_a} - \mathrm{e}^{-\mathrm{I}2\omega_0 k T_a}\right) \tag{1.16b}$$

If we now only use discrete values of the form

$$\omega_0 = \frac{2\pi l}{N T_a} \quad \text{with} \quad l = 1, 2, \ldots \tag{1.17}$$

for the angular frequency ω_0, then using the relationship

$$\sum_{k=0}^{N-1} z^{-k} = \frac{1 - z^{-N}}{1 - z^{-1}} \tag{1.18}$$

we obtain the following expression for the sum

$$\sum_{k=0}^{N-1}\mathrm{e}^{-\mathrm{I}2\omega_0 k T_a} = \sum_{k=0}^{N-1}\mathrm{e}^{-\mathrm{I}\frac{4\pi l k}{N}} = \frac{1-\mathrm{e}^{-\mathrm{I}4\pi l}}{1-\mathrm{e}^{-\mathrm{I}\frac{4\pi l}{N}}} = \begin{cases} N & \text{for } l = \frac{r}{2}N,\ r = \pm 1, \pm 2, \ldots \\ 0 & \text{for } l = \pm 1, \pm 2, \ldots \end{cases} \tag{1.19}$$

The analogous result is obtained, of course, if I is replaced by $-\mathrm{I}$. For ω_0 according to (1.17), the expressions (1.16) are therefore calculated as

$$\sum_{k=0}^{N-1}\cos(2\omega_0 k T_a) = \begin{cases} N & \text{for } l = \frac{r}{2}N,\ r = \pm 1, \pm 2, \ldots \\ 0 & \text{for } l = \pm 1, \pm 2, \ldots \end{cases} \tag{1.20a}$$

$$\sum_{k=0}^{N-1}\sin(2\omega_0 k T_a) = 0 \tag{1.20b}$$

If the test frequency ω_0 is chosen according to (1.17) such that $l < N/2$, then the expression in (1.20a) is also zero, and the optimal solution $\hat{Y}_c$ is calculated from (1.15) in the form

$$\hat{Y}_c = \frac{2}{N}\sum_{k=0}^{N-1} y_k \cos\left(\frac{2\pi l}{N}k\right)\ . \tag{1.21}$$

Note 1.1. Note that, for a given sampling time T_a, the maximum possible frequency of a uniquely representable harmonic function must be strictly less than the Nyquist frequency $\omega_{\max} = \pi/T_a$. It is immediately apparent that the value $l = N/2$ inserted into ω_0 from (1.17) corresponds to this Nyquist frequency.

Exercise 1.1. Show that the optimal $\hat{Y}_s$ is calculated as follows:

$$\hat{Y}_s = \frac{2}{N} \sum_{k=0}^{N-1} y_k \sin\left(\frac{2\pi l}{N} k\right) \tag{1.22}$$

This can now be summarized as follows: If the system from Figure 1.3 is excited with the harmonic sequence

$$(u_k) = (U \sin(\omega_0 k T_a)) \quad \text{with} \quad \omega_0 = \frac{2\pi l}{N T_a},\ l = 1, 2, \ldots, \frac{N}{2} - 1 \tag{1.23}$$

then, from the N measured values y_k, $k = 0, \ldots, N-1$, the discrete frequency response $G\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)$ at the frequencies $\omega_0 = \frac{2\pi l}{N T_a}$, $l = 1, 2, \ldots, \frac{N}{2} - 1$ can be approximated via the relationships (see (1.7),(1.8))

$$\hat{Y}_c = U \left|\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right| \sin\left(\arg\left(\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right)\right) \tag{1.24a}$$

$$\hat{Y}_s = U \left|\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right| \cos\left(\arg\left(\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right)\right) \tag{1.24b}$$

as follows

$$\left|\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right| = \frac{\sqrt{\hat{Y}_s^2 + \hat{Y}_c^2}}{U} \tag{1.25a}$$

$$\arg\left(\hat{G}\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)\right) = \arctan\left(\frac{\hat{Y}_c}{\hat{Y}_s}\right) \tag{1.25b}$$

with $\hat{Y}_c$ and $\hat{Y}_s$ from (1.21) and (1.22), respectively.

The above relationships (1.21) and (1.22) are closely related to the *discrete Fourier transform (DFT)* of the sequences (u_k) and (y_k). As a reminder, the discrete Fourier transform $F_n(\omega)$ of a sequence (f_k) is

$$F_n(\omega) = \sum_{k=0}^{N-1} f_k \mathrm{e}^{-\mathrm{I}\omega k T_a} \quad \text{with} \quad \omega = \frac{2\pi n}{N T_a}, \quad n = 0, 1, \ldots, N-1 \tag{1.26}$$

and the inverse discrete Fourier transform (IDFT) is

$$f_k = \frac{1}{N} \sum_{n=0}^{N-1} F_n \mathrm{e}^{\mathrm{I}\omega k T_a} \quad \text{with} \quad \omega = \frac{2\pi n}{N T_a}, \quad k = 0, 1, \ldots, N-1 \tag{1.27}$$

The DFT of the input sequence (u_k) according to (1.23) and the DFT of the measured output sequence (y_k) are, considering (1.21) and (1.22) with $\omega_0 = \frac{2\pi l}{NT_a}$,

$$\begin{aligned} U_n &= \sum_{k=0}^{N-1} u_k \mathrm{e}^{-\mathrm{I}\omega k T_a} = \frac{U}{2\mathrm{I}} \sum_{k=0}^{N-1} \left(\mathrm{e}^{\mathrm{I}\omega_0 k T_a} - \mathrm{e}^{-\mathrm{I}\omega_0 k T_a}\right) \mathrm{e}^{-\mathrm{I}\omega k T_a} \\ &= \frac{U}{2\mathrm{I}} \sum_{k=0}^{N-1} \left(\mathrm{e}^{\mathrm{I}\frac{2\pi k}{N}(l-n)} - \mathrm{e}^{-\mathrm{I}\frac{2\pi k}{N}(l+n)}\right) = \begin{cases} \frac{NU}{2\mathrm{I}} & \text{for } l = n \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1.28}$$

and

$$Y_n = \sum_{k=0}^{N-1} y_k \mathrm{e}^{-\mathrm{I}\omega k T_a} = \sum_{k=0}^{N-1} y_k \left(\cos\left(\frac{2\pi n}{N} k\right) - \mathrm{I}\sin\left(\frac{2\pi n}{N} k\right)\right) = \frac{N}{2}\left(\hat{Y}_c - \mathrm{I}\hat{Y}_s\right) . \tag{1.29}$$

Exercise 1.2. Prove the relationship in (1.28).

Thus, it can be seen that the estimation of the discrete frequency response $G\left(\mathrm{e}^{\mathrm{I}\omega_0 T_a}\right)$ at the frequencies $\omega_0 = \frac{2\pi l}{NT_a}$, $l = 1, 2, \ldots, \frac{N}{2} - 1$ according to (1.25) can be very easily calculated using the discrete Fourier transforms of the input and output sequences according to (1.28) and (1.29) – namely,

$$\frac{Y_n(\omega_0)}{U_n(\omega_0)} = \frac{\mathrm{DFT}((y_k))}{\mathrm{DFT}((u_k))} = \frac{\frac{N}{2}\left(\hat{Y}_c - \mathrm{I}\hat{Y}_s\right)}{\frac{NU}{2\mathrm{I}}} = \frac{\sqrt{\hat{Y}_c^2 + \hat{Y}_s^2}}{U} \mathrm{e}^{\mathrm{I}\arctan\frac{\hat{Y}_c}{\hat{Y}_s}} . \tag{1.30}$$

For practical application, it is of course advisable to perform the calculation of the discrete Fourier transforms using the *more efficient FFT (Fast Fourier Transform)* algorithm.

1.2.3 Frequency Response Analysis - ETFE

The relationship (1.30) now suggests choosing as the input sequence (u_k) not only a harmonic signal with constant frequency ω_0, but a signal containing several frequencies, thereby making use of the *superposition principle of linear systems*. The estimation of the frequency response according to relationship (1.30) is then called *empirical transfer function estimate (ETFE)* in the English-language literature – see also the Matlab command of the same name. The estimation is called empirical because, apart from the assumption of *linearity and time invariance of the system*, no further assumptions about the model structure are made. If, for certain frequencies ω, the Fourier transform $U(\omega) = 0$ due to the excitation (u_k), then the ETFE is not defined at these frequencies. Examples of suitable input sequences are impulse sequences, the 3-2-1 step, chirp signals, or PRBS signals.

A linear chirp signal with trapezoidal windowing is given by

$$u_k = U_0 + r_k \sin\left(\omega_{start} k T_a + \frac{(\omega_{end} - \omega_{start})}{NT_a} \frac{(kT_a)^2}{2}\right) \tag{1.31a}$$

$$r_k = U \mathrm{sat}\left(\frac{10k}{N}\right) \mathrm{sat}\left(\frac{10(N-k)}{N}\right) \tag{1.31b}$$

with

$$\mathrm{sat}(x) = \begin{cases} 1 & \text{for } x \geq 1 \\ x & \text{for } -1 < x < 1 \\ -1 & \text{for } x \leq -1 \end{cases} \tag{1.32}$$

and $k = 0, 1, \ldots, N-1$, for the sampling time T_a, the number of sampling points N, the constant U_0 to compensate for an offset of the amplitude U, and the lower and upper chirp frequencies ω_{start} and ω_{end}.

Exercise 1.3. Plot the time response of the chirp signal (1.31) and the response of the corresponding discrete Fourier transform for $N = 256$, $U = 1$, $\omega_{start} = 0.1$, and $\omega_{end} = 5/2$, $T_a = 0.5\,\mathrm{s}$ in Matlab.

So-called PRBS (Pseudo Random Binary Signal) sequences are frequently used in identification tasks because they exhibit similar properties to white noise. A PRBS signal of order O_p can be determined by the difference equation

$$p_k = \mathrm{mod}\Big(a_1 p_{k-1} + a_2 p_{k-2} + \ldots + a_{O_p} p_{k-O_p}, 2\Big) \tag{1.33}$$

with the coefficients $a_j \in \{0, 1\}$, $j = 1, \ldots, O_p$, which are shown for different orders in Table 1.1. Note that the PRBS signal repeats every $2^{O_p} - 1$ sampling steps.

Order O_p	$a_j \neq 0$ for the following j
2	$1, 2$
3	$2, 3$
4	$1, 4$
5	$2, 5$
6	$1, 6$
7	$3, 7$
8	$1, 2, 7, 8$
9	$4, 9$
10	$7, 10$
11	$9, 11$

Table 1.1: Coefficients of the PRBS signal.

To influence the frequency characteristics of the PRBS signal, oversampling of the signal p_k with a factor $P_p \geq 1$ is often useful. Given a sampling time T_a, the values of p_k are upsampled P_p-times, i.e.,

$$u((P_p k + j)T_a) = U p_k, \quad j = 0, \ldots, P_p - 1, \quad k = 0, \ldots, 2^{O_p} - 2 \tag{1.34}$$

with the amplitude U of the signal. It can be shown that this results in a low-pass filtering of the PRBS signal in the frequency domain. In the literature, an oversampling

of $P_p = 4$ is typically recommended. With an order O_p and an oversampling P_p, a number of $N = (2^{O_p} - 1)P_p$ sample values are available.

Exercise 1.4. Plot the time response of the PRBS signal (1.34) and the response of the corresponding discrete Fourier transform for $O_p = 10$, $U = 1$, and a sampling time $T_a = 0.5$s in Matlab. Investigate the influence of different values for the oversampling P_p and initialize the difference equation (1.34) with $p_0 = 1$.

Exercise 1.5. Assume that the sequences (u_k) and (y_k) consist of N equidistant sampling points with the sampling time T_a. Show that the minimum resolvable frequency is given by $\omega_{\min} = \frac{2\pi}{NT_a}$ and the maximum resolvable frequency by $\omega_{\max} = \frac{\pi}{T_a}$.

Exercise 1.6. Using the `fft` routine from Matlab, write a program to calculate the frequency response of a linear, time-invariant sampled-data system with the transfer function $G(z)$ according to (1.30). The input parameters should be the sampled sequences of the input and output variables (u_k) and (y_k), the sampling time T_a, and the number of measurement points N. The result should be the discrete frequencies $\omega = \frac{2\pi}{NT_a}n$, $\left|G\left(\mathrm{e}^{\mathrm{I}\omega T_a}\right)\right|$, and $\arg\left(G\left(\mathrm{e}^{\mathrm{I}\omega T_a}\right)\right)$ for $n = 1, \ldots, \frac{N}{2} - 1$. Test the program using the system

$$G(s) = \frac{1}{s^2 + 0.25s + 1}$$

using a chirp signal according to (1.31), (1.32).

The results of the ETFE are generally very good for deterministic input signals, especially for those frequencies that are sufficiently well excited by the input signal. For stochastic input signals, additional measures, such as smoothing the ETFE using suitable window functions (Hamming, Bartlett, Kaiser, etc.), must usually be taken to obtain usable results. These details will not be discussed in this lecture; interested students are referred to the literature listed at the end, in particular the book by L. Ljung [1.1].

In practical applications, it has proven very helpful to incorporate *a priori knowledge* about the process into the identification process. This can be done, for example, by choosing the input signal in such a way that it specifically excites the system in the frequency range of interest. Another possibility is to incorporate known parts of the transfer function to be identified. To this end, the transfer function $G(z)$ is factored into a known and an unknown part in the form

$$G(z) = \underbrace{\frac{z_b(z)}{n_b(z)}}_{\text{known}} \underbrace{G_1(z)}_{\text{unknown}} \tag{1.35}$$

and then the identification task for $G_1(z)$ is solved according to Figure 1.4.

This method, also known as *Clary's method*, has proven to be very effective in the identification of certain mechanical systems. In these systems, for example, it is known that the s-transfer function $G(s)$ has a double pole at $s_i = 0$. This means that the z-transfer function to be identified has a double pole at $z_i = 1$ according to the relationship

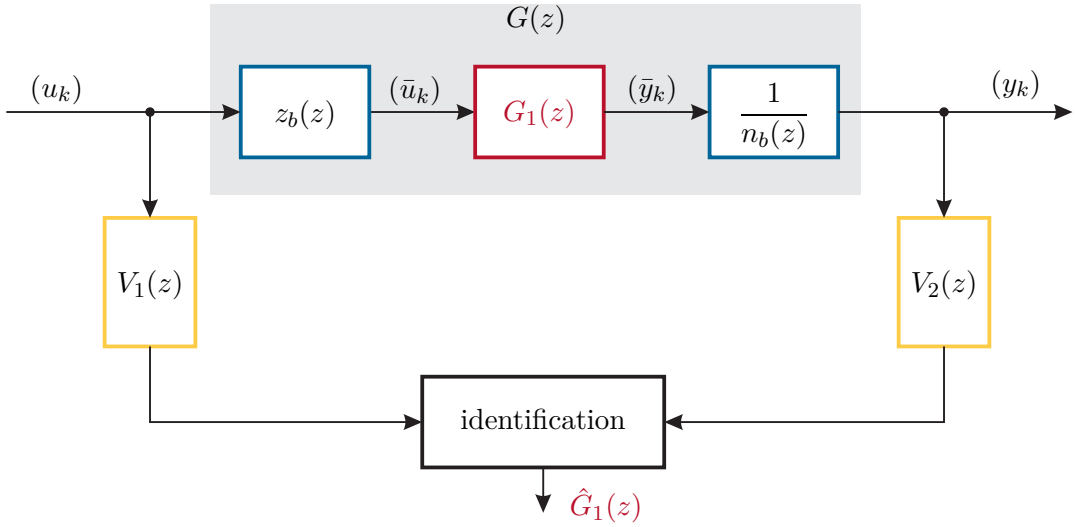


Figure 1.4: Identification considering known parts of the transfer function (Clary's method).

$z = \exp(sT_a)$ with the sampling time T_a. This direct assignment of the poles of s- and z-transfer functions is known not to be transferable to their zeros. In particular, the zeros outside the unit circle usually have only a small influence on the frequency response in the frequency range of interest, but are extremely difficult to identify. A non-exact but effective procedure is to assign the value $z_j = -1$ to these zeros z_j outside the unit circle. In order for the sequences (u_k) and (y_k) to be prepared for identification according to Figure 1.4, the individual prefilters $V_1(z)$ and $V_2(z)$ must be proper, i.e., the numerator degree must be less than or equal to the denominator degree. To ensure this, the known part of the transfer function from (1.35) is written in the form

$$\frac{z_b(z)}{n_b(z)} = \frac{z_b(z)}{z^n}\frac{z^n}{n_b(z)} \quad \text{with} \quad n = \max(\deg(z_b(z)), \deg(n_b(z))). \tag{1.36}$$

The prefilters are then

$$V_1(z) = \frac{z_b(z)}{z^n} \quad \text{and} \quad V_2(z) = \frac{n_b(z)}{z^n}\,. \tag{1.37}$$

Exercise 1.7. Identify the system transfer function

$$G(s) = \frac{1}{s^2}\frac{1}{s^2 + 0.25s + 1} \tag{1.38}$$

using Clary's method for the sampling time $T_a = 0.5\,\mathrm{s}$. Assume that you know from the system that it possesses a double integrator. Use the chirp signal from Exercise 1.3 as the input signal.

1.3 Parametric Methods

As the name implies, these methods use models with a finite number of parameters. These parameters are then determined within the framework of the identification task such that the model agrees with the system to be identified as well as possible in the sense of a quality criterion. The literature contains a vast number of different model structures, which is why a classification of these models is carried out in the first step. The following classification is essentially based on the Matlab System Identification Toolbox.

1.3.1 Model Structures

The starting point of the considerations is again the system from Figure 1.3. For the further steps, it is necessary to characterize the stochastic disturbance (noise) (v_k) in a suitable form. A relatively simple approach is to model (v_k) as the output sequence of a linear time-invariant system with the transfer function $H(z)$ and *white noise* (w_k) (sequence of independent random variables) with a given probability density function, i.e.,

$$v_k = \sum_{i=-\infty}^{k} h_{k-i} w_i = \sum_{i=0}^{\infty} h_i w_{k-i} \quad \text{with} \quad (h_k) = \mathcal{Z}^{-1}\{H(z)\} \; . \tag{1.39}$$

This approach is sufficient for most practical applications; however, not every possible stochastic disturbance (v_k) can be characterized in this way. Note that the character of the stochastic disturbance signal (v_k) can be specifically influenced by the choice of the probability density function of (w_k). For example, it is plausible that for one and the same system with the transfer function $H(z)$, a white noise $(w_{2,k})$ with the probability density function

$$\begin{aligned} w_{2,k} &= 0 \quad \text{with probability } 1-\mu \\ w_{2,k} &= r \quad \text{with probability } \mu \end{aligned} \tag{1.40}$$

for a very small μ and a uniformly distributed random variable $r \in (-1, 1)$ gives a completely different picture for $(v_{2,k})$ than a white noise $(w_{1,k})$ with the probability density function $w_{1,k} = r$.

Note the time respoonse of $(v_{1,k})$ and $(v_{2,k})$ for $H(z) = \mathbf{Z}\{H(s)\}$, with

$$H(s) = \frac{1}{\frac{s^2}{(2\pi 10)^2} + 2\frac{0.2s}{2\pi 10} + 1}, \tag{1.41}$$

in Figure 1.5, where $T_a = 10\text{ms}$ and $\mu = 0.07$ were used.

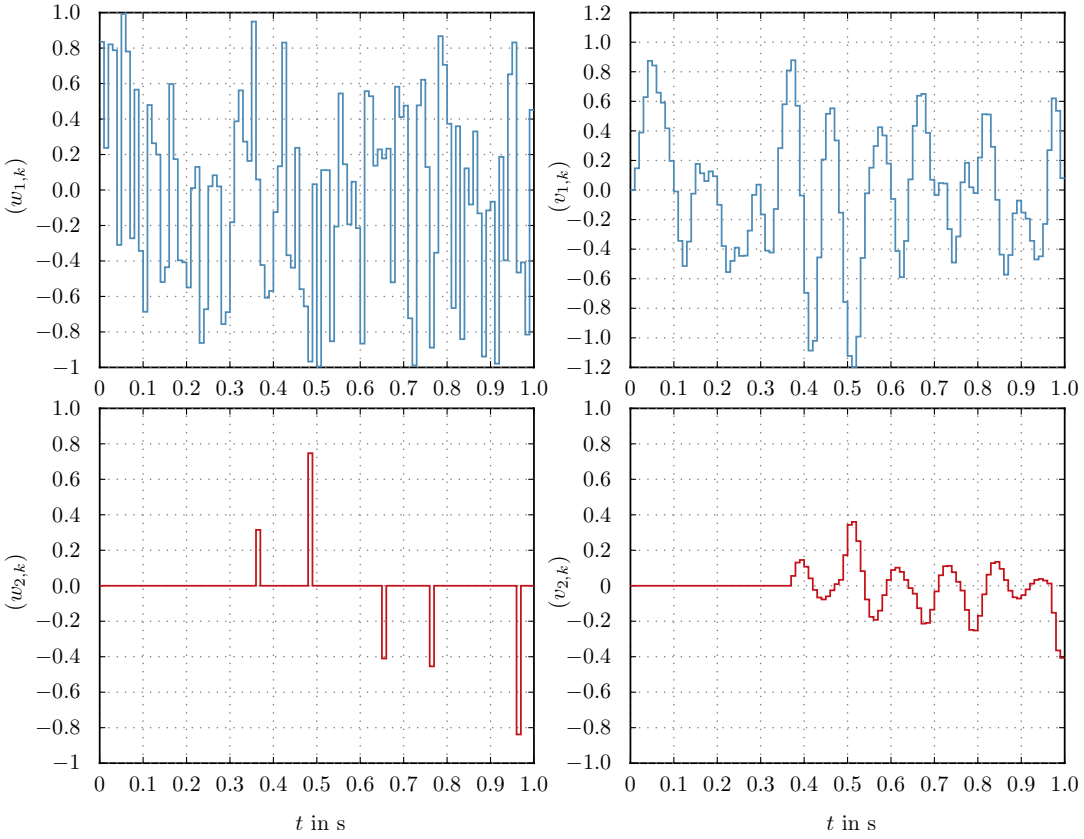


Figure 1.5: Responses $(v_{1,k})$ and $(v_{2,k})$ of the second-order linear time-invariant discrete-time system (1.41) to white noise $(w_{1,k})$ and $(w_{2,k})$ with different probability density functions.

Combining (1.1) with (1.39), the output variable y_k from Figure 1.3 is

$$y_k = \sum_{i=0}^{\infty} g_i u_{k-i} + \sum_{i=0}^{\infty} h_i w_{k-i} \tag{1.42}$$

or, introducing the shift operator δ with $u_{k+1} = \delta u_k$ or $u_{k-1} = \delta^{-1} u_k$,

$$y_k = \sum_{i=0}^{\infty} g_i \delta^{-i} u_k + \sum_{i=0}^{\infty} h_i \delta^{-i} w_k \ . \tag{1.43}$$

To save writing effort later on, we also write (1.43) as

$$y_k = G(\delta) u_k + H(\delta) w_k \tag{1.44}$$

with the transfer operators

$$G(\delta) = \sum_{i=0}^{\infty} g_i \delta^{-i} \quad \text{and} \quad H(\delta) = \sum_{i=0}^{\infty} h_i \delta^{-i} \ . \tag{1.45}$$

Note that the expressions for the transfer operators $G(\delta)$ and $H(\delta)$ are the same as the z-transfer functions of linear time-invariant sampled-data systems with the impulse responses (g_k) and (h_k). However, if $G(\delta)$ and $H(\delta)$ were interpreted as z-transfer functions, the notation of (1.44) would not be permissible.

By representing $G(\delta)$ and $H(\delta)$ in the form of rational transfer operators with their associated numerator and denominator polynomials and combining all common poles of $G(\delta)$ and $H(\delta)$ in the polynomial $A(\delta)$, (1.44) becomes (see Figure 1.6)

$$A(\delta) y_k = \frac{B(\delta)}{F(\delta)} u_k + \frac{C(\delta)}{D(\delta)} w_k \ . \tag{1.46}$$

Based on (1.46), a classification of the various model structures can now be carried out.

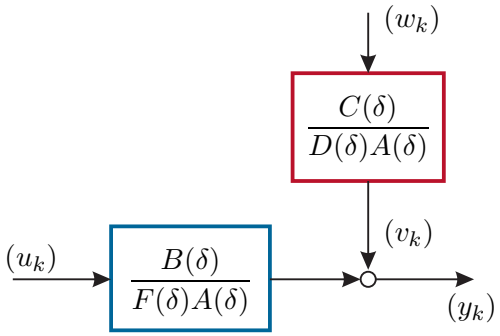


Figure 1.6: On the model structures of parametric identification.

ARMA Model

In econometrics, one often deals with models that do not have a deterministic input variable, i.e., $(u_k) = (0)$. In these so-called *time series models*, a distinction is made between the following special cases:

The model

$$y_k = \frac{1}{D(\delta)} w_k \tag{1.47}$$

or for $D(\delta) = d_0 + d_1\delta^{-1} + \ldots + d_n\delta^{-n}$ with $d_0 = 1$

$$y_k = -d_1 y_{k-1} - d_2 y_{k-2} - \ldots - d_n y_{k-n} + w_k \tag{1.48}$$

is called an *autoregressive model* or *AR model*. Equation (1.48) can also be written in the form

$$y_k = \mathbf{s}_k^{\mathrm{T}} \mathbf{p} \tag{1.49}$$

with the *parameter vector* $\mathbf{p}$ and the *data vector* $\mathbf{s}_k$ in the form

$$\mathbf{p}^{\mathrm{T}} = \begin{bmatrix} -d_1 & -d_2 & \ldots & -d_n & 1 \end{bmatrix} \tag{1.50a}$$

$$\mathbf{s}_k^{\mathrm{T}} = \begin{bmatrix} y_{k-1} & y_{k-2} & \ldots & y_{k-n} & w_k \end{bmatrix} \tag{1.50b}$$

In statistics, models that are linear in the parameters $\mathbf{p}$, such as (1.48), are also called *linear regression models*. If the data vector $\mathbf{s}_k$ contains only past values y_j, $j < k$, of the quantity y_k to be calculated, then this model is called *autoregressive*.

A model of the form

$$y_k = C(\delta) w_k \tag{1.51}$$

or for $C(\delta) = c_0 + c_1\delta^{-1} + \ldots + c_m\delta^{-m}$ with $c_0 = 1$

$$y_k = w_k + c_1 w_{k-1} + c_2 w_{k-2} + \ldots + c_m w_{k-m} \tag{1.52}$$

is called a *moving average model* or *MA model*.

Exercise 1.8. Show that the impulse response sequence (g_k) of the z-transfer function

$$G(z) = c_0 + c_1 z^{-1} + \ldots + c_m z^{-m} \tag{1.53}$$

is calculated as follows:

$$(g_k) = (c_0, c_1, \ldots, c_m, 0, 0, \ldots) \tag{1.54}$$

For this reason, a transfer function of the form (1.53) is also called a FIR (finite impulse response) model. Analogously, a z-transfer function of the form

$$G(z) = \frac{1}{d_0 + d_1 z^{-1} + \ldots + d_n z^{-n}} \tag{1.55}$$

is called an IIR (infinite impulse response) model.

Combining (1.47) and (1.51), we obtain the so-called *ARMA (autoregressive moving average) model*

$$y_k = \frac{C(\delta)}{D(\delta)} w_k \ . \tag{1.56}$$

For $C(\delta) = c_0 + c_1\delta^{-1} + \ldots + c_m\delta^{-m}$ and $D(\delta) = d_0 + d_1\delta^{-1} + \ldots + d_n\delta^{-n}$, $d_0 = 1$, the output variable y_k is

$$y_k = c_0 w_k + c_1 w_{k-1} + c_2 w_{k-2} + \ldots + c_m w_{k-m} - d_1 y_{k-1} - d_2 y_{k-2} - \ldots - d_n y_{k-n} \ . \tag{1.57}$$

ARX Model

If the AR model (1.47) is extended to include the influence of the deterministic input variable u_k (*exogenous input*), then the *ARX (autoregressive with exogenous input) model* is obtained:

$$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{1}{A(\delta)} w_k \ . \tag{1.58}$$

As can be seen from (1.58), in this case the denominator polynomials of the transfer operators $G(\delta)$ and $H(\delta)$ according to (1.44) are identical. A mathematical justification for this choice of structure will be given later in the context of least-squares identification. As will be shown, the main advantage of this model structure lies in the fact that the parameters are linear in the estimation error and can therefore be very easily estimated using linear least-squares methods.

ARMAX Model

Analogously to the ARX model, the *ARMAX (autoregressive moving average with exogenous input) model* is given by

$$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{C(\delta)}{A(\delta)} w_k, \tag{1.59}$$

where again the denominator polynomials of the transfer operators from the deterministic and stochastic inputs u_k and w_k to the output y_k are equal. It should be mentioned at this point that there are a large number of other models in the literature, but they are all composed in a similar way to those discussed so far and are special cases of (1.46). See Table 1.2 below.

Model structure	Model equation
MA	$y_k = C(\delta) w_k$
AR	$y_k = \frac{1}{D(\delta)} w_k$
ARMA	$y_k = \frac{C(\delta)}{D(\delta)} w_k$
ARX	$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{1}{A(\delta)} w_k$
ARMAX	$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{C(\delta)}{A(\delta)} w_k$
ARARX	$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{1}{D(\delta)A(\delta)} w_k$
ARARMAX	$y_k = \frac{B(\delta)}{A(\delta)} u_k + \frac{C(\delta)}{D(\delta)A(\delta)} w_k$
OE (output error)	$y_k = \frac{B(\delta)}{F(\delta)} u_k + w_k$
BJ (Box-Jenkins)	$y_k = \frac{B(\delta)}{F(\delta)} u_k + \frac{C(\delta)}{D(\delta)} w_k$

Table 1.2: Model structures and model equations.

1.3.2 Least Squares Method

Given is the *overdetermined* linear system of equations

$$\mathbf{y} = \mathbf{S}\mathbf{p} \tag{1.60}$$

with the $(m \times n)$-matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, the m-dimensional vector $\mathbf{y} \in \mathbb{R}^m$, and the n-dimensional vector of unknowns $\mathbf{p} \in \mathbb{R}^n$. For $m > n$ and $\operatorname{rank}(\mathbf{S}) = n \neq \operatorname{rank}([\mathbf{S}, \mathbf{y}])$, the system of equations (1.60) has no solution for $\mathbf{p}$. We now seek the solution $\mathbf{p}_0$ that minimizes the quadratic error

$$\min_{\mathbf{p}} \|\mathbf{e}\|_2^2 \quad \text{with} \quad \mathbf{e} = \mathbf{y} - \mathbf{S}\mathbf{p}. \tag{1.61}$$

Setting the derivative of $\|\mathbf{e}\|_2^2$ with respect to $\mathbf{p}$ equal to zero

$$\frac{\partial}{\partial \mathbf{p}} \mathbf{e}^{\mathrm{T}} \mathbf{e} = \frac{\partial}{\partial \mathbf{p}} \underbrace{(\mathbf{y} - \mathbf{S}\mathbf{p})^{\mathrm{T}} (\mathbf{y} - \mathbf{S}\mathbf{p})}_{\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{p} + \mathbf{p}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{p}} = -2\mathbf{y}^{\mathrm{T}}\mathbf{S} + 2\mathbf{p}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{0}^{\mathrm{T}}, \tag{1.62}$$

the optimal solution $\mathbf{p} = \mathbf{p}_0$ in the sense of (1.61) is

$$\mathbf{p}_0 = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1} \mathbf{S}^{\mathrm{T}} \mathbf{y} \ . \tag{1.63}$$

The expression $\mathbf{S}^{\dagger} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1} \mathbf{S}^{\mathrm{T}}$ is also called the *pseudoinverse* of the matrix $\mathbf{S}$ (see the Matlab commands `\` and `pinv`). It can be seen that for the regularity of $\mathbf{S}^{\mathrm{T}}\mathbf{S}$, the matrix $\mathbf{S}$ must have full column rank, i.e., $\operatorname{rank}(\mathbf{S}) = n$.

Exercise 1.9 (Polynomial approximation in the least squares sense). Given are N measurement points that describe the relationship $y_j = f(x_j)$, $j = 1, \ldots, N$. Assume an n-th order polynomial with $n + 1 < N$ of the form

$$f(x) = p_0 + p_1 x + p_2 x^2 + \ldots + p_{n-1} x^{n-1} + p_n x^n$$

for the function $f(x)$ and determine the polynomial coefficients $p_0, p_1, \ldots, p_n$ using the least squares method. Test your approach using the function $g(x) = \tanh(x/10)$ for $x_j = -20 + j$, $j = 0, \ldots, 40$, and different orders n of the polynomial.

Exercise 1.10. The mathematical model of a externaly excited DC motor is

$$L_a \frac{d}{dt} i_a = u_a - R_a i_a - k_a \omega$$
$$J_r \frac{d}{dt} \omega = k_a i_a - d_v \omega - d_c \text{sign}(\omega)$$

with the armature current i_a, the angular velocity ω, the armature voltage u_a, the armature inductance L_a, the armature resistance R_a, the motor constant k_a, the moment of inertia J_r, and the friction parameters d_v and d_c. Determine the parameters k_a, R_a, d_v, and d_c from stationary measurements of u_a, i_a, and ω using the least squares method. Test your algorithm using the measurement data `data.mat` and `data_rausch.mat`, in which the measured values of u_a, i_a, and ω are stored with and without measurement noise. Compare your identification results with the nominal parameters $R_a = 1.373\,\Omega$, $k_a = 0.0652\,\text{V s/rad}$, $d_c = 0.0188\,\text{N m}$, and $d_v = 43.3 \cdot 10^{-6}\,\text{N m s/rad}$.

Data for Exercise 1.10:
https://www.acin.tuwien.ac.at/file/teaching/master/Regelungssysteme-1/Daten_Aufgabe_1_10.zip.

The optimal solution $\mathbf{p}_0$ according to (1.63) of the optimization problem (1.61) allows the following geometrical interpretation: The n linearly independent column vectors of the matrix $\mathbf{S}$ span an n-dimensional subspace $\mathcal{U}$ in $\mathbb{R}^m$. The solvability of the system of equations (1.60) is equivalent to the question of whether a linear combination (described by the entries in the vector $\mathbf{p}$) of the column vectors of $\mathbf{S}$ exists such that the vector $\mathbf{y}$ can be represented. Thus, the system of equations (1.60) is uniquely solvable if $\mathbf{y}$ lies in this subspace $\mathcal{U}$, i.e., $\text{rank}(\mathbf{S}) = \text{rank}([\mathbf{S}, \mathbf{y}])$. If $\mathbf{y}$ does not lie in the subspace $\mathcal{U}$, then the error between the quantity $\mathbf{y}_0 = \mathbf{S}\mathbf{p}_0 = \mathbf{S}\left(\mathbf{S}^\mathrm{T}\mathbf{S}\right)^{-1}\mathbf{S}^\mathrm{T}\mathbf{y}$ (calculated using the optimal solution $\mathbf{p}_0$ estimated in the least squares sense) and the actual quantity $\mathbf{y}$ is

$$\mathbf{e}_0 = \mathbf{y} - \mathbf{y}_0 = \mathbf{y} - \mathbf{S}\left(\mathbf{S}^\mathrm{T}\mathbf{S}\right)^{-1}\mathbf{S}^\mathrm{T}\mathbf{y} \,. \tag{1.64}$$

It can now be seen that this error $\mathbf{e}_0$ is orthogonal to the subspace $\mathcal{U}$, since

$$\mathbf{S}^\mathrm{T}\mathbf{e}_0 = \mathbf{S}^\mathrm{T}\mathbf{y} - \mathbf{S}^\mathrm{T}\mathbf{S}\left(\mathbf{S}^\mathrm{T}\mathbf{S}\right)^{-1}\mathbf{S}^\mathrm{T}\mathbf{y} = \mathbf{0} \,. \tag{1.65}$$

Figure 1.7 illustrates this geometrically for $m = 3$ and $n = 2$. The finding that the error

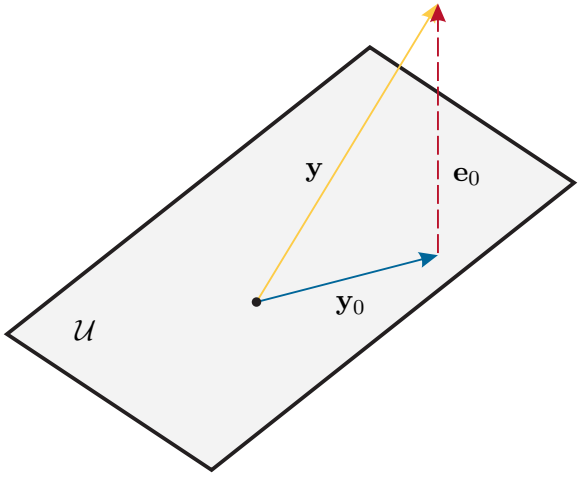


Figure 1.7: On the least squares method.

$\mathbf{e}_0$ is orthogonal to the subspace $\mathcal{U}$ is a special case of the so-called *projection theorem in a Hilbert space*.

Projection Theorem in a Hilbert Space

To illustrate the projection theorem, the concepts of vector space, normed vector space, and Hilbert space will be briefly reviewed below.

Definition 1.1 (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 $+$ is a commutative group, i.e., for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ the following holds:

(1)	$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$	Commutativity
(2)	$\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$	Associativity
(3)	$\mathbf{0} + \mathbf{x} = \mathbf{x}$	Neutral element
(4)	$\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$	Inverse element

(2) Scalar multiplication $\cdot$ with scalars a, $b \in K$ satisfies:

(1)	$a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$	Distributivity
(2)	$(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$	Distributivity
(3)	$(ab)\mathbf{x} = a(b\mathbf{x})$	Associativity
(4)	$1\mathbf{x} = \mathbf{x} \quad , \quad 0\mathbf{x} = \mathbf{0}$	

Definition 1.2 (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 $\|\mathbf{x}\| : \mathcal{X} \to \mathbb{R}_+$, which assigns to each $\mathbf{x} \in \mathcal{X}$ a real-valued number $\|\mathbf{x}\|$, the so-called *norm* of $\mathbf{x}$, and satisfies the following norm axioms:

(1)	$\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathcal{X}$	Non-negativity
(2)	$\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$	
(3)	$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$	Triangle inequality
(4)	$\|\alpha\mathbf{x}\| = \|\alpha\|\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathcal{X}$ and all $\alpha \in K$	

Definition 1.3 (Pre-Hilbert Space)**.** Let $\mathcal{X}$ be a linear vector space with the scalar field K. A mapping $\langle \mathbf{x}, \mathbf{y} \rangle : \mathcal{X} \times \mathcal{X} \to K$, which assigns a scalar to any two elements $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, is called an *inner product* if it satisfies the following conditions:

(1) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ Bilinearity
(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) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \quad \text{and} \quad \langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$

with $a \in K$.

A complete pre-Hilbert space is called a Hilbert space, cf. e.g. [1.2]. Two vectors $\mathbf{x}$ and $\mathbf{y}$ from a Hilbert space $\mathcal{H}$ are called *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. A non-empty set $\mathcal{U}$ is called a *subspace* of $\mathcal{H}$ if for all linear combinations of vectors $\mathbf{x}$ and $\mathbf{y}$ from $\mathcal{U}$ it holds that $a\mathbf{x} + b\mathbf{y} \in \mathcal{U}$, with scalars a and b. Furthermore, the subspace $\mathcal{U}$ is called *closed* if the limit of every convergent sequence in $\mathcal{U}$ is also in $\mathcal{U}$.

Exercise 1.11. Show that $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ defines a norm according to Definition 1.2.

The following projection theorem now gives a necessary and sufficient condition for solving the following problem: Given is a vector $\mathbf{y}$ in a Hilbert space $\mathcal{H}$. The vector $\mathbf{y}_0$ from a subspace $\mathcal{U}$ of $\mathcal{H}$ is sought that minimizes the norm $\|\mathbf{y} - \mathbf{y}_0\|$.

Theorem 1.1 (Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a closed subspace of $\mathcal{H}$. For every vector $\mathbf{y} \in \mathcal{H}$, there exists a unique vector $\mathbf{y}_0 \in \mathcal{U}$ such that*

$$\|\mathbf{y} - \mathbf{y}_0\| \leq \|\mathbf{y} - \mathbf{x}\| \tag{1.66}$$

for all $\mathbf{x} \in \mathcal{U}$. The vector $\mathbf{y}_0$ is the unique minimizing vector if and only if

$$\langle \mathbf{y} - \mathbf{y}_0, \mathbf{x} \rangle = 0 \tag{1.67}$$

for all $\mathbf{x} \in \mathcal{U}$. This means that the error $\mathbf{y} - \mathbf{y}_0$ must be orthogonal to all vectors of the subspace $\mathcal{U}$.

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

In the following, the projection theorem will be used to solve the optimization problem (1.61). Given is the Hilbert space $\mathcal{H} = \mathbb{R}^m$ with the inner product

$$\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^{\mathrm{T}} \mathbf{z} = \sum_{j=1}^{m} x_j z_j \tag{1.68}$$

with

$$\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix} \tag{1.69a}$$

$$\mathbf{z}^{\mathrm{T}} = \begin{bmatrix} z_1 & z_2 & \dots & z_m \end{bmatrix} \tag{1.69b}$$

and the closed subspace $\mathcal{U}$ of $\mathcal{H}$, which is spanned by the column vectors $\mathbf{s}_k$, $k = 1, \dots, n$ of the matrix $\mathbf{S}$, i.e., $\mathcal{U} = \operatorname{span}\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$. In the sense of (1.61), we are looking for

that $\mathbf{S}\mathbf{p}_0 = \mathbf{y}_0 \in \mathcal{U}$ which, for a given $\mathbf{y} \in \mathcal{H}$, minimizes the squared norm of the error $\|\mathbf{e}\|_2^2 = \|\mathbf{y} - \mathbf{y}_0\|_2^2$. According to Theorem 1.1, it must therefore hold that

$$\langle \mathbf{y} - \mathbf{S}\mathbf{p}_0, \mathbf{s}_j \rangle = \langle \mathbf{y} - \mathbf{s}_1 p_{0,1} - \mathbf{s}_2 p_{0,2} - \ldots - \mathbf{s}_n p_{0,n}, \mathbf{s}_j \rangle = 0 \quad \text{for all } j = 1, \ldots, n \ . \tag{1.70}$$

In matrix notation, using the properties of the inner product, we obtain

$$\underbrace{\begin{bmatrix} \langle \mathbf{s}_1, \mathbf{s}_1 \rangle & \ldots & \langle \mathbf{s}_n, \mathbf{s}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{s}_1, \mathbf{s}_n \rangle & \ldots & \langle \mathbf{s}_n, \mathbf{s}_n \rangle \end{bmatrix}}_{\mathbf{G} = \mathbf{S}^{\mathrm{T}}\mathbf{S}} \underbrace{\begin{bmatrix} p_{0,1} \\ \vdots \\ p_{0,n} \end{bmatrix}}_{\mathbf{p}_0} = \underbrace{\begin{bmatrix} \langle \mathbf{y}, \mathbf{s}_1 \rangle \\ \vdots \\ \langle \mathbf{y}, \mathbf{s}_n \rangle \end{bmatrix}}_{\mathbf{S}^{\mathrm{T}}\mathbf{y}} \tag{1.71}$$

and thus immediately the solution for $\mathbf{p}_0$ from (1.63). The matrix $\mathbf{G}$ from (1.71) is also called the *Gramian matrix*.

Exercise 1.12. Show that the Gramian matrix from (1.71) is non-singular if and only if the vectors $\mathbf{s}_k$, $k = 1, \ldots, n$, are linearly independent.

Exercise 1.13. Calculate the solution of the quadratic minimization problem from Section 1.2.2 using Theorem 1.1.

