Model-based identification and estimation

Lecture Notes

Dr. Alexander Schaum

Summer term 2020

Model-based identification and estimation

Lecture Notes, Summer term 2020

Dr. Alexander Schaum

Chair of Automatic Control
Kiel University
Institute for Elektrical Engineering and Information Technology
Technical Faculty
Kaiserstraße 2
D–24143 Kiel

alsc@tf.uni–kiel.de
http://www.control.tf.uni–kiel.de

Contents

1 General aspects on linear time-invariant systems

This chapter gives a short review and summary of important aspects of linear time-invariant systems.

In the first part the focus is put on continuous-time systems, their solutions, stability properties and the analysis via the Laplace transform. It is supposed that the reader is more or less familiar with these concepts, so that the presentation here is hold quite short. Some examples are considered which should help to recall and visualize these aspects.

In the second part discrete-time systems are considered. Besides aspects about their solution and stability properties, particular focus is put on the relation between continuous and discrete-time systems and their solutions. The aspects that are considered in this text are only those which are important later for the purpose of identification and estimation.

1.1 Continuous-time system

When thinking about mathematical modeling of physical systems, the laws of physics typically guide us to continuous-time models based on Newton's laws, thermodynamics, electromagnetism, etc. [Ari78]. Continuous-time systems are based on a continuous change over time which itself is considered as a continuously changing variable. These model then typically lead to differential equations whose solutions aspects and stability properties are recalled in the next sections.

1.1.1 Solutions of continuous-time linear systems

Consider the continuous-time linear time-invariant system

$$\dot{\boldsymbol{x}} = A\boldsymbol{x}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{1.1}$$

with $A \in \mathbb{R}^{n\times n}$. The solution $\boldsymbol{x}(t)$ can be written as

$$\boldsymbol{x}(t) = S(t)\boldsymbol{x}_0 \tag{1.2}$$

with the matrix exponential (the so-called fundamental solution)

$$S(t) = \exp(At) = \sum_{i=0}^{\infty} \frac{A^i t^i}{i!}, \quad \boldsymbol{x}(t) = S(t)\boldsymbol{x}_0. \tag{1.3}$$

It follows that $\dfrac{dS(t)}{dt} = AS(t)$ so that

$$\frac{d\boldsymbol{x}(t)}{dt} = \frac{dS(t)}{dt}\boldsymbol{x}_0 = AS(t)\boldsymbol{x}_0 = A\boldsymbol{x}(t),$$

showing that (1.2) really is a solution of (1.11).

The characteristic equation of A is given by

$$\det(\lambda I - A) = \sum_{i=0}^{n} a_i \lambda^i = 0. \tag{1.4}$$

By Cailey-Hamilton's theorem [Fis09; Dym07; Kai80] it follows that $\sum_{i=0}^{n} a_i A^i = 0$. Thus, there exist constants c_{mi} such that

$$\forall m \geq N : A^m = \sum_{i=0}^{N-1} c_{mi} A^i.$$

Accordingly, the series in (1.3) can be expressed as a sum of the first n powers of A

$$S(t) = \exp(At) = \sum_{i=0}^{n} \kappa_i(t) A^i, \tag{1.5}$$

with adequately chosen coefficient functions κ_i. The exact determination of the κ_i as infinite series and the proof of their convergence goes beyond the scope and purpose of the present text and can be reviewed somewhere else (see e.g. [Har64; Tes12]).

The matrix function $S(t)$, i.e. the fundamental solution of (1.11) satisfies the following properties:

(i) $S(0) = I$

(ii) $S(t_1)S(t_0) = S(t_1 + t_0)$

(iii) $S(t - t_0)^{-1} = S(t_0 - t)$

which correspond to the so-called flow axioms.

For the case of interest of input-output systems of the form

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \tag{1.6}$$

with $\boldsymbol{u}(t) \in \mathbb{R}^p$, multiply both sides with $\exp(-At)$ to obtain

$$e^{-At}\dot{\boldsymbol{x}}(t) = e^{-At}A\boldsymbol{x}(t) + e^{-At}B\boldsymbol{u}(t) = -\left(\frac{d}{dt}\left(e^{-At}\right)\right)\boldsymbol{x}(t) + e^{-At}B\boldsymbol{u}(t)$$

or equivalently, recalling the product formula for differentiation

$$\frac{d}{dt}\left(e^{-At}\boldsymbol{x}(t)\right) = e^{-At}B\boldsymbol{u}(t).$$

Integrate from 0 to t

$$\int_0^t \frac{d}{d\tau}\left(e^{-A\tau}\boldsymbol{x}(\tau)\right)d\tau = e^{-A\tau}\boldsymbol{x}(\tau)\Big|_0^t = e^{-At}\boldsymbol{x}(t) - \boldsymbol{x}_0 = \int_0^t e^{-A\tau}B\boldsymbol{u}(\tau)d\tau,$$

rearrange, and multiply with $S(t) = \exp(At)$ to obtain

$$\boldsymbol{x}(t) = \exp(At)\boldsymbol{x}_0 + \int_0^t \exp(A(t-\tau))B\boldsymbol{u}(\tau)d\tau,$$

or equivalently

$$\boldsymbol{x}(t) = S(t)\boldsymbol{x}_0 + \int_0^t S(t-\tau)B\boldsymbol{u}(\tau)d\tau, \quad S(t) = e^{At}. \tag{1.7}$$

This relation is known as the variation of constants formula.

Example 1.1. *Consider the scalar system*

$$\dot{x}(t) = \lambda x(t), \quad x(0) = x_0.$$

The unique solution is given by

$$x(t) = S(t)x_0, \quad \text{with } S(t) = e^{\lambda t},$$

and for $\lambda < 0$ tends to zero for $t \to \infty$.

Now, consider the input-output system

$$\begin{aligned}\dot{x}(t) &= \lambda x(t) + bu(t), \quad x(0) = x_0\\ y(t) &= x(t).\end{aligned}$$

By the variation of constants formula (1.7) *it follows that*

$$\begin{aligned}y(t) = x(t) &= S(t)x_0 + \int_0^t S(t-\tau)bu(\tau)d\tau\\ &= e^{\lambda t}x_0 + \int_0^t e^{\lambda(t-\tau)}bu(\tau)d\tau\\ &= e^{\lambda t}\left(x_0 + \int_0^t e^{-\lambda\tau}bu(\tau)d\tau\right).\end{aligned}$$

Considering e.g. $u(t) = h(t)$ (Heaviside step function) and $x_0 = 0$ it follows that

$$y(t) = e^{\lambda t}b\int_0^t e^{-\lambda\tau}d\tau = -\frac{b}{\lambda}\left(1 - e^{\lambda t}\right).$$

The step-response for $\lambda = -1, b = 1.5$ is shown in Fig. 1.1.

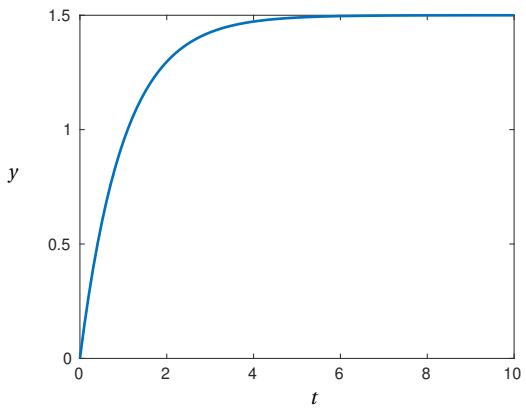


Figure 1.1: Step response for system (1.1) with $\lambda = -1, b = 1.5$.

Example 1.2. *Consider the system*

$$\dot{x}_1(t) = \lambda_1 x_1(t) + x_2(t), \qquad x_1(0) = x_{10} \tag{1.8a}$$
$$\dot{x}_2(t) = \lambda_2 x_2(t) + bu, \qquad x_2(0) = x_{20} \tag{1.8b}$$
$$y(t) = x_1(t) \tag{1.8c}$$

Following the preceding example, the solution for x_2 is given by

$$x_2(t) = e^{\lambda_2 t}\left(x_{20} + \int_0^t e^{-\lambda_2 \tau} bu(\tau)\, d\tau\right)$$

and for $x_{20} = 0$ and $u(t) = h(t)$ reads $x_2(t) = -\frac{b}{\lambda_2}\left(1 - e^{\lambda_2 t}\right)$, $t \geq 0$.