Remark: (to Exercise 1.13) The set of all N-tuples $\xi = \{\xi_0, \xi_2, \ldots, \xi_{N-1}\}$ with the inner product

$$\langle x, z \rangle = \frac{1}{N} \sum_{k=0}^{N-1} x_k z_k \quad \text{with} \quad x, z \in \mathcal{H}$$

is used as the Hilbert space $\mathcal{H}$. The set of all N-tuples $\{C\cos(\omega_0 k T_a) + S\sin(\omega_0 k T_a)\}$, $k = 0, \ldots, N-1$ with arbitrary but constant coefficients C and S and fixed angular frequency ω_0 forms a closed subspace $\mathcal{U}$ of $\mathcal{H}$. The quadratic minimization problem from Section 1.2.2 can now be formulated such that we are looking for that $\hat{y} \in \mathcal{U}$ which minimizes the norm (compare with (1.13))

$$\|y - \hat{y}\|^2 = \frac{1}{N} \sum_{k=0}^{N-1} (y_k - \hat{y}_k)^2$$

with

$$\hat{y}_k = \hat{Y}_c \cos(\omega_0 k T_a) + \hat{Y}_s \sin(\omega_0 k T_a)$$

for a given $y \in \mathcal{H}$. According to Theorem 1.1, this is the case if and only if

$$\langle y - \hat{y}, x \rangle = 0 \quad \text{for all } x \in \mathcal{U}$$

or, for the two special cases $x = \cos(\omega_0 k T_a)$ and $x = \sin(\omega_0 k T_a)$, we obtain the

conditions

$$\frac{1}{N}\sum_{k=0}^{N-1}\Big(y_k - \hat{Y}_c\cos(\omega_0 kT_a) - \hat{Y}_s\sin(\omega_0 kT_a)\Big)\cos(\omega_0 kT_a) = 0$$
$$\frac{1}{N}\sum_{k=0}^{N-1}\Big(y_k - \hat{Y}_c\cos(\omega_0 kT_a) - \hat{Y}_s\sin(\omega_0 kT_a)\Big)\sin(\omega_0 kT_a) = 0\ .$$

1.3.3 Least-Squares Identification

The starting point is the ARX model (1.58) without stochastic disturbance, i.e., $(w_k) = (0)$, of the form

$$y_k = \frac{B(\delta)}{A(\delta)}u_k = \delta^{-d}\frac{b_0 + b_1\delta^{-1} + \ldots + b_m\delta^{-m}}{1 + a_1\delta^{-1} + \ldots + a_n\delta^{-n}}u_k \quad \text{with} \quad d \geq 0. \tag{1.72}$$

For a transfer function $G(z)$ with numerator degree m and denominator degree n, d is calculated as $d = n - m$. Note that dead times of the form $z^{-\rho}$, $\rho > 0$, can also be taken into account with the formulation (1.72). The value of the output sequence (y_k) at the k-th sampling time is calculated from (1.72) as

$$y_k = -a_1 y_{k-1} - \ldots - a_n y_{k-n} + b_0 u_{k-d} + b_1 u_{k-d-1} + \ldots + b_m u_{k-d-m} \tag{1.73}$$

or in vector notation

$$y_k = \underbrace{\begin{bmatrix} -y_{k-1} & -y_{k-2} & \ldots & -y_{k-n} & u_{k-d} & u_{k-d-1} & \ldots & u_{k-d-m} \end{bmatrix}}_{\mathbf{s}_k^{\mathrm{T}}} \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b_0 \\ \vdots \\ b_m \end{bmatrix}}_{\mathbf{p}} \tag{1.74}$$

with the *data vector* $\mathbf{s}_k$ and the *parameter vector* $\mathbf{p}$. In the identification task, the parameter vector $\mathbf{p}$ is now estimated such that a model error, yet to be defined, is minimized. If the model error is chosen as the so-called *generalized equation error* according to Figure 1.8, we obtain the relationship

$$e_k = \hat{A}(\delta)y_k - \hat{B}(\delta)u_k \tag{1.75}$$

with

$$\hat{A}(\delta) = 1 + \hat{a}_1\delta^{-1} + \ldots + \hat{a}_n\delta^{-n} \tag{1.76a}$$
$$\hat{B}(\delta) = \hat{b}_0\delta^{-d} + \hat{b}_1\delta^{-1-d} + \ldots + \hat{b}_m\delta^{-m-d} \tag{1.76b}$$

and the estimated coefficients of the numerator and denominator polynomials $\hat{b}_j$, $j = 0, \ldots, m$ and $\hat{a}_k$, $k = 1, \ldots, n$. Substituting (1.76) into (1.75) and combining the coefficients

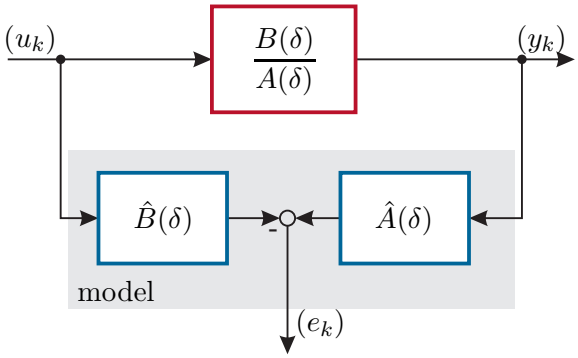


Figure 1.8: On the generalized equation error.

to be estimated in the *parameter estimate vector* $\hat{\mathbf{p}} = \begin{bmatrix}\hat{a}_1 & \dots & \hat{a}_n & \hat{b}_0 & \dots & \hat{b}_m\end{bmatrix}^{\mathrm{T}}$, we obtain

$$e_k = y_k - \mathbf{s}_k^{\mathrm{T}}\hat{\mathbf{p}} \tag{1.77}$$

with the *data vector* $\mathbf{s}_k$ according to (1.74). From (1.77), it can be immediately seen that the estimate $\hat{y}_k$ of y_k is calculated as $\hat{y}_k = \mathbf{s}_k^{\mathrm{T}}\hat{\mathbf{p}}$. By combining $j = 0, \dots, N$ measurements, (1.77) can be extended as follows:

$$\underbrace{\begin{bmatrix} e_0 \\ \vdots \\ e_N \end{bmatrix}}_{\mathbf{e}_N} = \underbrace{\begin{bmatrix} y_0 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}_N} - \underbrace{\begin{bmatrix} \mathbf{s}_0^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_N^{\mathrm{T}} \end{bmatrix}}_{\mathbf{S}_N} \hat{\mathbf{p}}_N \tag{1.78}$$

where the index N of the parameter estimate vector $\hat{\mathbf{p}}$ indicates that $N+1$ measurements are used to estimate $\mathbf{p}$. For $N > n+m$ with m and n according to (1.72) and $\mathrm{rank}(\mathbf{S}_N) \neq \mathrm{rank}([\mathbf{S}_N, \mathbf{y}_N])$, the system of equations (1.78) has no solution for $\mathbf{e}_N = \mathbf{0}$. Following the considerations of Section 1.3.2, the solution of (1.78) in the least squares sense (see (1.63)) is

$$\hat{\mathbf{p}}_N = \left(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\right)^{-1}\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N \ . \tag{1.79}$$

Remark: Regarding the choice of notation, it should be noted that $j = 0$ does not necessarily refer to the starting time of the measurement $k = 0$, but to the time k at which the entry u_{k-d} in the data vector $\mathbf{s}_k^{\mathrm{T}}$ is first non-zero.

It should be emphasized again that only the specific choice of the generalized equation error according to Figure 1.8 leads to a parametrically linear estimation problem. If, for example, the so-called *output error* shown in Figure 1.9

$$e_k = y_k - \hat{y}_k = y_k - \frac{\hat{B}(\delta)}{\hat{A}(\delta)}u_k = y_k - \delta^{-d}\frac{\hat{b}_0 + \hat{b}_1\delta^{-1} + \ldots + \hat{b}_m\delta^{-m}}{1 + \hat{a}_1\delta^{-1} + \ldots + \hat{a}_n\delta^{-n}}u_k \tag{1.80}$$

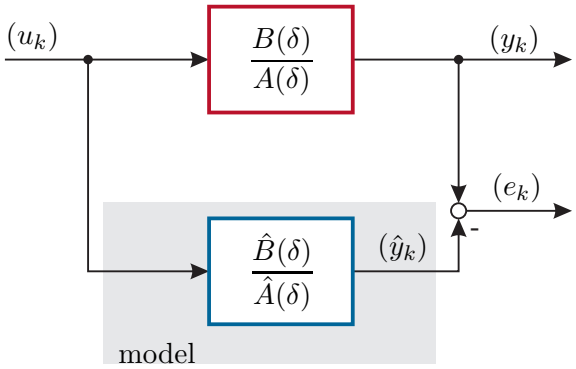


Figure 1.9: On the output error.

were used, then from

$$\hat{y}_k = -\hat{a}_1\hat{y}_{k-1} - \ldots - \hat{a}_n\hat{y}_{k-n} + \hat{b}_0 u_{k-d} + \hat{b}_1 u_{k-d-1} + \ldots + \hat{b}_m u_{k-d-m} \tag{1.81}$$

it can already be seen for $k = 0, \ldots, d+1$ that

$$\begin{aligned} \hat{y}_k &= 0 \quad \text{for } k = 0, 1, \ldots, d-1 \\ \hat{y}_d &= \hat{b}_0 u_0 \\ \hat{y}_{d+1} &= -\hat{a}_1\hat{y}_d + \hat{b}_0 u_1 + \hat{b}_1 u_0 = \left(\hat{b}_1 - \hat{a}_1\hat{b}_0\right)u_0 + \hat{b}_0 u_1, \end{aligned} \tag{1.82}$$

that in this case it is a parametrically nonlinear estimation problem, which is orders of magnitude more difficult to solve.

Before specifying how a stochastic disturbance (v_k) affects the result of the parameter estimation $\hat{\mathbf{p}}$ from (1.79), two fundamental properties of parameter estimation methods are defined.

Definition 1.4 (Unbiased Estimate)**.** If an estimate for an arbitrary number N of measurements yields a systematic error

$$\mathrm{E}(\hat{\mathbf{p}}_N - \mathbf{p}) = \mathrm{E}(\hat{\mathbf{p}}_N) - \mathbf{p} = \mathbf{b} \neq \mathbf{0} \tag{1.83}$$

then this error is called *bias*. For an *unbiased* estimate, therefore,

$$\mathrm{E}(\hat{\mathbf{p}}_N) = \mathbf{p}\ . \tag{1.84}$$

Definition 1.5 (Consistent Estimate)**.** An estimate is called *consistent* if the estimate becomes more accurate the larger the number N of measurements, i.e., if

$$\lim_{N\to\infty} \mathrm{E}(\hat{\mathbf{p}}_N) = \mathbf{p}\ . \tag{1.85}$$

An estimate is called *consistent in the mean square* if, in addition to (1.85), the condition

$$\lim_{N\to\infty} \operatorname{cov}(\hat{\mathbf{p}}_N) = \lim_{N\to\infty} \mathrm{E}\Big([\hat{\mathbf{p}}_N - \mathbf{p}][\hat{\mathbf{p}}_N - \mathbf{p}]^{\mathrm{T}}\Big) = \mathbf{0} \tag{1.86}$$

is satisfied.

Remark: Note that a consistent estimate says nothing about the quality of the estimate for finite N. It is therefore possible that a consistent estimate is biased for finite N.

To analyze the influence of a stochastic disturbance, it is assumed that the measured output variable y_k is composed of the undisturbed output $\bar{y}_k$ and a stochastic disturbance v_k in the form $y_k = \bar{y}_k + v_k$, see Figure 1.3. The undisturbed output is calculated according to (1.74) as

$$\bar{y}_k = \underbrace{\begin{bmatrix} -\bar{y}_{k-1} & -\bar{y}_{k-2} & \dots & -\bar{y}_{k-n} & u_{k-d} & u_{k-d-1} & \dots & u_{k-d-m} \end{bmatrix}}_{\bar{\mathbf{s}}_k^{\mathrm{T}}} \underbrace{\begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b_0 \\ \vdots \\ b_m \end{bmatrix}}_{\mathbf{p}} . \tag{1.87}$$

Replacing the entries $\bar{y}_j$ in (1.87) by $\bar{y}_j = y_j - v_j$, we obtain

$$\bar{y}_k = \mathbf{s}_k^{\mathrm{T}}\mathbf{p} + \mathbf{n}_k^{\mathrm{T}}\mathbf{p} \tag{1.88}$$

with

$$\mathbf{s}_k^{\mathrm{T}} = \begin{bmatrix} -y_{k-1} & -y_{k-2} & \dots & -y_{k-n} & u_{k-d} & u_{k-d-1} & \dots & u_{k-d-m} \end{bmatrix} \tag{1.89a}$$

$$\mathbf{n}_k^{\mathrm{T}} = \begin{bmatrix} v_{k-1} & v_{k-2} & \dots & v_{k-n} & 0 & 0 & \dots & 0 \end{bmatrix} . \tag{1.89b}$$

By combining $k = 0, \dots, N$ measurements and using $y_k = \bar{y}_k + v_k$, the following formulation can be found:

$$\underbrace{\begin{bmatrix} y_0 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}_N} = \underbrace{\begin{bmatrix} \mathbf{s}_0^{\mathrm{T}} \\ \vdots \\ \mathbf{s}_N^{\mathrm{T}} \end{bmatrix}}_{\mathbf{S}_N} \mathbf{p} + \underbrace{\begin{bmatrix} \mathbf{n}_0^{\mathrm{T}} \\ \vdots \\ \mathbf{n}_N^{\mathrm{T}} \end{bmatrix}}_{\mathbf{N}_N} \mathbf{p} + \underbrace{\begin{bmatrix} v_0 \\ \vdots \\ v_N \end{bmatrix}}_{\mathbf{v}_N} . \tag{1.90}$$

If we now perform a least-squares identification based on the known part $\mathbf{y}_N = \mathbf{S}_N\mathbf{p}$, we obtain

$$\hat{\mathbf{p}} = \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1}\mathbf{S}_N^{\mathrm{T}}\mathbf{y} = \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1}\mathbf{S}_N^{\mathrm{T}}(\mathbf{S}_N\mathbf{p} + \mathbf{N}_N\mathbf{p} + \mathbf{v}_N) \tag{1.91}$$

or, after a short calculation,

$$\hat{\mathbf{p}} = \mathbf{p} + \left(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\right)^{-1}\mathbf{S}_N^{\mathrm{T}}(\mathbf{N}_N\mathbf{p} + \mathbf{v}_N) \ . \tag{1.92}$$

The expected value of the estimation error is thus

$$\mathrm{E}(\hat{\mathbf{p}}) - \mathbf{p} = \mathrm{E}\left(\left(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\right)^{-1}\mathbf{S}_N^{\mathrm{T}}(\mathbf{N}_N\mathbf{p} + \mathbf{v}_N)\right) = \mathbf{b}, \tag{1.93}$$

with the bias $\mathbf{b}$. The estimate $\hat{\mathbf{p}}$ of $\mathbf{p}$ is unbiased if and only if

$$\mathbf{b} = \mathbf{0} \tag{1.94}$$

Furthermore, the estimate is obviously consistent if

$$\lim_{N\to\infty} \mathbf{b} = \mathbf{0} \tag{1.95}$$

is satisfied. The following theorem now provides information on the requirements that must be placed on the stochastic disturbance v_k and the model structure for these two conditions to be satisfied.

Theorem 1.2 (Unbiased and Consistent Least-Squares Identification)**.** *The least-squares identification of the parameter* $\mathbf{p}$ *for the identification problem according to* (1.90) *or according to Figure 1.3 is unbiased and consistent if the stochastic disturbance* v_k *satisfies the* Yule-Walker equation *of an autoregressive signal process of the form*

$$v_k + a_1 v_{k-1} + a_2 v_{k-2} + \ldots + a_n v_{k-n} = w_k, \tag{1.96}$$

with zero-mean white noise w_k *and the coefficients* a_j, $j = 1, \ldots, n$, *of the denominator of the transfer function to be identified.*

That is, the disturbance signal v_k must have been generated by filtering white noise w_k through a filter with the transfer function $1/A(\delta)$. This structure corresponds exactly to the model structure of an ARX model defined in Section 1.3.1, see Figure 1.10. It can also be shown that in this case the least-squares identification is consistent in the mean square. For a proof of Theorem 1.2, the reader is referred to the literature, in particular [1.3].

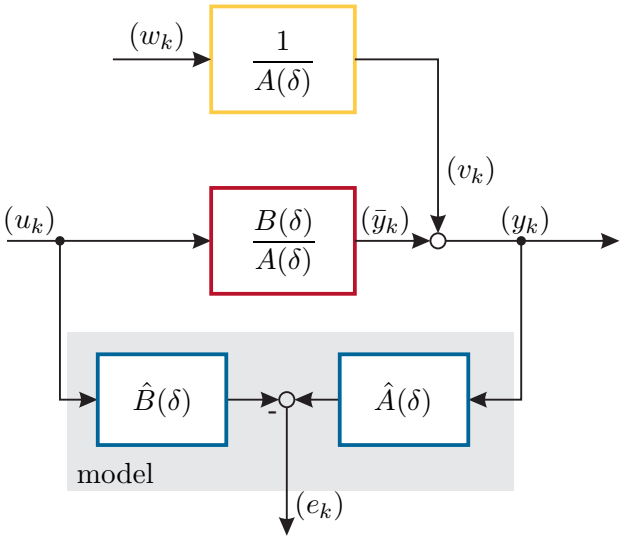


Figure 1.10: On least-squares identification with stochastic disturbance.

1.3.4 Recursive Least-Squares (RLS) Identification

The parameter estimation (1.79) is not suitable for online operation due to the ever-increasing dimensions of the measurement vector $\mathbf{y}_N$ and the matrix $\mathbf{S}_N$. In the following, a *recursive method* is given based on (1.79), which improves the estimate $\hat{\mathbf{p}}$ of the parameter vector $\mathbf{p}$ with each new measurement. According to (1.79), the optimal estimate for $N+1$ measurements is

$$\hat{\mathbf{p}}_N = \left(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\right)^{-1}\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N \tag{1.97}$$

and for $N+2$ measurements

$$\hat{\mathbf{p}}_{N+1} = \left(\mathbf{S}_{N+1}^{\mathrm{T}}\mathbf{S}_{N+1}\right)^{-1}\mathbf{S}_{N+1}^{\mathrm{T}}\mathbf{y}_{N+1}\ . \tag{1.98}$$

Partitioning the data matrix $\mathbf{S}_{N+1}$ and the measurement vector $\mathbf{y}_{N+1}$ in the form

$$\mathbf{S}_{N+1} = \begin{bmatrix}\mathbf{S}_N \\ \mathbf{s}_{N+1}^{\mathrm{T}}\end{bmatrix} \quad \text{and} \quad \mathbf{y}_{N+1} = \begin{bmatrix}\mathbf{y}_N \\ y_{N+1}\end{bmatrix}, \tag{1.99}$$

then $\hat{\mathbf{p}}_{N+1}$ is

$$\begin{aligned}\hat{\mathbf{p}}_{N+1} &= \left(\begin{bmatrix}\mathbf{S}_N^{\mathrm{T}} & \mathbf{s}_{N+1}\end{bmatrix}\begin{bmatrix}\mathbf{S}_N \\ \mathbf{s}_{N+1}^{\mathrm{T}}\end{bmatrix}\right)^{-1}\left(\begin{bmatrix}\mathbf{S}_N^{\mathrm{T}} & \mathbf{s}_{N+1}\end{bmatrix}\begin{bmatrix}\mathbf{y}_N \\ y_{N+1}\end{bmatrix}\right) \\ &= \left(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N + \mathbf{s}_{N+1}\mathbf{s}_{N+1}^{\mathrm{T}}\right)^{-1}\left(\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N + \mathbf{s}_{N+1}y_{N+1}\right)\ .\end{aligned} \tag{1.100}$$

Exercise 1.14. Show the validity of (1.100).

For further calculation, the following auxiliary theorem on *matrix inversion* is required:

Theorem 1.3 (On Matrix Inversion). *If* $\mathbf{A}$, $\mathbf{C}$, *and* $(\mathbf{A}+\mathbf{BCD})$ *are non-singular square matrices, then*

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\Big(\mathbf{C}^{-1}+\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\Big)^{-1}\mathbf{D}\mathbf{A}^{-1} \ . \tag{1.101}$$

Applying Theorem 1.3 to $\mathbf{F} = \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N + \mathbf{s}_{N+1}\mathbf{s}_{N+1}^{\mathrm{T}}\Big)$ with $\mathbf{A} = \mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N$, $\mathbf{B} = \mathbf{s}_{N+1}$, $\mathbf{C} = 1$, and $\mathbf{D} = \mathbf{s}_{N+1}^{\mathrm{T}}$, we obtain

$$\mathbf{F}^{-1} = \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1} - \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1}\mathbf{s}_{N+1}\Big(1+\mathbf{s}_{N+1}^{\mathrm{T}}\Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1}\mathbf{s}_{N+1}\Big)^{-1}\mathbf{s}_{N+1}^{\mathrm{T}}\Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1} \ . \tag{1.102}$$

With the abbreviations

$$\mathbf{P}_N = \Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N\Big)^{-1} \tag{1.103a}$$

$$\mathbf{P}_{N+1} = \Big(\mathbf{S}_{N+1}^{\mathrm{T}}\mathbf{S}_{N+1}\Big)^{-1} \tag{1.103b}$$

$$\mathbf{k}_{N+1} = \frac{\mathbf{P}_N\mathbf{s}_{N+1}}{\Big(1+\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N\mathbf{s}_{N+1}\Big)} \tag{1.103c}$$

according to (1.102) we have

$$\mathbf{P}_{N+1} = \mathbf{P}_N - \mathbf{k}_{N+1}\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N \tag{1.104}$$

and

$$\mathbf{P}_{N+1}\mathbf{s}_{N+1} = \frac{\mathbf{P}_N\mathbf{s}_{N+1}\Big(1+\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N\mathbf{s}_{N+1}\Big) - \mathbf{P}_N\mathbf{s}_{N+1}\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N\mathbf{s}_{N+1}}{\Big(1+\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N\mathbf{s}_{N+1}\Big)} = \mathbf{k}_{N+1} \ . \tag{1.105}$$

Substituting the relationships (1.103)–(1.105) into (1.100), we obtain

$$\hat{\mathbf{p}}_{N+1} = \mathbf{P}_{N+1}\Big(\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N + \mathbf{s}_{N+1}y_{N+1}\Big) = \mathbf{P}_N\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N - \mathbf{k}_{N+1}\mathbf{s}_{N+1}^{\mathrm{T}}\mathbf{P}_N\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N + \mathbf{k}_{N+1}y_{N+1} \tag{1.106}$$

or with (1.97)

$$\hat{\mathbf{p}}_{N+1} = \hat{\mathbf{p}}_N + \mathbf{k}_{N+1}\Big(y_{N+1} - \mathbf{s}_{N+1}^{\mathrm{T}}\hat{\mathbf{p}}_N\Big) \ . \tag{1.107}$$

The recursive least-squares identification can thus be summarized as follows:

(1) Suitable initial values $\hat{\mathbf{p}}_{-1}$ and $\mathbf{P}_{-1}$ are chosen (see discussion below).

(2) For the sampling times $j = 0, 1, \ldots$, y_j is measured and the data vector is set up according to (1.74):

$$\mathbf{s}_j^{\mathrm{T}} = \begin{bmatrix} -y_{j-1} & -y_{j-2} & \ldots & -y_{j-n} & u_{j-d} & u_{j-d-1} & \ldots & u_{j-d-m} \end{bmatrix} \tag{1.108}$$

The parameter vector $\mathbf{p}$ can then be estimated online using $j+1$ measurements with the iteration rule

$$\mathbf{k}_j = \frac{\mathbf{P}_{j-1}\mathbf{s}_j}{\left(1 + \mathbf{s}_j^{\mathrm{T}}\mathbf{P}_{j-1}\mathbf{s}_j\right)} \tag{1.109a}$$

$$\mathbf{P}_j = \mathbf{P}_{j-1} - \mathbf{k}_j\mathbf{s}_j^{\mathrm{T}}\mathbf{P}_{j-1} \tag{1.109b}$$

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \mathbf{k}_j\left(y_j - \mathbf{s}_j^{\mathrm{T}}\hat{\mathbf{p}}_{j-1}\right) \tag{1.109c}$$

In the last step, the question of *suitable initial values* $\hat{\mathbf{p}}_{-1}$ and $\mathbf{P}_{-1}$ must still be clarified. To this end, the iteration rule for $\mathbf{P}_k^{-1}$ (see (1.100) and (1.103)) is first considered:

$$\mathbf{P}_{k+1}^{-1} = \mathbf{P}_k^{-1} + \mathbf{s}_{k+1}\mathbf{s}_{k+1}^{\mathrm{T}} \tag{1.110}$$

and this is iterated for $k = -1, 0, 1, \ldots, N, \ldots$

$$\mathbf{P}_0^{-1} = \mathbf{P}_{-1}^{-1} + \mathbf{s}_0\mathbf{s}_0^{\mathrm{T}} \tag{1.111a}$$

$$\mathbf{P}_1^{-1} = \mathbf{P}_{-1}^{-1} + \mathbf{s}_0\mathbf{s}_0^{\mathrm{T}} + \mathbf{s}_1\mathbf{s}_1^{\mathrm{T}} \tag{1.111b}$$

$$\vdots \quad \vdots$$

$$\mathbf{P}_N^{-1} = \mathbf{P}_{-1}^{-1} + \sum_{j=0}^{N}\mathbf{s}_j\mathbf{s}_j^{\mathrm{T}} = \mathbf{P}_{-1}^{-1} + \mathbf{S}_N^{\mathrm{T}}\mathbf{S}_N \ . \tag{1.111c}$$

It can be seen that with the choice

$$\mathbf{P}_{-1} = \alpha\mathbf{E} \tag{1.112}$$

for large values of α, $\lim_{\alpha\to\infty}\mathbf{P}_{-1}^{-1} = \mathbf{0}$ holds, and thus (1.111) agrees with the corresponding expression for the non-recursive estimate from (1.79). For large values of α in (1.112), $\mathbf{P}_{-1}^{-1}$ therefore has a negligible influence on the recursively calculated $\mathbf{P}_N$. The iteration rule for $\hat{\mathbf{p}}_k$ is, using (1.105) and (1.107),

$$\hat{\mathbf{p}}_{k+1} = \hat{\mathbf{p}}_k + \mathbf{P}_{k+1}\mathbf{s}_{k+1}\left(y_{k+1} - \mathbf{s}_{k+1}^{\mathrm{T}}\hat{\mathbf{p}}_k\right) \ . \tag{1.113}$$

If this iteration is now carried out for $k = -1, 0, 1, \ldots, N, \ldots$, then using (1.111) we obtain

$$\hat{\mathbf{p}}_0 = \hat{\mathbf{p}}_{-1} + \mathbf{P}_0\mathbf{s}_0\left(y_0 - \mathbf{s}_0^{\mathrm{T}}\hat{\mathbf{p}}_{-1}\right) = \mathbf{P}_0\left(\mathbf{s}_0 y_0 + \underbrace{\left(\mathbf{P}_0^{-1} - \mathbf{s}_0\mathbf{s}_0^{\mathrm{T}}\right)}_{\mathbf{P}_{-1}^{-1}}\hat{\mathbf{p}}_{-1}\right) \tag{1.114a}$$

$$\hat{\mathbf{p}}_1 = \mathbf{P}_1\left(\mathbf{s}_1 y_1 + \mathbf{P}_0^{-1}\hat{\mathbf{p}}_0\right) = \mathbf{P}_1\left(\mathbf{s}_1 y_1 + \mathbf{s}_0 y_0 + \mathbf{P}_{-1}^{-1}\hat{\mathbf{p}}_{-1}\right) \tag{1.114b}$$

$$\vdots \quad \vdots$$

$$\hat{\mathbf{p}}_N = \mathbf{P}_N\left(\sum_{j=0}^{N}\mathbf{s}_j y_j + \mathbf{P}_{-1}^{-1}\hat{\mathbf{p}}_{-1}\right) = \mathbf{P}_N\left(\mathbf{S}_N^{\mathrm{T}}\mathbf{y}_N + \mathbf{P}_{-1}^{-1}\hat{\mathbf{p}}_{-1}\right) \ . \tag{1.114c}$$

As can be seen from (1.114), the choice of $\mathbf{P}_{-1}^{-1}$ from (1.112) for large α implies that the initial value $\hat{\mathbf{p}}_{-1}$ can be chosen freely and yet $\hat{\mathbf{p}}_N$ from (1.114) coincides with the result of the non-recursive estimate from (1.79). For simplicity, $\hat{\mathbf{p}}_{-1} = \mathbf{0}$ is therefore often chosen.

1.3.5 Weighted Least Squares Method

In the weighted least squares method, we seek a solution to the overdetermined linear system of equations (see (1.60))

$$\mathbf{y}_N = \mathbf{S}_N \mathbf{p} \tag{1.115}$$

such that the *weighted quadratic error*

$$\sum_{j=0}^{N} \alpha_j e_j^2 \quad , \quad e_j = y_j - \mathbf{s}_j^{\mathrm{T}} \mathbf{p} \tag{1.116}$$

with the sequence of positive weighting coefficients α_j ($\alpha_j > 0$ for all j) is minimized with respect to $\mathbf{p}$. It can be seen that by choosing

$$\tilde{y}_j = \sqrt{\alpha_j} y_j \quad \text{and} \quad \tilde{\mathbf{s}}_j^{\mathrm{T}} = \sqrt{\alpha_j} \mathbf{s}_j^{\mathrm{T}} \tag{1.117}$$

the optimization problem (1.116) can be transformed into the classical least-squares problem according to (1.61)

$$\min_{\mathbf{p}} \sum_{j=0}^{N} \tilde{e}_j^2 = \min_{\mathbf{p}} \|\tilde{\mathbf{e}}_N\|_2^2 \quad \text{with} \quad \tilde{\mathbf{e}}_N = \tilde{\mathbf{y}}_N - \tilde{\mathbf{S}}_N \mathbf{p} \tag{1.118}$$

The corresponding recursive estimator according to (1.109) is therefore

$$\tilde{\mathbf{k}}_j = \frac{\tilde{\mathbf{P}}_{j-1} \tilde{\mathbf{s}}_j}{\left(1 + \tilde{\mathbf{s}}_j^{\mathrm{T}} \tilde{\mathbf{P}}_{j-1} \tilde{\mathbf{s}}_j\right)} \tag{1.119a}$$

$$\tilde{\mathbf{P}}_j = \tilde{\mathbf{P}}_{j-1} - \tilde{\mathbf{k}}_j \tilde{\mathbf{s}}_j^{\mathrm{T}} \tilde{\mathbf{P}}_{j-1} \tag{1.119b}$$

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \tilde{\mathbf{k}}_j \left(\tilde{y}_j - \tilde{\mathbf{s}}_j^{\mathrm{T}} \hat{\mathbf{p}}_{j-1}\right) . \tag{1.119c}$$

Substituting the relationship (1.117) and the transformation

$$\tilde{\mathbf{k}}_j = \frac{\mathbf{k}_j}{\sqrt{\alpha_j}} \quad \text{and} \quad \tilde{\mathbf{P}}_j = \mathbf{P}_j \tag{1.120}$$

into (1.119), we obtain

$$\mathbf{k}_j = \frac{\mathbf{P}_{j-1} \mathbf{s}_j}{\left(\frac{1}{\alpha_j} + \mathbf{s}_j^{\mathrm{T}} \mathbf{P}_{j-1} \mathbf{s}_j\right)} \tag{1.121a}$$

$$\mathbf{P}_j = \mathbf{P}_{j-1} - \mathbf{k}_j \mathbf{s}_j^{\mathrm{T}} \mathbf{P}_{j-1} \tag{1.121b}$$

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \mathbf{k}_j \left(y_j - \mathbf{s}_j^{\mathrm{T}} \hat{\mathbf{p}}_{j-1}\right) . \tag{1.121c}$$

A special choice for α_j is given by

$$\alpha_j = q^{N-j} \quad \text{with} \quad 0 < q \leq 1 \tag{1.122}$$

Substituting (1.122) into the cost function of (1.116)

$$\sum_{j=0}^{N} q^{N-j} e_j^2 = \|\mathbf{e}_N\|_Q^2 = \mathbf{e}_N^{\mathrm{T}} \mathbf{Q}_N \mathbf{e}_N \quad \text{with} \quad \mathbf{Q}_N = \begin{bmatrix} q^N & 0 & \cdots & 0 & 0 \\ 0 & q^{N-1} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & q & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{1.123}$$

it is evident that with this special choice of α_j, the equation errors are weighted less the further back they lie. This is also referred to as an *exponentially decaying memory* with the *memory factor* q. For this memory factor q, values in the range of $0.9 < q < 0.995$ have proven useful in practical application. Note that this method is also suitable for the identification of *slowly time-varying systems*, where the speed of the parameter changes of the system naturally determines the choice of q, i.e., the speed of forgetting old measurement values.

Exercise 1.15. In your opinion, where does the name exponentially decaying memory come from?

The recursive least-squares estimator with exponentially decaying memory follows directly from (1.121) with $\alpha_j = q^{N-j}$ and the transformation $\mathbf{P}_j \to \mathbf{P}_j / q^{N-j}$ to

$$\mathbf{k}_j = \frac{\mathbf{P}_{j-1}\mathbf{s}_j}{\left(q + \mathbf{s}_j^{\mathrm{T}} \mathbf{P}_{j-1} \mathbf{s}_j\right)} \tag{1.124a}$$

$$\mathbf{P}_j = \left(\mathbf{P}_{j-1} - \mathbf{k}_j \mathbf{s}_j^{\mathrm{T}} \mathbf{P}_{j-1}\right) \frac{1}{q} \tag{1.124b}$$

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \mathbf{k}_j \left(y_j - \mathbf{s}_j^{\mathrm{T}} \hat{\mathbf{p}}_{j-1}\right) . \tag{1.124c}$$

Exercise 1.16. Verify the correctness of (1.124).

Instead of the diagonal matrix $\mathbf{Q}_N$ in (1.123), any arbitrary positive definite *weighting matrix* $\mathbf{Q}_N > 0$ could be chosen. In this case, not the Euclidean norm of the error $\|\mathbf{e}_N\|_2^2$, but a weighted Euclidean norm of the form

$$\|\mathbf{e}_N\|_Q^2 = \mathbf{e}_N^{\mathrm{T}} \mathbf{Q}_N \mathbf{e}_N \tag{1.125}$$

is minimized. It is now easy to see that this problem

$$\min_{\mathbf{p}} \|\mathbf{e}_N\|_Q^2 \quad \text{with} \quad \mathbf{e}_N = \mathbf{y}_N - \mathbf{S}_N \mathbf{p} \tag{1.126}$$

by applying the projection theorem of Theorem 1.1 with the inner product

$$\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{z} \tag{1.127}$$

leads to the solution

$$\hat{\mathbf{p}}_N = \left(\mathbf{S}_N^{\mathrm{T}} \mathbf{Q}_N \mathbf{S}_N\right)^{-1} \mathbf{S}_N^{\mathrm{T}} \mathbf{Q}_N \mathbf{y}_N \tag{1.128}$$

Exercise 1.17. Show the validity of (1.128). Furthermore, check whether the error $\mathbf{e}_N = \mathbf{y}_N - \mathbf{S}_N\hat{\mathbf{p}}_N$ is orthogonal to the optimal solution (1.128) in the sense of the inner product (1.127).

Exercise 1.18. Show that every positive definite matrix $\mathbf{Q}$ has exactly one positive definite square root $\mathbf{W}$, i.e., $\mathbf{Q} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$.

Remark: (to Exercise 1.18) Use the fact that every symmetric matrix $\mathbf{Q}$ can be brought to diagonal form $\mathbf{D}$ by a similarity transformation, so

$$\mathbf{D} = \mathbf{T}\mathbf{Q}\mathbf{T}^{\mathrm{T}} \quad \text{with} \quad \mathbf{T}\mathbf{T}^{\mathrm{T}} = \mathbf{E} \; .$$

Exercise 1.19. Prove that (1.125) is a norm in the sense of Definition 1.2.

Exercise 1.20. Given is the continuous transfer function

$$G(s) = \frac{K}{(sT_1 + 1)(sT_2 + 1)}$$

with the parameters $K, T_1, T_2 > 0$. Determine a state-space representation for $G(s)$ and implement it in Matlab/Simulink for a time-varying parameter T_2 and constant parameters K and T_1. Subsequently, determine the corresponding z-transfer function for $G(s)$ for a sampling time of $T_a = 0.25\,\mathrm{s}$. Identify the coefficients a_j and b_j of the denominator and numerator polynomials of the discrete-time transfer function using the recursive least-squares algorithm. Choose a multiple 3-2-1 step as the input signal and for the parameters $K = 1$, $T_1 = 7.5\,\mathrm{s}$, and a step change of T_2 from $5\,\mathrm{s}$ to $2.5\,\mathrm{s}$. Investigate the influence of the forgetting factor q and compare the result for $q = 1$ with the result of the off-line least-squares procedure.

1.3.6 Least-Mean Squares (LMS) Identification

Another way to solve the system of equations (1.71)

$$\left(\sum_{j=0}^{N} \mathbf{s}_j \mathbf{s}_j^{\mathrm{T}}\right) \hat{\mathbf{p}}_N = \mathbf{S}_N^{\mathrm{T}} \mathbf{S}_N \hat{\mathbf{p}}_N = \mathbf{S}_N^{\mathrm{T}} \mathbf{y}_N = \sum_{j=0}^{N} \mathbf{s}_j y_j \tag{1.129}$$

online, is given by the so-called *Least-Mean Squares (LMS) algorithm*, also known as the *stochastic gradient method*. The parameter vector $\mathbf{p}$ is recursively estimated in the form

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \mu_j \mathbf{s}_j \underbrace{\left(y_j - \mathbf{s}_j^{\mathrm{T}} \hat{\mathbf{p}}_{j-1}\right)}_{e_j} \tag{1.130}$$

with a suitably chosen initial value $\hat{\mathbf{p}}_{-1}$ and the estimation error at the j-th time instant e_j. Using the *(time-varying) parameter* μ_j ($\mu_j \geq 0$ for all $j \geq 0$), it is possible to adjust the convergence speed and the sensitivity to noise to a certain extent. Typically, a constant value $\mu_j = \bar{\mu}$ is chosen for μ_j, whereby the parameter estimator (1.130) is able to react faster to changes in the parameter vector $\mathbf{p}$ (time-varying system) for larger values of $\bar{\mu}$ and is less sensitive to measurement noise for smaller values of $\bar{\mu}$. To make the convergence speed independent of the signal level of the entries in $\mathbf{s}_j$, it is common to choose the (time-varying) parameter μ_j in the LMS algorithm (1.130) in the form

$$\mu_j = \frac{\bar{\mu}}{\mathbf{s}_j^{\mathrm{T}} \mathbf{s}_j} \tag{1.131}$$

or

$$\mu_j = \frac{\bar{\mu}}{l_j} \quad \text{with} \quad l_{j+1} = l_j + \bar{\mu}\left(\mathbf{s}_j^{\mathrm{T}} \mathbf{s}_j - l_j\right) \tag{1.132}$$

with the initial value $l_{-1} = \mathbf{s}_{-1}^{\mathrm{T}} \mathbf{s}_{-1} > 0$. Without proof, it should be noted that the convergence of (1.130) can be guaranteed for sufficiently small $\bar{\mu}$.

Note 1.2. The LMS algorithm is very frequently used in applications for the adaptive filtering of signals (*adaptive noise cancellation*), such as echo compensation of transmitted signals. Consider the arrangement in Figure 1.11, where the measured signal sequence (d_k) consists of a useful signal component (r_k) to be identified and a non-measurable noise component $G_1(\delta)m_k$ in the form

$$d_k = G_1(\delta) m_k + r_k \tag{1.133}$$

with the noise signal (m_k) and the unknown transfer operator $G_1(\delta)$. Furthermore, it is assumed that the noise signal (m_k) and the useful signal (r_k) are uncorrelated. The noise signal (m_k) is not directly measured but can be indirectly captured via the signal (n_k) with

$$n_k = G_2(\delta) m_k \tag{1.134}$$

and the unknown transfer operator $G_2(\delta)$. If the two transfer operators $G_1(\delta)$ and $G_2(\delta)$ were known, then the useful signal could be reconstructed from knowledge of

(d_k) and (n_k) in the form

$$r_k = d_k - \underbrace{G_1(\delta)G_2^{-1}(\delta)}_{\tilde{G}(\delta)} n_k \tag{1.135}$$

Note that $\tilde{G}(\delta) = G_1(\delta)G_2^{-1}(\delta)$ is generally a non-causal transfer operator.

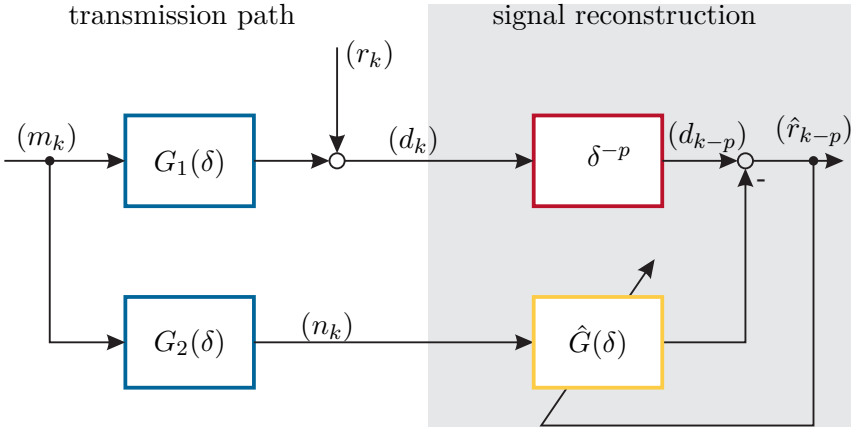


Figure 1.11: On adaptive signal filtering with the LMS algorithm.

Exercise 1.21. Under what conditions on $G_1(\delta)$ and $G_2(\delta)$ is $\tilde{G}(\delta)$ a causal transfer operator?

For this reason, a positive integer number $p > 0$ is chosen such that the transfer operator

$$G(\delta) = \tilde{G}(\delta)\delta^{-p} \tag{1.136}$$

is always causal. Replacing $\tilde{G}(\delta)$ in (1.135) by $G(\delta)$ according to (1.136), we obtain

$$r_k = d_k - G(\delta)\delta^p n_k \tag{1.137a}$$
$$\delta^{-p} r_k = \delta^{-p} d_k - G(\delta) n_k \tag{1.137b}$$
$$r_{k-p} = d_{k-p} - G(\delta) n_k \ . \tag{1.137c}$$

Since the transfer operator $G(\delta)$ in (1.137) is unknown, a transfer operator $\hat{G}(\delta)$ to be identified is substituted for $G(\delta)$. In most applications, an MA model (FIR filter) according to (1.47), (1.48) of the form

$$\hat{G}(\delta) = \hat{C}(\delta) = \hat{c}_0 + \hat{c}_1\delta^{-1} + \ldots + \hat{c}_q\delta^{-q} \tag{1.138}$$

is used in combination with the LMS algorithm. Of course, the transfer operator $G(\delta)$ is only describable by an MA model in exceptional cases, which is why a very high model order q of the MA model is required for good reconstruction of the useful signal (r_k). Substituting the expression $\hat{G}(\delta)$ from (1.138) for $G(\delta)$ in (1.137), the

estimate of the useful signal $\hat{r}_{k-p}$ is calculated as

$$\hat{r}_{k-p} = d_{k-p} - \underbrace{\begin{bmatrix} n_k & n_{k-1} & n_{k-2} & \dots & n_{k-q} \end{bmatrix}}_{\mathbf{s}_k^{\mathrm{T}}} \underbrace{\begin{bmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_q \end{bmatrix}}_{\hat{\mathbf{p}}} \tag{1.139}$$

with the estimate $\hat{\mathbf{p}}$ of the parameter vector and the data vector $\mathbf{s}_k^{\mathrm{T}}$. Applying the LMS algorithm (1.130), the parameter estimator becomes

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{j-1} + \mu_j \mathbf{s}_j \underbrace{\left(d_{j-p} - \mathbf{s}_j^{\mathrm{T}} \hat{\mathbf{p}}_{j-1}\right)}_{\hat{r}_{j-p}} \tag{1.140}$$

with the parameter μ_j according to (1.131) or (1.132) and a suitably chosen initial value $\hat{\mathbf{p}}_{-1}$. Figure 1.11 provides a graphical illustration of the algorithm.

Exercise 1.22. Given is a system according to Figure 1.11 with the measured signal sequence (d_k), in which a periodic useful signal $(s_k) = (20\sin(3t - \pi/4))$ is hidden. The noise signal (m_k) is white noise, the sampling time is 0.1 s, and the two transfer operators are

$$G_1(\delta) = 10\frac{\delta + 2}{0.8 + \delta + \delta^2} \quad \text{and} \quad G_2(\delta) = \frac{1 - 2\delta}{(\delta + 0.2)^2(\delta - 0.8)} \ .$$

Design an online algorithm according to the LMS method for extracting the useful signal using an MA model. Perform these calculations in Matlab and also check the results in the frequency domain using the FFT. Specify different orders of the MA models and change the signal delay p according to (1.136). How do these specifications and changes affect the result?

1.4 References

[1.1] L. Ljung, *System Identification*. New Jersey, USA: Prentice Hall, 1999.

[1.2] A. Kugi, "Regelungssysteme," in *Skriptum zur Vorlesung, Institut für Automatisierungstechnik, TU Wien*, vol. 1, Vienna, Austria, Sep. 2011.

[1.3] R. Isermann, *Identifikation dynamischer Systeme 1 und 2*, 2nd ed. Berlin, Deutschland: Springer, 1992.

[1.4] G. Franklin, J. Powell, and M. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, USA: Addison–Weseley, 1998.

[1.5] M. Gevers and G. Li, *Parametrization in Control, Estimation and Filtering Problems*. London, UK: Springer, 1993.

[1.6] E. Kreyszig, *Statistische Methoden und ihre Anwendungen*, 7th ed. Göttingen, Deutschland: Vandenhoeck & Ruprecht, 1998.

[1.7] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, USA: MIT Press, 1987.

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

[1.9] O. Nelles, *Nonlinear System Identification*. Berlin, Deutschland: Springer, 2001.

[1.10] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, USA: McGraw-Hill, 2002.

2 Optimal Estimators

This chapter addresses the question of how the state $\mathbf{x}_{k+m}$ of a dynamic system can be estimated from an input sequence (u_k) and a measured output sequence (y_k). Depending on the value of m, the estimation process is referred to as

(1) Smoothing for $m < 0$

(2) Filtering for $m = 0$, or

(3) Prediction for $m > 0$

Figure 2.1 provides a graphical illustration of these three cases.

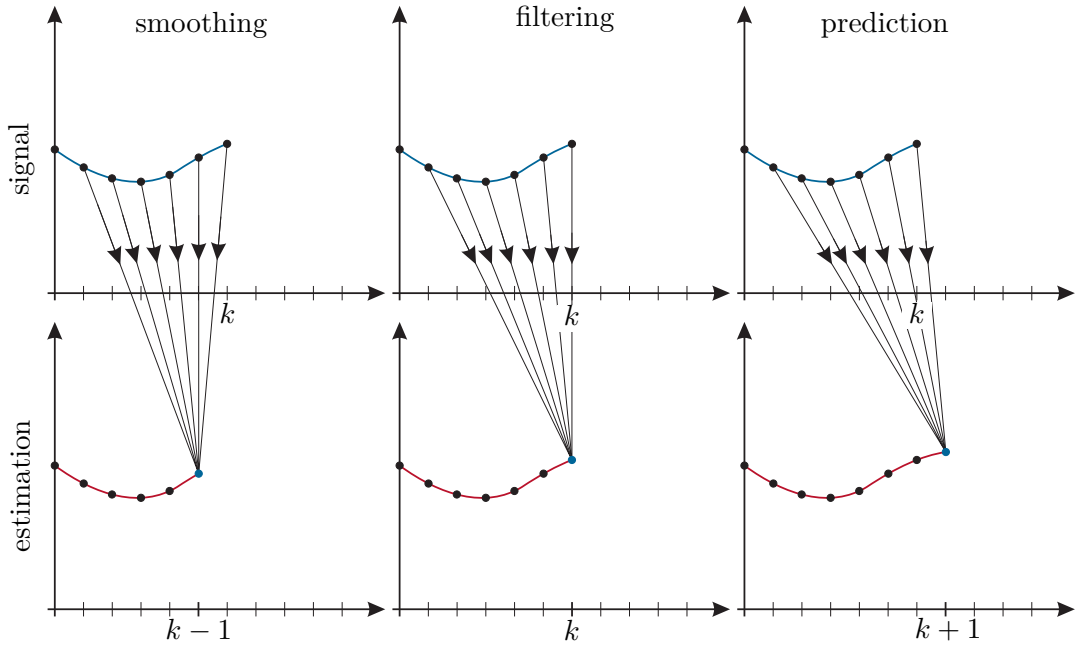


Figure 2.1: Concerning the terms smoothing, filtering, and prediction.