The solution for x_1 correspondingly is given by

$$\begin{aligned} x_1(t) &= e^{\lambda_1 t} x_{10} + \int_0^t e^{\lambda_1 (t-\tau)} x_2(\tau)\, d\tau \\ &= e^{\lambda_1 t}\left(x_{10} - \frac{b}{\lambda_2}\int_0^t e^{-\lambda_1 \tau}\left(1 - e^{\lambda_2 \tau}\right) d\tau\right). \end{aligned}$$

With $x_{10} = 0$ this yields the step-response

$$y(t) = \frac{b}{\lambda_1 \lambda_2}\left(1 + \frac{\lambda_2}{\lambda_1 - \lambda_2} e^{\lambda_1 t} - \frac{\lambda_1}{\lambda_1 - \lambda_2} e^{\lambda_2 t}\right).$$

In particular, for $\lambda_1 = \lambda_2$ it follows by the rule of l'Hôpital-Bernoulli that [a]

$$y(t) = \frac{b}{\lambda_1^2}\left(1 - (\lambda_1 t - 1) e^{\lambda_1 t}\right).$$

The step response for $\lambda_1 = -5, \quad \lambda_2 = -0.4, \quad b = 1.5$ is shown in Fig. 1.2.

[a] Note that the same result is obtained when in the integral formulation above $\lambda_1 = \lambda_2$ is set.

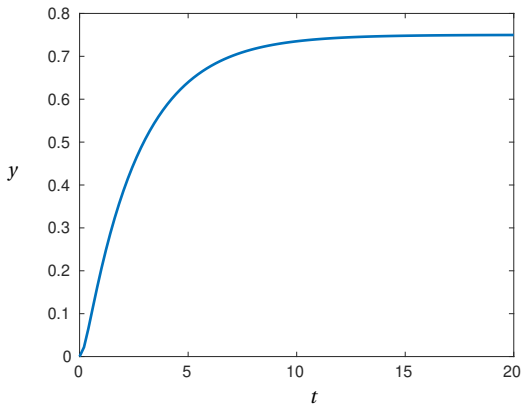


Figure 1.2: Step response for system (1.8) with $\lambda_1 = -5, \lambda_2 = -0.4, b = 1.5$.

Interestingly, comparing the behavior of x_1 in Fig. 1.2 and x in Fig. 1.1 it results that the effect of the second order behavior is only present at the beginning of the step response and after a short

initial transient difference the asymptotic behavior qualitatively (but not quantitatively) looks quite similar.

1.1.2 Stability of continous-time systems

For a constant input $\boldsymbol{u} = \bar{\boldsymbol{u}}$ the point $\bar{\boldsymbol{x}}$ is called an equilibrium point of the system

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x} + B\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

if it holds that $\mathbf{0} = A\bar{\boldsymbol{x}} + B\bar{\boldsymbol{u}}$.

Definition 1.1

The equilibrium point $\bar{\boldsymbol{x}}$ is stable if for any $\epsilon > 0$ there exists a $\delta > 0$ such that for any initial condition $\boldsymbol{x}_0$ in a δ-neighborhood of $\bar{\boldsymbol{x}}$ the solution $\boldsymbol{x}(t;\boldsymbol{x}_0)$ is contained in an ϵ-neighborhood of $\bar{\boldsymbol{x}}$, i.e.

$$\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}\| \leq \delta \quad \Rightarrow \quad \|\boldsymbol{x}(t;\boldsymbol{x}_0) - \bar{\boldsymbol{x}}\| \leq \epsilon.$$

If $\bar{\boldsymbol{x}}$ is not stable it is called unstable.

Definition 1.2

The equilibrium point $\bar{\boldsymbol{x}}$ is asymptotically stable if it is stable and attractive, meaning that

$$\lim_{t\to\infty} \|\boldsymbol{x}(t;\boldsymbol{x}_0) - \bar{\boldsymbol{x}}\| = 0.$$