Further considerations will be limited to the last case (3), as this is the most interesting for control engineering applications. As a result of the following considerations, an optimal state observer, the so-called *Kalman filter*, will be determined, which minimizes a quadratic performance criterion. However, this requires an intermediate step of extending the results of the least-squares estimation from the previous chapter. Since the expected value and the covariance of random numbers are used repeatedly in this chapter, these two concepts will be explained for normally distributed and uniformly distributed random numbers.

Note 2.1 (Normally Distributed Random Variables). A scalar normally distributed random variable x is defined by the probability density function (Gaussian distribution)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}, \tag{2.1}$$

with the mean (expected value) m and the variance σ^2. The mean (expected value, first moment) and the variance (second central moment) are, as explained in Appendix A, given by

$$\mathrm{E}(x) = m = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \tag{2.2a}$$

$$\mathrm{E}((x - \mathrm{E}(x))^2) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mathrm{E}(x))^2 f(x)\,\mathrm{d}x \tag{2.2b}$$

Figure 2.2 shows the probability density function for different parameterizations of a normally distributed random number.

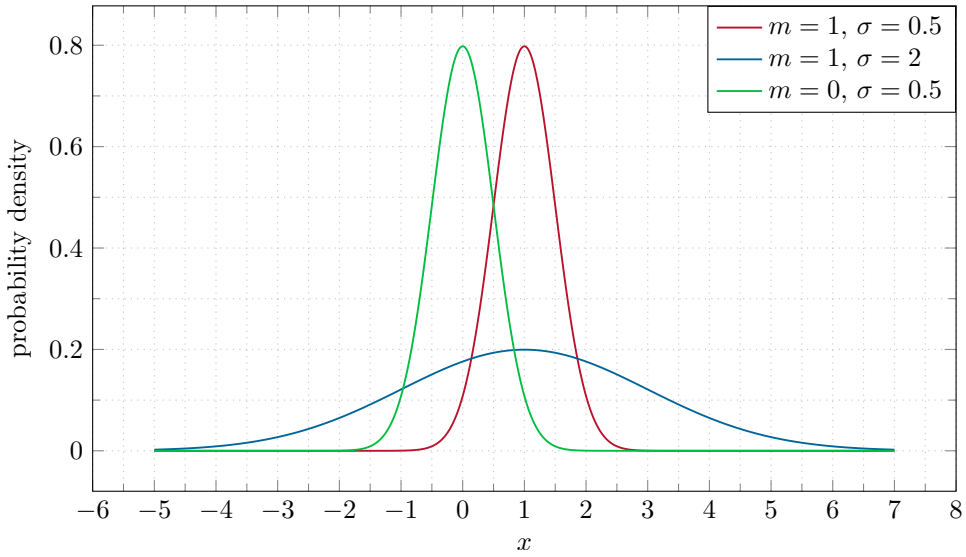


Figure 2.2: Probability density function for different normally distributed random variables.

The probability that a random number x lies in the interval δ around the expected value $\mathrm{E}(x) = m$ is calculated as

$$\mathrm{P}(m-\delta < x \leq m+\delta) = \int_{m-\delta}^{m+\delta} f(x)\,\mathrm{d}x = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\sigma}\right), \tag{2.3}$$

cf. Exercise A.1 in Appendix A. For example, if $\delta = \sigma$, a random number x lies in this interval with a probability of 0.68. Furthermore, a random number x lies in the interval $\delta = 2\sigma$ with a probability of 0.95. The variance thus represents a measure of the scatter of the random numbers around the expected value. The joint probability density function $f(\mathbf{x})$ of an n-dimensional normally distributed random vector $\mathbf{x}$ is calculated in the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbf{Q})}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}-\mathbf{m})}, \tag{2.4}$$

with the expected value $\mathbf{m}$ and the covariance matrix $\mathbf{Q}$,

$$\mathbf{m} = \mathrm{E}(\mathbf{x}) \tag{2.5a}$$

$$\mathbf{Q} = \mathrm{E}((\mathbf{x}-\mathrm{E}(\mathbf{x}))(\mathbf{x}-\mathrm{E}(\mathbf{x}))^{\mathrm{T}})\ . \tag{2.5b}$$

For the special case $n = 2$, the probability that a random vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ lies in an ellipse of the form

$$(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}-\mathbf{m}) = C^2 \tag{2.6}$$

is given by

$$\mathrm{P} = 1 - \mathrm{e}^{-\frac{C^2}{2}} \tag{2.7}$$

The proof of this statement can be found, for example, in [2.1].

Thus, with the help of the covariance matrix $\mathbf{Q}$, the region (i.e., the ellipses for $n = 2$) can be determined in which a random vector $\mathbf{x}$ lies with a probability P. Figure 2.3 shows the distribution of 3000 random vectors $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^{\mathrm{T}}$, with the normally distributed, uncorrelated random variables x_1 ($\mathrm{E}(x_1) = m_1 = 1$, $\sigma_1 = 2$) and x_2 ($\mathrm{E}(x_2) = m_2 = 2$, $\sigma_2 = 1$). Furthermore, the ellipses for $\mathrm{P} = 0.5$ and $\mathrm{P} = 0.95$, and the expected value $\mathrm{E}(\mathbf{x}) = \mathbf{m}$ are shown.

This consideration for $n = 2$ can be generalized to n-dimensional normally distributed random variables. In this case, the n-dimensional ellipsoid $(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{x}-\mathbf{m}) = C^2$ describes a measure of the distribution of the random vector. The probability P that a random vector $\mathbf{x}$ lies in this ellipsoid is given by

$$1 - \mathrm{P} = \frac{n}{2^{n/2}\Gamma\left(\frac{n}{2}+1\right)} \int_C^{\infty} \xi^{n-1}\mathrm{e}^{-\frac{\xi^2}{2}}\,\mathrm{d}\xi, \tag{2.8}$$

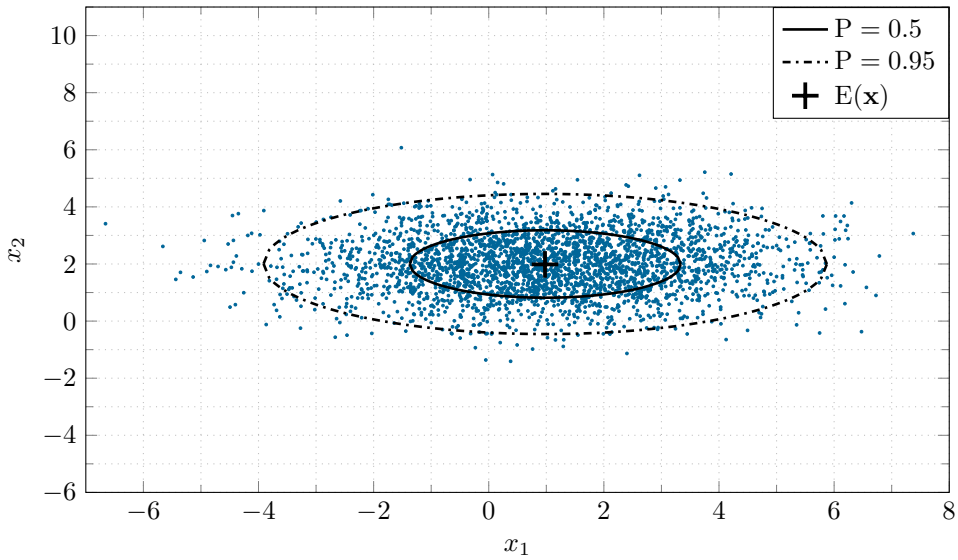


Figure 2.3: Graphical representation of the expected value $\mathrm{E}(\mathbf{x})$ and the covariance matrix $\mathbf{Q}$ for normally distributed random vectors $\mathbf{x}$.

with the Gamma function Γ, see [2.1].

An exact relationship between the ellipsoids defined by the covariance matrix $\mathbf{Q}$ and the probability that a random vector $\mathbf{x}$ lies in this region is only defined for normally distributed random variables. For other distributions, these ellipsoids are only a more or less accurate approximation. Nevertheless, the covariance matrix $\mathbf{Q}$ is also a meaningful measure for estimating the distribution of the random vectors for these distributions, which is why the corresponding ellipses or ellipsoids are often displayed here as well.

Note 2.2. A scalar random variable x uniformly distributed in the interval $[a, b]$ is defined by the probability density function

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

The corresponding expected value m and variance σ^2 are calculated as

$$\mathrm{E}(x) = m = \frac{a+b}{2} \tag{2.10a}$$

$$\mathrm{E}((x - \mathrm{E}(x))^2) = \sigma^2 = \frac{1}{12}(b-a)^2 \ . \tag{2.10b}$$

Figure 2.4 shows the probability density functions for different uniformly distributed random variables.

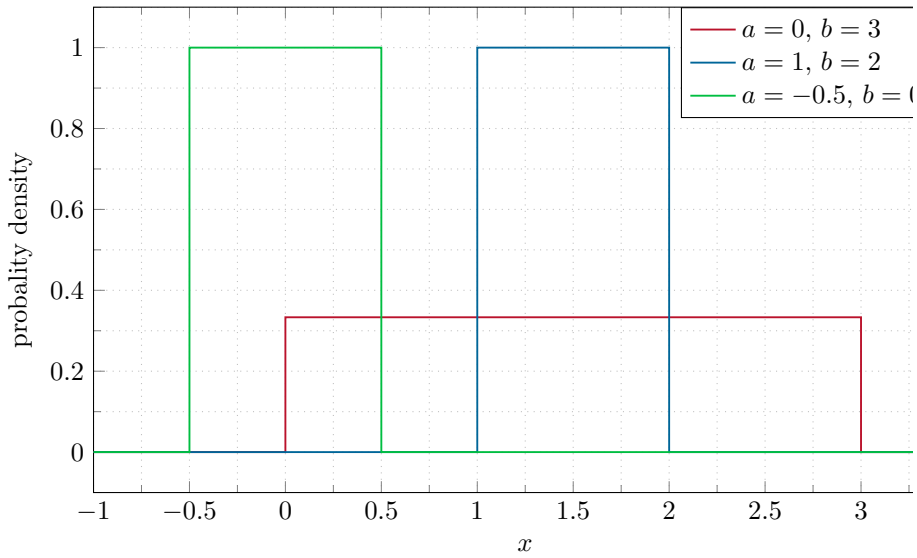


Figure 2.4: Probability density function for different uniformly distributed random variables.

2.1 Gauss-Markov Estimation

Consider the overdetermined linear equation system of (1.60) (compare also (1.87)-(1.90)) augmented by the stochastic disturbance $\mathbf{v}$

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{2.11}$$

with the known $(m \times n)$-matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, the n-dimensional vector of unknowns $\mathbf{p} \in \mathbb{R}^n$, and the m-dimensional measurement vector $\mathbf{y} \in \mathbb{R}^m$. We now assume that the stochastic disturbance $\mathbf{v}$ has the following stochastic properties:

$$\mathrm{E}(\mathbf{v}) = \mathbf{0} \quad \text{and} \quad \mathrm{cov}(\mathbf{v}) = \mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \mathbf{Q} \quad \text{with} \quad \mathbf{Q} > 0\ . \tag{2.12}$$

It should be noted here that $\mathbf{v}$ can also be interpreted as a measurement error. We are now looking for a *linear estimator* of the form

$$\hat{\mathbf{p}} = \mathbf{K}\mathbf{y} \tag{2.13}$$

with a constant $(n \times m)$-matrix $\mathbf{K} \in \mathbb{R}^{n \times m}$. Since $\mathbf{y}$ is the sum of a constant vector $\mathbf{S}\mathbf{p}$ and a stochastic vector (vector with stochastic entries) $\mathbf{v}$, $\mathbf{y}$, the estimated parameter $\hat{\mathbf{p}}$ according to (2.13), and the parameter error $\mathbf{e} = \mathbf{p} - \hat{\mathbf{p}}$ are stochastic quantities. Therefore, it does not make sense to determine the matrix $\mathbf{K}$ such that $\|\mathbf{e}\|_2^2$ is minimized, but the task

$$\min_{\mathbf{K}} \mathrm{E}(\|\mathbf{e}\|_2^2) = \min_{\mathbf{K}} \mathrm{E}((\mathbf{p} - \mathbf{K}\mathbf{y})^{\mathrm{T}}(\mathbf{p} - \mathbf{K}\mathbf{y})) \tag{2.14}$$

must be solved.

If we now substitute the relationship (2.11) into (2.14), then we obtain, taking into

account (2.12) and the results from Appendix A (see Exercise A.3):

$$\min_{\mathbf{K}} \mathrm{E}((\mathbf{p}-\mathbf{KSp}-\mathbf{Kv})^{\mathrm{T}}(\mathbf{p}-\mathbf{KSp}-\mathbf{Kv})) =$$

$$\min_{\mathbf{K}} \left\{ \mathrm{E}((\mathbf{p}-\mathbf{KSp})^{\mathrm{T}}(\mathbf{p}-\mathbf{KSp})) - 2\underbrace{\mathrm{E}((\mathbf{p}-\mathbf{KSp})^{\mathrm{T}}\mathbf{Kv})}_{=0} + \underbrace{\mathrm{E}(\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\mathbf{Kv})}_{=\mathrm{tr}(\mathrm{E}(\mathbf{Kvv}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}))} \right\} =$$

$$\min_{\mathbf{K}} \left\{ \|\mathbf{p}-\mathbf{KSp}\|_2^2 + \mathrm{tr}(\mathbf{KQK}^{\mathrm{T}}) \right\} . \tag{2.15}$$

The matrix $\mathbf{K}$ that minimizes the expression (2.15) is obviously a function of the unknown parameter vector $\mathbf{p}$. Therefore, the solution of this minimization problem is also not suitable for providing an estimator for $\mathbf{p}$ according to (2.13). To circumvent this problem, a substitute problem is solved in the following: It can be seen that with the choice

$$\mathbf{KS} = \mathbf{E} \tag{2.16}$$

and $\mathbf{E}$ as the identity matrix, the solution of the minimization problem (2.15) is independent of the parameter $\mathbf{p}$. This *constraint* (2.16) may seem arbitrary at first glance, but if we calculate the expected value of $\hat{\mathbf{p}}$ of the linear estimator (2.13), then it follows that

$$\mathrm{E}(\hat{\mathbf{p}}) = \mathrm{E}(\mathbf{Ky}) = \mathrm{E}(\mathbf{KSp}+\mathbf{Kv}) = \mathbf{KS}\underbrace{\mathrm{E}(\mathbf{p})}_{=\mathbf{p}} + \mathbf{K}\underbrace{\mathrm{E}(\mathbf{v})}_{=\mathbf{0}} . \tag{2.17}$$

I.e., the constraint (2.16) implies that $\mathrm{E}(\hat{\mathbf{p}}) = \mathbf{p}$ and thus the estimator is *unbiased*.

The problem

$$\min_{\mathbf{K}} \left\{ \mathrm{tr}(\mathbf{KQK}^{\mathrm{T}}) \right\} \quad \text{subject to} \quad \mathbf{KS} = \mathbf{E} \tag{2.18}$$

is now equivalent to solving n separate optimization problems of the form

$$\min_{\mathbf{k}_j} \mathbf{k}_j^{\mathrm{T}}\mathbf{Q}\mathbf{k}_j \quad \text{subject to} \quad \mathbf{k}_j^{\mathrm{T}}\mathbf{S} = \mathbf{e}_j^{\mathrm{T}} \quad \text{for} \quad j = 1,\ldots,n . \tag{2.19}$$

To show this, write the matrix $\mathbf{K}$ and the identity matrix $\mathbf{E}$ in the form

$$\mathbf{K} = \begin{bmatrix} \mathbf{k}_1^{\mathrm{T}} \\ \mathbf{k}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{k}_n^{\mathrm{T}} \end{bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{bmatrix} \mathbf{e}_1^{\mathrm{T}} \\ \mathbf{e}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{e}_n^{\mathrm{T}} \end{bmatrix} \tag{2.20}$$

and substitute this into (2.18):

$$\min_{\mathbf{K}} \left\{ \mathrm{tr}(\mathbf{KQK}^{\mathrm{T}}) \right\} = \min_{\mathbf{K}} \left\{ \sum_{j=1}^{n} \mathbf{k}_j^{\mathrm{T}}\mathbf{Q}\mathbf{k}_j \right\} \quad \text{subject to} \quad \mathbf{k}_j^{\mathrm{T}}\mathbf{S} = \mathbf{e}_j^{\mathrm{T}} \tag{2.21}$$

for $j = 1,\ldots,n$. Since the j-th summand in (2.21) only depends on $\mathbf{k}_j$, the minimization problem of (2.21) can be replaced by n minimization problems according to (2.19).

For the following, the solution of a problem of type (2.19) for a fixed j is therefore of interest. If we now consider the Hilbert space $\mathcal{H} = \mathbb{R}^m$ with the inner product $\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{z} = \sum_{j=1}^{m} x_j z_j$, then we see that the column vectors $\mathbf{k}_k$, $k = 1, \ldots, n$, of the matrix $\mathbf{K}$, i.e., $\operatorname{span}\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$, which satisfy the constraints of (2.19), do not form a subspace of the Hilbert space $\mathcal{H}$ and thus the projection theorem of Theorem 1.1 is not directly applicable.

> *Exercise* 2.1. Show that the sum $\mathbf{k}_j + \mathbf{k}_k$ does not satisfy the constraint of (2.21), even if $\mathbf{k}_j$ and $\mathbf{k}_k$ individually satisfy this constraint.

2.1.1 Quadratic Minimization with Affine Constraints

To solve the above problem, an extension of the projection theorem of Theorem 1.1 is required:

Theorem 2.1 (Extension of the Projection Theorem)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a closed subspace of $\mathcal{H}$. The translational shift of $\mathcal{U}$ in the form $\mathcal{A} = \mathbf{x} + \mathcal{U}$ for a fixed $\mathbf{x} \in \mathcal{H}$ is called a* linear variety *or* affine subspace*. Then there exists a unique vector $\mathbf{x}_0 \in \mathcal{A}$ of minimal norm, and this is orthogonal to $\mathcal{U}$ (see Figure 2.5).*

> *Proof of Theorem 2.1.* Shift the affine subspace $\mathcal{A}$ by $-\mathbf{x}$ so that it becomes a closed subspace and then apply the projection theorem of Theorem 1.1. Note that the optimal solution $\mathbf{x}_0$ is *not* orthogonal to the affine subspace $\mathcal{A}$ but orthogonal to $\mathcal{U}$. □

Before the problem of (2.19) can be solved, some theoretical foundations should be explained:

Definition 2.1 (Orthogonal Complement)**.** If $\mathcal{U}$ is a subspace of a Hilbert space $\mathcal{H}$ with the inner product $\langle \cdot, \cdot \rangle$, then the set of all vectors orthogonal to $\mathcal{U}$ is called the *orthogonal complement of* $\mathcal{U}$, and we write $\mathcal{U}^{\perp}$ for it.

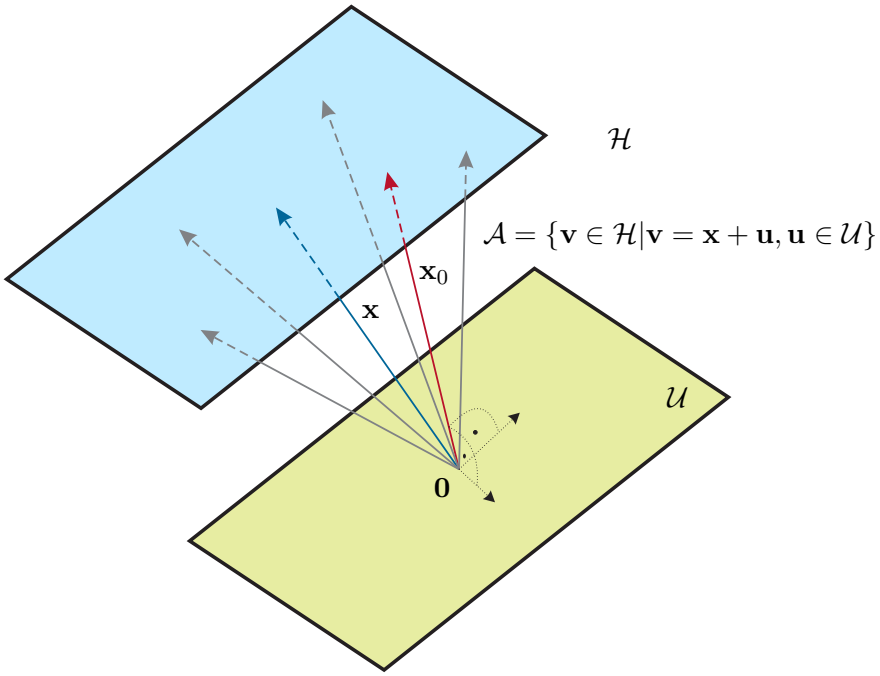


Figure 2.5: On the projection theorem for affine subspaces.

For the subspaces $\mathcal{U}$ and $\mathcal{V}$ of a Hilbert space, the following properties now hold:

(1) The orthogonal complement $\mathcal{U}^\perp$ is a closed subspace,

(2) $\mathcal{U} \subseteq \mathcal{U}^{\perp\perp}$,

(3) If $\mathcal{U} \subset \mathcal{V}$, then $\mathcal{V}^\perp \subset \mathcal{U}^\perp$,

(4) $\mathcal{U}^{\perp\perp\perp} = \mathcal{U}^\perp$, and

(5) $\mathcal{U}^{\perp\perp}$ is the smallest closed subspace containing $\mathcal{U}$.

Proof of (1). Since the linear combination of orthogonal vectors is again orthogonal, it follows immediately that $\mathcal{U}^\perp$ is a subspace. The closedness of $\mathcal{U}^\perp$ follows from the fact that, due to the continuity of the inner product $\langle \cdot, \cdot \rangle$, for the limit $\mathbf{x}$ of a convergent sequence $(\mathbf{x}_k)$ in $\mathcal{U}^\perp$, we have

$$\mathbf{0} = \langle \mathbf{y}, \mathbf{x}_k \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \tag{2.22}$$

for all $\mathbf{y} \in \mathcal{U}$ and thus also $\mathbf{x} \in \mathcal{U}^\perp$. □

Exercise 2.2. Prove the above properties (2) to (4).

Definition 2.2 (Direct Sum)**.** A vector space $\mathcal{X}$ is called the direct sum of two subspaces $\mathcal{U}$ and $\mathcal{V}$, and we write $\mathcal{X} = \mathcal{U} \oplus \mathcal{V}$, if every vector $\mathbf{x} \in \mathcal{X}$ can be *uniquely* represented as the sum $\mathbf{x} = \mathbf{u} + \mathbf{v}$ with $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$.

Without proof, the following theorem holds as a corollary of the projection theorem of Theorem 1.1:

Theorem 2.2. *If $\mathcal{U}$ is a closed linear subspace of a Hilbert space $\mathcal{H}$, then $\mathcal{H} = \mathcal{U} \oplus \mathcal{U}^\perp$ and $\mathcal{U} = \mathcal{U}^{\perp\perp}$.*

The following theorem can now be given for the solution of the minimization problem with affine constraints from (2.19):

Theorem 2.3 (Minimization with Affine Constraints)**.** *Let $\mathcal{H}$ be a Hilbert space with the linearly independent vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Among all possible vectors $\mathbf{x} \in \mathcal{H}$ that satisfy the affine equation system*

$$\begin{aligned} \langle \mathbf{x}, \mathbf{s}_1 \rangle &= c_1 \\ \langle \mathbf{x}, \mathbf{s}_2 \rangle &= c_2 \\ &\vdots \\ \langle \mathbf{x}, \mathbf{s}_n \rangle &= c_n \end{aligned} \tag{2.23}$$

with the constant coefficients $c_1, c_2, \ldots, c_n$, let the vector $\mathbf{x}_0$ have the minimal norm. Then $\mathbf{x}_0$ can be written in the form

$$\mathbf{x}_0 = \sum_{j=1}^{n} p_{0,j} \mathbf{s}_j = \mathbf{S}\mathbf{p}_0 \tag{2.24}$$

where $\mathbf{p}_0^T = \begin{bmatrix} p_{0,1} & p_{0,2} & \ldots & p_{0,n} \end{bmatrix}$ is calculated from the relationship

$$\underbrace{\begin{bmatrix} \langle \mathbf{s}_1, \mathbf{s}_1 \rangle & \cdots & \langle \mathbf{s}_n, \mathbf{s}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{s}_1, \mathbf{s}_n \rangle & \cdots & \langle \mathbf{s}_n, \mathbf{s}_n \rangle \end{bmatrix}}_{\mathbf{G} = \mathbf{S}^\mathrm{T}\mathbf{S}} \underbrace{\begin{bmatrix} p_{0,1} \\ \vdots \\ p_{0,n} \end{bmatrix}}_{\mathbf{p}_0} = \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}}_{\mathbf{c}} \tag{2.25}$$

with the Gramian matrix $\mathbf{G}$.

Proof of Theorem 2.3. Let $\mathcal{U} = \operatorname{span}\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ be a closed linear subspace of the Hilbert space $\mathcal{H}$. The set of all possible vectors $\mathbf{x} \in \mathcal{H}$ that satisfy the affine equation system (2.23) form an affine subspace of $\mathcal{H}$, namely the translational shift of $\mathcal{U}^\perp$ by a vector $\boldsymbol{\gamma}$. Since the orthogonal complement $\mathcal{U}^\perp$ is a closed subspace, we can apply Theorem 2.1 and thus know that the optimal solution $\mathbf{x}_0$ exists, is unique, and is orthogonal to $\mathcal{U}^\perp$. However, this implies $\mathbf{x}_0 \in \mathcal{U}^{\perp\perp}$, and due to the closedness of $\mathcal{U}$ and Theorem 2.2, we obtain $\mathcal{U}^{\perp\perp} = \mathcal{U}$. Since $\mathbf{x}_0 \in \mathcal{U}$ holds, $\mathbf{x}_0$ must be representable

as a linear combination of the $\mathbf{s}_j$, $j = 1, \ldots, n$, according to (2.24). Substituting (2.24) into the affine constraint (2.23), we obtain the result (2.25). □

Applying Theorem 2.3 to the minimization problem (2.19) with $\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \ldots & \mathbf{s}_n \end{bmatrix}$ and the inner product $\langle \mathbf{x}, \mathbf{z} \rangle_Q = \mathbf{x}^{\mathrm{T}} \mathbf{Q} \mathbf{z}$ in the Hilbert space $\mathcal{H} = \mathbb{R}^m$, i.e.,

$$\min_{\mathbf{k}_j} \langle \mathbf{k}_j, \mathbf{k}_j \rangle_Q \quad \text{subject to} \quad \left\langle \mathbf{k}_j, \mathbf{Q}^{-1} \mathbf{s}_l \right\rangle_Q = \delta_{jl} = \begin{cases} 1 & \text{for } j = l \\ 0 & \text{otherwise} \end{cases} \tag{2.26}$$

for $j = 1, \ldots, n$, then, replacing $\mathbf{s}_l$ with $\mathbf{Q}^{-1}\mathbf{s}_l$ in (2.24) and (2.25), we obtain the result

$$\mathbf{k}_{j,0} = \mathbf{Q}^{-1}\mathbf{S}\mathbf{p}_0 \quad \text{and} \quad (\mathbf{Q}^{-1}\mathbf{S})^{\mathrm{T}}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{S}\mathbf{p}_0 = \mathbf{e}_j \tag{2.27}$$

or

$$\mathbf{k}_{j,0} = \mathbf{Q}^{-1}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{e}_j \ . \tag{2.28}$$

According to (2.20), the optimal solution $\mathbf{K}_0$ for the matrix $\mathbf{K}$ is calculated as

$$\mathbf{K}_0^{\mathrm{T}} = \mathbf{Q}^{-1}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1} \tag{2.29}$$

and thus the desired linear *Gauss-Markov estimator* according to (2.13) is

$$\hat{\mathbf{p}} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{y} \ . \tag{2.30}$$

Comparing (2.30) with (1.128), we see that the result is identical to the result of the least-squares identification with weighted least squares with the weighting matrix $\mathbf{Q}^{-1}$, where $\mathbf{Q} = \mathrm{cov}(\mathbf{v})$ holds.

For the expected value of the estimation error $\mathbf{e} = \mathbf{p} - \hat{\mathbf{p}}$, we obtain $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ according to (2.17), and the covariance matrix of the estimation error is given by (2.16):

$$\begin{aligned} \mathrm{E}(\mathbf{e}\mathbf{e}^{\mathrm{T}}) = \mathrm{E}((\mathbf{p} - \mathbf{K}\mathbf{y})(\mathbf{p} - \mathbf{K}\mathbf{y})^{\mathrm{T}}) \quad &= \mathrm{E}\Big([\mathbf{p} - \mathbf{K}(\mathbf{S}\mathbf{p} + \mathbf{v})][\mathbf{p} - \mathbf{K}(\mathbf{S}\mathbf{p} + \mathbf{v})]^{\mathrm{T}}\Big) \\ = \mathrm{E}\Big([\mathbf{K}\mathbf{v}][\mathbf{K}\mathbf{v}]^{\mathrm{T}}\Big) &= \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}} \end{aligned} \tag{2.31}$$

or, with (2.29),

$$\mathrm{E}(\mathbf{e}\mathbf{e}^{\mathrm{T}}) = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1} = (\mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S})^{-1} \ . \tag{2.32}$$

In the literature, the linear estimator (2.30) is often also called *BLUE (best linear unbiased estimate)*.

If the covariance matrix $\mathrm{cov}(\mathbf{v}) = \mathrm{E}(\mathbf{v}\mathbf{v}^{\mathrm{T}}) = \mathbf{Q}$ is a diagonal matrix of the form $\mathbf{Q} = \mathrm{diag}(q_0, q_1, \ldots)$ with $q_j > 0$ for all $j \geq 0$, then for the Gauss-Markov estimator (2.30), according to the recursive method of weighted least squares (1.121), with $\alpha_j = 1/q_j$, a recursive version can be immediately given. This also clarifies the question of what the optimal choice of the sequence of positive weighting coefficients α_j ($\alpha_j > 0$ for all j) looks like in (1.121).

2.2 Minimum-Variance Estimation

It has been assumed so far that no information is available about the n-dimensional vector of unknowns $\mathbf{p}$ in (2.11). However, in some cases, *a priori information about* $\mathbf{p}$ in the form of stochastic characteristics (expected value, covariance matrix) is known. Therefore, it is assumed that for the system of equations (compare also (2.11))

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{2.33}$$

with the stochastic perturbation $\mathbf{v}$, the known $(m \times n)$-matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, the n-dimensional random vector $\mathbf{p} \in \mathbb{R}^n$, and the m-dimensional measurement vector $\mathbf{y} \in \mathbb{R}^m$, the following holds:

$$\begin{aligned} \mathrm{E}(\mathbf{v}) = \mathbf{0}, \qquad & \mathrm{cov}(\mathbf{v}) = \mathrm{E}\Big(\mathbf{v}\mathbf{v}^{\mathrm{T}}\Big) = \mathbf{Q} \qquad \text{with} \quad \mathbf{Q} \geq 0 \\ \mathrm{E}(\mathbf{p}) = \mathbf{0}, \qquad & \mathrm{cov}(\mathbf{p}) = \mathrm{E}\Big(\mathbf{p}\mathbf{p}^{\mathrm{T}}\Big) = \mathbf{R} \qquad \text{with} \quad \mathbf{R} \geq 0 \\ & \mathrm{E}\Big(\mathbf{p}\mathbf{v}^{\mathrm{T}}\Big) = \mathbf{N} \ . \end{aligned} \tag{2.34}$$

Furthermore, it is assumed that the matrix $\Big(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}} + \mathbf{Q} + \mathbf{S}\mathbf{N} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\Big)$ is non-singular. A linear estimator is sought

$$\hat{\mathbf{p}} = \mathbf{K}\mathbf{y} \tag{2.35}$$

with a constant $(n \times m)$-matrix $\mathbf{K} \in \mathbb{R}^{n \times m}$ such that the following minimization problem

$$\min_{\mathbf{K}} \mathrm{E}\Big(\|\mathbf{e}\|_2^2\Big) = \min_{\mathbf{K}} \mathrm{E}\Big([\mathbf{p} - \mathbf{K}\mathbf{y}]^{\mathrm{T}}[\mathbf{p} - \mathbf{K}\mathbf{y}]\Big) \tag{2.36}$$

is solved. By expanding (2.36) and using the relationship $\mathrm{spur}(\mathbf{K}\mathbf{S}\mathbf{R}) = \mathrm{spur}\Big(\mathbf{R}(\mathbf{K}\mathbf{S})^{\mathrm{T}}\Big)$ we obtain

$$\begin{aligned} &\min_{\mathbf{K}} \mathrm{E}\Big([\mathbf{p} - \mathbf{K}\mathbf{y}]^{\mathrm{T}}[\mathbf{p} - \mathbf{K}\mathbf{y}]\Big) = \\ &\min_{\mathbf{K}} \left\{ \underbrace{\mathrm{E}\Big([\mathbf{p} - \mathbf{K}\mathbf{S}\mathbf{p}]^{\mathrm{T}}[\mathbf{p} - \mathbf{K}\mathbf{S}\mathbf{p}]\Big)}_{\mathrm{spur}\left(\mathrm{E}\left([\mathbf{E}-\mathbf{K}\mathbf{S}]\mathbf{p}\mathbf{p}^{\mathrm{T}}[\mathbf{E}-\mathbf{K}\mathbf{S}]^{\mathrm{T}}\right)\right)} - 2 \underbrace{\mathrm{E}\Big([\mathbf{p} - \mathbf{K}\mathbf{S}\mathbf{p}]^{\mathrm{T}}\mathbf{K}\mathbf{v}\Big)}_{\mathrm{spur}\left(\mathrm{E}\left([\mathbf{E}-\mathbf{K}\mathbf{S}]\mathbf{p}\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)\right)} + \underbrace{\mathrm{E}\Big(\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\mathbf{K}\mathbf{v}\Big)}_{\mathrm{spur}\left(\mathrm{E}\left(\mathbf{K}\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)\right)} \right\} = \\ &\min_{\mathbf{K}} \Big\{ \mathrm{spur}\Big([\mathbf{E} - \mathbf{K}\mathbf{S}]\mathbf{R}[\mathbf{E} - \mathbf{K}\mathbf{S}]^{\mathrm{T}} - 2\mathbf{K}\mathbf{N}^{\mathrm{T}} - \mathbf{K}\Big[\mathbf{S}\mathbf{N} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\Big]\mathbf{K}^{\mathrm{T}} + \mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}} \Big) \Big\} = \\ &\min_{\mathbf{K}} \Big\{ \mathrm{spur}\Big(\mathbf{K}\Big(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}} + \mathbf{Q} + \mathbf{S}\mathbf{N} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\Big)\mathbf{K}^{\mathrm{T}} - 2\mathbf{K}\Big(\mathbf{S}\mathbf{R} + \mathbf{N}^{\mathrm{T}}\Big) \Big) \Big\} \ . \end{aligned} \tag{2.37}$$

Writing the matrix $\mathbf{K}$ and the identity matrix $\mathbf{E}$ as in (2.20)

$$\mathbf{K} = \begin{bmatrix} \mathbf{k}_1^{\mathrm{T}} \\ \mathbf{k}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{k}_n^{\mathrm{T}} \end{bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{bmatrix} \tag{2.38}$$

then (2.37) becomes

$$\min_{\mathbf{K}}\left\{\sum_{j=1}^{n}\left(\mathbf{k}_j^{\mathrm{T}}\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)\mathbf{k}_j-2\mathbf{k}_j^{\mathrm{T}}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)\mathbf{e}_j\right)\right\}. \tag{2.39}$$

Comparing the minimization problem (2.39) with that of (1.61), i.e.,

$$\min_{\mathbf{p}}(\mathbf{y}-\mathbf{S}\mathbf{p})^{\mathrm{T}}(\mathbf{y}-\mathbf{S}\mathbf{p})=\min_{\mathbf{p}}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y}-2\mathbf{p}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{y}+\mathbf{p}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{p}\right), \tag{2.40}$$

it is seen that the two problems are equivalent. Thus, the solution of (2.39) can be directly given by the solution of (2.40) (compare (1.63))

$$\mathbf{p}_0=\left(\mathbf{S}^{\mathrm{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y} \tag{2.41}$$

by setting $\mathbf{S}^{\mathrm{T}}\mathbf{S}=\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}$ and $\mathbf{S}^{\mathrm{T}}\mathbf{y}=\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)\mathbf{e}_j$ in (2.41). This yields the optimal solution $\mathbf{K}_0$ for the matrix $\mathbf{K}$ of (2.35) as

$$\mathbf{K}_0^{\mathrm{T}}=\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)^{-1}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right) \tag{2.42}$$

and the sought linear *minimum-variance estimator* according to (2.35) is

$$\hat{\mathbf{p}}=\left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\right)\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)^{-1}\mathbf{y}\,. \tag{2.43}$$

Note that at the beginning of this section it was assumed that $\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)$ is non-singular.

The covariance matrix of the estimation error is calculated as

$$\begin{aligned}
\mathrm{cov}(\mathbf{e}) &= \mathrm{E}\left([\mathbf{p}-\mathbf{K}(\mathbf{S}\mathbf{p}+\mathbf{v})][\mathbf{p}-\mathbf{K}(\mathbf{S}\mathbf{p}+\mathbf{v})]^{\mathrm{T}}\right)=\mathrm{E}\left([\mathbf{E}-\mathbf{K}\mathbf{S}]\mathbf{p}\mathbf{p}^{\mathrm{T}}[\mathbf{E}-\mathbf{K}\mathbf{S}]^{\mathrm{T}}\right)-\\
&\quad \mathrm{E}\left(\mathbf{K}\mathbf{v}\mathbf{p}^{\mathrm{T}}\left(\mathbf{E}-\mathbf{S}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)+(\mathbf{E}-\mathbf{K}\mathbf{S})\mathbf{p}\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)+\mathrm{E}\left(\mathbf{K}\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)\\
&= (\mathbf{E}-\mathbf{K}\mathbf{S})\mathbf{R}(\mathbf{E}-\mathbf{K}\mathbf{S})^{\mathrm{T}}-\mathbf{K}\mathbf{N}^{\mathrm{T}}\left(\mathbf{E}-\mathbf{S}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\right)-(\mathbf{E}-\mathbf{K}\mathbf{S})\mathbf{N}\mathbf{K}^{\mathrm{T}}+\mathbf{K}\mathbf{Q}\mathbf{K}^{\mathrm{T}}\\
&= \mathbf{R}-\mathbf{K}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)-\left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\right)\mathbf{K}^{\mathrm{T}}+\mathbf{K}\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)\mathbf{K}^{\mathrm{T}}\\
&= \mathbf{R}-\left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\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)^{-1}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)-\\
&\quad \left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\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)^{-1}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)+\left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\right)\\
&\quad \left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)^{-1}\left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)\\
&\quad \left(\mathbf{S}\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{Q}+\mathbf{S}\mathbf{N}+\mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\right)^{-1}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right)\\
&= \mathbf{R}-\left(\mathbf{R}\mathbf{S}^{\mathrm{T}}+\mathbf{N}\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)^{-1}\left(\mathbf{S}\mathbf{R}+\mathbf{N}^{\mathrm{T}}\right).
\end{aligned} \tag{2.44}$$

Exercise 2.3. If you have the time, verify (2.44).

In contrast to the Gauss-Markov estimator (2.30), the minimum-variance estimator (2.43) yields meaningful results even if fewer measurements m than unknowns n are available, provided that the matrix $\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}$ is non-singular. The reason for this property is obviously that, in the minimum-variance estimator, stochastic information about the parameter vector $\mathbf{p}$ is available, and thus a meaningful estimation is possible with fewer measurements, or even with no measurements for $\mathbf{Q} > 0$.

Applying now the matrix inversion lemma, Theorem 1.3,

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\Big(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\Big)^{-1}\mathbf{DA}^{-1} \tag{2.45}$$

to the covariance matrix of the estimation error $\mathrm{cov}(\mathbf{e})$ from (2.44) with $\mathbf{A} = \mathbf{R}^{-1}$, $\mathbf{B} = \mathbf{S}^{\mathrm{T}} + \mathbf{R}^{-1}\mathbf{N}$, $\mathbf{C} = \Big(\mathbf{Q} - \mathbf{N}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{N}\Big)^{-1}$ and $\mathbf{D} = \mathbf{S} + \mathbf{N}^{\mathrm{T}}\mathbf{R}^{-1}$, we obtain

$$\begin{aligned}\mathrm{cov}(\mathbf{e}) &= \mathbf{R} - \Big(\mathbf{RS}^{\mathrm{T}} + \mathbf{N}\Big)\Big(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\Big)^{-1}\Big(\mathbf{SR} + \mathbf{N}^{\mathrm{T}}\Big) \\ &= \Big(\mathbf{R}^{-1} + \Big(\mathbf{S}^{\mathrm{T}} + \mathbf{R}^{-1}\mathbf{N}\Big)\Big(\mathbf{Q} - \mathbf{N}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{N}\Big)^{-1}\Big(\mathbf{S} + \mathbf{N}^{\mathrm{T}}\mathbf{R}^{-1}\Big)\Big)^{-1}.\end{aligned} \tag{2.46}$$

Furthermore, it is easy to verify that the minimum-variance estimator (2.43) can be written in the form

$$\hat{\mathbf{p}} = \mathrm{cov}(\mathbf{e})\Big(\mathbf{S}^{\mathrm{T}} + \mathbf{R}^{-1}\mathbf{N}\Big)\Big(\mathbf{Q} - \mathbf{N}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{N}\Big)^{-1}\mathbf{y}. \tag{2.47}$$

Exercise 2.4. Show the validity of (2.47).

Equation (2.47) now shows that for $\mathbf{R}^{-1} = \mathbf{0}$, i.e., infinitely high variance of the parameter vector $\mathbf{p}$ – thus no meaningful a priori information for $\mathbf{p}$ is available – and $\mathbf{Q} > 0$, the minimum-variance estimator (2.43) or (2.47) transitions into the Gauss-Markov estimator (2.30).

The results (2.43) and (2.44) in combination with the relationships

$$\mathrm{E}\Big(\mathbf{py}^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{pp}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} + \mathbf{pv}^{\mathrm{T}}\Big) = \Big(\mathbf{RS}^{\mathrm{T}} + \mathbf{N}\Big) \tag{2.48}$$

and

$$\mathrm{E}\Big(\mathbf{yy}^{\mathrm{T}}\Big) = \mathrm{E}\Big([\mathbf{Sp} + \mathbf{v}][\mathbf{Sp} + \mathbf{v}]^{\mathrm{T}}\Big) = \Big(\mathbf{SRS}^{\mathrm{T}} + \mathbf{Q} + \mathbf{SN} + \mathbf{N}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\Big) \tag{2.49}$$

can be summarized in the following theorem:

Theorem 2.4 (Minimum-Variance Estimator)**.** *For the system of equations* (2.33)

$$\mathbf{y} = \mathbf{Sp} + \mathbf{v} \tag{2.50}$$

with the stochastic quantities $\mathbf{p}$*,* $\mathbf{v}$ *and* $\mathbf{y}$*, it is assumed that* $\mathrm{E}\Big(\mathbf{yy}^{\mathrm{T}}\Big)$ *is invertible. The optimal linear estimate* $\hat{\mathbf{p}}$ *of* $\mathbf{p}$*, which minimizes the expected value of the quadratic*

error $\mathrm{E}\Big([\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}[\mathbf{p}-\hat{\mathbf{p}}]\Big)$, *is given by*

$$\hat{\mathbf{p}} = \mathrm{E}\Big(\mathbf{p}\mathbf{y}^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\mathbf{y}\mathbf{y}^{\mathrm{T}}\Big)\Big]^{-1}\mathbf{y} \tag{2.51}$$

with the associated error covariance matrix

$$\begin{aligned}\mathrm{cov}(\mathbf{e}) &= \mathrm{E}\Big([\mathbf{p}-\hat{\mathbf{p}}][\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{p}\mathbf{p}^{\mathrm{T}}\Big) - \mathrm{E}\Big(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}}\Big)\\ &= \mathrm{E}\Big(\mathbf{p}\mathbf{p}^{\mathrm{T}}\Big) - \mathrm{E}\Big(\mathbf{p}\mathbf{y}^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\mathbf{y}\mathbf{y}^{\mathrm{T}}\Big)\Big]^{-1}\mathrm{E}\Big(\mathbf{y}\mathbf{p}^{\mathrm{T}}\Big).\end{aligned} \tag{2.52}$$

Note the similarity of (2.51) to the optimal solution in the sense of least squares from (1.63).

Exercise 2.5. Show the validity of the relationship (2.52). Further show that

$$\mathrm{E}\Big([\mathbf{p}-\hat{\mathbf{p}}][\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{p}[\mathbf{p}-\hat{\mathbf{p}}]^{\mathrm{T}}\Big) \quad \text{or} \quad \mathrm{E}\Big(\hat{\mathbf{p}}\hat{\mathbf{p}}^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{p}\hat{\mathbf{p}}^{\mathrm{T}}\Big)\ . \tag{2.53}$$

Remark: (to Exercise 2.5) Simply substitute the expression from (2.51) for $\hat{\mathbf{p}}$.

Exercise 2.6. Show that the relationships (2.43) and (2.44) can be derived directly using Theorem 2.4.

Exercise 2.7. Assume that the expected values $\mathrm{E}(\mathbf{y})$ and $\mathrm{E}(\mathbf{p})$ are not zero as in (2.34), but $\mathrm{E}(\mathbf{y}) = \mathbf{y}_0 \neq \mathbf{0}$ 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 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*

$$\mathbf{z} = \mathbf{C}\mathbf{p} \tag{2.54}$$

with the optimal estimator

$$\hat{\mathbf{z}} = \mathbf{K}_{\mathbf{z}}\mathbf{y} \tag{2.55}$$

based on the measurements

$$\mathbf{y} = \mathbf{S}\mathbf{p} + \mathbf{v} \tag{2.56}$$

we understand the solution of the minimization problem

$$\min_{\mathbf{K}_{\mathbf{z}}} \mathrm{E}\Big([\mathbf{z}-\hat{\mathbf{z}}]^{\mathrm{T}}[\mathbf{z}-\hat{\mathbf{z}}]\Big)\ . \tag{2.57}$$

The following theorem now applies:

Theorem 2.5 (Minimum-Variance Estimator of a Linear Function)**.** *The linear minimum-variance estimate* (2.55) *of a linear function* $\mathbf{Cp}$ *based on the measurements* (2.56) *is equivalent to the linear function of the minimum-variance estimate* $\hat{\mathbf{p}}$ *itself, i.e., the best estimate of* $\mathbf{Cp}$ *is* $\mathbf{C}\hat{\mathbf{p}}$*.*

Exercise 2.8. Prove Theorem 2.5.