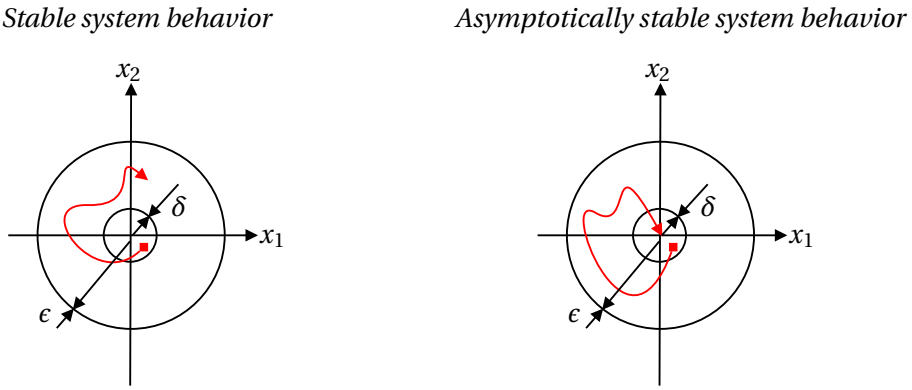


Figure 1.3: Illustration in $\mathbb{R}^2$ of a trajectory for a stable system (left) and an asymptotically stable system (right) for the case that $\bar{\boldsymbol{x}} = \mathbf{0}$.

Necessary and sufficient conditions for the (asymptotic) stability of linear time-invariant systems are stated in the next theorem.

Theorem 1.1

The equilibrium point $\bar{\boldsymbol{x}}$ is stable if and only if the eigenvalues λ_i, $i = 1,\ldots,n$ of A satisfy $\Re(\lambda_i) \leq 0$. It is asymptotically stable if and only if $\Re(\lambda_i) < 0$ for all $i = 1,\ldots,n$.

Example 1.3. *Consider again the system (cp. Example 1.1)*

$$\dot{x} = \lambda x + bu, x(0) = x_0$$

For $u = 0$ the unique equilibrium point is given by $\bar{x} = 0$, which according to Theorem 1.1 is

- *stable for $\lambda \leq 0$*
- *asymptotically stable for $\lambda < 0$.*

For a constant input $u \in \mathbb{R}$ the equilibrium points are determined by the equation

$$0 = \lambda \bar{x} + bu.$$

For $\lambda \neq 0$ it follows that $\bar{x} = \frac{b}{\lambda} u$ and the asymptotic stability is ensured for $\lambda < 0$. Figure 1.1 verifies this property for $\lambda = -1, b = 1.5$.

Example 1.4. *Consider again the system (cp. Example 1.8)*

$$\dot{x}_1 = \lambda_1 x_1 + x_2, \quad x_1(0) = x_{10}$$
$$\dot{x}_2 = \lambda_2 x_2 + bu, \quad x_2(0) = x_{20}$$

The dynamic matrix is given by

$$A = \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{bmatrix}$$

with eigenvalues[a] λ_1, λ_2. Accordingly, for $u = 0$ the equilibrium point $\bar{x} = 0$ is

- *stable for $\lambda_1, \lambda_2 \leq 0$*
- *asymptotically stable for $\lambda_1, \lambda_2 < 0$.*

For a constant input $u \in \mathbb{R}$ and $\lambda_1, \lambda_2 \neq 0$ the equilibrium must satisfy

$$\bar{x}_1 = \frac{1}{\lambda_1} \bar{x}_2, \quad \bar{x}_2 = \frac{b}{\lambda_2} u.$$

This equilibrium is asymptotically stable for $\lambda_1, \lambda_2 < 0$. An illustration of this property is shown in Figure 1.2 (showing only x_1) for the case that $\lambda_1 = -5, \lambda_2 = -0.4, b = 1.5$.

[a] As the matrix is triangular the eigenvalues coincide with the diagonal elements.

1.1.3 The Laplace-transform and its application to continuous-time linear systems

The Laplace-transform [Kai80; Det88] is the unilateral transformation defined by

$$\hat{f}(s) = \mathscr{L}\{f\}(s) = \int_0^\infty f(t) e^{-st} \mathrm{d}t$$

Here $s = \sigma + j\omega$ represents a complex frequency. The inverse Laplace transform is given by

$$f(t) = \mathscr{L}^{-1}\{\hat{f}\} = \frac{1}{2\pi j} \int_{\sigma - j\infty}^{\sigma + j\infty} e^{st} \hat{f}(s) \mathrm{d}s.$$

This transform has several very useful properties, summarized next.