2.2.1 Recursive Minimum-Variance Estimation

In the next step, we will investigate how the optimal estimate $\hat{\mathbf{p}}$ from (2.47) can be improved by adding new measurements. This is of particular importance for on-line applications. The method is again 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 $\mathcal{U}_1 \oplus \mathcal{U}_2^* = \mathcal{U}_1 + \mathcal{U}_2$. If $\mathcal{U}_2$ is spanned by a finite number of vectors, then the differences between these vectors and their projections onto $\mathcal{U}_1$ span the subspace $\mathcal{U}_2^*$. Figure 2.6 illustrates this situation.

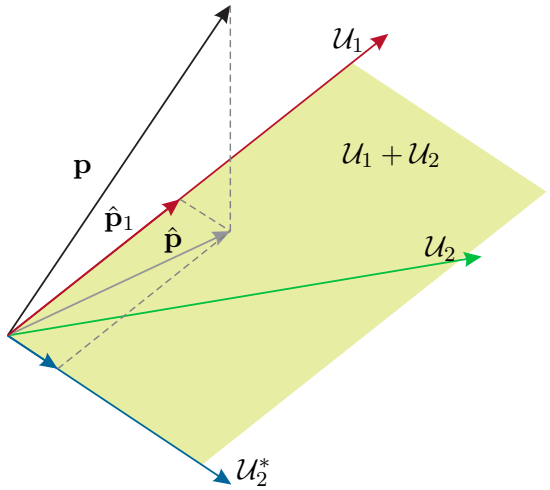


Figure 2.6: On the projection onto the sum of orthogonal subspaces.

Thus, the following theorem can be stated:

Theorem 2.6 (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* $\mathcal{U}_1$*. 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* $\mathcal{U}_1 + \mathcal{U}_2$ *is*

$$\hat{\mathbf{p}} = \hat{\mathbf{p}}_1 + \mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_2 \ . \tag{2.58}$$

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* $\mathcal{U}_2^*$ *(the subspace generated by* $\tilde{\mathbf{y}}_2$*).*

Proof of Theorem 2.6. It is easily verified that $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{U}_1 \oplus \mathcal{U}_2^*$ and that $\mathcal{U}_2^*$ is orthogonal to $\mathcal{U}_1$. Equation (2.58) then follows from the fact that the projection onto a sum of subspaces is equal to the sum of the projections onto the individual subspaces, provided that these are orthogonal. □

The result of Theorem 2.6 can also be interpreted as follows: If $\hat{\mathbf{p}}_1$ denotes the optimal estimate based on measurements that span the subspace $\mathcal{U}_1$, then when receiving new measurements that span the subspace $\mathcal{U}_2$, only that part needs to be considered 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 (2.33)

$$\mathbf{y}_1 = \mathbf{S}_1\mathbf{p} + \mathbf{v}_1 \ . \tag{2.59}$$

Furthermore, $\hat{\mathbf{p}}_1 = \mathrm{E}\Big(\mathbf{p}\mathbf{y}_1^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\mathbf{y}_1\mathbf{y}_1^{\mathrm{T}}\Big)\Big]^{-1}\mathbf{y}_1$ denotes the optimal minimum-variance estimate of $\mathbf{p}$ according to (2.47) or (2.51) based on $\dim(\mathbf{y}_1)$ measurements with the error covariance matrix

$$\mathrm{cov}(\mathbf{p} - \hat{\mathbf{p}}_1) = \mathrm{E}\Big([\mathbf{p} - \hat{\mathbf{p}}_1][\mathbf{p} - \hat{\mathbf{p}}_1]^{\mathrm{T}}\Big) = \mathbf{P}_1 \ . \tag{2.60}$$

The question now arises how to improve the estimate of $\mathbf{p}$ by adding new measurements

$$\mathbf{y}_2 = \mathbf{S}_2\mathbf{p} + \mathbf{v}_2 \tag{2.61}$$

For the stochastic perturbation $\mathbf{v}_2$ and the random parameter vector $\mathbf{p}$, let

$$\begin{aligned} \mathrm{E}(\mathbf{v}_2) = \mathbf{0}, \qquad & \mathrm{cov}(\mathbf{v}_2) = \mathrm{E}\Big(\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\Big) = \mathbf{Q}_2 \quad \text{with} \quad \mathbf{Q}_2 \geq 0 \\ \mathrm{E}(\mathbf{p}) = \mathbf{0}, \qquad & \mathrm{E}\Big(\mathbf{p}\mathbf{v}_2^{\mathrm{T}}\Big) = \mathbf{N}_2 \ . \end{aligned} \tag{2.62}$$

Furthermore, it is reasonable to assume that the perturbation $\mathbf{v}_2$ is not correlated with past measurements $\mathbf{y}_1$, and thus

$$\mathrm{E}\Big(\mathbf{v}_2\mathbf{y}_1^{\mathrm{T}}\Big) = \mathbf{0} \quad \text{or} \quad \mathrm{E}\Big(\mathbf{v}_2\hat{\mathbf{p}}_1^{\mathrm{T}}\Big) = \mathbf{0} \ . \tag{2.63}$$

The best estimate $\hat{\mathbf{y}}_2$ of $\mathbf{y}_2$ based on the past measurements $\mathbf{y}_1$ is

$$\hat{\mathbf{y}}_2 = \mathbf{S}_2\hat{\mathbf{p}}_1 \ . \tag{2.64}$$

Thus, according to Theorem 2.6 with $\tilde{\mathbf{y}}_2 = \mathbf{y}_2 - \hat{\mathbf{y}}_2$, the improved estimate $\hat{\mathbf{p}}_2$ is

$$\hat{\mathbf{p}}_2 = \hat{\mathbf{p}}_1 + \mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_2 \tag{2.65}$$

with

$$\mathrm{E}\Big(\mathbf{p}\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{p}(\mathbf{p}-\hat{\mathbf{p}}_1)^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}} + \mathbf{p}\mathbf{v}_2^{\mathrm{T}}\Big) \overset{(2.53)}{=} \mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{N}_2 \tag{2.66}$$

and

$$\mathrm{E}\Big(\tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^{\mathrm{T}}\Big) = \mathrm{E}\Big([\mathbf{S}_2(\mathbf{p}-\hat{\mathbf{p}}_1)+\mathbf{v}_2][\mathbf{S}_2(\mathbf{p}-\hat{\mathbf{p}}_1)+\mathbf{v}_2]^{\mathrm{T}}\Big) = \mathbf{S}_2\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{Q}_2 + \mathbf{S}_2\mathbf{N}_2 + \mathbf{N}_2^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}} \ . \tag{2.67}$$

Exercise 2.9. Show that the error covariance matrix, analogous to (2.44), can be calculated in the form

$$\begin{aligned}\mathbf{P}_2 &= \mathrm{cov}(\mathbf{p}-\hat{\mathbf{p}}_2)\\ &= \mathbf{P}_1 - \Big(\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{N}_2\Big)\Big(\mathbf{S}_2\mathbf{P}_1\mathbf{S}_2^{\mathrm{T}} + \mathbf{Q}_2 + \mathbf{S}_2\mathbf{N}_2 + \mathbf{N}_2^{\mathrm{T}}\mathbf{S}_2^{\mathrm{T}}\Big)^{-1}\Big(\mathbf{S}_2\mathbf{P}_1 + \mathbf{N}_2^{\mathrm{T}}\Big)\end{aligned} \tag{2.68}$$

This yields the *recursive minimum-variance estimator*

$$\hat{\mathbf{p}}_k = \hat{\mathbf{p}}_{k-1} + \Big(\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{N}_k\Big)\Big(\mathbf{S}_k\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{Q}_k + \mathbf{S}_k\mathbf{N}_k + \mathbf{N}_k^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\Big)^{-1}(\mathbf{y}_k - \mathbf{S}_k\hat{\mathbf{p}}_{k-1}) \tag{2.69}$$

with

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \Big(\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{N}_k\Big)\Big(\mathbf{S}_k\mathbf{P}_{k-1}\mathbf{S}_k^{\mathrm{T}} + \mathbf{Q}_k + \mathbf{S}_k\mathbf{N}_k + \mathbf{N}_k^{\mathrm{T}}\mathbf{S}_k^{\mathrm{T}}\Big)^{-1}\Big(\mathbf{S}_k\mathbf{P}_{k-1} + \mathbf{N}_k^{\mathrm{T}}\Big) \tag{2.70}$$

and the initial values $\mathbf{P}_{-1}$ and $\hat{\mathbf{p}}_{-1}$.

Assuming now 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 (2.69), (2.70), the recursive minimum-variance estimator becomes

$$\mathbf{k}_k = \frac{\mathbf{P}_{k-1}\mathbf{s}_k + \mathbf{n}_k}{(q_k + 2\mathbf{s}_k^{\mathrm{T}}\mathbf{n}_k + \mathbf{s}_k^{\mathrm{T}}\mathbf{P}_{k-1}\mathbf{s}_k)} \tag{2.71a}$$

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{k}_k\Big(\mathbf{s}_k^{\mathrm{T}}\mathbf{P}_{k-1} + \mathbf{n}_k^{\mathrm{T}}\Big) \tag{2.71b}$$

$$\hat{\mathbf{p}}_k = \hat{\mathbf{p}}_{k-1} + \mathbf{k}_k\Big(y_k - \mathbf{s}_k^{\mathrm{T}}\hat{\mathbf{p}}_{k-1}\Big) \ . \tag{2.71c}$$

Note also in this context the analogy to the recursive weighted least squares method (1.121) for $q_k = 1/\alpha_k$ and $\mathbf{n}_k = \mathbf{0}$.

2.3 The Kalman Filter

Building on the previous considerations, especially the recursive minimum variance estimation, the next step is to derive the Kalman filter, an *optimal observer* in the sense of control theory. For the fundamentals of observer theory, refer to Chapter 8 of the Automation script. Numerous versions of the Kalman filter exist in the literature. In this lecture, we will initially consider a linear, time-invariant, discrete-time system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.72a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{2.72b}$$

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 $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$, and $\mathbf{D} \in \mathbb{R}^{q\times p}$. It should be noted here, however, that the Kalman filter can also be designed for linear, time-variant, and continuous-time systems. The following assumptions apply:

(1) For the disturbance $\mathbf{w}$ and the measurement noise $\mathbf{v}$, it is assumed that

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{v}_k\mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{R}\delta_{kj} \tag{2.73a}$$

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}\delta_{kj} \tag{2.73b}$$

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{0} \tag{2.73c}$$

with $\mathbf{Q} \geq 0$, $\mathbf{R} > 0$, and the Kronecker delta $\delta_{kj} = 1$ 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

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\left([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\right) = \mathbf{P}_0 \geq 0 \tag{2.74}$$

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 uncorrelated with the initial value $\mathbf{x}_0$, i.e.,

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0} \tag{2.75a}$$

$$\mathrm{E}\left(\mathbf{v}_l\mathbf{x}_0^{\mathrm{T}}\right) = \mathbf{0}\ . \tag{2.75b}$$

However, due to

$$\mathbf{x}_j = \mathbf{\Phi}^j\mathbf{x}_0 + \sum_{l=0}^{j-1} \mathbf{\Phi}^l(\mathbf{\Gamma}\mathbf{u}_{j-1-l} + \mathbf{G}\mathbf{w}_{j-1-l}) \tag{2.76}$$

and (2.73), the relationship

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{x}_j^{\mathrm{T}}\right) = \mathbf{0} \quad \text{for } k \geq j \tag{2.77a}$$

$$\mathrm{E}\left(\mathbf{v}_l\mathbf{x}_j^{\mathrm{T}}\right) = \mathbf{0} \quad \text{for all } l, j\ . \tag{2.77b}$$

also holds.

For further considerations, the following notation is introduced:

Definition 2.3. The optimal estimate of $\mathbf{x}_k$ considering $0, \dots, j$ measurements is abbreviated as $\hat{\mathbf{x}}(k|j)$.

Theorem 2.7 (Kalman Filter). *The optimal estimate $\hat{\mathbf{x}}(k+1|k)$ of the state $\mathbf{x}_{k+1}$ of the system (2.72) considering $l = 0, \dots, k$ measurements is calculated according to the iteration rule*

$$\begin{aligned}\hat{\mathbf{x}}(k+1|k) &= \mathbf{\Phi}\hat{\mathbf{x}}(k|k-1) + \mathbf{\Gamma}\mathbf{u}_k + \\ &\mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k)\end{aligned} \tag{2.78}$$

with the covariance matrix of the estimation error

$$\begin{aligned}\mathbf{P}(k+1|k) &= \mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} \\ &- \mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}}\end{aligned} \tag{2.79}$$

and the initial values $\hat{\mathbf{x}}(0|-1) = \mathbf{m}_0$ and $\mathbf{P}(0|-1) = \mathbf{P}_0$.

Proof of Theorem 2.7. Assume that the measurements $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{k-1}$ have been used for the optimal estimate $\hat{\mathbf{x}}(k|k-1)$ with the error covariance matrix

$$\mathbf{P}(k|k-1) = \mathrm{E}\Big([\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k-1)]^{\mathrm{T}}\Big) \tag{2.80}$$

At time k, the measurement $\mathbf{y}_k$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{2.81}$$

is now used to improve the estimate of $\mathbf{x}_k$. According to Theorem 2.6, the estimate $\hat{\mathbf{x}}(k|k)$ of $\mathbf{x}_k$ is

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathrm{E}\Big(\mathbf{x}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big)\Big[\mathrm{E}\Big(\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big)\Big]^{-1}\tilde{\mathbf{y}}_k \tag{2.82a}$$

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k \tag{2.82b}$$

or with (2.53) in

$$\mathrm{E}\Big(\mathbf{x}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big) = \mathrm{E}\Big(\mathbf{x}_k(\mathbf{C}\mathbf{x}_k + \mathbf{v}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1))^{\mathrm{T}}\Big) = \mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} \tag{2.83}$$

and

$$\mathrm{E}\Big(\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k^{\mathrm{T}}\Big) = \mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R} \tag{2.84}$$

it follows

$$\begin{aligned}\hat{\mathbf{x}}(k|k) &= \hat{\mathbf{x}}(k|k-1) + \\ &\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}(k|k-1) - \mathbf{D}\mathbf{u}_k)\ .\end{aligned} \tag{2.85}$$

Thus, the error covariance matrix can be written in the form (compare (2.44), (2.70))

$$\begin{aligned}\mathbf{P}(k|k) &= \mathrm{E}\Big([\mathbf{x}_k - \hat{\mathbf{x}}(k|k)][\mathbf{x}_k - \hat{\mathbf{x}}(k|k)]^{\mathrm{T}}\Big) = \\ &\mathbf{P}(k|k-1) - \mathbf{P}(k|k-1)\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}(k|k-1)\end{aligned} \tag{2.86}$$

According to Theorem 2.5, the optimal estimate of $\mathbf{\Phi}\mathbf{x}_k$ is equal to the optimal estimate $\hat{\mathbf{x}}_k$ of $\mathbf{x}_k$ multiplied by $\mathbf{\Phi}$, so it holds that

$$\hat{\mathbf{x}}(k+1|k) = \mathbf{\Phi}\hat{\mathbf{x}}(k|k) + \mathbf{\Gamma}\mathbf{u}_k \ . \tag{2.87}$$

For the covariance matrix of the estimation error, we obtain

$$\begin{aligned}\mathbf{P}(k+1|k) &= \mathrm{E}\Big([\mathbf{x}_{k+1} - \hat{\mathbf{x}}(k+1|k)][\mathbf{x}_{k+1} - \hat{\mathbf{x}}(k+1|k)]^{\mathrm{T}}\Big) \\ &= \mathrm{E}\Big([\mathbf{\Phi}(\mathbf{x}_k - \hat{\mathbf{x}}(k|k)) + \mathbf{G}\mathbf{w}_k][\mathbf{\Phi}(\mathbf{x}_k - \hat{\mathbf{x}}(k|k)) + \mathbf{G}\mathbf{w}_k]^{\mathrm{T}}\Big) \\ &= \mathbf{\Phi}\mathbf{P}(k|k)\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}\end{aligned} \tag{2.88}$$

Combining (2.85)–(2.88) directly yields the result of Theorem 2.7. □

The composition of the error covariance matrix (2.79) should be interpreted at this point: The term $\mathbf{\Phi}\mathbf{P}(k|k-1)\mathbf{\Phi}^{\mathrm{T}}$ describes the change in 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 a negative sign describes how the error variance is reduced by adding the information of new measurements.

2.3.1 The Kalman Filter as an Optimal Observer

Introducing 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 $\mathbf{P}_k = \mathbf{P}(k|k-1)$, (2.78) and (2.79) can also be represented in the compact form

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}_k(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k) \tag{2.89}$$

with

$$\hat{\mathbf{K}}_k = \mathbf{\Phi}\mathbf{P}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1} \tag{2.90}$$

and

$$\mathbf{P}_{k+1} = \mathbf{\Phi}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} - \mathbf{\Phi}\mathbf{P}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} \tag{2.91}$$

Equation (2.91) is also called the *discrete Riccati equation*. Comparing (2.89) with a Luenberger observer, it is seen that the Kalman filter is an 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

$$\begin{aligned} \mathrm{E}(\mathbf{e}_{k+1}) &= \mathrm{E}\Big(\mathbf{\Phi}\mathbf{x}_k + \mathbf{G}\mathbf{w}_k - \mathbf{\Phi}\hat{\mathbf{x}}_k - \hat{\mathbf{K}}_k(\mathbf{C}\mathbf{x}_k + \mathbf{v}_k - \mathbf{C}\hat{\mathbf{x}}_k)\Big) \\ &= \mathrm{E}\Big(\Big(\mathbf{\Phi} - \hat{\mathbf{K}}_k\mathbf{C}\Big)(\mathbf{x}_k - \hat{\mathbf{x}}_k) + \mathbf{G}\mathbf{w}_k - \hat{\mathbf{K}}_k\mathbf{v}_k\Big) \\ &= \Big(\mathbf{\Phi} - \hat{\mathbf{K}}_k\mathbf{C}\Big)\,\mathrm{E}(\mathbf{e}_k)\ . \end{aligned} \tag{2.92}$$

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, it can be seen from (2.90) and (2.91) 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 without knowledge of the measurements $\mathbf{y}_k$ for all $k \geq 0$.

If no previous measurement values about the process are available, one typically sets $\hat{\mathbf{x}}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \alpha\mathbf{E}$ for $\alpha \gg 1$. If the observer runs for a very long time, the problem can be treated mathematically as if it were in operation indefinitely. It turns out that for infinitely long 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

$$\hat{\mathbf{K}}_\infty = \mathbf{\Phi}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1} \tag{2.93}$$

with $\mathbf{P}_\infty$ as the solution of the so-called *discrete algebraic Riccati equation*

$$\mathbf{P}_\infty = \mathbf{\Phi}\mathbf{P}_\infty\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} - \mathbf{\Phi}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\mathbf{P}_\infty\mathbf{C}^{\mathrm{T}} + \mathbf{R}\Big)^{-1}\mathbf{C}\mathbf{P}_\infty\mathbf{\Phi}^{\mathrm{T}}\ . \tag{2.94}$$

Equation (2.94) 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 met:

(1) The pair $(\mathbf{C}, \mathbf{\Phi})$ is *detectable*, i.e., all eigenvalues outside the unit circle are observable.

(2) The pair $\Big(\mathbf{\Phi}, \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}}\Big)$ 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 (2.94) is also called a *stabilizing solution*. Since for this stabilizing solution all eigenvalues of $\Big(\mathbf{\Phi} - \hat{\mathbf{K}}_\infty\mathbf{C}\Big)$ lie in the open interior of the unit circle, the expected value of the observation error decreases according to (2.92), and $\lim_{k\to\infty}\mathrm{E}(\mathbf{e}_k) = \mathbf{0}$ holds. The solution $\mathbf{P}_\infty$ of the discrete algebraic Riccati equation (2.94) can be easily obtained by iterating the discrete Riccati equation (2.91) starting from the initial value $\mathbf{P}_0$ until $\mathbf{P}_k$ changes only insignificantly in terms of a norm. Although the iteration rule generally converges very quickly to a stationary value, in practice, as is also the case in MATLAB, the algebraic Riccati equation (2.94) is solved more efficiently numerically via an eigenvector decomposition, see the MATLAB commands `care` or `dare`.

Exercise 2.10. The motion of a satellite about 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 torque M_c as the control variable, the torque M_d as the disturbance, and the angle φ. Determine the corresponding discrete-time system of the form (2.72) for the sampling time $T_a = 1\mathrm{s}$, $I = 1$, and the output variable φ. Assume that the measurement of the angle φ is superimposed with the measurement noise v and that the disturbance torque M_d corresponds to the process disturbance w. Let

$$\mathrm{E}(w) = 0 \qquad \mathrm{E}\left(w^2\right) = q$$
$$\mathrm{E}(v) = 0 \qquad \mathrm{E}\left(v^2\right) = 0.1\ .$$

Calculate and plot the elements of $\mathbf{P}_k$ of the Kalman filter according to the iteration rule (2.91) for the initial value $\mathbf{P}_0 = \mathbf{E}$ and $q = \{0.1, 0.01, 0.001\}$. Implement the Kalman filter in MATLAB/SIMULINK.

Remark: The relationship (2.93) with the constant observer gain matrix $\hat{\mathbf{K}}_\infty$ specifies an *optimal full observer* that is usable *for both single-variable and multivariable systems*. In contrast to observer design using the pole placement method (note the Ackermann formula according to Chapter 8 of the Automation kecture notes), no poles of the error system need to be chosen for the Kalman filter, which can be very difficult, especially in the multivariable case. The behavior of the error system is instead 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 $\mathbf{v}$, a very often interpretable approach based on the (noise) characteristics of the sensor can be found. Furthermore, a distinction can be made between reliable and less reliable measurements via the weighting of the entries of the covariance matrix $\mathbf{R}$. If a *measurement is less reliable*, the *corresponding entry* on the main diagonal *of the covariance matrix is chosen to be very large*. This assumed large variance of the measurement causes the observer to weight this measurement less in the state estimation compared to the other measurements. In the practical application of the Kalman filter, it is even common to switch the covariance matrix during operation if one or more sensors deliver implausible measurements or if an operating range is reached where it is known in advance that certain sensors no longer provide reliable information.

These assumptions generally do not apply to the process disturbance $\mathbf{w}$ and thus to the covariance matrix $\mathbf{Q}$. The assumption that $\mathbf{w}$ is white noise is usually not true. One might now get the idea to choose 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 usually done by trial and error.

Remark: In the MATLAB CONTROL SYSTEMS TOOLBOX, a slightly more general system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k, \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.95a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{H}\mathbf{w}_k + \mathbf{v}_k, \tag{2.95b}$$

with the additional term $\mathbf{H}\mathbf{w}_k$ in the measured output, is considered. Furthermore, instead of $\mathrm{E}\left(\mathbf{w}_k\mathbf{v}_k^{\mathrm{T}}\right) = \mathbf{0}$, it is allowed that

$$\mathrm{E}\left(\mathbf{w}_k\mathbf{v}_k^{\mathrm{T}}\right) = \mathbf{N}\delta_{kj} \neq \mathbf{0} \tag{2.96}$$

holds. For this case, the optimal state estimation is calculated as

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}_k(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k) \tag{2.97}$$

with

$$\hat{\mathbf{K}}_k = \left(\mathbf{\Phi}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{H}^{\mathrm{T}} + \mathbf{G}\mathbf{N}\right)\left(\mathbf{C}\mathbf{P}_k\mathbf{C}^{\mathrm{T}} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}} + \mathbf{R} + \mathbf{H}\mathbf{N} + \mathbf{N}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\right)^{-1} \tag{2.98}$$

and

$$\mathbf{P}_{k+1} = \mathbf{\Phi}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} + \mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{T}} - \hat{\mathbf{K}}_k\left(\mathbf{C}\mathbf{P}_k\mathbf{\Phi}^{\mathrm{T}} + \mathbf{H}\mathbf{Q}\mathbf{G}^{\mathrm{T}} + \mathbf{N}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\right). \tag{2.99}$$

The proof can be found in the literature, in particular [2.2].

2.3.2 Frequency-domain properties of the stationary Kalman filter

Assume that the measurement noise $\mathbf{v}_k$ from (2.72), i.e.,

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.100a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{2.100b}$$

is generated by a dynamic system of the form

$$\mathbf{z}_{k+1} = \mathbf{\Phi}_{\mathbf{z}}\mathbf{z}_k + \mathbf{n}_k \tag{2.101a}$$
$$\mathbf{v}_k = \mathbf{C}_{\mathbf{z}}\mathbf{z}_k + \mathbf{m}_k \tag{2.101b}$$

with the stochastic disturbances $\mathbf{n}_k$ and $\mathbf{m}_k$. Substituting $\mathbf{v}_k$ from (2.101) into (2.100) yields the extended system in the form

$$\underbrace{\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{z}_{k+1}\end{bmatrix}}_{\tilde{\mathbf{x}}_{k+1}} = \underbrace{\begin{bmatrix}\boldsymbol{\Phi} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Phi}_{\mathbf{z}}\end{bmatrix}}_{\tilde{\boldsymbol{\Phi}}} \underbrace{\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix}}_{\tilde{\mathbf{x}}_k} + \underbrace{\begin{bmatrix}\boldsymbol{\Gamma}\\ \mathbf{0}\end{bmatrix}}_{\tilde{\boldsymbol{\Gamma}}} \mathbf{u}_k + \underbrace{\begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}\end{bmatrix}}_{\tilde{\mathbf{G}}} \underbrace{\begin{bmatrix}\mathbf{w}_k\\ \mathbf{n}_k\end{bmatrix}}_{\tilde{\mathbf{w}}_k} \qquad \begin{matrix}\mathbf{x}(0) = \mathbf{x}_0\\ \mathbf{z}(0) = \mathbf{z}_0\end{matrix} \tag{2.102a}$$

$$\mathbf{y}_k = \underbrace{\begin{bmatrix}\mathbf{C} & \mathbf{C}_{\mathbf{z}}\end{bmatrix}}_{\tilde{\mathbf{C}}} \underbrace{\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix}}_{\tilde{\mathbf{x}}_k} + \mathbf{D}\mathbf{u}_k + \mathbf{m}_k \ . \tag{2.102b}$$

The stationary Kalman filter according to (2.89) and (2.93) for the extended system (2.102) is then

$$\begin{bmatrix}\hat{\mathbf{x}}_{k+1}\\ \hat{\mathbf{z}}_{k+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Phi} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Phi}_{\mathbf{z}}\end{bmatrix} \begin{bmatrix}\hat{\mathbf{x}}_k\\ \hat{\mathbf{z}}_k\end{bmatrix} + \begin{bmatrix}\boldsymbol{\Gamma}\\ \mathbf{0}\end{bmatrix} \mathbf{u}_k + \underbrace{\begin{bmatrix}\hat{\mathbf{K}}_{\mathbf{x}}\\ \hat{\mathbf{K}}_{\mathbf{z}}\end{bmatrix}}_{\hat{\mathbf{K}}_\infty} (\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{C}_{\mathbf{z}}\hat{\mathbf{z}}_k - \mathbf{D}\mathbf{u}_k) \tag{2.103}$$

or

$$\begin{bmatrix}\hat{\mathbf{x}}_{k+1}\\ \hat{\mathbf{z}}_{k+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Phi} - \hat{\mathbf{K}}_{\mathbf{x}}\mathbf{C} & -\hat{\mathbf{K}}_{\mathbf{x}}\mathbf{C}_{\mathbf{z}}\\ -\hat{\mathbf{K}}_{\mathbf{z}}\mathbf{C} & \boldsymbol{\Phi}_{\mathbf{z}} - \hat{\mathbf{K}}_{\mathbf{z}}\mathbf{C}_{\mathbf{z}}\end{bmatrix} \begin{bmatrix}\hat{\mathbf{x}}_k\\ \hat{\mathbf{z}}_k\end{bmatrix} + \begin{bmatrix}\hat{\mathbf{K}}_{\mathbf{x}}\\ \hat{\mathbf{K}}_{\mathbf{z}}\end{bmatrix} \mathbf{y}_k + \begin{bmatrix}\boldsymbol{\Gamma} - \hat{\mathbf{K}}_{\mathbf{x}}\mathbf{D}\\ -\hat{\mathbf{K}}_{\mathbf{z}}\mathbf{D}\end{bmatrix} \mathbf{u}_k \ . \tag{2.104}$$

Since the system is linear and thus the superposition principle applies, the input $\mathbf{u}_k = \mathbf{0}$ is set in the following. The z-transfer matrix from the input $\mathbf{y}_k$ to the output $\hat{\mathbf{x}}_k$ is

$$\mathbf{G}(z) = \frac{\hat{\mathbf{x}}_z(z)}{\mathbf{y}_z(z)} = \begin{bmatrix}\mathbf{E} & \mathbf{0}\end{bmatrix} \begin{bmatrix}z\mathbf{E} - \boldsymbol{\Phi} + \hat{\mathbf{K}}_{\mathbf{x}}\mathbf{C} & \hat{\mathbf{K}}_{\mathbf{x}}\mathbf{C}_{\mathbf{z}}\\ \hat{\mathbf{K}}_{\mathbf{z}}\mathbf{C} & z\mathbf{E} - \boldsymbol{\Phi}_{\mathbf{z}} + \hat{\mathbf{K}}_{\mathbf{z}}\mathbf{C}_{\mathbf{z}}\end{bmatrix}^{-1} \begin{bmatrix}\hat{\mathbf{K}}_{\mathbf{x}}\\ \hat{\mathbf{K}}_{\mathbf{z}}\end{bmatrix} . \tag{2.105}$$

The following theorem now applies:

Theorem 2.8 (Transmission zeros of the stationary Kalman filter)**.** *The transmission zeros of* (2.105) *are given by the relationship*

$$\det(z\mathbf{E} - \boldsymbol{\Phi}_{\mathbf{z}}) = 0 \tag{2.106}$$

Proof. Before Theorem 2.8 is proven, the concept of a transmission zero must be briefly discussed. The zeros of the transfer function $G(z)$ of a single-input single-output system are usually characterized as the roots of the numerator polynomial of $G(z)$. In the case of a multivariable system with a transfer matrix $\mathbf{G}(z)$, this is no longer as simple. A strict definition is given as follows: The zeros of the transfer matrix $\mathbf{G}(z)$ are the roots of the numerator polynomials of the Smith-McMillan form of $\mathbf{G}(z)$. For the definition of the Smith-McMillan form, refer to the literature cited at the end. A physical interpretation of a transmission zero is now given. Consider

the linear, time-invariant, discrete-time system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.107a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k \tag{2.107b}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional input $\mathbf{u} \in \mathbb{R}^p$, the q-dimensional output $\mathbf{y} \in \mathbb{R}^q$, and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, and $\mathbf{C} \in \mathbb{R}^{q\times n}$. A complex number z_j is a transmission zero if, for an input of the form $(\mathbf{u}_k) = \mathbf{u}_0\left(z_j^k\right)$, $\mathbf{u}_0 \neq \mathbf{0}$, there exists an initial state $\mathbf{x}_0 \neq \mathbf{0}$ such that the output $(\mathbf{y}_k)$ vanishes identically for all times $k \geq 0$. This property is also known in the literature as *transmission-blocking*. In the z-domain, the output $\mathbf{y}_z(z) = \mathcal{Z}\{(\mathbf{y}_k)\}$ of (2.107) in response to the input

$$\mathbf{u}_z(z) = \mathcal{Z}\{(\mathbf{u}_k)\} = \mathcal{Z}\left\{\mathbf{u}_0\left(z_j^k\right)\right\} = \mathbf{u}_0\frac{z}{z - z_j} \tag{2.108}$$

with the initial value $\mathbf{x}(0) = \mathbf{x}_0$ is calculated as

$$\mathbf{y}_z(z) = \mathbf{C}(z\mathbf{E} - \mathbf{\Phi})^{-1}\left(z\mathbf{x}_0 + \mathbf{\Gamma}\mathbf{u}_0\frac{z}{z - z_j}\right). \tag{2.109}$$

Using the resolvent identity

$$(z\mathbf{E} - \mathbf{\Phi})^{-1} - (z_j\mathbf{E} - \mathbf{\Phi})^{-1} = (z\mathbf{E} - \mathbf{\Phi})^{-1}(z_j - z)(z_j\mathbf{E} - \mathbf{\Phi})^{-1} \tag{2.110}$$

(2.109) can be rewritten as

$$\begin{aligned}\mathbf{y}_z(z) = \mathbf{C}\Big((z_j\mathbf{E} - \mathbf{\Phi})^{-1} + (z\mathbf{E} - \mathbf{\Phi})^{-1}(z_j - z)(z_j\mathbf{E} - \mathbf{\Phi})^{-1}\Big)\frac{\mathbf{\Gamma}\mathbf{u}_0 z}{z - z_j} + \\ \mathbf{C}(z\mathbf{E} - \mathbf{\Phi})^{-1}z\mathbf{x}_0 \\ = \mathbf{C}(z\mathbf{E} - \mathbf{\Phi})^{-1}z\Big(\mathbf{x}_0 - (z_j\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}\mathbf{u}_0\Big) + \mathbf{C}(z_j\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}\mathbf{u}_0\frac{z}{z - z_j}\end{aligned} \tag{2.111}$$

Exercise 2.11. Prove the identity of (2.110).

From (2.111), it can be seen that $\mathbf{y}_z(z)$ vanishes identically only if the equations

$$\mathbf{x}_0 - (z_j\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}\mathbf{u}_0 = \mathbf{0} \tag{2.112a}$$

$$\mathbf{C}(z_j\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma}\mathbf{u}_0 = \mathbf{0} \tag{2.112b}$$

are satisfied. In summary, it can be stated that for a transmission zero z_j, non-trivial vectors $\mathbf{u}_0$ and $\mathbf{x}_0$ exist that satisfy the system of equations

$$\begin{bmatrix} (z_j\mathbf{E} - \mathbf{\Phi}) & -\mathbf{\Gamma} \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{2.113}$$

and have the property that the output $(\mathbf{y}_k)$ of (2.107) for the input $(\mathbf{u}_k) = \mathbf{u}_0\left(z_j^k\right)$ and the initial value $\mathbf{x}_0$ vanishes identically for all times $k \geq 0$. Applying this result to the transfer matrix $\mathbf{G}(z)$ of (2.105) with the input $\mathbf{y}_k$ and the output $\hat{\mathbf{x}}_k$, the conditions (2.113) in this case are

$$\begin{bmatrix} z_j\mathbf{E} - \mathbf{\Phi} + \hat{\mathbf{K}}_\mathbf{x}\mathbf{C} & \hat{\mathbf{K}}_\mathbf{x}\mathbf{C}_\mathbf{z} \\ \hat{\mathbf{K}}_\mathbf{z}\mathbf{C} & z_j\mathbf{E} - \mathbf{\Phi}_\mathbf{z} + \hat{\mathbf{K}}_\mathbf{z}\mathbf{C}_\mathbf{z} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_0 \\ \hat{\mathbf{z}}_0 \end{bmatrix} - \begin{bmatrix} \hat{\mathbf{K}}_\mathbf{x} \\ \hat{\mathbf{K}}_\mathbf{z} \end{bmatrix} \mathbf{y}_0 = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{2.114a}$$

$$\begin{bmatrix} \mathbf{E} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_0 \\ \hat{\mathbf{z}}_0 \end{bmatrix} = \mathbf{0} \ . \tag{2.114b}$$

Exercise 2.12. Derive the relationships (2.114).

From (2.114b), we obtain $\hat{\mathbf{x}}_0 = \mathbf{0}$, and thus

$$\begin{aligned} \hat{\mathbf{K}}_\mathbf{x}(\mathbf{C}_\mathbf{z}\hat{\mathbf{z}}_0 - \mathbf{y}_0) &= \mathbf{0} \\ \left(z_j\mathbf{E} - \mathbf{\Phi}_\mathbf{z} + \hat{\mathbf{K}}_\mathbf{z}\mathbf{C}_\mathbf{z}\right)\hat{\mathbf{z}}_0 - \hat{\mathbf{K}}_\mathbf{z}\mathbf{y}_0 &= \mathbf{0} \ . \end{aligned} \tag{2.115}$$

Assuming the full column rank of $\hat{\mathbf{K}}_\mathbf{x}$, $\mathbf{C}_\mathbf{z}\hat{\mathbf{z}}_0 - \mathbf{y}_0 = \mathbf{0}$ follows, and thus $(z_j\mathbf{E} - \mathbf{\Phi}_\mathbf{z})\hat{\mathbf{z}}_0 = \mathbf{0}$. It is now seen that (2.115) has non-trivial solutions for $\mathbf{y}_0$ and $\hat{\mathbf{z}}_0$ if and only if $\det(z_j\mathbf{E} - \mathbf{\Phi}_\mathbf{z}) = 0$ holds, which proves Theorem 2.8. □

Remark: This result shows that the stationary Kalman filter (2.103) has zeros at the poles of the disturbance model (2.101). To now design a Kalman filter that blocks certain frequencies, simply choose a disturbance model with poles at these frequencies. In general, it can be said that the greater the energy of the disturbance at the respective frequencies, the more the Kalman filter suppresses these frequencies.

Exercise 2.13. Design a Kalman filter to estimate the rotational speed ω of a permanent magnet DC motor based on a measurement of the rotation angle φ. For a certain choice of parameters and neglecting the dynamics of the electrical subsystem, the

model of the DC motor is obtained in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\varphi\\ \omega\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & -1\end{bmatrix}\begin{bmatrix}\varphi\\ \omega\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}u$$
$$y = \varphi = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}\varphi\\ \omega\end{bmatrix}.$$

It is now desired that resonance frequencies with an angular frequency ω_0, resulting from the mechanical load, are suppressed by the Kalman filter. The corresponding disturbance model can be modeled in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}z_1\\ z_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ -\omega_0^2 & -2\xi\omega_0\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix} + \begin{bmatrix}0\\ \omega_0^2\end{bmatrix}n$$
$$v = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}z_1\\ z_2\end{bmatrix}$$

with white noise n as the input signal. Design a Kalman filter for these specifications for a sampling time $T_a = 0.05$ s and investigate the influence of different values for ξ in the range $0.01 < \xi < 0.7$. Choose $\omega_0 = 3$, $\mathbf{G} = \mathbf{E}$, $\mathbf{Q} = \mathbf{E}$, and $\mathbf{R} = 1$. Plot the Bode diagram of the Kalman filter and test your Kalman filter via simulation.

2.4 The Extended Kalman Filter

Before discussing the Extended Kalman Filter as an observer for nonlinear systems, the Kalman Filter from Section 2.3 for linear time-variant sampled systems of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}_k\mathbf{x}_k + \mathbf{\Gamma}_k\mathbf{u}_k + \mathbf{G}_k\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.116a}$$
$$\mathbf{y}_k = \mathbf{C}_k\mathbf{x}_k + \mathbf{D}_k\mathbf{u}_k + \mathbf{v}_k \tag{2.116b}$$

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 time-variant matrices $\mathbf{\Phi}_k \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma}_k \in \mathbb{R}^{n\times p}$, $\mathbf{G}_k \in \mathbb{R}^{n\times r}$, $\mathbf{C}_k \in \mathbb{R}^{q\times n}$, and $\mathbf{D}_k \in \mathbb{R}^{q\times p}$ will be written down. Analogous to Section 2.3, the following assumptions are made:

(1) For the disturbance $\mathbf{w}$ and the measurement noise $\mathbf{v}$, it is assumed that

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{v}_k\mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{R}_k\delta_{kj} \tag{2.117a}$$
$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}_k\delta_{kj} \tag{2.117b}$$
$$\mathrm{E}\left(\mathbf{w}_k\mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{0} \tag{2.117c}$$

with $\mathbf{Q}_k \geq 0$ and $\mathbf{R}_k > 0$ and the Kronecker delta $\delta_{kj} = 1$ 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

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\Big([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\Big) = \mathbf{P}_0 \geq 0 \tag{2.118}$$

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 uncorrelated with the initial value $\mathbf{x}_0$, i.e., it holds that

$$\mathrm{E}\Big(\mathbf{w}_k \mathbf{x}_0^{\mathrm{T}}\Big) = \mathbf{0} \tag{2.119a}$$

$$\mathrm{E}\Big(\mathbf{v}_l \mathbf{x}_0^{\mathrm{T}}\Big) = \mathbf{0}\ . \tag{2.119b}$$

The derivation of the Kalman filter for the system (2.116) proceeds in a completely analogous manner as in Section 2.3 and is given for $k \geq 0$, compare (2.89)–(2.91),

$$\hat{\mathbf{K}}_k = \mathbf{\Phi}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} \Big(\mathbf{C}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \tag{2.120a}$$

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}_k \hat{\mathbf{x}}_k + \mathbf{\Gamma}_k \mathbf{u}_k + \hat{\mathbf{K}}_k (\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_k - \mathbf{D}_k \mathbf{u}_k) \tag{2.120b}$$

$$\mathbf{P}_{k+1} = \mathbf{\Phi}_k \mathbf{P}_k \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^{\mathrm{T}} - \mathbf{\Phi}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} \Big(\mathbf{C}_k \mathbf{P}_k \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \mathbf{C}_k \mathbf{P}_k \mathbf{\Phi}_k^{\mathrm{T}}\ . \tag{2.120c}$$

If the initial value $\mathbf{x}_0$ is known, then we set $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ and $\mathbf{P}_0 = \mathbf{0}$, and in the case that no information about the initial value is available, $\hat{\mathbf{x}}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \alpha \mathbf{E}$ with $\alpha \gg 1$ is typically chosen. It should be noted that this representation can also be generalized for $\mathbf{H} \neq \mathbf{0}$ and $\mathrm{E}\Big(\mathbf{w}_k \mathbf{v}_j^{\mathrm{T}}\Big) \neq \mathbf{0}$ analogously to the discussions in the last section.

In the literature, it is often common to represent the Kalman filter in a slightly different form. The optimal estimation of the state $\mathbf{x}_k$ and the error covariance matrix $\mathbf{P}_k$ considering $0, \ldots, k-1$ measurements, compare Definition 2.3,

$$\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}(k|k-1) \tag{2.121a}$$

$$\mathbf{P}_k^- = \mathbf{P}(k|k-1) = \mathrm{E}\Big(\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^-\Big]\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^-\Big]^{\mathrm{T}}\Big) \tag{2.121b}$$

is referred to as the *a priori estimate*, and the optimal estimation of $\mathbf{x}_k$ and $\mathbf{P}_k$ considering $0, \ldots, k$ measurements

$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}(k|k) \tag{2.122a}$$

$$\mathbf{P}_k^+ = \mathbf{P}(k|k) = \mathrm{E}\Big(\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^+\Big]\Big[\mathbf{x}_k - \hat{\mathbf{x}}_k^+\Big]^{\mathrm{T}}\Big) \tag{2.122b}$$

is referred to as the *a posteriori estimate*. (2.120) can thus be written in the equivalent form

Kalman Gain Matrix:	$\hat{\mathbf{L}}_k = \mathbf{P}_k^- \mathbf{C}_k^{\mathrm{T}} \left(\mathbf{C}_k \mathbf{P}_k^- \mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k \right)^{-1}$	(2.123a)
State Estimate Update:	$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k \left(\mathbf{y}_k - \mathbf{C}_k \hat{\mathbf{x}}_k^- - \mathbf{D}_k \mathbf{u}_k \right)$	(2.123b)
Error Covariance Update:	$\mathbf{P}_k^+ = \left(\mathbf{E} - \hat{\mathbf{L}}_k \mathbf{C}_k \right) \mathbf{P}_k^-$	(2.123c)
State Extrapolation (2.87):	$\hat{\mathbf{x}}_{k+1}^- = \mathbf{\Phi}_k \hat{\mathbf{x}}_k^+ + \mathbf{\Gamma}_k \mathbf{u}_k$	(2.123d)
Error Covariance Extrapolation (2.88):	$\mathbf{P}_{k+1}^- = \mathbf{\Phi}_k \mathbf{P}_k^+ \mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k \mathbf{Q}_k \mathbf{G}_k^{\mathrm{T}}$	(2.123e)

for $k \geq 0$ and the initial values $\hat{\mathbf{x}}_0^- = \hat{\mathbf{x}}_0$ and $\mathbf{P}_0^- = \mathbf{P}_0$.

Exercise 2.14. Show the equivalence of the relations (2.120) and (2.123). For this, perform the following substitutions in (2.123): $\hat{\mathbf{x}}_{k+1}^- = \hat{\mathbf{x}}_{k+1}$, $\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}_k$, $\mathbf{P}_{k+1}^- = \mathbf{P}_{k+1}$, $\mathbf{P}_k^- = \mathbf{P}_k$, 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, continuous-time multivariable system of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{w}, t) \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.124a}$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}, \mathbf{u}, \mathbf{v}, t) \tag{2.124b}$$

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 $\mathbf{w} \in \mathbb{R}^r$, and the measurement noise $\mathbf{v}$. Since the Kalman filter is normally implemented in a digital computer, the control variables are applied to the process via a zero-order hold (D/A converter) with the sampling time T_a, and the measured variables are sampled with the sampling time T_a via an A/D converter, the corresponding sampled system to (2.124)

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k) \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{2.125a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k) \tag{2.125b}$$

must be calculated. From the lecture on Automation (Section 6.2.1), it is known that the exact solution of (2.124) is necessary to determine the sampled system.

Remark: The solution of a nonlinear system of differential equations of the form (2.124) is known to be only possible in special cases. Therefore, an approximate solution using an integration method is sought below. It is assumed that the control variable $\mathbf{u}(t)$ and the disturbance $\mathbf{w}(t)$ are constant for the sampling interval $kT_a \leq t < (k+1)T_a$, i.e., $\mathbf{u}(t) = \mathbf{u}(kT_a) = \mathbf{u}_k$ and $\mathbf{w}(t) = \mathbf{w}(kT_a) = \mathbf{w}_k$, and the differential equation (2.124a) is integrated over the sampling interval

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \int_{kT_a}^{(k+1)T_a} \mathbf{f}(\mathbf{x}(t), \mathbf{u}_k, \mathbf{w}_k, t)\, \mathrm{d}t \tag{2.126}$$

with $\mathbf{x}_{k+1} = \mathbf{x}((k+1)T_a)$ and $\mathbf{x}_k = \mathbf{x}(kT_a)$. The approximation of the integral in (2.126) can be done in different ways. In the following, only two possible solutions

will be given:

(1) *Euler method*

$$\int_{kT_a}^{(k+1)T_a} \mathbf{f}(\mathbf{x}(t), \mathbf{u}_k, \mathbf{w}_k, t)\,\mathrm{d}t = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k, kT_a)T_a \tag{2.127}$$

(2) *Fourth-order Runge-Kutta method*

$$\int_{kT_a}^{(k+1)T_a} \mathbf{f}(\mathbf{x}(t), \mathbf{u}_k, \mathbf{w}_k, t)\,\mathrm{d}t = \frac{\Delta\mathbf{x}_1 + 2\Delta\mathbf{x}_2 + 2\Delta\mathbf{x}_3 + \Delta\mathbf{x}_4}{6} \tag{2.128}$$

with

$$\begin{aligned}
\Delta\mathbf{x}_1 &= \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k, kT_a)T_a \\
\Delta\mathbf{x}_2 &= \mathbf{f}\Big(\mathbf{x}_k + \frac{\Delta\mathbf{x}_1}{2}, \mathbf{u}_k, \mathbf{w}_k, kT_a + \frac{T_a}{2}\Big)T_a \\
\Delta\mathbf{x}_3 &= \mathbf{f}\Big(\mathbf{x}_k + \frac{\Delta\mathbf{x}_2}{2}, \mathbf{u}_k, \mathbf{w}_k, kT_a + \frac{T_a}{2}\Big)T_a \\
\Delta\mathbf{x}_4 &= \mathbf{f}(\mathbf{x}_k + \Delta\mathbf{x}_3, \mathbf{u}_k, \mathbf{w}_k, (k+1)T_a)T_a\ .
\end{aligned} \tag{2.129}$$

Here, for $\Delta\mathbf{x}_4$ in (2.129), the left-hand limit $\lim_{t\to(k+1)T_a} \mathbf{u}(t) = \mathbf{u}_k$ and $\lim_{t\to(k+1)T_a} \mathbf{w}(t) = \mathbf{w}_k$ was used.

The output equation (2.125b) of the sampled system is obtained very easily for both cases by evaluating the output equation (2.124b) of the continuous-time multivariable system for $t = kT_a$

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k, kT_a) = \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k)\ . \tag{2.130}$$

Assuming that a discrete-time system of the form (2.125) is given by the relations (2.126)–(2.130), the *idea of the Extended Kalman Filter* is based on the fact that a Taylor series expansion is performed for the right-hand side of (2.125a) around the point $\mathbf{x}_k = \hat{\mathbf{x}}_k^+$, $\mathbf{u}_k = \mathbf{u}_k$ and $\mathbf{w}_k = \mathbf{0}$ and truncated after the linear term, i.e.,

$$\mathbf{x}_{k+1} = \mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) + \frac{\partial}{\partial\mathbf{x}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big)\Big(\mathbf{x}_k - \hat{\mathbf{x}}_k^+\Big) + \frac{\partial}{\partial\mathbf{w}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big)\mathbf{w}_k\ . \tag{2.131}$$

Analogously, the right-hand side of the output equation (2.125b) is developed into 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

$$\mathbf{y}_k = \mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) + \frac{\partial}{\partial\mathbf{x}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big)\Big(\mathbf{x}_k - \hat{\mathbf{x}}_k^-\Big) + \frac{\partial}{\partial\mathbf{v}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big)\mathbf{v}_k\ . \tag{2.132}$$

Note that the following simplified notation is used consistently here:

$$\frac{\partial}{\partial\mathbf{x}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) = \frac{\partial}{\partial\mathbf{x}_k}\mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k)\Big|_{\mathbf{x}_k=\hat{\mathbf{x}}_k^-, \mathbf{u}_k=\mathbf{u}_k, \mathbf{w}_k=\mathbf{0}} \tag{2.133}$$

The relations (2.131) and (2.132) can be written more compactly for further considerations in the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}_k \mathbf{x}_k + \bar{\mathbf{u}}_k + \mathbf{G}_k \mathbf{w}_k \tag{2.134a}$$
$$\mathbf{y}_k = \mathbf{C}_k \mathbf{x}_k + \breve{\mathbf{u}}_k + \breve{\mathbf{v}}_k \tag{2.134b}$$

with

$$\begin{aligned} \mathbf{\Phi}_k &= \frac{\partial}{\partial \mathbf{x}_k} \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right) & \bar{\mathbf{u}}_k &= \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right) - \mathbf{\Phi}_k \hat{\mathbf{x}}_k^+ \\ \mathbf{G}_k &= \frac{\partial}{\partial \mathbf{w}_k} \mathbf{F}_k\left(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\right) & \mathbf{C}_k &= \frac{\partial}{\partial \mathbf{x}_k} \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right) \\ \breve{\mathbf{u}}_k &= \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right) - \mathbf{C}_k \hat{\mathbf{x}}_k^- & \breve{\mathbf{v}}_k &= \frac{\partial}{\partial \mathbf{v}_k} \mathbf{h}_k\left(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\right) \mathbf{v}_k \end{aligned} \tag{2.135}$$

It is now obvious that the structure of the system (2.134) directly allows the application of the Kalman filter according to (2.123). The necessary calculation steps for the implementation are summarized again below.

(1) For the nonlinear, time-variant, continuous-time multivariable system (2.124), calculate a discrete-time system of the form (2.125).

(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) It is again assumed for the disturbance $\mathbf{w}_k$ and the measurement noise $\breve{\mathbf{v}}_k$ in (2.134) that

$$\mathrm{E}(\breve{\mathbf{v}}_k) = \mathbf{0} \qquad \mathrm{E}\left(\breve{\mathbf{v}}_k \breve{\mathbf{v}}_j^{\mathrm{T}}\right) = \mathbf{R}_k \delta_{kj} \tag{2.136a}$$
$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k \mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}_k \delta_{kj} \tag{2.136b}$$
$$\mathrm{E}\left(\mathbf{w}_k \breve{\mathbf{v}}_j^{\mathrm{T}}\right) = \mathbf{0} \tag{2.136c}$$

with $\mathbf{Q}_k \geq 0$ and $\mathbf{R}_k > 0$ and the Kronecker delta $\delta_{kj} = 1$ for $k = j$ and $\delta_{kj} = 0$ otherwise.

(4) The iteration equations of the Extended Kalman Filter are then for $k \geq 0$

$$\mathbf{C}_k = \frac{\partial}{\partial \mathbf{x}_k}\mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big) \tag{2.137a}$$

$$\hat{\mathbf{L}}_k = \mathbf{P}_k^-\mathbf{C}_k^{\mathrm{T}}\Big(\mathbf{C}_k\mathbf{P}_k^-\mathbf{C}_k^{\mathrm{T}} + \mathbf{R}_k\Big)^{-1} \tag{2.137b}$$

$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k\Big(\mathbf{y}_k - \mathbf{C}_k\hat{\mathbf{x}}_k^- - \breve{\mathbf{u}}_k\Big) = \hat{\mathbf{x}}_k^- + \hat{\mathbf{L}}_k\Big(\mathbf{y}_k - \mathbf{h}_k\Big(\hat{\mathbf{x}}_k^-, \mathbf{u}_k, \mathbf{0}\Big)\Big) \tag{2.137c}$$

$$\mathbf{P}_k^+ = \Big(\mathbf{E} - \hat{\mathbf{L}}_k\mathbf{C}_k\Big)\mathbf{P}_k^- \tag{2.137d}$$

$$\mathbf{\Phi}_k = \frac{\partial}{\partial \mathbf{x}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{2.137e}$$

$$\mathbf{G}_k = \frac{\partial}{\partial \mathbf{w}_k}\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{2.137f}$$

$$\hat{\mathbf{x}}_{k+1}^- = \mathbf{\Phi}_k\hat{\mathbf{x}}_k^+ + \underbrace{\mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) - \mathbf{\Phi}_k\hat{\mathbf{x}}_k^+}_{\bar{\mathbf{u}}_k} = \mathbf{F}_k\Big(\hat{\mathbf{x}}_k^+, \mathbf{u}_k, \mathbf{0}\Big) \tag{2.137g}$$

$$\mathbf{P}_{k+1}^- = \mathbf{\Phi}_k\mathbf{P}_k^+\mathbf{\Phi}_k^{\mathrm{T}} + \mathbf{G}_k\mathbf{Q}_k\mathbf{G}_k^{\mathrm{T}}\ . \tag{2.137h}$$

Remark: In the Extended Kalman Filter design, it is assumed that the linearized transformation of the mean and covariance corresponds with good accuracy to the mean and covariance of the nonlinear transformation. This assumption is generally not true, which is why the *Unscented Kalman Filter* considered in the next section is often used to improve the observer design for nonlinear systems.

Exercise 2.15. The mathematical model

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_1 &= x_2 + w_1\\ \frac{\mathrm{d}}{\mathrm{d}t}x_2 &= \frac{1}{2}\rho_0\exp\Big(-\frac{x_1}{k}\Big)C_w\frac{A}{m}x_2^2 - g + w_2\end{aligned}$$

describes the free fall of a body of 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 acceleration due to gravity. The process noise is given by the stochastic variables w_1 and w_2. The altitude x_1 can be determined 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, in addition to the altitude x_1 and velocity x_2, the constant drag coefficient C_w. Extend the system of differential equations to include the state $x_3 = C_w$ with

$$\frac{\mathrm{d}}{\mathrm{d}t}x_3 = 0 + w_3$$

and the process noise component w_3. Use the Euler method to determine the discrete-time system. Assume that the nominal values or initial conditions of the variables C_w, x_1, and x_2 are normally distributed. The corresponding values of the means and variances can be found in the following Table 2.1. For the simulation, assume 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
$x_1(0)$	$39 \cdot 10^3\,\mathrm{m}$	$1 \cdot 10^4\,\mathrm{m}^2$
$x_2(0)$	$0\,\mathrm{m/s}$	$1\,\mathrm{m}^2/\mathrm{s}^2$

Table 2.1: Means and variances of the parameters and initial conditions.

Exercise 2.16. An Extended Kalman Filter is to be used for position determination of a vehicle in a two-dimensional space (o-axis: East coordinate, n-axis: North coordinate). Several measuring stations with coordinates (O_i, N_i), $i = 1, \ldots, M$ measure the distance to the vehicle. The acceleration of the vehicle in the North and East directions is modeled by white noise. The system of difference equations

$$\underbrace{\begin{bmatrix} o_{k+1} \\ n_{k+1} \\ o_{v,k+1} \\ n_{v,k+1} \end{bmatrix}}_{\mathbf{x}_{k+1}} = \underbrace{\begin{bmatrix} 1 & 0 & T_a & 0 \\ 0 & 1 & 0 & T_a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{\Phi}} \underbrace{\begin{bmatrix} o_k \\ n_k \\ o_{v,k} \\ n_{v,k} \end{bmatrix}}_{\mathbf{x}_k} + \underbrace{\begin{bmatrix} w_{1,k} \\ w_{2,k} \\ w_{3,k} \\ w_{4,k} \end{bmatrix}}_{\mathbf{w}_k}$$

describes the vehicle behavior, where o_k and n_k and $o_{v,k}$ and $n_{v,k}$ denote the coordinates and velocities of the vehicle relative to the origin of a fixed coordinate system in the East and North directions at time kT_a with sampling time T_a, and $w_{j,k}$, $j = 1, \ldots, 4$ are the components of the process noise. Furthermore, the distance measurements of the vehicle from the stations are given by

$$y_{i,k} = \sqrt{(n_k - N_i)^2 + (o_k - O_i)^2} + v_{i,k}, \quad i = 1, \ldots, M$$

with the measurement noise $v_{i,k}$, $i = 1, \ldots, M$. Assume that all stochastic variables $(w_{j,k})$ and $(v_{i,k})$ are normally distributed, uncorrelated, and zero-mean. The sampling time is given as $T_a = 0.1\,\mathrm{s}$. For the covariance matrix of the process noise, it applies that

$$\mathrm{E}\left(\mathbf{w}_k \mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}\delta_{kj} \quad \text{with} \quad \mathbf{Q} = \mathrm{diag}(0, 0, 4, 4)$$

and the covariance of the measurement noise is

$$\mathrm{E}(v_{i,k} v_{i,j}) = R_i \delta_{kj} \quad \text{with} \quad R_i = 1, \quad i = 1, \ldots, M\ .$$

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 to estimate the states. Vary the number and position of the measuring stations.

2.5 The Unscented Kalman Filter

The Extended Kalman Filter (EKF), discussed in the last chapter, is the (industrial) standard for state estimation of nonlinear dynamic systems. However, the EKF has the disadvantage that it can provide unreliable estimations (of the expected value and the error covariance) if the system exhibits pronounced nonlinearities. This unreliability results from the linearization of the nonlinear system dynamics, which is used to calculate the expected value and the covariance of the state. To illustrate this problem in more detail, the next section shows how the expected value and the covariance of a random variable change through a nonlinear transformation.

2.5.1 Expected Value and Covariance of Nonlinear Transformations

In the following, the example of a transformation from cylindrical coordinates to Cartesian coordinates is considered

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{h}(\mathbf{x}) = \begin{bmatrix} x_1 \cos(x_2) \\ x_1 \sin(x_2) \end{bmatrix}. \tag{2.138}$$

It is assumed that the random vector $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ is defined by the uncorrelated, uniformly distributed random variables x_1 (interval: $a_1 = m_{x1} - \delta_{x1} = 1 - 0.01$, $b_1 = m_{x1} + \delta_{x1} = 1 + 0.01$) and x_2 (interval: $a_2 = m_{x2} - \delta_{x2} = \frac{\pi}{2} - 0.35$, $b_2 = m_{x2} + \delta_{x2} = \frac{\pi}{2} + 0.35$). Figure 2.7 shows 10000 random vectors generated according to this distribution. The expected value (mean) $\mathbf{m}_x$ is $\mathbf{m}_x = [1, \pi/2]^{\mathrm{T}}$ and the covariance matrix is calculated as

$$\mathbf{Q}_x = \begin{bmatrix} \frac{1}{3}10^{-4} & 0 \\ 0 & \frac{49}{12}10^{-2} \end{bmatrix}. \tag{2.139}$$

In Figure 2.7, the covariance matrix is represented by the ellipse $(\mathbf{x} - \mathbf{m}_x)^{\mathrm{T}} \mathbf{Q}_x^{-1} (\mathbf{x} - \mathbf{m}_x) = 1$.

The expected value $\mathrm{E}(\mathbf{y}) = \mathbf{m}_y$ of the quantity $\mathbf{y} = \mathbf{h}(\mathbf{x})$ calculated using the nonlinear transformation results from

$$\begin{bmatrix} m_{y1} \\ m_{y2} \end{bmatrix} = \begin{bmatrix} \mathrm{E}(h_1(\mathbf{x})) \\ \mathrm{E}(h_2(\mathbf{x})) \end{bmatrix} = \begin{bmatrix} \mathrm{E}(x_1 \cos(x_2)) \\ \mathrm{E}(x_1 \sin(x_2)) \end{bmatrix}. \tag{2.140}$$

Considering the independence of the random variables x_1 and x_2 and introducing the decomposition $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{m}_x$, we obtain for the first row of (2.140)

$$\mathrm{E}(x_1 \cos(x_2)) = \mathrm{E}(x_1)\, \mathrm{E}(\cos(\tilde{x}_2 + m_{x2})) \ . \tag{2.141}$$

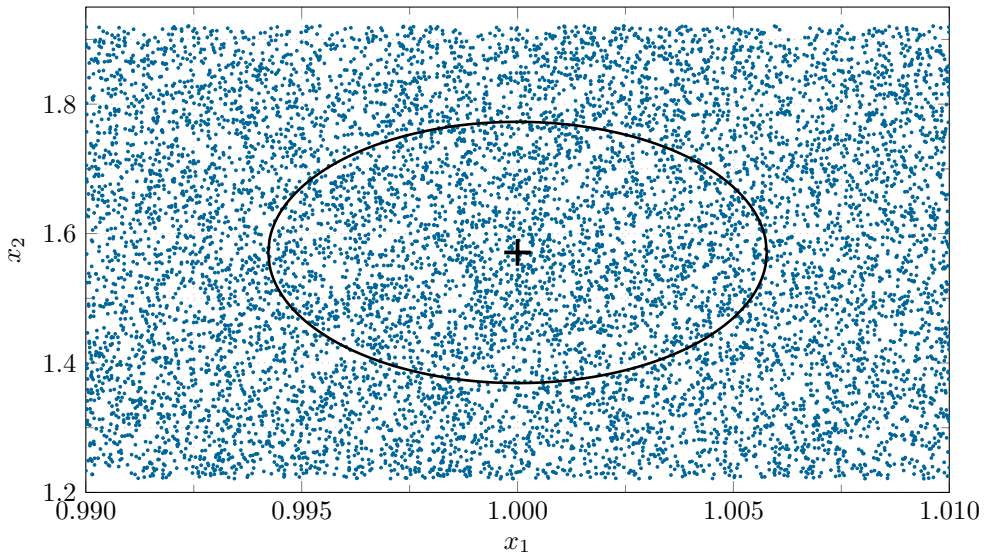


Figure 2.7: Distribution of the random vectors $\mathbf{x}$ with associated expected value $\mathbf{m}_x$ and covariance matrix $\mathbf{Q}_x$.

It holds that $\mathrm{E}(x_1) = m_{x1} = 1$. Splitting the cos-term in (2.141) and substituting the expected value $m_{x2} = \pi/2$, we obtain

$$\mathrm{E}(\cos(\tilde{x}_2 + m_{x2})) = \mathrm{E}(\cos(\tilde{x}_2)\cos(m_{x2}) - \sin(\tilde{x}_2)\sin(m_{x2})) = -\mathrm{E}(\sin(\tilde{x}_2)) \ . \tag{2.142}$$

Since the expected value of a skew-symmetric function of a random variable with a symmetric probability density function vanishes, $\mathrm{E}(\sin(\tilde{x}_2)) = 0$ and thus $m_{y1} = 0$. To show this last step, consider

$$\mathrm{E}(\sin(\tilde{x}_2)) = \int_{-\delta_{x2}}^{\delta_{x2}} \sin(\tilde{x}_2)\frac{1}{2\delta_{x2}}\,\mathrm{d}\tilde{x}_2 = \frac{1}{2\delta_{x2}}(-\cos(\delta_{x2}) + \cos(-\delta_{x2})) = 0 \ . \tag{2.143}$$

In an analogous way, the second row of (2.140) can be calculated as $\mathrm{E}(x_1\sin(x_2)) = \sin(\delta_{x2})/\delta_{x2}$. In summary, the exact expected value $\mathbf{m}_y$ of the random variable $\mathbf{y}$ is obtained in the form

$$\mathbf{m}_y = \begin{bmatrix} 0 \\ \frac{\sin(\delta_{x2})}{\delta_{x2}} \end{bmatrix}. \tag{2.144}$$

The exact calculation of the expected value $\mathbf{m}_y$ is only possible for a few special cases. Note that in this example, a simple uniform distribution and a simple nonlinear transformation were assumed to be able to calculate an analytical solution. An obvious way to approximately calculate the expected value is to approximate the nonlinear transformation $\mathbf{h}(\mathbf{x})$ by a Taylor series, which was determined around the expected value $\mathbf{m}_x$ of the random vector $\mathbf{x}$.

$$\mathbf{h}(\mathbf{x}) = \mathbf{h}(\mathbf{m}_x) + \mathrm{D}_{\tilde{\mathbf{x}}}\mathbf{h} + \frac{1}{2!}\mathrm{D}^2_{\tilde{\mathbf{x}}}\mathbf{h} + \frac{1}{3!}\mathrm{D}^3_{\tilde{\mathbf{x}}}\mathbf{h} + \ldots, \tag{2.145}$$

where the notation

$$\mathrm{D}_{\tilde{\mathbf{x}}}^{k}\mathbf{h} = \left(\sum_{i=1}^{n} \tilde{x}_i \frac{\partial}{\partial x_i}\right)^{k} \mathbf{h}(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{m}_x} \tag{2.146}$$

is used. Here, n denotes the dimension of the random variable $\mathbf{x}$.

If a first-order Taylor series approximation of $\mathbf{h}(\mathbf{x})$ is used to calculate the expected value, we obtain

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{h}_l(\mathbf{x}) = \mathbf{h}(\mathbf{m}_x) + \tilde{x}_1 \frac{\partial \mathbf{h}(\mathbf{x})}{\partial x_1}\bigg|_{\mathbf{x}=\mathbf{m}_x} + \tilde{x}_2 \frac{\partial \mathbf{h}(\mathbf{x})}{\partial x_2}\bigg|_{\mathbf{x}=\mathbf{m}_x}. \tag{2.147}$$

The associated expected value is then calculated as

$$\mathrm{E}(\mathbf{h}_l(\mathbf{x})) = \mathbf{h}(\mathbf{m}_x) + \frac{\partial \mathbf{h}(\mathbf{x})}{\partial x_1}\bigg|_{\mathbf{x}=\mathbf{m}_x} \mathrm{E}(\tilde{x}_1) + \frac{\partial \mathbf{h}(\mathbf{x})}{\partial x_2}\bigg|_{\mathbf{x}=\mathbf{m}_x} \mathrm{E}(\tilde{x}_2) = \mathbf{h}(\mathbf{m}_x) \tag{2.148}$$

and thus

$$\mathrm{E}(\mathbf{h}_l(\mathbf{x})) = \mathbf{m}_{yl} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.149}$$

A comparison with the exact result (2.144) shows that the errors increase with increasing δ_{x2}. Figure 2.8 shows a comparison of the exact expected value $\mathbf{m}_y$ with the expected value $\mathbf{m}_{yl}$, which was determined based on the linearization. Since this linearization is also inherently included in the calculation of an EKF, a deviation from the exact result must also be expected for an EKF. This deviation increases with increasing nonlinearity of the system and with increasing covariance of the random variables.

One way to reduce the error due to the linearization is to use a second-order approximation $\mathbf{h}_q(\mathbf{x})$ of $\mathbf{h}(\mathbf{x})$. In this case, the approximate expected value is

$$\mathrm{E}(\mathbf{h}_q(\mathbf{x})) = \mathbf{m}_{yq} = \begin{bmatrix} 0 \\ 1 - \frac{\sigma_{x2}^2}{2} \end{bmatrix}, \tag{2.150}$$

with the variance σ_{x2} of x_2. In Figure 2.8, the expected improvement of the approximation of $\mathrm{E}(\mathbf{h}(\mathbf{x}))$ by $\mathrm{E}(\mathbf{h}_q(\mathbf{x}))$ compared to the linear approximation $\mathrm{E}(\mathbf{h}_l(\mathbf{x}))$ can be seen.

Exercise 2.17. Show the result $\mathrm{E}(\mathbf{h}_q(\mathbf{x}))$ from (2.150).

Thus, an arbitrarily accurate approximation of $\mathrm{E}(\mathbf{h}(\mathbf{x}))$ could be determined by a sufficiently high approximation order of $\mathbf{h}(\mathbf{x})$. However, this approach has two major disadvantages:

1. A very high approximation order may be necessary to approximate a general nonlinear function. This leads to a significant increase in computational effort.

2. In the calculation of the expected value with an approximation of order k, the central moments up to order k of the random vector $\mathbf{x}$ are necessary.

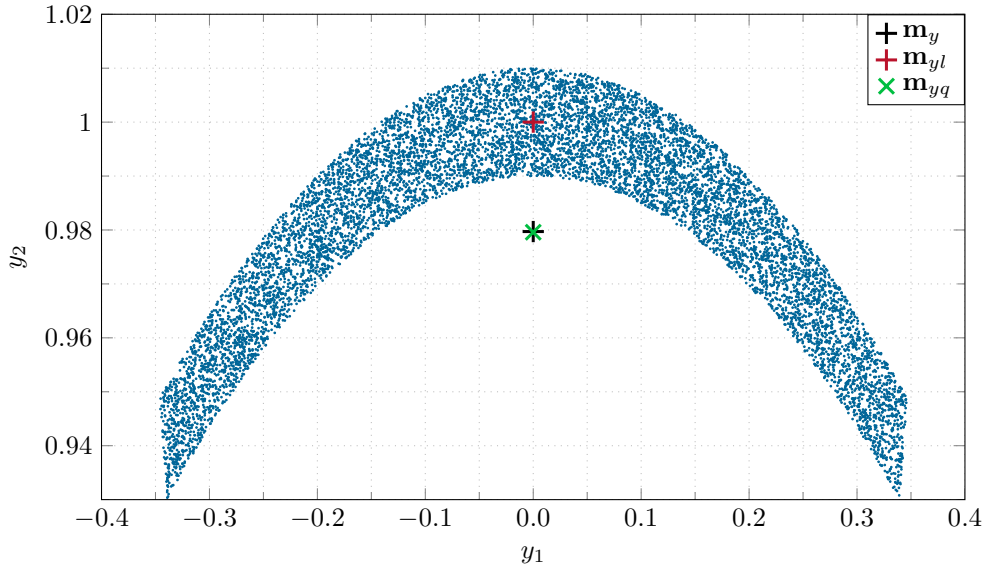


Figure 2.8: Distribution of the random vectors $\mathbf{y} = \mathbf{h}(\mathbf{x})$ with the expected value $\mathrm{E}(\mathbf{y}) = \mathbf{m}_y$, the approximation $\mathbf{m}_{yl}$ based on a 1st-order Taylor series and the approximation $\mathbf{m}_{yq}$ based on a 2nd-order Taylor series.

Therefore, this approach is only conditionally useful for practical implementation.

To characterize a (normally distributed) random variable $\mathbf{x}$, the covariance matrix $\mathrm{E}\Big((\mathbf{x}-\mathbf{m}_x)(\mathbf{x}-\mathbf{m}_x)^{\mathrm{T}}\Big) = \mathbf{Q}_x$ is also necessary. Therefore, it is interesting to investigate how the covariance matrix changes due to the nonlinear transformation. For the example considered, it holds that

$$\begin{aligned}\mathbf{Q}_y &= \mathrm{E}\left(\begin{bmatrix} x_1\cos(x_2) \\ x_1\sin(x_2) - \frac{\sin(\delta_{x2})}{\delta_{x2}} \end{bmatrix}\begin{bmatrix} x_1\cos(x_2) & x_1\sin(x_2) - \frac{\sin(\delta_{x2})}{\delta_{x2}} \end{bmatrix}\right) \\ &= \mathrm{E}\left(\begin{bmatrix} x_1^2\cos(x_2)^2 & x_1^2\cos(x_2)\sin(x_2) - x_1\cos(x_2)\frac{\sin(\delta_{x2})}{\delta_{x2}} \\ x_1^2\cos(x_2)\sin(x_2) - x_1\cos(x_2)\frac{\sin(\delta_{x2})}{\delta_{x2}} & \left(x_1\sin(x_2) - \frac{\sin(\delta_{x2})}{\delta_{x2}}\right)^2 \end{bmatrix}\right),\end{aligned} \tag{2.151}$$

which, after a short calculation, leads to

$$\mathbf{Q}_y = \begin{bmatrix} \frac{1}{2}(1+\sigma_{x1}^2)\left(1 - \frac{\sin(2\delta_{x2})}{2\delta_{x2}}\right) & 0 \\ 0 & \frac{1}{2}(1+\sigma_{x1}^2)\left(1 + \frac{\sin(2\delta_{x2})}{2\delta_{x2}}\right) - \left(\frac{\sin(\delta_{x2})}{\delta_{x2}}\right)^2 \end{bmatrix} \tag{2.152}$$

Exercise 2.18. Verify the solution (2.152).

To estimate the transformed covariance matrix $\mathbf{Q}_y$ using the first-order Taylor series approximation, the expression

$$\mathbf{y}_l - \mathrm{E}(\mathbf{y}_l) = \mathrm{D}_{\tilde{\mathbf{x}}}\mathbf{h} = \tilde{x}_1 \left.\frac{\partial \mathbf{h}}{\partial x_1}\right|_{\mathbf{x}=\mathbf{m}_x} + \tilde{x}_2 \left.\frac{\partial \mathbf{h}}{\partial x_2}\right|_{\mathbf{x}=\mathbf{m}_x}. \tag{2.153}$$

is used. This yields the approximation $\mathbf{Q}_{yl}$ of the covariance matrix as

$$\begin{aligned}\mathbf{Q}_{yl} &= \mathrm{E}\left(\left[\tilde{x}_1 \left.\frac{\partial \mathbf{h}}{\partial x_1}\right|_{\mathbf{x}=\mathbf{m}_x} + \tilde{x}_2 \left.\frac{\partial \mathbf{h}}{\partial x_2}\right|_{\mathbf{x}=\mathbf{m}_x}\right]\left[\tilde{x}_1 \left.\frac{\partial \mathbf{h}}{\partial x_1}\right|_{\mathbf{x}=\mathbf{m}_x} + \tilde{x}_2 \left.\frac{\partial \mathbf{h}}{\partial x_2}\right|_{\mathbf{x}=\mathbf{m}_x}\right]^{\mathrm{T}}\right) \\ &= \left[\left.\frac{\partial \mathbf{h}}{\partial x_1}\right|_{\mathbf{x}=\mathbf{m}_x} \quad \left.\frac{\partial \mathbf{h}}{\partial x_2}\right|_{\mathbf{x}=\mathbf{m}_x}\right] \mathrm{E}\left(\begin{bmatrix}\tilde{x}_1^2 & \tilde{x}_1\tilde{x}_2 \\ \tilde{x}_1\tilde{x}_2 & \tilde{x}_2^2\end{bmatrix}\right) \begin{bmatrix}\left.\frac{\partial \mathbf{h}}{\partial x_1}\right|_{\mathbf{x}=\mathbf{m}_x}^{\mathrm{T}} \\ \left.\frac{\partial \mathbf{h}}{\partial x_2}\right|_{\mathbf{x}=\mathbf{m}_x}^{\mathrm{T}}\end{bmatrix} = \mathbf{H}\mathbf{Q}_x\mathbf{H}^{\mathrm{T}}\end{aligned} \tag{2.154}$$

or for the example considered

$$\mathbf{Q}_{yl} = \begin{bmatrix}\sigma_{x2}^2 & 0 \\ 0 & \sigma_{x1}^2\end{bmatrix}. \tag{2.155}$$

Figure 2.9 shows a comparison of the covariance matrix $\mathbf{Q}_y$ and the approximation $\mathbf{Q}_{yl}$ in the form of the ellipses defined by them for $C = 1$.

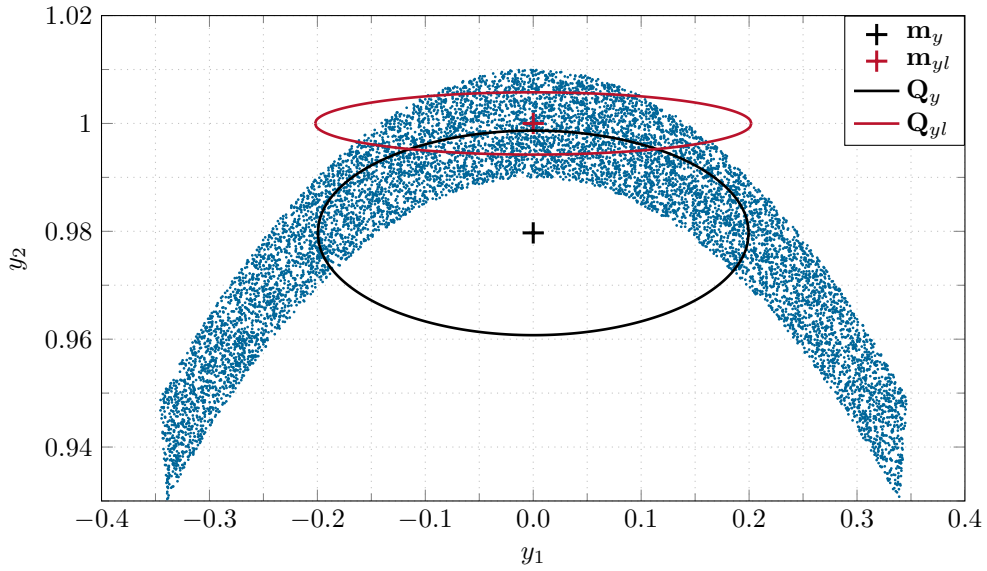


Figure 2.9: Distribution of the random vectors $\mathbf{y} = \mathbf{h}(\mathbf{x})$ with expected value $\mathrm{E}(\mathbf{y}) = \mathbf{m}_y$ and covariance matrix $\mathbf{Q}_y$, and the approximations $\mathbf{m}_{yl}$ and $\mathbf{Q}_{yl}$ based on a 1st-order Taylor series.

Theorem 2.9 (Approximation Order). *The expected value $\mathbf{m}_y$ and the covariance matrix $\mathbf{Q}_y$ of a random variable $\mathbf{y}$, which is calculated by a nonlinear transformation of the form $\mathbf{y} = \mathbf{h}(\mathbf{x})$, are given by*

$$\mathbf{m}_y = \mathbf{h}(\mathbf{m}_x) + \mathrm{E}\left(\mathrm{D}_{\tilde{\mathbf{x}}}\mathbf{h} + \frac{1}{2!}\mathrm{D}_{\tilde{\mathbf{x}}}^2\mathbf{h} + \frac{1}{3!}\mathrm{D}_{\tilde{\mathbf{x}}}^3\mathbf{h} + \frac{1}{4!}\mathrm{D}_{\tilde{\mathbf{x}}}^4\mathbf{h} + \ldots\right) \tag{2.156}$$

and

$$\mathbf{Q}_y = \mathbf{H}\mathbf{Q}_x\mathbf{H}^{\mathrm{T}} + \mathrm{E}\Big(\frac{1}{3!}\mathrm{D}_{\tilde{\mathbf{x}}}\mathbf{h}\Big(\mathrm{D}^3_{\tilde{\mathbf{x}}}\mathbf{h}\Big)^{\mathrm{T}} + \frac{1}{4}\mathrm{D}^2_{\tilde{\mathbf{x}}}\mathbf{h}\Big(\mathrm{D}^2_{\tilde{\mathbf{x}}}\mathbf{h}\Big)^{\mathrm{T}} + \frac{1}{3!}\mathrm{D}^3_{\tilde{\mathbf{x}}}\mathbf{h}(\mathrm{D}_{\tilde{\mathbf{x}}}\mathbf{h})^{\mathrm{T}}\Big) - \mathrm{E}\Big(\mathrm{D}^2_{\tilde{\mathbf{x}}}\mathbf{h}\Big)\,\mathrm{E}\Big(\mathrm{D}^2_{\tilde{\mathbf{x}}}\mathbf{h}\Big)^{\mathrm{T}} + \dots, \tag{2.157}$$

with $\mathbf{H} = \partial\mathbf{h}/\partial\mathbf{x}|_{\mathbf{x}=\mathbf{m}_x}$. *To calculate the expected value* $\mathbf{m}_y$ *with an approximation accuracy of order* m, *the partial derivatives of* $\mathbf{h}$ *and the central moments of* $\mathbf{x}$ *up to the* m*-th order are necessary. Furthermore, the term of the* m*-th order of the series of the covariance matrix can only be determined if the derivatives of* $\mathbf{h}$ *and the central moments of* $\mathbf{x}$ *up to the* $2m$*-th order are known.*

Since the EKF is largely based on this linearization, a better form of approximation of the expected value and the covariance matrix is necessary for systems with pronounced nonlinearity.

2.5.2 The Unscented Transformation

The unscented transformation is based on the fact that it is often easier to approximate the distribution of a random variable than a general nonlinear function or transformation. In the unscented transformation, a set $\mathcal{S}$ of sigma points $\boldsymbol{\xi}_i$ is chosen such that their expected value $\mathbf{m}_{xu}$ and the covariance matrix $\mathbf{Q}_{xu}$ correspond to those of the original random vector $\mathbf{x}$. Applying a nonlinear transformation $\mathbf{y} = \mathbf{h}(\mathbf{x})$ to each of these sigma points $\boldsymbol{\xi}_i$ yields the transformed sigma points $\boldsymbol{\eta}_i = \mathbf{h}(\boldsymbol{\xi}_i)$. The statistical properties of the transformed points $\boldsymbol{\eta}_i$ are then an approximation of the exact statistical properties of the nonlinear transformation.

The set of sigma points $\mathcal{S}$ is defined by the vectors $\boldsymbol{\xi}_i$ and the associated weights W_i, i.e., $\mathcal{S} = \{\boldsymbol{\xi}_i, W_i | i = 1, \dots, p\}$. The choice of sigma points is not unique, i.e., for a random vector $\mathbf{x}$ with expected value $\mathbf{m}_x$ and covariance matrix $\mathbf{Q}_x$, there are several possible realizations of sigma points. To obtain an unbiased estimate, the weights W_i must satisfy the constraint

$$\sum_{i=1}^{p} W_i = 1 \tag{2.158}$$

In the literature, for a random vector $\mathbf{x} \in \mathbb{R}^n$, a set $\mathcal{S}$ of $2n$ points lying on the ellipsoid $(\mathbf{x} - \mathbf{m}_x)\mathbf{Q}_x^{-1}(\mathbf{x} - \mathbf{m}_x)^{\mathrm{T}} = n$ is often used. The sigma points for a random vector $\mathbf{x}$ are therefore defined by

$$\boldsymbol{\xi}_i = \begin{cases} \mathbf{m}_x + \left(\sqrt{n\mathbf{Q}_x}\right)_i^{\mathrm{T}} & \text{for } i = 1, \dots, n \\ \mathbf{m}_x - \left(\sqrt{n\mathbf{Q}_x}\right)_{i-n}^{\mathrm{T}} & \text{for } i = n+1, \dots, 2n \end{cases} \tag{2.159}$$

with the dimension n, the expected value $\mathbf{m}_x$, and the covariance matrix $\mathbf{Q}_x$ of the random vector $\mathbf{x}$. Furthermore, $\left(\sqrt{n\mathbf{Q}_x}\right)_i$ denotes the i-th row of the square root of the matrix $n\mathbf{Q}_x$. The associated weights are given by

$$W_i = \frac{1}{2n} \ . \tag{2.160}$$

Remark: The square root $\mathbf{R}$ of a matrix $\mathbf{Q}$ can be determined very efficiently using the Cholesky decomposition (Matlab command `chol`). Then it holds that $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{Q}$.

Exercise 2.19. Show that the mean and covariance matrix of the sigma points defined by (2.159) and (2.160) correspond to the mean $\mathbf{m}_x$ and the covariance matrix $\mathbf{Q}_x$ of the random vector $\mathbf{x}$.

The procedure for determining the approximate expected value $\mathbf{m}_{yu}$ and the covariance matrix $\mathbf{Q}_{yu}$ using the unscented transformation consists of the following steps:

1. The nonlinear transformation $\mathbf{h}(\mathbf{x})$ is applied to each sigma point $\boldsymbol{\xi}_i$, resulting in the transformed sigma points $\boldsymbol{\eta}_i$

$$\boldsymbol{\eta}_i = \mathbf{h}(\boldsymbol{\xi}_i) \ . \tag{2.161}$$

2. The expected value $\mathbf{m}_{yu}$ results from the weighted sum of the transformed sigma points

$$\mathbf{m}_{yu} = \sum_{i=1}^{2n} W_i \boldsymbol{\eta}_i \ . \tag{2.162}$$

3. The covariance matrix $\mathbf{Q}_{yu}$ of the transformed sigma points is calculated as

$$\mathbf{Q}_{yu} = \sum_{i=1}^{2n} W_i (\boldsymbol{\eta}_i - \mathbf{m}_{yu})(\boldsymbol{\eta}_i - \mathbf{m}_{yu})^{\mathrm{T}} \ . \tag{2.163}$$

As documented in the literature, n sigma points would already suffice to correctly approximate the expected value and the covariance matrix. However, the symmetric choice of the $2n$ sigma points according to (2.159) also ensures that the third central moment is exactly fulfilled.

The approximation order of the unscented transformation is 2, i.e., the mean and the covariance matrix are correctly reproduced up to the second term. Note that for the calculation of the mean and the covariance, no higher moments of $\mathbf{x}$ or partial derivatives of the nonlinear transformation are necessary. The latter property allows the unscented transformation to be applied to non-continuously differentiable (even non-continuous) nonlinear transformations.

In the following, the application of the unscented transformation to the nonlinear transformation $\mathbf{y} = \mathbf{h}(\mathbf{x})$ from (2.138) in Section 2.5.1 is considered. The $2n = 4$ sigma points $\boldsymbol{\xi}_i$ according to (2.159) are shown in Figure 2.10 together with the exact mean $\mathbf{m}_x$ and the covariance matrix $\mathbf{Q}_x$ characterized by the ellipse for $C = \sqrt{2}$. As expected, the sigma points $\boldsymbol{\xi}_i$ are symmetrically distributed on this ellipse. Figure 2.11 shows the transformed sigma points $\boldsymbol{\eta}_i = \mathbf{h}(\boldsymbol{\xi}_i)$. Furthermore, the mean $\mathbf{m}_{yu}$ approximated according to (2.162) and the approximated covariance matrix $\mathbf{Q}_{yu}$ according to (2.163)

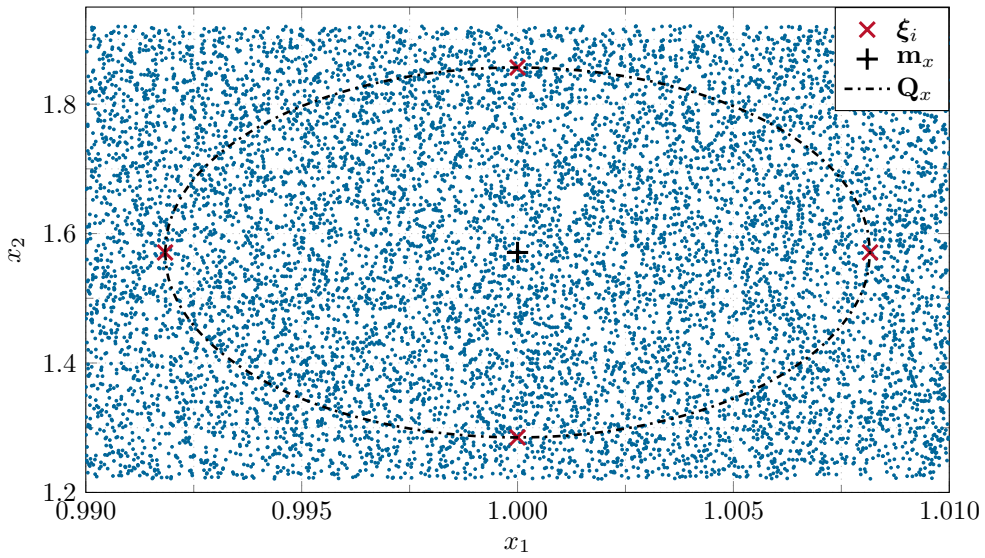


Figure 2.10: Uniformly distributed random variable $\mathbf{x}$ according to (2.138) with mean $\mathbf{m}_x$, covariance matrix $\mathbf{Q}_x$, and sigma points $\boldsymbol{\xi}_i$.

are shown. Compared to the covariance matrix $\mathbf{Q}_{yl}$ calculated using linearization from the last section 2.5.1, a drastic improvement in the approximation results, cf. Figure 2.9.

Note 2.3. As already noted, the choice of sigma points is not unique. In addition to the sigma points presented in this script based on the unscented transformation, so-called simplex sigma points or spherical sigma points are also often used in the literature, see, e.g., [2.3]. These are characterized by a somewhat simpler calculation, but have disadvantages in terms of approximation accuracy. If the random vector $\mathbf{x}$ is normally distributed, the extended choice of sigma points of the form

$$\boldsymbol{\xi}_0 = \mathbf{m}_x \tag{2.164a}$$

$$\boldsymbol{\xi}_i = \begin{cases} \mathbf{m}_x + \left(\sqrt{\frac{n}{1-\lambda}\mathbf{Q}_x}\right)_i^{\mathrm{T}} & \text{for } i = 1, \dots, n \\ \mathbf{m}_x - \left(\sqrt{\frac{n}{1-\lambda}\mathbf{Q}_x}\right)_{i-n}^{\mathrm{T}} & \text{for } i = n+1, \dots, 2n \end{cases} \tag{2.164b}$$

with the scalar parameter λ often proves useful. The associated weights are given by

$$W_0 = \lambda \tag{2.165a}$$

$$W_i = \frac{1-\lambda}{2n} \,. \tag{2.165b}$$

It can be shown that the choice $\lambda = 1 - n/3$ is optimal for normally distributed random vectors $\mathbf{x}$ [2.4]. The rationale for choosing the extended sigma points is based on an analysis of the influence of higher-order moments on the approximation of the

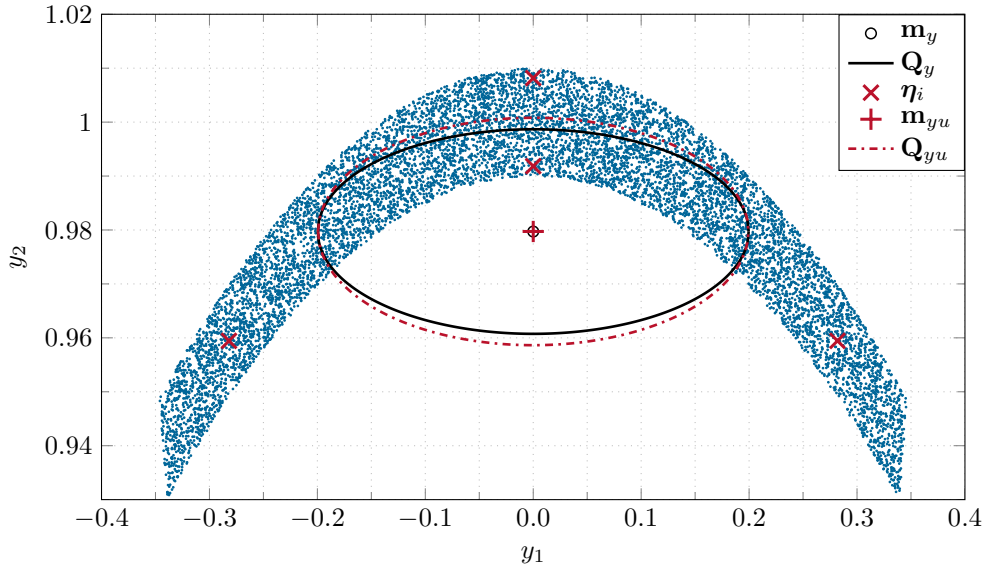


Figure 2.11: Transformed random variable $\mathbf{y} = \mathbf{h}(\mathbf{x})$ according to (2.138) with exact mean $\mathbf{m}_y$, exact covariance matrix $\mathbf{Q}_y$, transformed sigma points $\boldsymbol{\eta}_i = \mathbf{h}(\boldsymbol{\xi}_i)$, approximated mean $\mathbf{m}_{yu}$, and approximated covariance matrix $\mathbf{Q}_{yu}$.

covariance matrix. For a detailed analysis and the determination of the optimal value of λ for normally distributed random vectors, the reader is referred to the literature, in particular [2.4], [2.3], [2.5].

To analyze the influence of the extended sigma points (2.164) and the weights (2.165) on the mean and the associated covariance matrix estimated using the unscented transformation, consider again the nonlinear transformation according to (2.138). It is now assumed that the random vector $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ is defined by normally distributed random numbers x_1 (expected value $m_{x1} = 1$, variance $\sigma_{x1} = 0.01$) and x_2 (expected value $m_{x2} = \pi/2$, variance $\sigma_{x2} = 0.35$). Figure 2.12 shows the distribution of $10,000$ random vectors $\mathbf{x}$, generated in Matlab using the command `randn`, with their mean $\mathbf{m}_x$ and covariance matrix $\mathbf{Q}_x$. Furthermore, the 4 sigma points $\boldsymbol{\xi}_i$ according to (2.159) and the 5 extended sigma points $\boldsymbol{\xi}_{ei}$ according to (2.164) are plotted. The optimal choice $\lambda = 1 - 2/3 = 1/3$ for normally distributed random vectors was made. It can be seen that the extended sigma points $\boldsymbol{\xi}_{ei}$ lie on an ellipse with an increased value of C. Equivalently, the choice $\lambda < 0$ would result in the sigma points lying on an ellipse with a reduced value of C.

If the nonlinear transformation $\mathbf{y} = \mathbf{h}(\mathbf{x})$ from (2.138) is applied to the sigma points, the transformed sigma points $\boldsymbol{\eta}_i$ and the transformed extended sigma points $\boldsymbol{\eta}_{ei}$ shown in Figure 2.13 are obtained. The advantage of the extended sigma points can be seen in the approximation of the covariance matrix $\mathbf{Q}_y$. Here, the extended sigma points $\boldsymbol{\xi}_{ei}$ from (2.164) with the associated weights W_i according to (2.165) provide a significant improvement in approximation accuracy compared to the original sigma points according to (2.159).

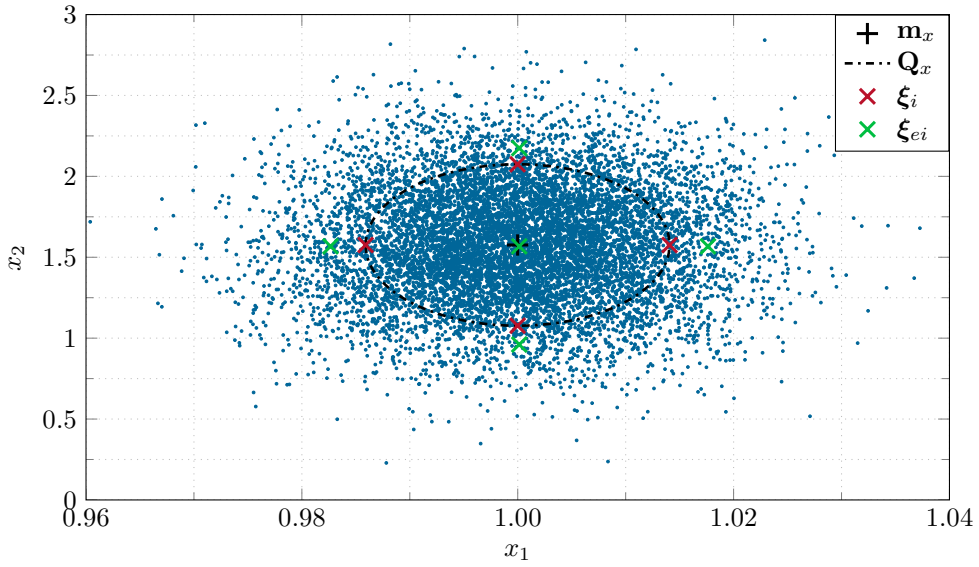


Figure 2.12: Normally distributed random vector $\mathbf{x}$ with mean $\mathbf{m}_x$ and covariance matrix $\mathbf{Q}_x$, and the sigma points $\boldsymbol{\xi}_i$ according to (2.159) and the extended sigma points $\boldsymbol{\xi}_{ei}$ according to (2.164) for $\lambda = 1/3$.

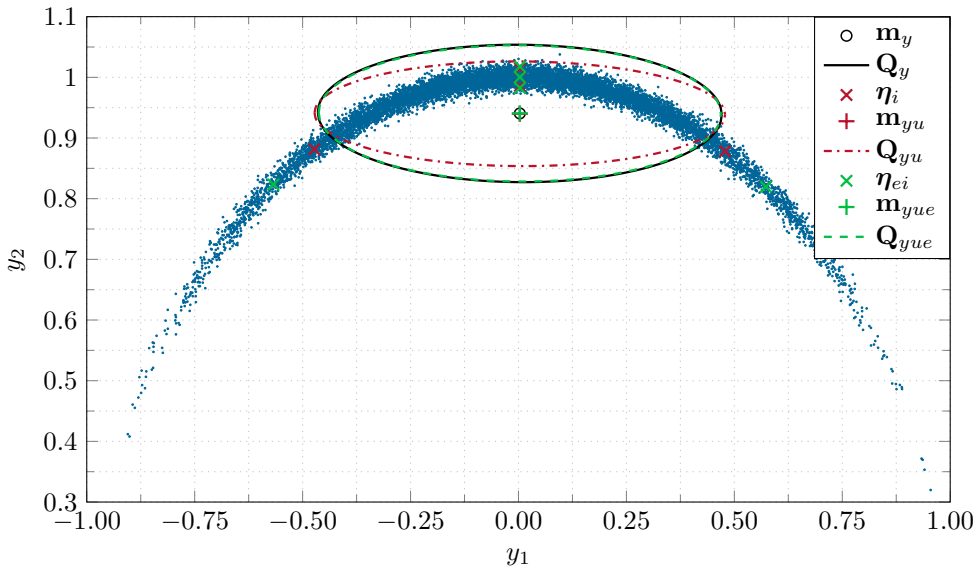


Figure 2.13: Transformed normally distributed random variable $\mathbf{y} = \mathbf{h}(\mathbf{x})$ with the exact mean $\mathbf{m}_y$ and the exact covariance matrix $\mathbf{Q}_y$. Furthermore, the approximation of the mean $\mathbf{m}_{yu}$ and the covariance matrix $\mathbf{Q}_{yu}$ based on the sigma points from (2.159) with the weights (2.160) and the approximation of the mean $\mathbf{m}_{yue}$ and the covariance matrix $\mathbf{Q}_{yue}$ based on the extended sigma points from (2.164) with the weights (2.165) are shown.

2.5.3 State Estimation of Dynamic Systems Using the Unscented Transformation

In this section, the unscented transformation is used for state estimation of nonlinear dynamic systems. In analogy to the extended Kalman filter from Section 2.4, a nonlinear, discrete-time dynamic system of the form

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k) \tag{2.166a}$$
$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k), \tag{2.166b}$$

with state $\mathbf{x}_k$, deterministic input $\mathbf{u}_k$, disturbance $\mathbf{w}_k$, and measurement noise $\mathbf{v}_k$, is considered. It is assumed that the statistical properties of $\mathbf{w}_k$ and $\mathbf{v}_k$ are known and given by

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{v}_k \mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{R}\delta_{kj} \tag{2.167a}$$
$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}\left(\mathbf{w}_k \mathbf{w}_j^{\mathrm{T}}\right) = \mathbf{Q}\delta_{kj} \tag{2.167b}$$
$$\mathrm{E}\left(\mathbf{w}_k \mathbf{v}_j^{\mathrm{T}}\right) = \mathbf{0}, \tag{2.167c}$$

with $\mathbf{Q} > 0$, $\mathbf{R} > 0$.

Furthermore, an estimate $\hat{\mathbf{x}}_0$ of the initial value $\mathbf{x}_0$ is given in the form of the expected value $\mathbf{m}_0$, i.e., $\hat{\mathbf{x}}_0 = \mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0$, and the covariance matrix of the estimation error $\mathbf{P}_0 = \mathrm{E}\left([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\right) \geq 0$ is assumed to be known. As with the Kalman filter and the extended Kalman filter, it is assumed that $\mathrm{E}\left(\mathbf{x}_0 \mathbf{w}_k^{\mathrm{T}}\right) = \mathbf{0}$ and $\mathrm{E}\left(\mathbf{x}_0 \mathbf{v}_k^{\mathrm{T}}\right) = \mathbf{0}$ hold.

For this system, a Kalman filter based on the unscented transformation, the so-called Unscented Kalman filter (also referred to as Sigma-Point Kalman filter in the literature), is now presented. Note that the disturbance $\mathbf{w}_k$ and the measurement noise $\mathbf{v}_k$ undergo a nonlinear transformation, so their statistical properties must also be captured using the unscented transformation. To this end, the augmented state $\mathbf{x}_k^a \in \mathbb{R}^{n^a}$ is defined as

$$\mathbf{x}_k^a = \begin{bmatrix} \mathbf{x}_k \\ \mathbf{w}_k \\ \mathbf{v}_k \end{bmatrix} \tag{2.168}$$

and the unscented transformation is calculated for this augmented state.

The following iteration of the Unscented Kalman filter can then be formulated; see [2.4], [2.5], [2.3] for a detailed derivation:

1. **Initialization:**

$$\hat{\mathbf{x}}_0 = \mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \tag{2.169a}$$

$$\mathbf{P}_0 = \mathrm{E}\Big([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\Big) \tag{2.169b}$$

$$\hat{\mathbf{x}}_0^{a+} = \mathrm{E}(\mathbf{x}_0^a) = \begin{bmatrix} \hat{\mathbf{x}}_0 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{2.169c}$$

$$\mathbf{P}_0^{a+} = \mathrm{E}\Big(\Big[\mathbf{x}_0^a - \hat{\mathbf{x}}_0^{a+}\Big]\Big[\mathbf{x}_0^a - \hat{\mathbf{x}}_0^{a+}\Big]^{\mathrm{T}}\Big) = \begin{bmatrix} \mathbf{P}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \tag{2.169d}$$

2. **Calculation of Sigma Points:**

$$\boldsymbol{\xi}_{0,k-1}^{a+} = \hat{\mathbf{x}}_{k-1}^{a+} \tag{2.170a}$$

$$\boldsymbol{\xi}_{i,k-1}^{a+} = \begin{cases} \hat{\mathbf{x}}_{k-1}^{a+} + \sqrt{\frac{n^a}{1-\lambda}}\Big(\sqrt{\mathbf{P}_{k-1}^{a+}}\Big)_i^{\mathrm{T}} & \text{for } i = 1, \ldots, n^a \\ \hat{\mathbf{x}}_{k-1}^{a+} - \sqrt{\frac{n^a}{1-\lambda}}\Big(\sqrt{\mathbf{P}_{k-1}^{a+}}\Big)_{i-n^a}^{\mathrm{T}} & \text{for } i = n^a+1, \ldots, 2n^a \end{cases} \tag{2.170b}$$

with

$$\boldsymbol{\xi}_{i,k-1}^{a+} = \begin{bmatrix} \boldsymbol{\xi}_{i,k-1}^{x+} \\ \boldsymbol{\xi}_{i,k-1}^{w+} \\ \boldsymbol{\xi}_{i,k-1}^{v+} \end{bmatrix}. \tag{2.171}$$

3. **State and Covariance Matrix Extrapolation:**

Predicted sigma points:

$$\boldsymbol{\xi}_{i,k}^{x-} = \mathbf{F}_{k-1}\Big(\boldsymbol{\xi}_{i,k-1}^{x+}, \mathbf{u}_{k-1}, \boldsymbol{\xi}_{i,k-1}^{w+}\Big) \tag{2.172}$$

Predicted mean and predicted error covariance matrix:

$$\hat{\mathbf{x}}_k^- = \sum_{i=0}^{2n^a} W_i \boldsymbol{\xi}_{i,k}^{x-} \tag{2.173a}$$

$$\mathbf{P}_k^- = \sum_{i=0}^{2n^a} W_i \Big(\boldsymbol{\xi}_{i,k}^{x-} - \hat{\mathbf{x}}_k^-\Big)\Big(\boldsymbol{\xi}_{i,k}^{x-} - \hat{\mathbf{x}}_k^-\Big)^{\mathrm{T}} \tag{2.173b}$$

Predicted output (measurement):

$$\boldsymbol{\eta}_{i,k}^{-} = \mathbf{h}_k\left(\boldsymbol{\xi}_{i,k}^{x-}, \mathbf{u}_k, \boldsymbol{\xi}_{i,k-1}^{v+}\right) \tag{2.174a}$$

$$\hat{\mathbf{y}}_k^{-} = \sum_{i=0}^{2n^a} W_i \boldsymbol{\eta}_{i,k}^{-} \tag{2.174b}$$

4. **Measurement Update:**

   Covariance matrix $\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k}$ between the predicted measurements and covariance matrix $\mathbf{P}_{\mathbf{x}_k\mathbf{y}_k}$ between predicted measurement and state:

$$\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k} = \sum_{i=0}^{2n^a} W_i \left(\boldsymbol{\eta}_{i,k}^{-} - \hat{\mathbf{y}}_k^{-}\right)\left(\boldsymbol{\eta}_{i,k}^{-} - \hat{\mathbf{y}}_k^{-}\right)^{\mathrm{T}} \tag{2.175a}$$

$$\mathbf{P}_{\mathbf{x}_k\mathbf{y}_k} = \sum_{i=0}^{2n^a} W_i \left(\boldsymbol{\xi}_{i,k}^{x-} - \hat{\mathbf{x}}_k^{-}\right)\left(\boldsymbol{\eta}_{i,k}^{-} - \hat{\mathbf{y}}_k^{-}\right)^{\mathrm{T}} \tag{2.175b}$$

   Measurement update of the estimated state and the error covariance:

$$\mathbf{K}_k = \mathbf{P}_{\mathbf{x}_k\mathbf{y}_k}(\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k})^{-1} \tag{2.176a}$$

$$\hat{\mathbf{x}}_k^{+} = \hat{\mathbf{x}}_k^{-} + \mathbf{K}_k\left(\mathbf{y}_k - \hat{\mathbf{y}}_k^{-}\right) \tag{2.176b}$$

$$\mathbf{P}_k^{+} = \mathbf{P}_k^{-} - \mathbf{K}_k\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k}\mathbf{K}_k^{\mathrm{T}} \tag{2.176c}$$

5. **Reset / Initialization for Next Iteration:**

$$\hat{\mathbf{x}}_k^{a+} = \begin{bmatrix} \hat{\mathbf{x}}_k^{+} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{2.177a}$$

$$\mathbf{P}_k^{a+} = \begin{bmatrix} \mathbf{P}_k^{+} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \tag{2.177b}$$

For the calculation of the Unscented Kalman filter, according to (2.170), $1 + 2(n + \dim(\mathbf{w}) + \dim(\mathbf{v}))$ sigma points are necessary. This can lead to a very large computational effort for a high system order. However, in many control engineering problems, it can be assumed that both the process noise $\mathbf{w}_k$ and the measurement noise $\mathbf{v}_k$ act additively on the system. Thus, the system (2.166) can be simplified to

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{w}_k \tag{2.178a}$$

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{v}_k, \tag{2.178b}$$

For this simplified system, the following simplified formulation of the Unscented Kalman filter can now be found, starting from the iteration equations for the general case (2.169)-(2.177):

1. **Initialization:**

$$\hat{\mathbf{x}}_0^+ = \mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \tag{2.179a}$$

$$\mathbf{P}_0^+ = \mathrm{E}\Big([\mathbf{x}_0 - \hat{\mathbf{x}}_0][\mathbf{x}_0 - \hat{\mathbf{x}}_0]^{\mathrm{T}}\Big) \tag{2.179b}$$

2. **Calculation of Sigma Points:**

$$\boldsymbol{\xi}_{0,k-1}^+ = \hat{\mathbf{x}}_{k-1}^+ \tag{2.180a}$$

$$\boldsymbol{\xi}_{i,k-1}^+ = \begin{cases} \hat{\mathbf{x}}_{k-1}^+ + \sqrt{\frac{n}{1-\lambda}}\Big(\sqrt{\mathbf{P}_{k-1}^+}\Big)_i^{\mathrm{T}} & \text{for } i = 1,\ldots,n \\ \hat{\mathbf{x}}_{k-1}^+ - \sqrt{\frac{n}{1-\lambda}}\Big(\sqrt{\mathbf{P}_{k-1}^+}\Big)_{i-n}^{\mathrm{T}} & \text{for } i = n+1,\ldots,2n \end{cases} \tag{2.180b}$$

3. **State and Covariance Matrix Extrapolation:**

Predicted sigma points:

$$\boldsymbol{\xi}_{i,k}^- = \mathbf{F}_{k-1}\Big(\boldsymbol{\xi}_{i,k-1}^+, \mathbf{u}_{k-1}\Big) \tag{2.181}$$

Predicted mean and predicted error covariance matrix:

$$\hat{\mathbf{x}}_k^- = \sum_{i=0}^{2n} W_i \boldsymbol{\xi}_{i,k}^- \tag{2.182a}$$

$$\mathbf{P}_k^- = \sum_{i=0}^{2n} W_i \Big(\boldsymbol{\xi}_{i,k}^- - \hat{\mathbf{x}}_k^-\Big)\Big(\boldsymbol{\xi}_{i,k}^- - \hat{\mathbf{x}}_k^-\Big)^{\mathrm{T}} + \mathbf{Q} \tag{2.182b}$$

Sigma point correction:

$$\boldsymbol{\xi}_{0,k}^- = \hat{\mathbf{x}}_k^- \tag{2.183a}$$

$$\boldsymbol{\xi}_{i,k}^- = \begin{cases} \hat{\mathbf{x}}_k^- + \sqrt{\frac{n}{1-\lambda}}\left(\sqrt{\mathbf{P}_k^-}\right)_i & \text{for } i = 1,\ldots,n \\ \hat{\mathbf{x}}_k^- - \sqrt{\frac{n}{1-\lambda}}\left(\sqrt{\mathbf{P}_k^-}\right)_{i-n} & \text{for } i = n+1,\ldots,2n \end{cases} \tag{2.183b}$$

Predicted output (measurement):

$$\boldsymbol{\eta}_{i,k}^- = \mathbf{h}_k\Big(\boldsymbol{\xi}_{i,k}^-, \mathbf{u}_k\Big) \tag{2.184a}$$

$$\hat{\mathbf{y}}_k^- = \sum_{i=0}^{2n} W_i \boldsymbol{\eta}_{i,k}^- \tag{2.184b}$$

4. **Measurement Update:**

   Covariance matrix $\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k}$ of the predicted measurements and covariance matrix $\mathbf{P}_{\mathbf{x}_k\mathbf{y}_k}$ between predicted measurement and state:

$$\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k} = \sum_{i=0}^{2n} W_i \left(\boldsymbol{\eta}_{i,k}^- - \hat{\mathbf{y}}_k^-\right)\left(\boldsymbol{\eta}_{i,k}^- - \hat{\mathbf{y}}_k^-\right)^{\mathrm{T}} + \mathbf{R} \tag{2.185a}$$

$$\mathbf{P}_{\mathbf{x}_k\mathbf{y}_k} = \sum_{i=0}^{2n} W_i \left(\boldsymbol{\xi}_{i,k}^- - \hat{\mathbf{x}}_k^-\right)\left(\boldsymbol{\eta}_{i,k}^- - \hat{\mathbf{y}}_k^-\right)^{\mathrm{T}} \tag{2.185b}$$

   Measurement update of the estimated state and the error covariance:

$$\mathbf{K}_k = \mathbf{P}_{\mathbf{x}_k\mathbf{y}_k}(\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k})^{-1} \tag{2.186a}$$

$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + \mathbf{K}_k\left(\mathbf{y}_k - \hat{\mathbf{y}}_k^-\right) \tag{2.186b}$$

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{K}_k\mathbf{P}_{\mathbf{y}_k\mathbf{y}_k}\mathbf{K}_k^{\mathrm{T}} \tag{2.186c}$$

Note 2.4. The sigma point correction in (2.183) is necessary because the influence of the process noise $\mathbf{w}_k$ is only considered in the predicted error covariance matrix $\mathbf{P}_k^-$ according to (2.182b) using the covariance matrix $\mathbf{Q}$. Due to this correction, the Cholesky decomposition of an $n \times n$ matrix must be calculated twice per iteration. This numerically expensive operation is often circumvented in practical implementation by using the original sigma points $\boldsymbol{\xi}_{i,k}^-$ from (2.181) in (2.184), (2.185). This results in reduced numerical effort; however, it must be checked in practical application whether the resulting errors are acceptable.

Another approach to correct the sigma points is to define an extended set of sigma points of the form

$$\boldsymbol{\xi}_{i,k}^{a-} = \boldsymbol{\xi}_{i,k}^- \quad \text{for } i = 0, \ldots, 2n \tag{2.187}$$

and

$$\boldsymbol{\xi}_{i+2n,k}^{a-} = \begin{cases} \boldsymbol{\xi}_{0,k}^- + \sqrt{\frac{2n}{1-\lambda}}(\sqrt{\mathbf{Q}})_i^{\mathrm{T}} & \text{for } i = 1, \ldots, n \\ \boldsymbol{\xi}_{0,k}^- - \sqrt{\frac{2n}{1-\lambda}}(\sqrt{\mathbf{Q}})_{i-n}^{\mathrm{T}} & \text{for } i = n+1, \ldots, 2n \end{cases}. \tag{2.188}$$

These extended sigma points $\boldsymbol{\xi}_{i,k}^{a-}$ and the weights W_i^a adapted to the new dimension $4n+1$ of $\boldsymbol{\xi}_{i,k}^{a-}$ are then used in the calculation of (2.184), (2.185). Since $\mathbf{Q}$ is a constant matrix, the calculation of the Cholesky decomposition in (2.188) is omitted, which reduces the numerical effort. On the other hand, $4n+1$ sigma points must now be considered in (2.184) and (2.185), which, compared to (2.183), again leads to an increase in numerical effort.

2.6 References

[2.1] H. van Trees, *Detection, Estimation and Modulation Theory: Part 1*. New York, USA: John Wiley & Sons, 2001.

[2.2] 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.

[2.3] D. Simon, *Optimal State Estimation*. New Jersey, USA: John Wiley & Sons, 2006.

[2.4] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[2.5] E. Wan and R. van der Merwe, "Kalman filtering and neural networks," in S. Haykin, Ed. New York, USA: John Wiley & Sons, 2001, ch. The Unscented Kalman Filter, pp. 221–280.

[2.6] G. Franklin, J. Powell, and M. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, USA: Addison–Weseley, 1998.

[2.7] L. Ljung, *System Identification*. New Jersey, USA: Prentice Hall, 1999.

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

[2.9] O. Nelles, *Nonlinear System Identification*. Berlin, Deutschland: Springer, 2001.

[2.10] R. Isermann, *Identifikation dynamischer Systeme 1 und 2*, 2nd ed. Berlin, Deutschland: Springer, 1992.

[2.11] K. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design*. New York, USA: Prentice Hall, 1997.

[2.12] A. Bryson and Y. Ho, *Applied Optimal Control*. Washington, USA: He, 1975.

[2.13] P. Dorato, C. Abdallah, and V. Cerone, *Linear Quadratic Control: An Introduction*. Florida, USA: Krieger Publishing Company, 2000.

[2.14] A. Gelb, *Applied Optimal Estimation*. Cambridge, USA: MIT Pre, 74.

3 Optimal State Feedback Controller

The goal of this chapter is the development of an *optimal state feedback controller* for linear, time-invariant systems and the combination of this state feedback controller with the optimal state observer from the last chapter. The starting point of the considerations is the linear, time-invariant, discrete-time system of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.1a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \tag{3.1b}$$

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 $\mathbf{y} \in \mathbb{R}^q$, 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 $\mathbf{D} \in \mathbb{R}^{q\times p}$. We seek a control sequence $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{N-1}$ that minimizes the cost functional

$$\begin{aligned} J(\mathbf{x}_0) &= \sum_{k=0}^{N-1} \left(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k \right) + \mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N \\ &= \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} + \mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N \end{aligned} \tag{3.2}$$

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 (3.2), this controller design is also known in the literature as the *LQR (Linear Quadratic Regulator) problem*. To solve this problem, the method of *Bellman's dynamic programming* is used.

3.1 Dynamic Programming after Bellman

The basis of dynamic programming is the *principle of optimality*:

Theorem 3.1 (Principle of Optimality)**.** *An optimal solution has the property that, starting at any point of this solution, the remaining solution is optimal in the sense of the problem to be solved, with the chosen point as the initial condition.*

Figure 3.1 illustrates Theorem 3.1. This idea is now used in the sense of *Bellman's dynamic programming* such that the optimization problem (3.2) is solved backwards starting from the final time point N. The value of the optimal control for time point N, i.e., $\mathbf{u}_{N-1}$, can be solved independently of the achieved 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. Repeating this procedure until $k = 0$, the optimal control strategy is found.

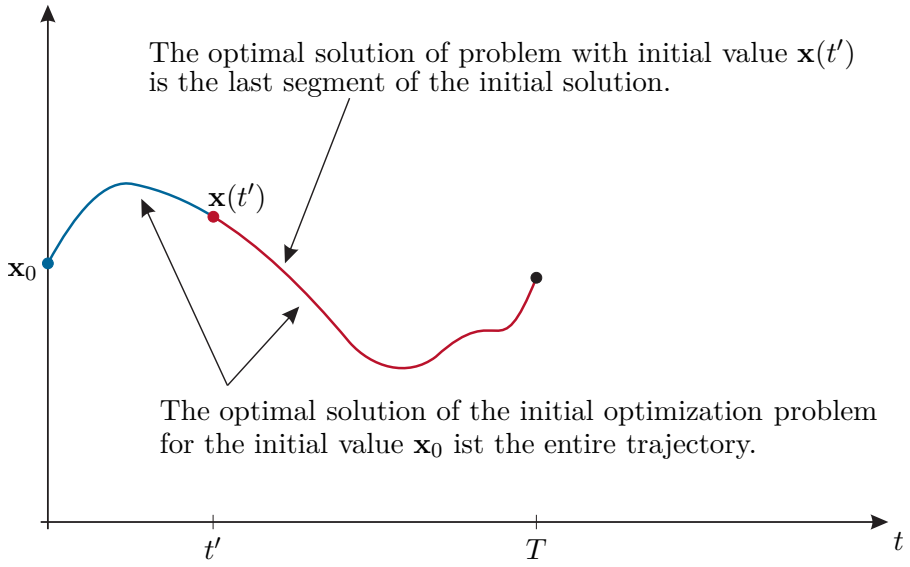


Figure 3.1: On the principle of optimality.

Since linearity of the system is not necessary for dynamic programming, the optimization problem

$$\min_{(\mathbf{u}_0,\ldots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) \quad \text{with} \quad J(\mathbf{x}_0) = \sum_{k=0}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k) + s(\mathbf{x}_N) \tag{3.3}$$

subject to the constraint

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \tag{3.4}$$

is investigated. As already mentioned, the optimization problem (3.3) is solved backwards starting from the final time point $k = N$. Since $J(\mathbf{x}_N)$ is independent of the input $\mathbf{u}$, it trivially holds that

$$J^*(\mathbf{x}_N) = s(\mathbf{x}_N), \tag{3.5}$$

where $J^*(\mathbf{x}_N)$ describes the optimal value of $J(\mathbf{x}_N)$. According to the principle of optimality, for the optimal control sequence $\mathbf{u}_0^*, \mathbf{u}_1^*, \ldots, \mathbf{u}_{N-1}^*$ with

$$J^*(\mathbf{x}_0) = \sum_{k=0}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + s(\mathbf{x}_N) \tag{3.6}$$

the following relationship also holds:

$$J^*(\mathbf{x}_0) = \sum_{k=0}^{l} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + \underbrace{\sum_{k=l+1}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k^*) + s(\mathbf{x}_N)}_{J^*(\mathbf{x}_{l+1})} \tag{3.7}$$

with

$$J^*(\mathbf{x}_{l+1}) = \min_{(\mathbf{u}_{l+1},\ldots,\mathbf{u}_{N-1})} J(\mathbf{x}_{l+1}) \tag{3.8}$$

and

$$J(\mathbf{x}_{l+1}) = \sum_{k=l+1}^{N-1} j_k(\mathbf{x}_k, \mathbf{u}_k) + s(\mathbf{x}_N) \tag{3.9}$$

subject to constraint (3.4). Note that (3.8) is solved with (3.9) for the initial value $\mathbf{x}_{l+1}$. If one now wants to go back one step based on (3.8) and determine the optimal value of the cost criterion $J^*(\mathbf{x}_l)$, then the principle of optimality yields the substitute problem

$$J^*(\mathbf{x}_l) = \min_{\mathbf{u}_l} \left(j_l(\mathbf{x}_l, \mathbf{u}_l) + J^*\Big(\underbrace{\mathbf{x}_{l+1}}_{\mathbf{f}(\mathbf{x}_l, \mathbf{u}_l)} \Big) \right) . \tag{3.10}$$

The minimum with respect to $\mathbf{u}_l$ in (3.10) can usually be determined from the relationship

$$\frac{\partial}{\partial \mathbf{u}_l}\{j_l(\mathbf{x}_l, \mathbf{u}_l) + J^*(\mathbf{f}(\mathbf{x}_l, \mathbf{u}_l))\} = \frac{\partial}{\partial \mathbf{u}_l} j_l(\mathbf{x}_l, \mathbf{u}_l) + \frac{\partial}{\partial \mathbf{z}} J^*(\mathbf{z}) \frac{\partial}{\partial \mathbf{u}_l} \mathbf{f}(\mathbf{x}_l, \mathbf{u}_l) = \mathbf{0} \tag{3.11}$$

Example 3.1. As a non-control engineering application, consider a simple assignment problem. Let an investment sum A be given, which is to be divided among N projects. Furthermore, it is assumed that assigning a sum u_k to project k yields a profit $g_k(u_k)$ for the project. The optimization problem to be solved is therefore

$$\max_{(u_0,\ldots,u_{N-1})} J(x_0) \quad \text{with} \quad J(x_0) = \sum_{k=0}^{N-1} g_k(u_k) \quad \text{subject to} \quad \sum_{k=0}^{N-1} u_k = A \; . \tag{3.12}$$

The problem can now be reformulated into an equivalent control problem of the form

$$\max_{(u_0,\ldots,u_{N-1})} J(x_0) \quad \text{with} \quad J(x_0) = \sum_{k=0}^{N-1} g_k(u_k) \tag{3.13}$$

subject to the constraint

$$x_{k+1} = x_k - u_k \quad \text{with} \quad x_0 = A \quad \text{and} \quad x_N = 0 \tag{3.14}$$

If, for example, $g_k(u_k) = \sqrt{u_k}$ is chosen, then using dynamic programming we obtain

$$\begin{aligned}
&J^*(x_N) = 0 \\
&J^*(x_{N-1}) = \max_{u_{N-1}} \{\sqrt{u_{N-1}}\} \overset{\text{subject to } x_{N-1} - u_{N-1} = 0}{=} \sqrt{x_{N-1}} && u^*_{N-1} = x_{N-1} \\
&J^*(x_{N-2}) = \max_{u_{N-2}} \{\sqrt{u_{N-2}} + \sqrt{x_{N-2} - u_{N-2}}\} = \sqrt{2x_{N-2}} && u^*_{N-2} = x_{N-2}/2 \\
&J^*(x_{N-3}) = \max_{u_{N-3}} \left\{\sqrt{u_{N-3}} + \sqrt{2(x_{N-3} - u_{N-3})}\right\} = \sqrt{3x_{N-3}} && u^*_{N-3} = x_{N-3}/3 \\
&\qquad \vdots && \qquad \vdots \\
&J^*(x_0) = \sqrt{N x_0} && u^*_0 = x_0/N \ .
\end{aligned} \tag{3.15}$$

Exercise 3.1. Interpret the result (3.15).

3.2 The LQR Problem

Applying the principle of dynamic programming to problem (3.2) with constraint (3.1), we obtain for $k = N$

$$J^*(\mathbf{x}_N) = \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{3.16}$$

and for $k = N - 1$

$$J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \left\{ \left(\mathbf{x}_{N-1}^{\mathrm{T}} \mathbf{Q} \mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{R} \mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{N} \mathbf{x}_{N-1} \right) + J^* \Big(\underbrace{\mathbf{x}_N}_{\mathbf{\Phi x}_{N-1} + \mathbf{\Gamma u}_{N-1}} \Big) \right\} \tag{3.17}$$

or

$$\begin{aligned}
J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \Big\{ &\left(\mathbf{x}_{N-1}^{\mathrm{T}} \mathbf{Q} \mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{R} \mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}} \mathbf{N} \mathbf{x}_{N-1} \right) + \\
&(\mathbf{\Phi x}_{N-1} + \mathbf{\Gamma u}_{N-1})^{\mathrm{T}} \mathbf{S} (\mathbf{\Phi x}_{N-1} + \mathbf{\Gamma u}_{N-1}) \Big\} \ .
\end{aligned} \tag{3.18}$$

Minimizing (3.18) with respect to $\mathbf{u}_{N-1}$ yields the optimal solution $\mathbf{u}^*_{N-1}$ of $\mathbf{u}_{N-1}$ as

$$\mathbf{u}^*_{N-1} = -\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} \ . \tag{3.19}$$

Substituting (3.19) into (3.18) yields

$$\begin{aligned}
&J^*(\mathbf{x}_{N-1}) \\
&= \mathbf{x}_{N-1}^{\mathrm{T}} \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} + (\mathbf{u}^*_{N-1})^{\mathrm{T}} \left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right) \mathbf{u}^*_{N-1} + 2(\mathbf{u}^*_{N-1})^{\mathrm{T}} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \mathbf{x}_{N-1} \\
&= \mathbf{x}_{N-1}^{\mathrm{T}} \Big\{ \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) - \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right)^{\mathrm{T}} \left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Gamma} \right)^{-1} \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{S} \mathbf{\Phi} \right) \Big\} \mathbf{x}_{N-1}
\end{aligned} \tag{3.20}$$

This directly allows the following theorem to be stated:

Theorem 3.2 (Linear Quadratic Regulator). *The unique solution of the optimization problem (3.2) for the linear, time-invariant, discrete-time system (3.1) with the symmetric positive semi-definite matrix* $\mathbf{S} = \mathbf{P}_N$*, the symmetric positive semi-definite matrix*

$$\mathbf{J} = \begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix} \tag{3.21}$$

and the positive definite matrix $\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{S}\mathbf{\Gamma}\right)$ *is given by the control law*

$$\mathbf{u}_k^* = \mathbf{K}_k \mathbf{x}_k \tag{3.22}$$

with

$$\mathbf{K}_k = -\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\right)^{-1}\left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\right) \tag{3.23}$$

and

$$\mathbf{P}_k = \left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\right) - \left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\right)^{\mathrm{T}}\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\right)^{-1}\left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\right) \tag{3.24}$$

The minimal value of the cost functional (3.2) is calculated as

$$\min_{(\mathbf{u}_0,\ldots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) = J^*(\mathbf{x}_0) = \mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0 \tag{3.25}$$

and $\mathbf{P}_k \geq 0$ *holds for all* $k = 0, 1, \ldots, N$*.*

Proof of Theorem 3.2. The control law (3.22), (3.23), the iteration rule (3.24), and the relationship (3.25) are obtained directly by repeated application of the dynamic programming iteration rule from equations (3.19) and (3.20). It remains to show that $\mathbf{P}_k$ is positive semi-definite for all $k = 0, 1, \ldots, N$. To do this, substitute $\mathbf{u}_{N-1}^* = \mathbf{K}_{N-1}\mathbf{x}_{N-1}$ into (3.20), which yields

$$\begin{aligned} J^*(\mathbf{x}_{N-1}) = \mathbf{x}_{N-1}^{\mathrm{T}}\Big(&\left(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\right) + \mathbf{K}_{N-1}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\right)\mathbf{K}_{N-1} \\ &+ 2\mathbf{K}_{N-1}^{\mathrm{T}}\left(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\right)\Big)\mathbf{x}_{N-1} = \mathbf{x}_{N-1}^{\mathrm{T}}\mathbf{P}_{N-1}\mathbf{x}_{N-1} \end{aligned} \tag{3.26}$$

and thus for $\mathbf{P}_k$ from (3.24)

$$\mathbf{P}_k = (\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_k)^{\mathrm{T}}\mathbf{P}_{k+1}(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_k) + \begin{bmatrix} \mathbf{E} & \mathbf{K}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{E} \\ \mathbf{K}_k \end{bmatrix}. \tag{3.27}$$

Since the matrices $\mathbf{P}_N = \mathbf{S}$ and $\mathbf{J}$ are positive semi-definite, the positive semi-definiteness of $\mathbf{P}_k$ for all $k = 0, 1, \ldots, N$ is also directly shown. □

As with the Kalman filter as an observer (see (2.91)), equation (3.24) is also a *discrete Riccati equation*, which is why the *time-varying state feedback controller* (3.22), (3.23) is

also called a *Riccati controller*. Note, however, that the discrete Riccati equation (3.24) runs *backward* in contrast to the Kalman filter! For a real-time implementation of the controller (3.22), (3.23), the final time N must therefore be known, and the matrices $\mathbf{P}_k$ and $\mathbf{K}_k$ must be pre-calculated.

If the final time $N \to \infty$, as with the Kalman filter (compare (2.93), (2.94)), a stationary solution $\mathbf{P}_s$ and $\mathbf{K}_s$ can be calculated from (3.22)–(3.24). The stationary solution $\mathbf{P}_s$ of the discrete Riccati equation (3.24) could now be determined by iterating from the initial value $\mathbf{P}_\infty = \alpha\mathbf{E}$ for $\alpha \gg 1$ until $\mathbf{P}_k$ changes only insignificantly in terms of a norm. Another solution is to solve the associated *discrete algebraic Riccati equation* for $\mathbf{P}_{k+1} = \mathbf{P}_k = \mathbf{P}_s$ in (3.24)

$$\mathbf{P}_s = \Big(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Phi}\Big) - \Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Phi}\Big)^{\mathrm{T}}\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Phi}\Big) \tag{3.28}$$

The *stationary Riccati controller*

$$\mathbf{u}_k^* = \mathbf{K}_s\mathbf{x}_k \tag{3.29a}$$

$$\mathbf{K}_s = -\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_s\mathbf{\Phi}\Big) \tag{3.29b}$$

has the structure of a classical state feedback controller according to Section 8 of the Automation script. The discrete algebraic Riccati equation (3.28) has a unique symmetric positive semi-definite solution $\mathbf{P}_s$ with the property that all eigenvalues of $(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_s)$ lie in the open interior of the unit circle if the following conditions are met:

(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

$$0 \le \mathbf{J} = \begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_\mathbf{J}^{\mathrm{T}} \\ \mathbf{D}_\mathbf{J}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{C_J} & \mathbf{D_J} \end{bmatrix} \tag{3.30}$$

is *detectable*, i.e., all eigenvalues outside the unit circle are observable via the output $\mathbf{C_J}$.

If it is now desired that all poles of the closed-loop system with the dynamic matrix $(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_s)$ not only lie inside the unit circle, but also inside a circle with radius $r < 1$ due to robustness considerations, then the controller design must be carried out for the equivalent system

$$\mathbf{x}_{k+1} = \tilde{\mathbf{\Phi}}\mathbf{x}_k + \tilde{\mathbf{\Gamma}}\mathbf{u}_k \tag{3.31}$$

with

$$\tilde{\mathbf{\Phi}} = \frac{1}{r}\mathbf{\Phi} \quad \text{and} \quad \tilde{\mathbf{\Gamma}} = \frac{1}{r}\mathbf{\Gamma} \tag{3.32}$$

Since the eigenvalues of the matrix $\Big(\tilde{\mathbf{\Phi}} + \tilde{\mathbf{\Gamma}}\mathbf{K}_s\Big)$ then lie inside the unit circle, it follows from (3.31) that the eigenvalues of $(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_s)$ lie inside a circle with radius r.

Exercise 3.2. Show that the solution of the optimization problem

$$J(\mathbf{x}_0) = \sum_{k=0}^{\infty} \frac{1}{r^{2k}} \left(\mathbf{x}_k^{\mathrm{T}} \mathbf{Q} \mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}} \mathbf{R} \mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}} \mathbf{N} \mathbf{x}_k \right) \tag{3.33}$$

with $0 < r < 1$ for the system (3.1) is given by

$$\mathbf{u}_k^* = \mathbf{K}_s \mathbf{x}_k \tag{3.34}$$

with

$$\mathbf{K}_s = -\left(r^2 \mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Gamma} \right)^{-1} \left(r^2 \mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Phi} \right) \tag{3.35}$$

and

$$r^2 \mathbf{P}_s = \left(r^2 \mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Phi} \right) - \left(r^2 \mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Phi} \right)^{\mathrm{T}} \left(r^2 \mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Gamma} \right)^{-1} \left(r^2 \mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P}_s \mathbf{\Phi} \right) \tag{3.36}$$

and that the eigenvalues of the closed-loop system's dynamic matrix $(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K}_s)$ lie inside a circle with radius $0 < r < 1$.

If in the cost functional (3.2) only the p-dimensional input $\mathbf{u}$ and the q-dimensional output $\mathbf{y}$ are to be weighted, i.e.,

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \left(\mathbf{y}_k^{\mathrm{T}} \mathbf{Q}_{\mathbf{y}} \mathbf{y}_k + \mathbf{u}_k^{\mathrm{T}} \mathbf{R} \mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}} \mathbf{N}_{\mathbf{y}} \mathbf{y}_k \right) + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{3.37a}$$

$$= \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{y}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q}_{\mathbf{y}} & \mathbf{N}_{\mathbf{y}}^{\mathrm{T}} \\ \mathbf{N}_{\mathbf{y}} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{y}_k \\ \mathbf{u}_k \end{bmatrix} + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N, \tag{3.37b}$$

then (3.37) can be transformed via the relationship $\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k$ and

$$\begin{aligned} &\sum_{k=0}^{N-1} \Big[\mathbf{x}_k^{\mathrm{T}} \underbrace{\mathbf{C}^{\mathrm{T}} \mathbf{Q}_{\mathbf{y}} \mathbf{C}}_{\tilde{\mathbf{Q}}} \mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}} \underbrace{\left(\mathbf{R} + \mathbf{D}^{\mathrm{T}} \mathbf{Q}_{\mathbf{y}} \mathbf{D} + \mathbf{N}_{\mathbf{y}} \mathbf{D} + \mathbf{D}^{\mathrm{T}} \mathbf{N}_{\mathbf{y}}^{\mathrm{T}} \right)}_{\tilde{\mathbf{R}}} \mathbf{u}_k \\ &\qquad + 2\mathbf{u}_k^{\mathrm{T}} \underbrace{\left(\mathbf{N}_{\mathbf{y}} + \mathbf{D}^{\mathrm{T}} \mathbf{Q}_{\mathbf{y}} \right) \mathbf{C}}_{\tilde{\mathbf{N}}} \mathbf{x}_k \Big] \\ &= \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{N}}^{\mathrm{T}} \\ \tilde{\mathbf{N}} & \tilde{\mathbf{R}} \end{bmatrix}}_{\tilde{\mathbf{J}}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \end{aligned} \tag{3.38}$$

into the form of (3.2)

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \tilde{\mathbf{Q}} & \tilde{\mathbf{N}}^{\mathrm{T}} \\ \tilde{\mathbf{N}} & \tilde{\mathbf{R}} \end{bmatrix}}_{\tilde{\mathbf{J}}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} + \mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N \tag{3.39}$$

and solved using Theorem 3.2.

Remark: Using the weighting matrix $\mathbf{J}$ of the cost functional (3.2) or (3.37), the behavior of the closed-loop system can be specifically influenced. As a general rule of thumb: the larger the entries of the matrix $\mathbf{R}$ (weighting of the control variables), the smaller the required control variables will be. Furthermore, by heavily weighting a specific state in $\mathbf{Q}$ or $\tilde{\mathbf{Q}}$, it can be ensured that this state decays very quickly to zero in the closed loop. Since this evaluation often proves to be very difficult in the discrete time domain, it is sensible to specify the cost functional (3.2) in continuous time and then translate it into discrete time. The idea is to evaluate the states and control variables in terms of "power" in the form

$$\int_0^T x^2(t)\mathrm{d}t \tag{3.40}$$

As a starting point, consider the linear, time-invariant, continuous-time system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{3.41a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \tag{3.41b}$$

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 corresponding discrete-time system (see Section 6 of the Automation script) is calculated for the sampling time T_a as

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \tag{3.42a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \tag{3.42b}$$

with

$$\mathbf{\Phi} = \exp(\mathbf{A}T_a) \quad \text{and} \quad \mathbf{\Gamma} = \int_0^{T_a} \exp(\mathbf{A}\tau)\mathrm{d}\tau\mathbf{B}\ . \tag{3.43}$$

Now choose a cost functional of the form

$$\begin{aligned} J(\mathbf{x}_0) &= \int_0^T \Big(\mathbf{x}^{\mathrm{T}}(t)\mathbf{Q}_c\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\mathbf{R}_c\mathbf{u}(t) + 2\mathbf{u}^{\mathrm{T}}(t)\mathbf{N}_c\mathbf{x}(t)\Big)\mathrm{d}t + \mathbf{x}^{\mathrm{T}}(T)\mathbf{S}_c\mathbf{x}(T) \\ &= \int_0^T \begin{bmatrix}\mathbf{x}^{\mathrm{T}}(t) & \mathbf{u}^{\mathrm{T}}(t)\end{bmatrix} \underbrace{\begin{bmatrix}\mathbf{Q}_c & \mathbf{N}_c^{\mathrm{T}} \\ \mathbf{N}_c & \mathbf{R}_c\end{bmatrix}}_{\mathbf{J}_c} \begin{bmatrix}\mathbf{x}(t) \\ \mathbf{u}(t)\end{bmatrix} \mathrm{d}t + \mathbf{x}^{\mathrm{T}}(T)\mathbf{S}_c\mathbf{x}(T) \end{aligned} \tag{3.44}$$

for the continuous-time system (3.41) with the symmetric, positive semi-definite weighting matrices $\mathbf{J}_c$ and $\mathbf{S}_c$ and the final time $T = NT_a$. To transform (3.44) into

the form of (3.2), first write (3.44) as follows

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \int_{kT_a}^{(k+1)T_a} \Big(\mathbf{x}^{\mathrm{T}}(t)\mathbf{Q}_c\mathbf{x}(t) + \mathbf{u}^{\mathrm{T}}(t)\mathbf{R}_c\mathbf{u}(t) + 2\mathbf{u}^{\mathrm{T}}(t)\mathbf{N}_c\mathbf{x}(t)\Big)\mathrm{d}t + \mathbf{x}^{\mathrm{T}}(NT_a)\mathbf{S}_c\mathbf{x}(NT_a) \tag{3.45}$$

Considering that the control variable $\mathbf{u}(t) = \mathbf{u}_k$ is constant in the sampling interval $kT_a \le t < (k+1)T_a$ and for the state $\mathbf{x}(t)$ it holds that

$$\mathbf{x}(t) = \underbrace{\exp(\mathbf{A}(t-kT_a))}_{\mathbf{\Phi}(t-kT_a)}\mathbf{x}_k + \underbrace{\int_{kT_a}^{t} \exp(\mathbf{A}(t-\tau))\mathrm{d}\tau\mathbf{B}}_{\mathbf{\Gamma}(t-kT_a)}\mathbf{u}_k, \quad kT_a \le t < (k+1)T_a, \tag{3.46}$$

then for (3.45) we obtain

$$J(\mathbf{x}_0) = \sum_{k=0}^{N-1} \Big(\mathbf{x}_k^T\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k\Big) + \mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N \tag{3.47}$$

with

$$\mathbf{Q} = \int_{kT_a}^{(k+1)T_a} \mathbf{\Phi}^{\mathrm{T}}(t-kT_a)\mathbf{Q}_c\mathbf{\Phi}(t-kT_a)\mathrm{d}t = \int_0^{T_a} \mathbf{\Phi}^{\mathrm{T}}(t)\mathbf{Q}_c\mathbf{\Phi}(t)\mathrm{d}t \tag{3.48a}$$

$$\mathbf{R} = \int_{kT_a}^{(k+1)T_a} \Big(\mathbf{\Gamma}^{\mathrm{T}}(t-kT_a)\mathbf{Q}_c\mathbf{\Gamma}(t-kT_a) + 2\mathbf{N}_c\mathbf{\Gamma}(t-kT_a) + \mathbf{R}_c\Big)\mathrm{d}t \tag{3.48b}$$

$$= \int_0^{T_a} \Big(\mathbf{\Gamma}^{\mathrm{T}}(t)\mathbf{Q}_c\mathbf{\Gamma}(t) + 2\mathbf{N}_c\mathbf{\Gamma}(t) + \mathbf{R}_c\Big)\mathrm{d}t$$

$$\mathbf{N} = \int_{kT_a}^{(k+1)T_a} \Big(\mathbf{\Gamma}^{\mathrm{T}}(t-kT_a)\mathbf{Q}_c + \mathbf{N}_c\Big)\mathbf{\Phi}(t-kT_a)\mathrm{d}t \tag{3.48c}$$

$$= \int_0^{T_a} \Big(\mathbf{\Gamma}^{\mathrm{T}}(t)\mathbf{Q}_c + \mathbf{N}_c\Big)\mathbf{\Phi}(t)\mathrm{d}t$$

$$\mathbf{S} = \mathbf{S}_c \tag{3.48d}$$

and

$$\mathbf{\Phi}(t) = \exp(\mathbf{A}t) \quad \text{and} \quad \mathbf{\Gamma}(t) = \int_0^t \exp(\mathbf{A}\tau)\mathrm{d}\tau\mathbf{B}\ . \tag{3.49}$$

Note that here, in general, the coupling matrix $\mathbf{N}$ is non-zero, even if $\mathbf{N}_c = \mathbf{0}$ holds.

Exercise 3.3. Compare the MATLAB commands `lqrd`, `dlqr`, and `dlqry`. What do these commands do? Look at the respective `help` text and then establish the connection to the theory presented so far.

3.3 The LQR Problem with Stochastic Disturbance

Model (3.1) according to (2.72) is now extended by the r-dimensional stochastic disturbance $\mathbf{w}$

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.50}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional deterministic input $\mathbf{u} \in \mathbb{R}^p$, and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, and $\mathbf{G} \in \mathbb{R}^{n\times r}$. The following assumptions now apply:

(1) For the disturbance $\mathbf{w}$, it is assumed that

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}) = \hat{\mathbf{Q}}\delta_{kj} \tag{3.51}$$

holds with the covariance matrix $\hat{\mathbf{Q}} \geq 0$ and the Kronecker delta $\delta_{kj} = 1$ for $k = j$ and $\delta_{kj} = 0$ otherwise.

(2) The expected value and the covariance matrix of the initial value $\mathbf{x}_0$ are given by

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{cov}(\mathbf{x}_0) = \mathrm{E}\Big((\mathbf{x}_0 - \mathbf{m}_0)(\mathbf{x}_0 - \mathbf{m}_0)^{\mathrm{T}}\Big) = \hat{\mathbf{P}}_0 \geq 0 \tag{3.52}$$

(3) The disturbance $\mathbf{w}_k$, $k \geq 0$, is uncorrelated with the initial value $\mathbf{x}_0$, i.e.,

$$\mathrm{E}(\mathbf{w}_k\mathbf{x}_0^{\mathrm{T}}) = \mathbf{0}\ . \tag{3.53}$$

Analogously to (2.76) and (2.77), because of

$$\mathbf{x}_j = \mathbf{\Phi}^j\mathbf{x}_0 + \sum_{l=0}^{j-1} \mathbf{\Phi}^l(\mathbf{\Gamma}\mathbf{u}_{j-1-l} + \mathbf{G}\mathbf{w}_{j-1-l}) \tag{3.54}$$

and (3.52) and (3.53), the relationship

$$\mathrm{E}(\mathbf{w}_k\mathbf{x}_j^{\mathrm{T}}) = \mathbf{0} \quad \text{for} \quad k \geq j \tag{3.55}$$

also follows.

Since $\mathbf{x}_k$, $k \geq 0$, is now a stochastic signal, in contrast to (3.1), the cost functional (3.2) must be replaced by

$$\begin{aligned} J(\mathbf{x}_0) &= \mathrm{E}\left(\sum_{k=0}^{N-1}\Big(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k\Big)\right) + \mathrm{E}(\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N) \\ &= \mathrm{E}\left(\sum_{k=0}^{N-1}\begin{bmatrix}\mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}}\end{bmatrix}\underbrace{\begin{bmatrix}\mathbf{Q} & \mathbf{N}^{\mathrm{T}}\\ \mathbf{N} & \mathbf{R}\end{bmatrix}}_{\mathbf{J}}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{u}_k\end{bmatrix}\right) + \mathrm{E}(\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N) \end{aligned} \tag{3.56}$$

To calculate the optimal control law, i.e., the minimization of (3.56) with respect to $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{N-1}$ subject to constraint (3.50), the method of dynamic programming is again used. For $k = N$,

$$J^*(\mathbf{x}_N) = \mathrm{E}(\mathbf{x}_N^{\mathrm{T}}\mathbf{S}\mathbf{x}_N) \tag{3.57}$$

holds, and for $k = N-1$ it follows that

$$J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \mathrm{E}\left(\mathbf{x}_{N-1}^{\mathrm{T}}\mathbf{Q}\mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}}\mathbf{R}\mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}}\mathbf{N}\mathbf{x}_{N-1} + J^*(\underbrace{\mathbf{x}_N}_{\mathbf{\Phi}\mathbf{x}_{N-1}+\mathbf{\Gamma}\mathbf{u}_{N-1}+\mathbf{G}\mathbf{w}_{N-1}}) \right) \tag{3.58}$$

with

$$J^*(\mathbf{x}_N) = \mathrm{E}\Big((\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1})^{\mathrm{T}}\mathbf{S}(\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1}) \Big) \ . \tag{3.59}$$

Minimizing (3.58) with respect to $\mathbf{u}_{N-1}$ yields the optimal solution $\mathbf{u}_{N-1}^*$ of $\mathbf{u}_{N-1}$ as

$$\mathbf{u}_{N-1}^* = \mathbf{K}_{N-1}\mathbf{x}_{N-1} \tag{3.60}$$

with

$$\mathbf{K}_{N-1} = -\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big), \tag{3.61}$$

where $\mathbf{S} = \mathbf{P}_N$ was set and the condition $\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\Big) > 0$ must be satisfied. Substituting (3.60) into (3.58) yields

$$\begin{aligned} J^*(\mathbf{x}_{N-1}) = {} & \mathrm{E}\Bigg(\mathbf{x}_{N-1}^{\mathrm{T}} \underbrace{\Big(\Big(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big) + \mathbf{K}_{N-1}^{\mathrm{T}}\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\Big)\mathbf{K}_{N-1} + 2\mathbf{K}_{N-1}^{\mathrm{T}}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big)\Big)}_{\mathbf{P}_{N-1}} \mathbf{x}_{N-1} \Bigg) \\ & + \underbrace{2\,\mathrm{E}\Big(\mathbf{w}_{N-1}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{P}_N(\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1})\Big)}_{2\mathrm{tr}(\mathbf{P}_N\mathbf{G}\,\mathrm{E}(\mathbf{w}_{N-1}\mathbf{x}_{N-1}^{\mathrm{T}})\mathbf{\Phi}^{\mathrm{T}}) + 2\,\mathrm{E}(\mathbf{w}_{N-1}^{\mathrm{T}})\mathbf{G}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\mathbf{u}_{N-1}} + \underbrace{\mathrm{E}\Big(\mathbf{w}_{N-1}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{P}_N\mathbf{G}\mathbf{w}_{N-1}\Big)}_{\mathrm{tr}(\mathbf{P}_N\mathbf{G}\,\mathrm{E}(\mathbf{w}_{N-1}\mathbf{w}_{N-1}^{\mathrm{T}})\mathbf{G}^{\mathrm{T}})} \ . \end{aligned} \tag{3.62}$$

This directly allows the following result to be stated: The optimal control law $\mathbf{u}_k^*$ for the system with stochastic disturbance (3.50) corresponds to that of the undisturbed system and is (see (3.22))

$$\mathbf{u}_k^* = \mathbf{K}_k\mathbf{x}_k \tag{3.63}$$

with

$$\mathbf{K}_k = -\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\Big) \tag{3.64a}$$

$$\mathbf{P}_k = \Big(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\Big) - \Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\Big)^{\mathrm{T}}\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi}\Big) \ . \tag{3.64b}$$

The minimal value of the cost functional (3.56) is calculated according to (3.62) as

$$\min_{(\mathbf{u}_0,\ldots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) = J^*(\mathbf{x}_0) = \mathrm{E}(\mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0) + \sum_{k=0}^{N-1} \mathrm{tr}(\mathbf{P}_{k+1}\mathbf{G}\hat{\mathbf{Q}}\mathbf{G}^{\mathrm{T}}) \ . \tag{3.65}$$

Exercise 3.4. Prove the validity of relationship (3.65). Furthermore, show that the expression $\mathrm{E}(\mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0)$ can be simplified as follows:

$$\mathrm{E}(\mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0) = \mathbf{m}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{m}_0 + \mathrm{tr}(\mathbf{P}_0\hat{\mathbf{P}}_0) \tag{3.66}$$

with

$$\hat{\mathbf{P}}_0 = \mathrm{cov}(\mathbf{x}_0) \quad \text{and} \quad \mathbf{m}_0 = \mathrm{E}(\mathbf{x}_0) \tag{3.67}$$

From (3.65), (3.66), and (3.67), it follows that the minimal value of the cost functional for the system with stochastic disturbance is greater than that of the undisturbed system (see (3.25)). Two additional terms are added: a term with $\hat{\mathbf{Q}}$ due to the stochastic disturbance $\mathbf{w}$ and a term with $\hat{\mathbf{P}}_0$ due to the stochastic nature of the initial value $\mathbf{x}_0$. In Theorem 3.2, it was shown that $\mathbf{P}_k \geq 0$ for all $k \geq 0$, provided that $\mathbf{S}$ and $\mathbf{J}$ from (3.56) are positive semi-definite. However, as can be seen from (3.65), the minimization problem with the cost functional (3.56) cannot be solved for $N \to \infty$. In order to still be able to calculate the stationary Riccati controller (3.29) even in the case of stochastic disturbance, the cost functional is set in the form

$$J(\mathbf{x}_0) = \lim_{N\to\infty} \frac{1}{N} \mathrm{E}\left(\sum_{k=0}^{N-1} \left(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k \right) \right) \tag{3.68}$$

Exercise 3.5. Show that the stationary Riccati controller (3.29) minimizes the cost functional (3.68) with the constraint (3.50) and that for the minimal value of the cost functional it holds that

$$J^*(\mathbf{x}_0) = \mathrm{tr}(\mathbf{P}_s\mathbf{G}\hat{\mathbf{Q}}\mathbf{G}^{\mathrm{T}})$$

with $\mathbf{P}_s$ as the solution of the discrete algebraic Riccati equation (3.28).

3.4 The LQG Control Problem

In the following, it is assumed that only the output $\mathbf{y}$ can be measured, while the state $\mathbf{x}$ is not available. To this end, consider the discrete-time, linear, time-invariant system underlying the Kalman filter (see (2.72)) of the form

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k + \mathbf{G}\mathbf{w}_k \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.69a}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{3.69b}$$

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 $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$, and $\mathbf{D} \in \mathbb{R}^{q\times p}$. The following assumptions now apply again:

(1) For the disturbance $\mathbf{w}$ and the measurement noise $\mathbf{v}$, it is assumed that

$$\mathrm{E}(\mathbf{v}_k) = \mathbf{0} \qquad \mathrm{E}(\mathbf{w}_k\mathbf{w}_j^{\mathrm{T}}) = \hat{\mathbf{Q}}\delta_{kj} \tag{3.70a}$$

$$\mathrm{E}(\mathbf{w}_k) = \mathbf{0} \qquad \mathrm{E}(\mathbf{v}_k\mathbf{v}_j^{\mathrm{T}}) = \hat{\mathbf{R}}\delta_{kj} \tag{3.70b}$$

$$\mathrm{E}(\mathbf{w}_k\mathbf{v}_j^{\mathrm{T}}) = \mathbf{0} \tag{3.70c}$$

with $\hat{\mathbf{Q}} \geq 0$ and $\hat{\mathbf{R}} > 0$ and the Kronecker delta $\delta_{kj} = 1$ 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

$$\mathrm{E}(\mathbf{x}_0) = \mathbf{m}_0 \qquad \mathrm{E}\Big((\mathbf{x}_0 - \hat{\mathbf{x}}_0)(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^{\mathrm{T}}\Big) = \hat{\mathbf{P}}_0 \geq 0 \tag{3.71}$$

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 uncorrelated with the initial value $\mathbf{x}_0$, i.e.,

$$\mathrm{E}(\mathbf{w}_k \mathbf{x}_0^{\mathrm{T}}) = \mathbf{0} \tag{3.72a}$$
$$\mathrm{E}(\mathbf{v}_l \mathbf{x}_0^{\mathrm{T}}) = \mathbf{0}\ . \tag{3.72b}$$

Because of (3.54) and (3.70), we also have the relationship

$$\mathrm{E}(\mathbf{w}_k \mathbf{x}_j^{\mathrm{T}}) = \mathbf{0} \quad \text{for} \quad k \geq j \tag{3.73a}$$
$$\mathrm{E}(\mathbf{v}_l \mathbf{x}_j^{\mathrm{T}}) = \mathbf{0} \quad \text{for all } l, j\ . \tag{3.73b}$$

The control task is now to minimize the cost functional

$$\begin{aligned} J(\mathbf{x}_0) &= \mathrm{E}\left(\sum_{k=0}^{N-1} \left(\mathbf{x}_k^{\mathrm{T}} \mathbf{Q} \mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}} \mathbf{R} \mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}} \mathbf{N} \mathbf{x}_k\right)\right) + \mathrm{E}(\mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N) \\ &= \mathrm{E}\left(\sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k^{\mathrm{T}} & \mathbf{u}_k^{\mathrm{T}} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{Q} & \mathbf{N}^{\mathrm{T}} \\ \mathbf{N} & \mathbf{R} \end{bmatrix}}_{\mathbf{J}} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}\right) + \mathrm{E}(\mathbf{x}_N^{\mathrm{T}} \mathbf{S} \mathbf{x}_N) \end{aligned} \tag{3.74}$$

for the positive semi-definite weighting matrices $\mathbf{J}$ and $\mathbf{S}$ with respect to $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{N-1}$ subject to constraint (3.69) such that the control law depends only on the measurable outputs $\mathbf{y}$. Denoting by $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}(k|k-1)$ the estimate of $\mathbf{x}_k$ using $0, 1, \ldots, k-1$ measurements (cf. Definition 2.3), it follows that

$$\begin{aligned} &\mathrm{E}(\mathbf{x}_k^{\mathrm{T}} \mathbf{Q} \mathbf{x}_k) = \mathrm{E}\Big((\hat{\mathbf{x}}_k - (\hat{\mathbf{x}}_k - \mathbf{x}_k))^{\mathrm{T}} \mathbf{Q} (\hat{\mathbf{x}}_k - (\hat{\mathbf{x}}_k - \mathbf{x}_k))\Big) \\ &= \mathrm{E}(\hat{\mathbf{x}}_k^{\mathrm{T}} \mathbf{Q} \hat{\mathbf{x}}_k) - 2\,\mathrm{E}(\hat{\mathbf{x}}_k^{\mathrm{T}} \mathbf{Q} (\hat{\mathbf{x}}_k - \mathbf{x}_k)) + \mathrm{E}\Big((\hat{\mathbf{x}}_k - \mathbf{x}_k)^{\mathrm{T}} \mathbf{Q} (\hat{\mathbf{x}}_k - \mathbf{x}_k)\Big) \\ &= \mathrm{E}(\hat{\mathbf{x}}_k^{\mathrm{T}} \mathbf{Q} \hat{\mathbf{x}}_k) - 2\mathrm{tr}\Big(\mathbf{Q}\,\mathrm{E}(\hat{\mathbf{x}}_k (\hat{\mathbf{x}}_k - \mathbf{x}_k)^{\mathrm{T}})\Big) + \mathrm{tr}\Big(\mathbf{Q}\,\mathrm{E}\Big((\hat{\mathbf{x}}_k - \mathbf{x}_k)(\hat{\mathbf{x}}_k - \mathbf{x}_k)^{\mathrm{T}}\Big)\Big)\ . \end{aligned} \tag{3.75}$$

In the following, only estimators are considered for which the relationship

$$\mathrm{E}(\hat{\mathbf{x}}_k (\hat{\mathbf{x}}_k - \mathbf{x}_k)^{\mathrm{T}}) = \mathbf{0} \tag{3.76}$$

holds, i.e., the second term in expression (3.75) becomes zero. According to Theorem 2.4 or Exercise 2.5, this is satisfied by the minimum-variance estimator and thus also by

the Gauss-Markov estimator and, of course, the Kalman filter, cf. Exercise 2.5. Using dynamic programming, we now obtain in a first step for $\mathbf{S} = \mathbf{P}_N$ the minimal value of the cost functional $J^*(\mathbf{x}_N)$ of $J(\mathbf{x}_N)$ as

$$J^*(\mathbf{x}_N) = \mathrm{E}(\mathbf{x}_N^{\mathrm{T}}\mathbf{P}_N\mathbf{x}_N) = \mathrm{tr}(\mathbf{P}_N\,\mathrm{E}(\mathbf{x}_N\mathbf{x}_N^{\mathrm{T}})) \tag{3.77}$$

or, with $\mathbf{x}_N = \mathbf{x}_N - \hat{\mathbf{x}}_N + \hat{\mathbf{x}}_N$,

$$J^*(\hat{\mathbf{x}}_N, \hat{\mathbf{P}}_N) = \mathrm{E}(\mathbf{x}_N^{\mathrm{T}}\mathbf{P}_N\mathbf{x}_N) = \mathrm{E}(\hat{\mathbf{x}}_N^{\mathrm{T}}\mathbf{P}_N\hat{\mathbf{x}}_N) + \mathrm{tr}(\mathbf{P}_N\hat{\mathbf{P}}_N) \tag{3.78}$$

with the covariance matrix of the estimation error

$$\hat{\mathbf{P}}_N = \mathrm{cov}(\hat{\mathbf{x}}_N - \mathbf{x}_N) = \mathrm{E}\Big((\hat{\mathbf{x}}_N - \mathbf{x}_N)(\hat{\mathbf{x}}_N - \mathbf{x}_N)^{\mathrm{T}}\Big)\,. \tag{3.79}$$

In the next step of dynamic programming, the following minimization problem must be solved:

$$J^*(\mathbf{x}_{N-1}) = \min_{\mathbf{u}_{N-1}} \mathrm{E}\left(\mathbf{x}_{N-1}^{\mathrm{T}}\mathbf{Q}\mathbf{x}_{N-1} + \mathbf{u}_{N-1}^{\mathrm{T}}\mathbf{R}\mathbf{u}_{N-1} + 2\mathbf{u}_{N-1}^{\mathrm{T}}\mathbf{N}\mathbf{x}_{N-1} + J^*(\underbrace{\mathbf{x}_N}_{\mathbf{\Phi}\mathbf{x}_{N-1}+\mathbf{\Gamma}\mathbf{u}_{N-1}+\mathbf{G}\mathbf{w}_{N-1}})\right) \tag{3.80}$$

with

$$J^*(\mathbf{x}_N) = \mathrm{E}\Big((\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1})^{\mathrm{T}}\mathbf{P}_N(\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1} + \mathbf{G}\mathbf{w}_{N-1})\Big) \tag{3.81}$$

Since $\mathbf{u}_{N-1}$ must not depend on $\mathbf{x}_{N-1}$, the minimization problem is solved with respect to an estimate $\hat{\mathbf{x}}_{N-1}$ by setting in (3.80), (3.81)

$$\mathbf{x}_{N-1} = \mathbf{x}_{N-1} - \hat{\mathbf{x}}_{N-1} + \hat{\mathbf{x}}_{N-1} \tag{3.82}$$

The control law with the controller gain matrix $\mathbf{K}_{N-1}$ is identical to (3.60) with (3.61) and is

$$\mathbf{u}_{N-1}^* = \mathbf{K}_{N-1}\hat{\mathbf{x}}_{N-1} \tag{3.83}$$

and

$$\mathbf{K}_{N-1} = -\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big) \tag{3.84}$$

with the minimal value of the cost functional

$$J^*(\mathbf{x}_{N-1}) = \mathrm{E}\left(\mathbf{x}_{N-1}^{\mathrm{T}}\underbrace{\Big(\Big(\mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big) + \mathbf{K}_{N-1}^{\mathrm{T}}\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\Big)\mathbf{K}_{N-1} + 2\mathbf{K}_{N-1}^{\mathrm{T}}\Big(\mathbf{N} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Phi}\Big)\Big)}_{\mathbf{P}_{N-1}}\mathbf{x}_{N-1}\right)$$
$$+ \underbrace{2\,\mathrm{E}\Big(\mathbf{w}_{N-1}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{P}_N(\mathbf{\Phi}\mathbf{x}_{N-1} + \mathbf{\Gamma}\mathbf{u}_{N-1})\Big)}_{2\mathrm{tr}(\mathbf{P}_N\mathbf{G}\,\mathrm{E}(\mathbf{w}_{N-1}\mathbf{x}_{N-1}^{\mathrm{T}})\mathbf{\Phi}^{\mathrm{T}}) + 2\,\mathrm{E}(\mathbf{w}_{N-1}^{\mathrm{T}})\mathbf{G}^{\mathrm{T}}\mathbf{P}_N\mathbf{\Gamma}\mathbf{u}_{N-1}} + \underbrace{\mathrm{E}\Big(\mathbf{w}_{N-1}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{P}_N\mathbf{G}\mathbf{w}_{N-1}\Big)}_{\mathrm{tr}(\mathbf{P}_N\mathbf{G}\,\mathrm{E}(\mathbf{w}_{N-1}\mathbf{w}_{N-1}^{\mathrm{T}})\mathbf{G}^{\mathrm{T}})}\,. \tag{3.85}$$

Finally, from the recursion, with $\mathrm{E}(\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}) = \hat{\mathbf{Q}}$ and (3.79), we obtain the minimal value of the cost functional for all control sequence values (compare the result with (3.65)):

$$\begin{aligned}\min_{(\mathbf{u}_0,\ldots,\mathbf{u}_{N-1})} J(\mathbf{x}_0) = J^*(\mathbf{x}_0) = \mathrm{E}\Big(\mathbf{x}_0^{\mathrm{T}}\mathbf{P}_0\mathbf{x}_0\Big) + \sum_{k=0}^{N-1} \mathrm{spur}\Big(\mathbf{P}_{k+1}\mathbf{G}\hat{\mathbf{Q}}\mathbf{G}^{\mathrm{T}}\Big) \\ + \sum_{k=0}^{N-1} \mathrm{spur}\Big(\hat{\mathbf{P}}_k\mathbf{K}_k^{\mathrm{T}}\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)\mathbf{K}_k\Big)\ .\end{aligned} \tag{3.86}$$

It can be seen that (3.86) is independent of the type of estimator, provided that it satisfies condition (3.76) and $\mathrm{E}(\mathbf{w}_j\hat{\mathbf{x}}_0^{\mathrm{T}})$. This property, together with the fact that the poles in the state observer and state controller design can be specified independently of each other, is also referred to as the *separation theorem for optimal state observer and state controller design*. Since the expression $\mathbf{K}_k^{\mathrm{T}}\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)\mathbf{K}_k$ is positive definite for all $k \geq 0$, the third term in (3.86) (or the last term in (3.85)) can also be written in the form

$$\sum_{k=0}^{N-1} \mathrm{tr}(\hat{\mathbf{P}}_k\mathbf{K}_k^{\mathrm{T}}\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)\mathbf{K}_k) = \sum_{k=0}^{N-1} \mathrm{E}\Big((\mathbf{x}_k - \hat{\mathbf{x}}_k)^{\mathrm{T}}\bar{\mathbf{K}}_k^{\mathrm{T}}\bar{\mathbf{K}}_k(\mathbf{x}_k - \hat{\mathbf{x}}_k)\Big) \tag{3.87}$$

with the Cholesky decomposition

$$\bar{\mathbf{K}}_k^{\mathrm{T}}\bar{\mathbf{K}}_k = \mathbf{K}_k^{\mathrm{T}}\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)\mathbf{K}_k \tag{3.88}$$

According to Theorem 2.5, the linear minimum-variance estimation of a linear function of a parameter vector is equivalent to the linear function of the minimum-variance estimation of the parameter vector itself, which is why the value of the cost functional (3.86) can be minimized by choosing a minimum-variance estimator. Since the Kalman filter is based on recursive minimum-variance estimation, the optimal LQG control problem is solved by combining an LQR state controller and a Kalman filter observer.

To this end, the *dynamic controller with optimal output feedback* for the system (3.69) according to (3.83)–(3.85) and Theorem 2.7 is summarized in the following form:

$$\hat{\mathbf{x}}_{k+1} = \boldsymbol{\Phi}\hat{\mathbf{x}}_k + \boldsymbol{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}_k(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k) \qquad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0 \tag{3.89a}$$

$$\mathbf{u}_k = \mathbf{K}_k\hat{\mathbf{x}}_k \tag{3.89b}$$

with

$$\mathbf{P}_k = \Big(\mathbf{Q} + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Phi}\Big) - \Big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Phi}\Big)^{\mathrm{T}}\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Phi}\Big) \tag{3.90a}$$

$$\mathbf{K}_k = -\Big(\mathbf{R} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Gamma}\Big)^{-1}\Big(\mathbf{N} + \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\boldsymbol{\Phi}\Big) \tag{3.90b}$$

$$\hat{\mathbf{P}}_{k+1} = \boldsymbol{\Phi}\hat{\mathbf{P}}_k\boldsymbol{\Phi}^{\mathrm{T}} + \mathbf{G}\hat{\mathbf{Q}}\mathbf{G}^{\mathrm{T}} - \boldsymbol{\Phi}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}} + \hat{\mathbf{R}}\Big)^{-1}\mathbf{C}\hat{\mathbf{P}}_k\boldsymbol{\Phi}^{\mathrm{T}} \tag{3.90c}$$

$$\hat{\mathbf{K}}_k = \boldsymbol{\Phi}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big(\mathbf{C}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}} + \hat{\mathbf{R}}\Big)^{-1}, \tag{3.90d}$$

the boundary condition $\mathbf{P}_N = \mathbf{S} \geq 0$, and the initial condition $\hat{\mathbf{P}}_0 \geq 0$.

It is easy to see that for $\mathbf{N} = \mathbf{0}$ and $\mathbf{G} = \mathbf{E}$ the *LQR state controller* and the *Kalman filter* are *dual* according to the definition in the Automation script. For (3.90), it then holds that

$$\mathbf{P}_k = \mathbf{Q} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\Big)^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi} \tag{3.91a}$$

$$\mathbf{K}_k = -\Big(\mathbf{R} + \mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Gamma}\Big)^{-1}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{P}_{k+1}\mathbf{\Phi} \tag{3.91b}$$

$$\hat{\mathbf{P}}_{k+1} = \hat{\mathbf{Q}} + \mathbf{\Phi}\hat{\mathbf{P}}_k\mathbf{\Phi}^{\mathrm{T}} - \mathbf{\Phi}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big(\hat{\mathbf{R}} + \mathbf{C}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big)^{-1}\mathbf{C}\hat{\mathbf{P}}_k\mathbf{\Phi}^{\mathrm{T}} \tag{3.91c}$$

$$\hat{\mathbf{K}}_k = \mathbf{\Phi}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big(\hat{\mathbf{R}} + \mathbf{C}\hat{\mathbf{P}}_k\mathbf{C}^{\mathrm{T}}\Big)^{-1}, \tag{3.91d}$$

i.e., the discrete Riccati equation of the state controller (3.91a) becomes the discrete Riccati equation of the Kalman filter (3.91c) for $\mathbf{\Phi}^{\mathrm{T}} = \mathbf{\Phi}$, $\mathbf{\Gamma} = \mathbf{C}^{\mathrm{T}}$, $\mathbf{P} = \hat{\mathbf{P}}$, $\mathbf{R} = \hat{\mathbf{R}}$, and $\mathbf{Q} = \hat{\mathbf{Q}}$. The only difference is that the Riccati equation of the state controller runs backward and that of the Kalman filter runs forward. Furthermore, for $\mathbf{\Phi}^{\mathrm{T}} = \mathbf{\Phi}$, $\mathbf{\Gamma} = \mathbf{C}^{\mathrm{T}}$, $\mathbf{P} = \hat{\mathbf{P}}$, and $\mathbf{R} = \hat{\mathbf{R}}$, the relationship $\mathbf{K}_k = \hat{\mathbf{K}}_k^{\mathrm{T}}$ holds.

As shown in Section 3.3 for the LQR controller with stochastic disturbance, the cost functional (3.74) is not meaningful for $N \to \infty$, which is why the stationary LQG problem (stationary LQR controller and stationary Kalman filter) is based on the performance criterion

$$\lim_{N\to\infty} \frac{1}{N}\,\mathrm{E}\left(\sum_{k=0}^{N-1}\Big(\mathbf{x}_k^{\mathrm{T}}\mathbf{Q}\mathbf{x}_k + \mathbf{u}_k^{\mathrm{T}}\mathbf{R}\mathbf{u}_k + 2\mathbf{u}_k^{\mathrm{T}}\mathbf{N}\mathbf{x}_k\Big)\right) \tag{3.92}$$

Note that the stationary Kalman filter and the stationary Riccati controller together

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k) \qquad \hat{\mathbf{x}}(0) = \mathbf{0} \tag{3.93a}$$

$$\mathbf{u}_k = \mathbf{K}\hat{\mathbf{x}}_k \tag{3.93b}$$

can also be written as a controller transfer matrix $\mathbf{R}_{LQG}(z)$ with the q-dimensional input $\mathbf{y}$ and the p-dimensional output $\mathbf{u}$ in the form

$$\mathbf{R}_{LQG}(z) = \frac{\mathbf{u}_z(z)}{\mathbf{y}_z(z)} = \mathbf{K}\Big(z\mathbf{E} - \Big(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K} - \hat{\mathbf{K}}(\mathbf{C} + \mathbf{D}\mathbf{K})\Big)\Big)^{-1}\hat{\mathbf{K}} \tag{3.94}$$

Exercise 3.6. Given is the simplified model of a music cassette drive according to Figure 3.2. The two DC motors can be controlled independently of each other, so that both the tape position x_3 (position of the read head) and the tensile force f_e can be controlled separately. The moment of inertia of the motors including the discs is given by $J = 6.375 \cdot 10^{-3}\,\mathrm{kgm}^2$, the radius of the discs is $r = 0.1\,\mathrm{m}$, the motor constant of the DC motors has the value $k_m = 0.544\,\mathrm{Nm/A}$, and the massless tape can be approximately modeled by a linear spring with the spring constant $c = 2.113 \cdot 10^3\,\mathrm{N/m}$ and a linear damper with the damping constant $d = 3.75\,\mathrm{Ns/m}$.

Calculate the mathematical model with the inputs $\mathbf{u}^{\mathrm{T}} = \begin{bmatrix} i_1 & i_2 \end{bmatrix}$ and the outputs $\mathbf{y}^{\mathrm{T}} = \begin{bmatrix} x_3 & f_e \end{bmatrix}$. Choose a suitable sampling time T_a and determine the corresponding discrete-time system. Design a Kalman filter and an LQR state controller using

MATLAB and simulate the result in MATLAB/SIMULINK.

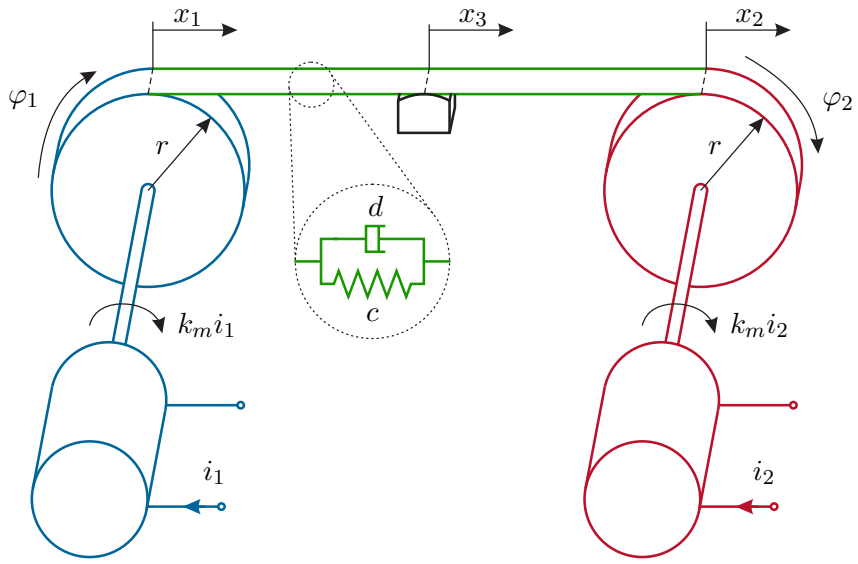


Figure 3.2: On the music cassette drive.

Remark: (for Exercise 3.6) The continuous-time mathematical model follows from the relationships

$$
\begin{aligned}
J\frac{\mathrm{d}^2}{\mathrm{d}t^2}\varphi_1 &= k_m i_1 + f_e r \\
J\frac{\mathrm{d}^2}{\mathrm{d}t^2}\varphi_2 &= k_m i_2 - f_e r \\
f_e &= c(x_2 - x_1) + d(\dot{x}_2 - \dot{x}_1) \\
x_3 &= \frac{x_1 + x_2}{2} \ .
\end{aligned}
$$

Choose the angles φ_1 and φ_2 and the corresponding angular velocities ω_1 and ω_2 as state variables.

Exercise 3.7. Figure 3.3 shows a simplified model of a magnetic bearing. The control task is to keep the mass m at a constant distance s (velocity $\dot{s} = w$) through the magnetic force f_{mag}. For the further calculation, the following parameters are given: mass $m = 500\,\text{kg}$, number of turns of the coil $N = 500$, resistance of the coil $R = 2\Omega$, permeability constant of air $\mu_0 = 4\pi 10^{-7}\,\frac{\text{Vs}}{\text{Am}}$, and the area of the air gap for calculating the magnetic force $A = 0.04\,\text{m}^2$.

Determine the nonlinear mathematical model in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{f}(\mathbf{x}, u)$$
$$y = s$$

with the state $\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} s & w & i \end{bmatrix}$, the input voltage u as input variable, and the position s as output variable. Linearize the mathematical model around the equilibrium point $(\mathbf{x}_0, u_0)$, which is determined by the stationary voltage $u_0 = 60\,\text{V}$. Design a continuous-time LQR state controller and a continuous-time Kalman filter using MATLAB and simulate the result in MATLAB/SIMULINK. Use the commands `ss`, `lqr`, `lqe`, `reg`, and `lqgreg` as well as `lqg` from the ROBUST CONTROL TOOLBOX.

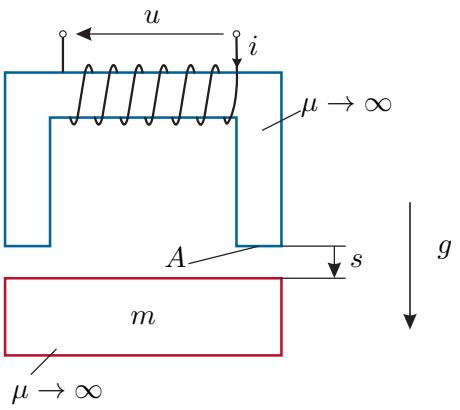


Figure 3.3: On the magnetic bearing.

Remark: (for Exercise 3.7) The nonlinear mathematical model is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} s \\ w \\ i \end{bmatrix} = \begin{bmatrix} w \\ g - \frac{ki^2}{2ms^2} \\ \frac{s}{k}\left(u - Ri + \frac{iwk}{s^2}\right) \end{bmatrix} \quad \text{with} \quad k = \frac{1}{2}A\mu_0 N^2 \ .$$

Exercise 3.8. Consider the temperature control system shown in Figure 3.4 with a constant mass flow rate $\dot{m}_{zu} = \dot{m}_{ab} = \dot{m}$. The temperature in the tank T_{tank} with the constant water quantity m_{tank} can be influenced via the mixing valve with the

output temperature T_m. For the temperature directly at the tank inlet T_{zu}, due to the neglect of heat losses in the pipeline,

$$T_{zu} = T_m(t - T_t)$$

holds with the dead time T_t caused by the transport process through the pipeline. Determine the mathematical model of the temperature control system with the state variable $x = T_{tank}$, the input variable $u = T_m$, and the output variable $y = T_{ab} = T_{tank}$. Furthermore, generally calculate the transport dead time T_t for a frictionless pipeline with inner diameter D and length L for given mass flow rate $\dot{m}$ and density ρ of the liquid. The energy E_{tank} stored in the water quantity m_{tank} with temperature T_{tank} and the specific heat capacity of water c_W is

$$E_{tank} = c_W m_{tank} T_{tank} \ .$$

Choose suitable parameters and implement the mathematical model in MATLAB/SIMULINK. For a suitable sampling time T_a, design a Kalman filter and an LQR state controller and simulate the closed-loop control system with the continuous-time plant in MATLAB/SIMULINK.

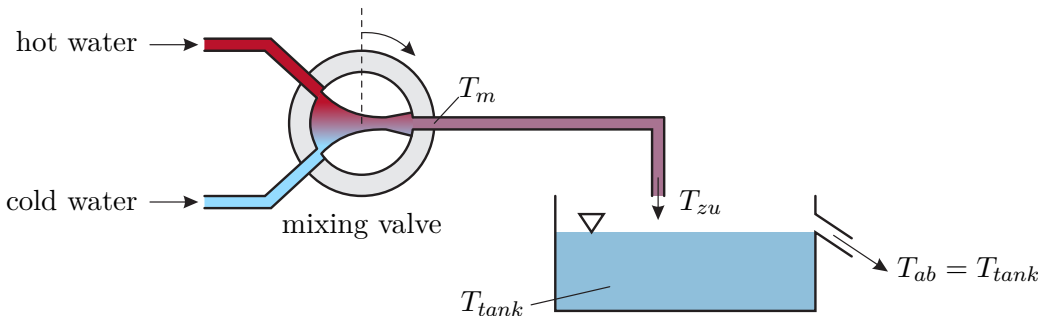


Figure 3.4: For Exercise 3.8 on the temperature control system.

3.5 Extended Concepts of State Control

3.5.1 Feedforward of the Estimated Disturbance

It was already shown in the Kalman filter (see Section 2.3.2) how deterministic disturbances can be systematically taken into account by adding a disturbance model. In the following, this concept is extended to the combined state controller and state observer design. Consider the linear, time-invariant, continuous-time system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{G}\mathbf{w} \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.95a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{3.95b}$$

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 deterministic disturbance $\mathbf{w} \in \mathbb{R}^r$, and the matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{G} \in \mathbb{R}^{n\times r}$, and $\mathbf{C} \in \mathbb{R}^{q\times n}$. It is assumed that the disturbance $\mathbf{w}$ can be described by the disturbance model

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z} = \mathbf{A}_\mathbf{z}\mathbf{z} \qquad \mathbf{z}(0) = \mathbf{z}_0 \tag{3.96a}$$

$$\mathbf{w} = \mathbf{C}_\mathbf{z}\mathbf{z} \tag{3.96b}$$

In a first step, the corresponding discrete-time system for the sampling time T_a is calculated for the extended system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{G}\mathbf{C}_\mathbf{z}\\ \mathbf{0} & \mathbf{A}_\mathbf{z}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} + \begin{bmatrix}\mathbf{B}\\ \mathbf{0}\end{bmatrix}\mathbf{u} \tag{3.97a}$$

$$\mathbf{y} = \begin{bmatrix}\mathbf{C} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} \tag{3.97b}$$

in the form

$$\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{z}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi} & \mathbf{\Phi}_{\mathbf{xz}}\\ \mathbf{0} & \mathbf{\Phi}_\mathbf{z}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}\\ \mathbf{0}\end{bmatrix}\mathbf{u}_k \tag{3.98a}$$

$$\mathbf{y}_k = \begin{bmatrix}\mathbf{C} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} \tag{3.98b}$$

Exercise 3.9. Show that the discrete-time system of (3.97) must have the structure of (3.98).

As can be seen, the state of the disturbance model $\mathbf{z}$ is not reachable via the input $\mathbf{u}$. If the system (3.98) is observable, the state $\mathbf{x}$ and the state of the disturbance model $\mathbf{z}$ can be observed via an observer

$$\begin{bmatrix}\hat{\mathbf{x}}_{k+1}\\ \hat{\mathbf{z}}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi} & \mathbf{\Phi}_{\mathbf{xz}}\\ \mathbf{0} & \mathbf{\Phi}_\mathbf{z}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{x}}_k\\ \hat{\mathbf{z}}_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}\\ \mathbf{0}\end{bmatrix}\mathbf{u}_k + \begin{bmatrix}\hat{\mathbf{K}}\\ \hat{\mathbf{K}}_\mathbf{z}\end{bmatrix}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) \tag{3.99}$$

In the next step, a state controller

$$\mathbf{u}_k = \mathbf{K}\mathbf{x}_k \tag{3.100}$$

is designed for the system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \tag{3.101}$$

and extended and implemented in the form

$$\mathbf{u}_k = \mathbf{K}\hat{\mathbf{x}}_k + \mathbf{K}_\mathbf{z}\hat{\mathbf{z}}_k \tag{3.102}$$

Thus, the closed-loop system (3.98), (3.99), and (3.102) is

$$\mathbf{x}_{k+1} = (\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K})\mathbf{x}_k + (\mathbf{\Phi}_{\mathbf{xz}} + \mathbf{\Gamma}\mathbf{K}_\mathbf{z})\mathbf{z}_k - \mathbf{\Gamma}\mathbf{K}\tilde{\mathbf{x}}_k - \mathbf{\Gamma}\mathbf{K}_\mathbf{z}\tilde{\mathbf{z}}_k \tag{3.103a}$$

$$\mathbf{z}_{k+1} = \mathbf{\Phi}_\mathbf{z}\mathbf{z}_k \tag{3.103b}$$

with the dynamics of the observation errors $\tilde{\mathbf{x}}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k$ and $\tilde{\mathbf{z}}_k = \mathbf{z}_k - \hat{\mathbf{z}}_k$

$$\begin{bmatrix} \tilde{\mathbf{x}}_{k+1} \\ \tilde{\mathbf{z}}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi} - \hat{\mathbf{K}}\mathbf{C} & \mathbf{\Phi}_{\mathbf{xz}} \\ -\hat{\mathbf{K}}_\mathbf{z}\mathbf{C} & \mathbf{\Phi}_\mathbf{z} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_k \\ \tilde{\mathbf{z}}_k \end{bmatrix} . \tag{3.104}$$

Note that the controller matrix $\mathbf{K}$ determines the dynamics with which an initial error $\mathbf{x}(0)$ is controlled to zero with vanishing disturbance, i.e., $\mathbf{w} = \mathbf{0}$. Furthermore, the observer matrices $\hat{\mathbf{K}}$ and $\hat{\mathbf{K}}_\mathbf{z}$ determine the error dynamics, and with the help of $\mathbf{K}_\mathbf{z}$ the influence of the disturbance $\mathbf{w}$ can be specifically suppressed. As can be seen from (3.103), the optimal choice for $\mathbf{K}_\mathbf{z}$, if possible, is given by

$$\mathbf{\Phi}_{\mathbf{xz}} + \mathbf{\Gamma}\mathbf{K}_\mathbf{z} = \mathbf{0} \tag{3.105}$$

This strategy for disturbance suppression is also called *feedforward of the estimated disturbance*. Figure 3.5 shows the corresponding block diagram.

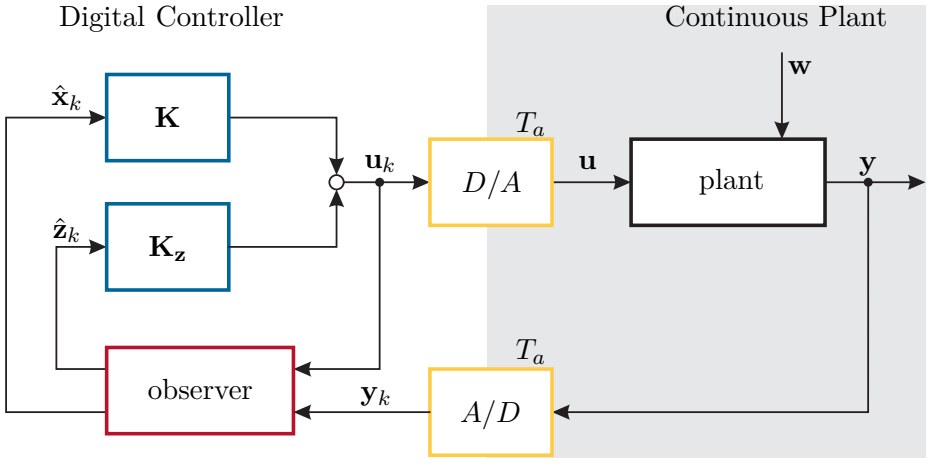


Figure 3.5: Block diagram for feedforward of the estimated disturbance.

3.5.2 State Controller and State Observer Design with Integral Action

Now assume that the disturbance $\mathbf{w}$ in (3.95) is constant but unknown. Then the corresponding disturbance model (3.96) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z} = \mathbf{0} \qquad \mathbf{z}(0) = \mathbf{w} \tag{3.106a}$$

$$\mathbf{w} = \mathbf{z} \tag{3.106b}$$

and the discrete-time system (3.98) becomes

$$\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{z}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi} & \mathbf{\Phi_{xz}}\\ \mathbf{0} & \mathbf{E}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}\\ \mathbf{0}\end{bmatrix}\mathbf{u}_k \tag{3.107a}$$

$$\mathbf{y}_k = \begin{bmatrix}\mathbf{C} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} \tag{3.107b}$$

with

$$\mathbf{\Phi} = \exp(\mathbf{A}T_a) \tag{3.108a}$$

$$\mathbf{\Phi_{xz}} = \int_0^{T_a} \exp(\mathbf{A}\tau)\mathrm{d}\tau\mathbf{G} \tag{3.108b}$$

$$\mathbf{\Gamma} = \int_0^{T_a} \exp(\mathbf{A}\tau)\mathrm{d}\tau\mathbf{B}\ . \tag{3.108c}$$

Exercise 3.10. Show the validity of (3.108).

Now assume that $\mathbf{\Phi_{xz}} = \mathbf{\Gamma}$ or $\mathbf{G} = \mathbf{B}$ holds, i.e., the disturbance $\mathbf{z}_k$ acts on the system before the input $\mathbf{u}_k$ in the sampled system, then (3.105) can be solved exactly. It holds that $\mathbf{K_z} = -\mathbf{E}$. Thus, the controller from (3.102) is

$$\mathbf{u}_k = \mathbf{K}\hat{\mathbf{x}}_k - \hat{\mathbf{z}}_k \tag{3.109}$$

and the observer from (3.99) has the form

$$\begin{bmatrix}\hat{\mathbf{x}}_{k+1}\\ \hat{\mathbf{z}}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi} & \mathbf{\Gamma}\\ \mathbf{0} & \mathbf{E}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{x}}_k\\ \hat{\mathbf{z}}_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}\\ \mathbf{0}\end{bmatrix}\mathbf{u}_k + \begin{bmatrix}\hat{\mathbf{K}}\\ \hat{\mathbf{K}}_\mathbf{z}\end{bmatrix}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k)\ . \tag{3.110}$$

Rewriting (3.109) and (3.110), we obtain

$$\hat{\mathbf{x}}_{k+1} = (\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K})\hat{\mathbf{x}}_k + \hat{\mathbf{K}}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) \tag{3.111a}$$

$$\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{z}}_k + \hat{\mathbf{K}}_\mathbf{z}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) \tag{3.111b}$$

$$\mathbf{u}_k = \mathbf{K}\hat{\mathbf{x}}_k - \hat{\mathbf{z}}_k, \tag{3.111c}$$

showing that the controller consists of the feedback of the estimated state $\hat{\mathbf{x}}_k$ with the controller matrix $\mathbf{K}$ and the integrated output error weighted by the matrix $\hat{\mathbf{K}}_\mathbf{z}$. Figure 3.6 shows the corresponding controller structure in the form of a block diagram. In summary, the design with integral action looks as follows:

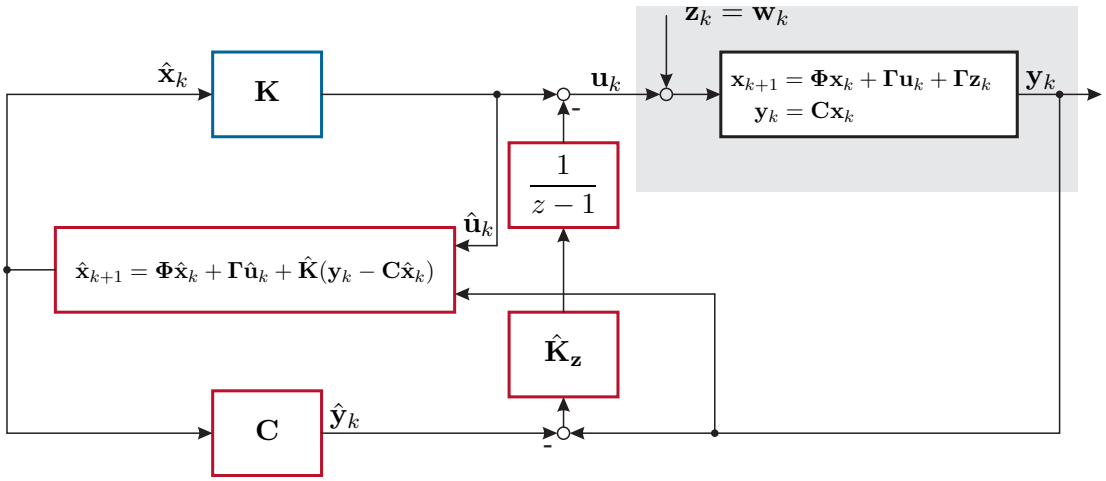


Figure 3.6: State controller/state observer design with integral action.

(1) Design a state controller of the form

$$\mathbf{u}_k = \mathbf{K}\mathbf{x}_k \ . \tag{3.112}$$

for the system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \tag{3.113a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k \tag{3.113b}$$

(2) In the second step, extend the system (3.113) by a constant, but unknown disturbance $\mathbf{w}$ acting at the system input, i.e.,

$$\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{z}_{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi} & \mathbf{\Gamma}\\ \mathbf{0} & \mathbf{E}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} + \begin{bmatrix}\mathbf{\Gamma}\\ \mathbf{0}\end{bmatrix}\mathbf{u}_k \qquad \mathbf{z}(0) = \mathbf{w} \tag{3.114a}$$

$$\mathbf{y}_k = \begin{bmatrix}\mathbf{C} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_k\\ \mathbf{z}_k\end{bmatrix} \tag{3.114b}$$

and design the observer gain matrices $\hat{\mathbf{K}}$ and $\hat{\mathbf{K}}_\mathbf{z}$ of a full observer (see (3.110)).

(3) The controller then follows according to (3.109) as

$$\mathbf{u}_k = \mathbf{K}\hat{\mathbf{x}}_k - \hat{\mathbf{z}}_k \ . \tag{3.115}$$

3.5.3 State Controller and State Observer Design with Setpoints

This section shows how setpoints can be systematically taken into account in state controller and state observer design. In the English-language literature, this problem is often referred to as the *servo problem*. It is assumed that the variables $\bar{\mathbf{y}}_k \in \mathbb{R}^p$

$$\bar{\mathbf{y}}_k = \mathbf{C}_\mathbf{r}\mathbf{x}_k \tag{3.116}$$

of the system

$$\mathbf{x}_{k+1} = \mathbf{\Phi}\mathbf{x}_k + \mathbf{\Gamma}\mathbf{u}_k \tag{3.117a}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k, \tag{3.117b}$$

with the n-dimensional state $\mathbf{x} \in \mathbb{R}^n$, the p-dimensional input $\mathbf{u} \in \mathbb{R}^p$, the q-dimensional output $\mathbf{y} \in \mathbb{R}^q$, and the matrices $\mathbf{\Phi} \in \mathbb{R}^{n\times n}$, $\mathbf{\Gamma} \in \mathbb{R}^{n\times p}$, and $\mathbf{C} \in \mathbb{R}^{q\times n}$, are to be controlled to a given stationary reference value $\mathbf{r}_s \in \mathbb{R}^p$

$$\bar{\mathbf{y}}_s = \mathbf{C}_\mathbf{r}\mathbf{x}_s = \mathbf{r}_s \tag{3.118}$$

In a first step, a state controller

$$\mathbf{u}_k = \mathbf{K}\mathbf{x}_k \tag{3.119}$$

is designed, for example, as a stationary Riccati controller, and in a second step it is extended in the form

$$\mathbf{u}_k = -\mathbf{K}(\mathbf{L}_\mathbf{r}\mathbf{r}_k - \mathbf{x}_k) + \mathbf{L}_\mathbf{u}\mathbf{r}_k \tag{3.120}$$

with $\mathbf{L}_\mathbf{r} \in \mathbb{R}^{n\times p}$ and $\mathbf{L}_\mathbf{u} \in \mathbb{R}^{p\times p}$. The matrices $\mathbf{L}_\mathbf{r}$ and $\mathbf{L}_\mathbf{u}$ are to satisfy the following conditions for the stationary state

$$\mathbf{L}_\mathbf{r}\mathbf{r}_s = \mathbf{x}_s \tag{3.121a}$$
$$\mathbf{L}_\mathbf{u}\mathbf{r}_s = \mathbf{u}_s \tag{3.121b}$$

From (3.117), it follows in the stationary state that

$$(\mathbf{E} - \mathbf{\Phi})\mathbf{x}_s - \mathbf{\Gamma}\mathbf{u}_s = \mathbf{0} \tag{3.122}$$

and substituting (3.121) into (3.118) and (3.122), we obtain

$$((\mathbf{E} - \mathbf{\Phi})\mathbf{L}_\mathbf{r} - \mathbf{\Gamma}\mathbf{L}_\mathbf{u})\mathbf{r}_s = \mathbf{0} \tag{3.123a}$$
$$\bar{\mathbf{y}}_s = \mathbf{C}_\mathbf{r}\mathbf{L}_\mathbf{r}\mathbf{r}_s = \mathbf{r}_s \tag{3.123b}$$

or, for $\mathbf{r}_s \neq \mathbf{0}$,

$$\underbrace{\begin{bmatrix} \mathbf{E} - \mathbf{\Phi} & -\mathbf{\Gamma} \\ \mathbf{C}_\mathbf{r} & \mathbf{0} \end{bmatrix}}_{\mathbf{X}} \begin{bmatrix} \mathbf{L}_\mathbf{r} \\ \mathbf{L}_\mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{E} \end{bmatrix} . \tag{3.124}$$

If the matrix $\mathbf{X}$ is invertible, then the matrices $\mathbf{L}_\mathbf{r}$ and $\mathbf{L}_\mathbf{u}$ can be calculated from (3.124). Figure 3.7 shows the corresponding controller structure.

It can be seen that (3.120) can also be simplified in the form

$$\mathbf{u}_k = \mathbf{K}\mathbf{x}_k + \mathbf{L}\mathbf{r}_k \quad \text{with} \quad \mathbf{L} = \mathbf{L}_\mathbf{u} - \mathbf{K}\mathbf{L}_\mathbf{r} \tag{3.125}$$

Now assume that the state $\mathbf{x}_k$ is not measurable. Then a state observer of the form

$$\hat{\mathbf{x}}_{k+1} = \mathbf{\Phi}\hat{\mathbf{x}}_k + \mathbf{\Gamma}\mathbf{u}_k + \hat{\mathbf{K}}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) \tag{3.126}$$

is additionally required. The closed-loop system (3.117), (3.125), and (3.126) is then given with the observation error $\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k$ as

$$\mathbf{x}_{k+1} = (\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{K})\mathbf{x}_k - \mathbf{\Gamma}\mathbf{K}\mathbf{e}_k + \mathbf{\Gamma}\mathbf{L}\mathbf{r}_k \tag{3.127a}$$
$$\mathbf{e}_{k+1} = \left(\mathbf{\Phi} - \hat{\mathbf{K}}\mathbf{C}\right)\mathbf{e}_k \tag{3.127b}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k \ . \tag{3.127c}$$

~~As expected, the observation error $\mathbf{e}_k$ is not reachable via the reference input $\mathbf{r}_k$.~~

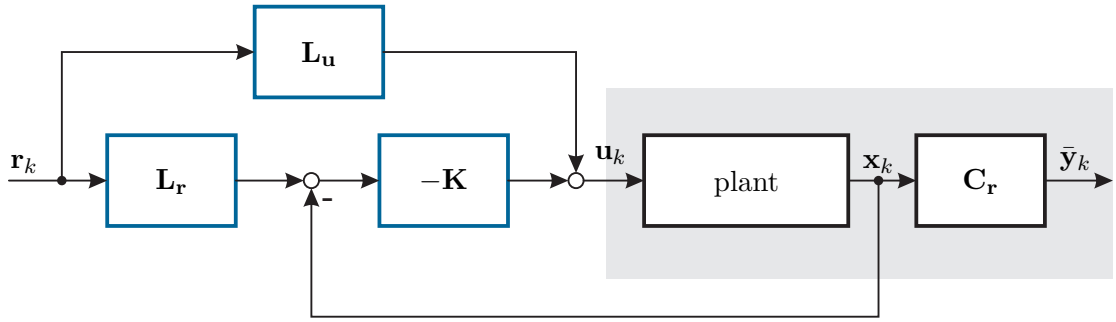


Figure 3.7: Controller structure for state control with setpoints.

Exercise 3.11. In an analogous way to that shown in Section 3.5.2, the state controller and state observer design with setpoints can be extended by an integral action. Show that in this case the control law is calculated as follows:

$$\begin{aligned}\hat{\mathbf{x}}_{k+1} &= (\boldsymbol{\Phi} + \boldsymbol{\Gamma}\mathbf{K})\hat{\mathbf{x}}_k + \hat{\mathbf{K}}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) + \boldsymbol{\Gamma}\mathbf{L}\mathbf{r}_k \\ \hat{\mathbf{z}}_{k+1} &= \hat{\mathbf{z}}_k + \hat{\mathbf{K}}_{\mathbf{z}}(\mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_k) \\ \mathbf{u}_k &= \mathbf{K}\hat{\mathbf{x}}_k - \hat{\mathbf{z}}_k + \mathbf{L}\mathbf{r}_k\end{aligned}$$

3.5.4 The Feedforward Concept

If the controlled system is to behave like a reference model of the form

$$\bar{\mathbf{x}}_{k+1} = \boldsymbol{\Phi}_m\bar{\mathbf{x}}_k + \boldsymbol{\Gamma}_m\mathbf{r}_k \tag{3.128a}$$

$$\bar{\mathbf{y}}_k = \mathbf{C}_m\bar{\mathbf{x}}_k \tag{3.128b}$$

then the following procedure can be chosen: Simulate the reference model (3.128) on the computer and set the control law as follows:

$$\mathbf{u}_k = -\mathbf{K}(\bar{\mathbf{x}}_k - \mathbf{x}_k) + \bar{\mathbf{u}}_k \tag{3.129}$$

The so-called *(control variable) feedforward signal* $\bar{\mathbf{u}}_k$ is determined such that, with ideal agreement between the reference model and the controlled system, i.e., $\bar{\mathbf{x}}_k = \mathbf{x}_k$, the outputs of the reference model and the controlled system also agree, i.e., $\bar{\mathbf{y}}_k = \mathbf{y}_k$. Calculating the feedforward signal in the multivariable case is generally relatively difficult. In the single-variable case, using the z-transfer functions of (3.117) and (3.128)

$$G(z) = \frac{y_z(z)}{u_z(z)} = \mathbf{c}^{\mathrm{T}}(z\mathbf{E} - \mathbf{\Phi})^{-1}\mathbf{\Gamma} \tag{3.130a}$$

$$G_m(z) = \frac{\bar{y}_z(z)}{r_z(z)} = \mathbf{c}_m^{\mathrm{T}}(z\mathbf{E} - \mathbf{\Phi}_m)^{-1}\mathbf{\Gamma}_m \tag{3.130b}$$

from the condition

$$G(z)u_z(z) = y_z(z) = \bar{y}_z(z) = G_m(z)r_z(z) \tag{3.131}$$

the z-transform $\bar{u}_z(z)$ of the feedforward signal $(\bar{u}_k)$ can be calculated as follows:

$$\bar{u}_z(z) = \frac{G_m(z)}{G(z)} r_z(z) \tag{3.132}$$

From (3.132), it can be seen that this is only possible if the degree difference of $G_m(z)$ is greater than or equal to the degree difference of $G(z)$, $G_m(z)$ is BIBO-stable, and all zeros outside the closed unit circle of $G(z)$ are also zeros of $G_m(z)$. If the numerator polynomials of $G(z)$ and $G_m(z)$ are identical and the orders of the denominator polynomials are equal, i.e.,

$$G(z) = \frac{z_G(z)}{n_G(z)} = \frac{b_0 + b_1 z + \cdots + b_{n-1}z^{n-1} + b_n z^n}{a_0 + a_1 z + \cdots + a_{n-1}z^{n-1} + a_n z^n} \tag{3.133a}$$

$$G_m(z) = \frac{z_{G_m}(z)}{n_{G_m}(z)} = V\frac{b_0 + b_1 z + \cdots + b_{n-1}z^{n-1} + b_n z^n}{\bar{a}_0 + \bar{a}_1 z + \cdots + \bar{a}_{n-1}z^{n-1} + z^n}, \tag{3.133b}$$

then $\bar{u}_z(z)$ simply follows as

$$\bar{u}_z(z) = V\frac{a_0 + a_1 z + \cdots + a_{n-1}z^{n-1} + a_n z^n}{\bar{a}_0 + \bar{a}_1 z + \cdots + \bar{a}_{n-1}z^{n-1} + z^n} r_z(z)\ . \tag{3.134}$$

The factor V is chosen, for example, such that $\lim_{z\to 1} G_m(z) = 1$ holds. If the reference model (3.128) is now in the first standard form, then the state realization of (3.134) in

the first standard form in the single-variable case is

$$\underbrace{\begin{bmatrix} \bar{x}_{1,k+1} \\ \bar{x}_{2,k+1} \\ \vdots \\ \bar{x}_{n-1,k+1} \\ \bar{x}_{n,k+1} \end{bmatrix}}_{\bar{\mathbf{x}}_{k+1}} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \\ -\bar{a}_0 & -\bar{a}_1 & \dots & -\bar{a}_{n-2} & -\bar{a}_{n-1} \end{bmatrix}}_{\mathbf{\Phi}_m} \underbrace{\begin{bmatrix} \bar{x}_{1,k} \\ \bar{x}_{2,k} \\ \vdots \\ \bar{x}_{n-1,k} \\ \bar{x}_{n,k} \end{bmatrix}}_{\bar{\mathbf{x}}_k} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{\mathbf{\Gamma}_m} r_k \tag{3.135a}$$

$$\bar{u}_k = \underbrace{V \begin{bmatrix} a_0 - \bar{a}_0 a_n & a_1 - \bar{a}_1 a_n & \dots & a_{n-1} - \bar{a}_{n-1} a_n \end{bmatrix}}_{\bar{\mathbf{c}}_m^T} \underbrace{\begin{bmatrix} \bar{x}_{1,k} \\ \bar{x}_{2,k} \\ \vdots \\ \bar{x}_{n-1,k} \\ \bar{x}_{n,k} \end{bmatrix}}_{\bar{\mathbf{x}}_k} + V a_n r_k \ . \tag{3.135b}$$

Now all the results obtained so far can be summarized in a common structure. A state controller and state observer with integral action and feedforward via a reference model in the case of a single-variable system consists of the *reference model* (see (3.128))

$$\bar{\mathbf{x}}_{k+1} = \mathbf{\Phi}_m \bar{\mathbf{x}}_k + \mathbf{\Gamma}_m r_k \tag{3.136a}$$

$$\bar{y}_k = \mathbf{c}_m^{\mathrm{T}} \bar{\mathbf{x}}_k \tag{3.136b}$$

the *state and disturbance observer* (see (3.110))

$$\begin{bmatrix} \hat{\mathbf{x}}_{k+1} \\ \hat{z}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi} & \mathbf{\Gamma} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}}_k \\ \hat{z}_k \end{bmatrix} + \begin{bmatrix} \mathbf{\Gamma} \\ 0 \end{bmatrix} u_k + \begin{bmatrix} \hat{\mathbf{k}} \\ \hat{k}_z \end{bmatrix} \left(y_k - \mathbf{c}^{\mathrm{T}} \hat{\mathbf{x}}_k \right) \tag{3.137}$$

and the *control law*, consisting of the *feedback component* $-\mathbf{k}^{\mathrm{T}}(\bar{\mathbf{x}}_k - \hat{\mathbf{x}}_k) - \hat{z}_k$ and the *feedforward component* $\bar{u}_k = \bar{\mathbf{c}}_m^{\mathrm{T}} \bar{\mathbf{x}}_k + V a_n r_k$ (see (3.119), (3.129), and (3.135)),

$$u_k = -\mathbf{k}^{\mathrm{T}}(\bar{\mathbf{x}}_k - \hat{\mathbf{x}}_k) - \hat{z}_k + \underbrace{\bar{\mathbf{c}}_m^{\mathrm{T}} \bar{\mathbf{x}}_k + V a_n r_k}_{\bar{u}_k} \tag{3.138}$$

Simplifying equations (3.136)–(3.138), we finally obtain

$$\bar{\mathbf{x}}_{k+1} = \mathbf{\Phi}_m \bar{\mathbf{x}}_k + \mathbf{\Gamma}_m r_k \tag{3.139a}$$

$$\hat{\mathbf{x}}_{k+1} = \left(\mathbf{\Phi} + \mathbf{\Gamma}\mathbf{k}^{\mathrm{T}} - \hat{\mathbf{k}}\mathbf{c}^{\mathrm{T}} \right) \hat{\mathbf{x}}_k + \mathbf{\Gamma} \left(\bar{\mathbf{c}}_m^{\mathrm{T}} - \mathbf{k}^{\mathrm{T}} \right) \bar{\mathbf{x}}_k + V a_n \mathbf{\Gamma} r_k + \hat{\mathbf{k}} y_k \tag{3.139b}$$

$$\hat{z}_{k+1} = \hat{z}_k + \hat{k}_z \left(y_k - \mathbf{c}^{\mathrm{T}} \hat{\mathbf{x}}_k \right) \tag{3.139c}$$

$$u_k = \left(\bar{\mathbf{c}}_m^{\mathrm{T}} - \mathbf{k}^{\mathrm{T}} \right) \bar{\mathbf{x}}_k + \mathbf{k}^{\mathrm{T}} \hat{\mathbf{x}}_k - \hat{z}_k + V a_n r_k \ . \tag{3.139d}$$

Figure 3.8 shows the structure of the control concept (3.139).

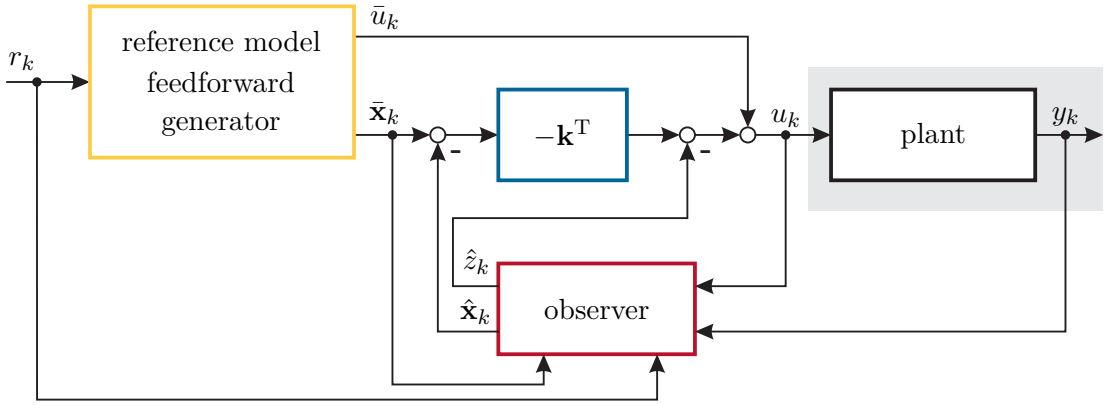


Figure 3.8: Control structure of the feedforward concept.

Exercise 3.12. Consider the transfer function of a double integrator

$$G(s) = \frac{1}{s^2} \ .$$

For the sampling time $T_a = 0.1\,\mathrm{s}$, determine the corresponding discrete-time model. Apply all control concepts presented in this chapter to this model, if possible. Consider a constant but unknown disturbance at the input and a sinusoidal input disturbance of the form

$$w(t) = A\sin(2t + \varphi)$$

with unknown amplitude A and unknown phase φ. Furthermore, jump-like and sinusoidal reference signals are to be specified, which the system can at least follow stationarily without control errors. Implement all control concepts in MATLAB/SIMULINK.

Exercise 3.13. Given is the mechanical system shown in Figure 3.9, driven by a externally excited DC motor. Determine the mathematical model under the assumption that the dynamics of the DC motor can be neglected. Use the angular velocities ω_1, ω_2, and the angle difference $\Delta\varphi = \varphi_1 - \varphi_2$ as state variables, the armature current i_a as input variable, and ω_2 as output variable. The armature circuit constant has the value $k_a = 1\,\mathrm{Nm/A}$, the moments of inertia of the two masses are $J_1 = 1.11\,\mathrm{kgm}^2$ and $J_2 = 10\,\mathrm{kgm}^2$, and the spring and damping constant of the torsion shaft are given by $c = 1\,\mathrm{Nm/rad}$ and $d = 0.1\,\mathrm{Nms/rad}$. For a sampling time $T_a = 0.5\,\mathrm{s}$, determine the corresponding sampled model. Apply all control concepts presented in this chapter to this sampled model, if possible, and test your results in MATLAB/SIMULINK.

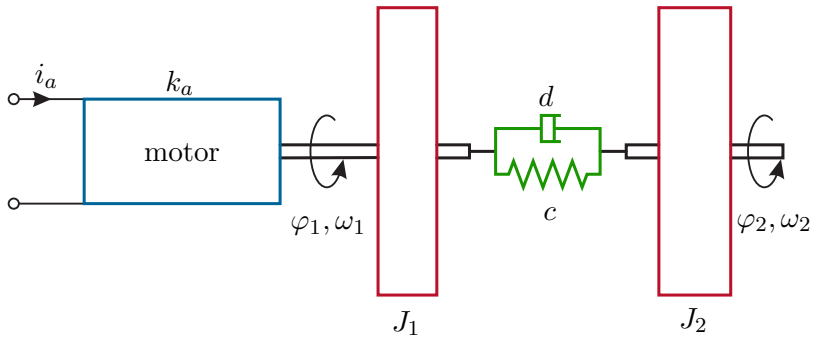


Figure 3.9: Mechanical system with torsion shaft.

3.6 References

[3.1] G. Franklin, J. Powell, and M. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, USA: Addison–Weseley, 1998.

[3.2] K. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design*. New York, USA: Prentice Hall, 1997.

[3.3] A. Bryson and Y. Ho, *Applied Optimal Control*. Washington, USA: He, 1975.

[3.4] P. Dorato, C. Abdallah, and V. Cerone, *Linear Quadratic Control: An Introduction*. Florida, USA: Krieger Publishing Company, 2000.

A Fundamentals of Stochastics

Theorem A.1 (Axioms of Probability)**.** *The following axioms of probability can be defined:*

(1) The probability $\mathrm{P}(A)$ *of an outcome* A *in an experiment is a uniquely determined non-negative real number, which can be at most equal to 1, thus it holds that*

$$0 \leq \mathrm{P}(A) \leq 1 \ . \tag{A.1}$$

(2) For a certain event A *of an experiment it holds that*

$$\mathrm{P}(A) = 1 \ . \tag{A.2}$$

For equivalent events B *and* C *in an experiment it holds that*

$$\mathrm{P}(B) = \mathrm{P}(C) \ . \tag{A.3}$$

(3) If two events B *and* C *in an experiment are mutually exclusive, then it holds that*

$$\mathrm{P}(B + C) = \mathrm{P}(B) + \mathrm{P}(C) \ . \tag{A.4}$$

Definition A.1 (Random Variable, Stochastic Variable)**.** A function X is called a random variable or stochastic variable if it is assigned to a random experiment and possesses the following properties:

(1) The values of X are real numbers and

(2) for every number a and every interval I on the number line, the probability of the event "X has the value a" or "X lies in the interval I" is in accordance with the axioms of probability.

In this context, a random experiment is an experiment in which the result of a single execution can be expressed by a single number.

If X is a *discrete random variable*, then the corresponding probabilities $p_1 = \mathrm{P}(X = x_1)$, $p_2 = \mathrm{P}(X = x_2), \ldots$ can be assigned to all values of X, hereinafter denoted by $x_1, x_2, \ldots$. The function

$$f(x) = \begin{cases} p_j & \text{for } x = x_j \\ 0 & \text{for all other } x \end{cases} \tag{A.5}$$

is then called the *probability function*. Since the random variable X always takes a value

x_j, it must also hold that the sum of all probabilities

$$\sum_j f(x_j) = 1 \tag{A.6}$$

is 1. The probability that the random variable X lies in the interval $a < X \leq b$ is then easily calculated in the form

$$\mathrm{P}(a < X \leq b) = \sum_{a<x_j\leq b} f(x_j) \ . \tag{A.7}$$

If the probability $\mathrm{P}(X \leq x)$ is plotted as a function of x, i.e., the probability of an experiment whose results are $X \leq x$, then one obtains the *probability distribution function*

$$F(x) = \mathrm{P}(X \leq x) = \sum_{x_j\leq x} f(x_j) \ . \tag{A.8}$$

For a *continuous random variable* X, the probability distribution function can be represented in integral form

$$F(x) = \int_{-\infty}^{x} f(v)\,\mathrm{d}v \quad \text{with} \quad F(\infty) = 1 \tag{A.9}$$

where the integrand $f(v)$ is a non-negative and, except for finitely many points, continuous function, which is also called the *probability density function.*

Exercise A.1. Show that for the continuous random variable X the relation

$$\mathrm{P}(a < X \leq b) = \int_a^b f(v)\,\mathrm{d}v = F(b) - F(a)$$

holds.

Figure A.1 shows the typical course of the probability distribution function of a discrete and a continuous random variable.

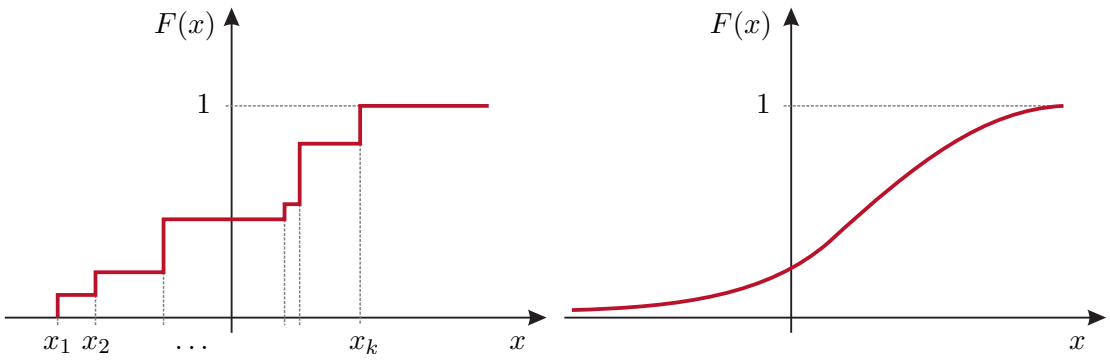


Figure A.1: On the probability distribution function.

Definition A.2 (Expected Value)**.** The *expected value* $\mathrm{E}(X)$ of a random variable X (also called mean value or 1st moment) is defined in the case of a discrete distribution as

$$\mathrm{E}(X) = \sum_j x_j f(x_j) \tag{A.10}$$

and in the case of a continuous distribution as

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\ . \tag{A.11}$$

For a function $g(X)$ of the random variable X, it generally holds that the expected value of the function $g(X)$ can be calculated in the form

$$\mathrm{E}(g(X)) = \sum_j g(x_j) f(x_j) \quad \text{or} \quad \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x \tag{A.12}$$

Exercise A.2. Show the validity of the following important equation

$$\mathrm{E}(\alpha g(X) + \beta h(X)) = \alpha\,\mathrm{E}(g(X)) + \beta\,\mathrm{E}(h(X))$$

with the constants α and β. Show that this also implies

$$\mathrm{E}(\mathrm{E}(X)) = \mathrm{E}(X) \quad \text{and} \quad \mathrm{E}(X\,\mathrm{E}(X)) = \mathrm{E}(X)^2$$

Definition A.3 (Variance)**.** The *variance* σ_X^2 of a random variable X measures the quadratic deviation from the expected value and is defined as

$$\sigma_X^2 = \mathrm{E}\Big([X - \mathrm{E}(X)]^2\Big) = \mathrm{E}\Big(X^2 - 2X\,\mathrm{E}(X) + \mathrm{E}(X)^2\Big) = \mathrm{E}\Big(X^2\Big) - \mathrm{E}(X)^2 \tag{A.13}$$

The variance is also referred to as the second central moment, and its positive square root σ_X as the standard deviation.

The previous considerations can now be easily extended to *multiple random variables*. For two discrete random variables X and Y, the probability distribution function takes the form

$$F(x,y) = \mathrm{P}(X \leq x, Y \leq y) = \sum_{x_k \leq x} \sum_{y_j \leq y} f(x_k, y_j) \tag{A.14}$$

with

$$f(x,y) = \begin{cases} p_{kj} & \text{for } x = x_k,\ y = y_j \\ 0 & \text{for all other } (x,y) \end{cases} \quad \text{with} \quad \sum_k \sum_j f(x_k, y_j) = 1 \tag{A.15}$$

and in the continuous case it holds that

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(v,w)\,\mathrm{d}v\,\mathrm{d}w \quad \text{with} \quad F(\infty,\infty) = 1\ . \tag{A.16}$$

The so-called *marginal distributions* are given by the following expressions

$$F_1(x) = \mathrm{P}(X \leq x, Y \text{ arbitrary}) = \sum_{x_k \leq x} \underbrace{\sum_j f(x_k, y_j)}_{f_1(x_k)} \tag{A.17}$$

or

$$F_1(x) = \mathrm{P}(X \leq x, Y \text{ arbitrary}) = \int_{-\infty}^{x} \underbrace{\int_{-\infty}^{\infty} f(v,w)\,\mathrm{d}v}_{f_1(w)}\,\mathrm{d}w \tag{A.18}$$

and

$$F_2(y) = \mathrm{P}(X \text{ arbitrary}, Y \leq y) = \sum_{y_j \leq y} \underbrace{\sum_k f(x_k, y_j)}_{f_2(y_j)} \tag{A.19}$$

or

$$F_2(y) = \mathrm{P}(X \text{ arbitrary}, Y \leq y) = \int_{-\infty}^{y} \underbrace{\int_{-\infty}^{\infty} f(v,w)\,\mathrm{d}w}_{f_2(v)}\,\mathrm{d}v \tag{A.20}$$

Definition A.4 (Independence of Random Variables)**.** Two random variables X and Y are independent if and only if for every pair of events of the form

$$a_1 < X \leq b_1 \quad \text{and} \quad a_2 < Y \leq b_2 \tag{A.21}$$

the relation

$$\mathrm{P}(a_1 < X \leq b_1, a_2 < Y \leq b_2) = \mathrm{P}(a_1 < X \leq b_1)\,\mathrm{P}(a_2 < Y \leq b_2) \tag{A.22}$$

and thus also

$$f(x,y) = f_1(x) f_2(y) \quad \text{and} \quad F(x,y) = F_1(x) F_2(y) \tag{A.23}$$

with $f_1(x)$, $f_2(y)$, $F_1(x)$ and $F_2(y)$ according to (A.17)–(A.20) holds.

Theorem A.2 (Expected Value and Covariance of Two Random Variables)**.** *For the expected value of a function $g(X,Y)$ of two random variables X and Y, it holds that*

$$\mathrm{E}(g(X,Y)) = \sum_k \sum_j g(x_k, y_j) f(x_k, y_j) \tag{A.24}$$

or

$$\mathrm{E}(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y \ . \tag{A.25}$$

The expected value $\mathrm{E}(X+Y)$ of a sum of random variables X and Y is calculated as

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) \ . \tag{A.26}$$

If two random variables X and Y are independent, then the following relation holds (shown here for the discrete case using (A.19))

$$\mathrm{E}(XY) = \sum_k \sum_j x_k y_j f(x_k, y_j) = \sum_k \sum_j x_k y_j f_1(x_k) f_2(y_j) = \mathrm{E}(X)\,\mathrm{E}(Y) \ . \tag{A.27}$$

If a random variable Z results from the sum of two random variables X and Y, then according to (A.13) the variance of Z is

$$\begin{aligned} \sigma_Z^2 &= \mathrm{E}\left(Z^2\right) - \mathrm{E}(Z)^2 = \mathrm{E}\left(X^2\right) + 2\,\mathrm{E}(XY) + \mathrm{E}\left(Y^2\right) \\ &\quad - \mathrm{E}(X)^2 - 2\,\mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(Y)^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \end{aligned} \tag{A.28}$$

with the so-called covariance

$$\sigma_{XY} = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \ . \tag{A.29}$$

It is easily verified that (A.29) *can also be written in the form*

$$\sigma_{XY} = \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]) \tag{A.30}$$

If the random variables X and Y are independent, then according to (A.27) $\sigma_{XY} = 0$.

Definition A.5 (Correlation Coefficient and Covariance Matrix)**.** The quotient

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{A.31}$$

is called the *correlation coefficient* between the random variables X and Y. If $r = 0$, then X and Y are *uncorrelated*. It can also be seen that two independent random variables X and Y must also be uncorrelated because of $\sigma_{XY} = 0$. The correlation coefficient $-1 \leq r \leq 1$ provides a measure of the *linear dependence* of X and Y. For a vector-valued random variable $\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix}$, it now holds that

$$\mathrm{E}\left(\mathbf{X}^{\mathrm{T}}\right) = \begin{bmatrix} \mathrm{E}(X_1) & \mathrm{E}(X_2) & \dots & \mathrm{E}(X_n) \end{bmatrix} \tag{A.32}$$

and the *covariance matrix* of the vector-valued random variable $\mathbf{X}$ is understood to be the matrix

$$\mathrm{cov}(\mathbf{X}) = \mathrm{E}\left([\mathbf{X} - \mathrm{E}(\mathbf{X})][\mathbf{X} - \mathrm{E}(\mathbf{X})]^{\mathrm{T}}\right) = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma_{X_n}^2 \end{bmatrix} . \tag{A.33}$$

It can be seen that the covariance matrix is symmetric.

Exercise A.3. Show that the relation

$$\mathrm{E}\left(\|\mathbf{X} - \mathrm{E}(\mathbf{X})\|_2^2\right) = \mathrm{E}\left([\mathbf{X} - \mathrm{E}(\mathbf{X})]^{\mathrm{T}}[\mathbf{X} - \mathrm{E}(\mathbf{X})]\right) = \mathrm{trace}(\mathrm{cov}(\mathbf{X}))$$

holds with $\mathrm{trace}(\mathbf{S}) = \sum_i s_{ii}$.

From what has been said so far, it follows that the covariance matrix of a vector-valued random variable $\mathbf{X}$, whose components X_i, X_j for $i \neq j = 1, \dots, n$ are uncorrelated, is a diagonal matrix.

For stochastic time signals originating from statistically identical signal sources, there exists not only a single realization $x_1(t)$ but a whole family (ensemble) of random time functions $\{x_j(t)\}$. This ensemble of random time functions or random sequences in the time-discrete case is called a *continuous-time* or *discrete-time stochastic process* $\mathrm{x}(t)$. A single realization $x_j(t)$ for a fixed j is also called a *sample function*. For every fixed point in time t, $\mathrm{x}(t)$ is a random variable with a probability distribution function (see (A.8))

$$F(x, t) = \mathrm{P}(\mathrm{x}(t) \leq x) \ . \tag{A.34}$$

The probability density function is then calculated according to (A.9) as

$$f(x,t) = \frac{\partial F(x,t)}{\partial x} \ . \tag{A.35}$$

Definition A.6 (Mean, Auto- and Cross-Correlation Function of a Stochastic Process)**.** The mean $\eta_{\mathrm{x}}(t)$ of a stochastic process $\mathrm{x}(t)$ is (see also (A.11))

$$\eta_{\mathrm{x}}(t) = \mathrm{E}(\mathrm{x}(t)) = \int_{-\infty}^{\infty} x f(x,t)\,\mathrm{d}x \ . \tag{A.36}$$

The *autocorrelation function* $\Phi_{\mathrm{xx}}(t_1,t_2)$ of a stochastic process $\mathrm{x}(t)$ is understood to be the expected value of the product $\mathrm{x}(t_1)\mathrm{x}(t_2)$

$$\Phi_{\mathrm{xx}}(t_1,t_2) = \mathrm{E}(\mathrm{x}(t_1)\mathrm{x}(t_2)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1,t_1) x_2 f(x_2,t_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \ . \tag{A.37}$$

The *cross-correlation function* $\Phi_{\mathrm{xy}}(t_1,t_2)$ of two stochastic processes $\mathrm{x}(t)$ and $\mathrm{y}(t)$ is given by the relation

$$\Phi_{\mathrm{xy}}(t_1,t_2) = \mathrm{E}(\mathrm{x}(t_1)\mathrm{y}(t_2)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,t_1) y f(y,t_2)\,\mathrm{d}x\,\mathrm{d}y. \tag{A.38}$$

Definition A.7 (Auto- and Cross-Covariance Function)**.** The *autocovariance function* $C_{\mathrm{xx}}(t_1,t_2)$ of a stochastic process $\mathrm{x}(t)$ is

$$C_{\mathrm{xx}}(t_1,t_2) = \mathrm{E}([\mathrm{x}(t_1)-\eta_{\mathrm{x}}(t_1)][\mathrm{x}(t_2)-\eta_{\mathrm{x}}(t_2)]) = \Phi_{\mathrm{xx}}(t_1,t_2) - \eta_{\mathrm{x}}(t_1)\eta_{\mathrm{x}}(t_2) \ . \tag{A.39}$$

The *cross-covariance function* $C_{\mathrm{xy}}(t_1,t_2)$ of two stochastic processes $\mathrm{x}(t)$ and $\mathrm{y}(t)$ is analogously given by

$$C_{\mathrm{xy}}(t_1,t_2) = \mathrm{E}([\mathrm{x}(t_1)-\eta_{\mathrm{x}}(t_1)][\mathrm{y}(t_2)-\eta_{\mathrm{y}}(t_2)]) = \Phi_{\mathrm{xy}}(t_1,t_2) - \eta_{\mathrm{x}}(t_1)\eta_{\mathrm{y}}(t_2) \ . \tag{A.40}$$

Exercise A.4. Show the validity of the right-hand identity of (A.39).

Definition A.8 (Stationary Stochastic Process)**.** A stochastic process $\mathrm{x}(t)$ is called *strictly stationary* if its statistical properties are invariant under time shifts, i.e., $\mathrm{x}(t)$ and $\mathrm{x}(t+c)$ have identical statistical properties for all c. The stochastic process $\mathrm{x}(t)$ is *stationary in a wider sense* if the mean $\eta_{\mathrm{x}}(t) = \eta_{\mathrm{x}}$ is constant and the autocorrelation function depends only on the time difference $\tau = t_2 - t_1$, i.e., $\Phi_{\mathrm{xx}}(\tau) = \mathrm{E}(\mathrm{x}(t)\mathrm{x}(t+\tau))$.

The functions (A.36)–(A.40) shown so far take into account all realizations, i.e., the entire ensemble, of a stochastic process. According to the so-called *ergodic hypothesis*, the functions (A.36)–(A.40) can also be written using only a single sample function if *infinitely long time intervals* are considered. Ergodic processes are also stationary, but the inverse does not generally hold. This results in the following expressions, with the continuous-time case on the left and the discrete-time case on the right:

(1) Mean value

$$\eta_x = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)\,\mathrm{d}t \quad \text{or} \quad \eta_x = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} x_k \tag{A.41}$$

(2) Autocorrelation function

$$\Phi_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)\,\mathrm{d}t \quad \text{or} \quad \Phi_{xx}(\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} x_k x_{k+\tau} \tag{A.42}$$

(3) Cross-correlation function

$$\Phi_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)y(t+\tau)\,\mathrm{d}t \quad \text{or} \quad \Phi_{xy}(\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} x_k y_{k+\tau} \tag{A.43}$$

Exercise A.5. Show that $\Phi_{xx}(\tau)$ satisfies the following properties:

1. $\Phi_{xx}(\tau) = \Phi_{xx}(-\tau)$
2. $\Phi_{xx}(0) = \mathrm{E}\left(x(t)^2\right)$
3. $\Phi_{xx}(\infty) = \mathrm{E}(x(t))^2$
4. $\Phi_{xx}(\tau) \leq \Phi_{xx}(0)$

Exercise A.6. Show that $\Phi_{xy}(\tau)$ satisfies the following properties:

1. $\Phi_{xy}(\tau) = \Phi_{yx}(-\tau)$
2. $\Phi_{xy}(0) = \mathrm{E}(x(t)y(t))$
3. $\Phi_{xy}(\infty) = \mathrm{E}(x(t))\,\mathrm{E}(y(t))$
4. $\Phi_{xy}(\tau) \leq \frac{1}{2}[\Phi_{xx}(0) + \Phi_{yy}(0)]$

Exercise A.7. What are the expressions for the autovariance and autocovariance functions in the ergodic case?

Definition A.9 (White Noise)**.** A stochastic process $\mathrm{v}(t)$ is called *(strict) white noise* if for all time points $t_1 \neq t_2$, $\mathrm{v}(t_1)$ and $\mathrm{v}(t_2)$ are statistically independent, or

$$C_{\mathrm{vv}}(\tau) = C_0\delta(\tau) \quad \text{with} \quad \delta(\tau) = \begin{cases} 1 & \text{for } \tau = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and } C_0 > 0\ . \tag{A.44}$$

It is further assumed that the mean $\eta_{\mathrm{v}} = 0$.

Remark: It can be shown that in the continuous-time case the average power of the white noise is infinite and therefore this process cannot be ideally realized. In contrast, in the discrete-time case, where the signal sequence values are at a finite distance from each other, the power remains finite, so that ideal white noise is also realizable.

A.1 References

[A.1] R. Isermann, *Identifikation dynamischer Systeme 1 und 2*, 2nd ed. Berlin, Deutschland: Springer, 1992.

[A.2] E. Kreyszig, *Statistische Methoden und ihre Anwendungen*, 7th ed. Göttingen, Deutschland: Vandenhoeck & Ruprecht, 1998.

[A.3] D. Luenberger, *Optimization by Vector Space Methods*. New York, USA: John Wiley & Sons, 1969.

[A.4] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York, USA: McGraw-Hill, 2002.