- Linearity:

$$\mathscr{L}\{a\,u(t)+b\,v(t)\}=a\mathscr{L}\{u(t)\}+b\mathscr{L}\{v(t)\}$$

- Time shifting:

$$\mathscr{L}\{u(t-\tau)\}=e^{-s\tau}\mathscr{L}\{u(t)\}$$

- Differentiation:

$$\mathscr{L}\left\{\frac{du(t)}{dt}\right\}=s\mathscr{L}\{u(t)\}-u(0)$$
$$\mathscr{L}\left\{\frac{d^n u(t)}{dt^n}\right\}=s^n\mathscr{L}\{u(t)\}-s^{n-1}u(0)-s^{n-2}u'(0)-\ldots-u^{(n-1)}(0),\quad n>1.$$

- Convolution:

$$\mathscr{L}\left\{\int_0^t g(t-\tau)u(\tau)d\tau\right\}=\mathscr{L}\{g(t)\}\,\mathscr{L}\{u(t)\}=\hat{g}(s)\hat{u}(s).$$

- Time scaling:

$$\mathscr{L}\{u(at)\}=\frac{1}{a}\mathscr{L}\{u(t)\}\left(\frac{s}{a}\right)=\frac{1}{a}\hat{u}\left(\frac{s}{a}\right).$$

- Frequency shifting:

$$\mathscr{L}\left\{u(t)e^{-at}\right\}=\mathscr{L}\{u(t)\}(s+a)=\hat{u}(s+a).$$

The transforms of the most important functions are listed in Table 1.1.

The properties of the Laplace transform are useful for the analysis of linear continuous-time systems.

Example 1.5. *Consider the first order system*

$$\dot{x}(t)=-\lambda x(t)+bu(t),\quad x(0)=x_0 \tag{1.10a}$$
$$y=x(t) \tag{1.10b}$$

In the frequency domain this equation reads

$$s\hat{x}(s)-x_0=-\lambda\hat{x}(s)+b\hat{u}(s),\quad \hat{y}(s)=\hat{x}(s)$$

with solution $\hat{x}(s)=\frac{1}{s+\lambda}x_0+\underbrace{\frac{b}{s+\lambda}}_{\hat{g}(s)}\hat{u}(s)$

The associated solution in the time domain can be easily obtained using the frequency shift and convolution properties and reads

$$x(t)=x_0e^{-\lambda t}+\int_0^t be^{-\lambda(t-\tau)}u(\tau)\mathrm{d}\tau$$

$u(t)$	$\hat{u}(s)$
$h(t)$	$\frac{1}{s}$
$e^{-at}h(t)$	$\frac{a}{s+a}$
$\cos(\omega_0 t)$	$\frac{s}{s^2+\omega_0^2}$
$\sin(\omega_0 t)$	$\frac{\omega_0}{s^2+\omega_0^2}$
$u(t)$	$\hat{u}(s)$
$\left(1-e^{-at}\right)h(t)$	$\frac{1}{s(s+a)}$
$e^{-at}\cos(\omega_0 t)$	$\frac{s+a}{(s+a)^2+\omega_0^2}$
$e^{-at}\sin(\omega_0 t)$	$\frac{\omega_0}{(s+a)^2+\omega_0^2}$
$h(t)t$	$\frac{1}{s^2}$

Table 1.1: Laplace transform of some important functions.

which is the same as obtained using the variation of constants formula. For the particular input $u(t) = a_0 \sin(\omega_0 t)$ the long-term behavior (i.e., after the decay of the transients) is given by

$$y(t) = a_0 |\hat{g}(j\omega_0)| \sin\left(\omega_0 t + arg(\hat{g}(j\omega_0))\right)$$

For the case at hand

$$\hat{g}(j\omega_0) = \frac{b}{j\omega_0 + \lambda}, \quad \left|\hat{g}(j\omega_0)\right| = \frac{b}{\omega_0^2 + \lambda^2}, \quad arg\left(\hat{g}(j\omega_0)\right) = \arctan\left(\frac{-\omega_0}{\lambda}\right)$$

The dependency of the amplitude $|\hat{g}(j\omega)|$ and the argument $arg(j\omega)$ can be summarized in the Bode- or Nyquist diagrams, respectively. For a case example with $b = 0.5, \lambda = 2$ the respective diagrams are shown in Figure 1.4.

1.2 Discrete-time systems

Given that most measurement methods yield only sampled data our observation of the world basically happens in discrete time. Accordingly, it is quite natural to describe systems in a discrete-time fashion. As discussed before most physical modeling approaches lead to continuous-time models, so a central question is how these different model descriptions can be matched. This question is answered first in this section, following two different approaches, namely (i) the exact discretization and (ii) the approximate (explicit) discretization.

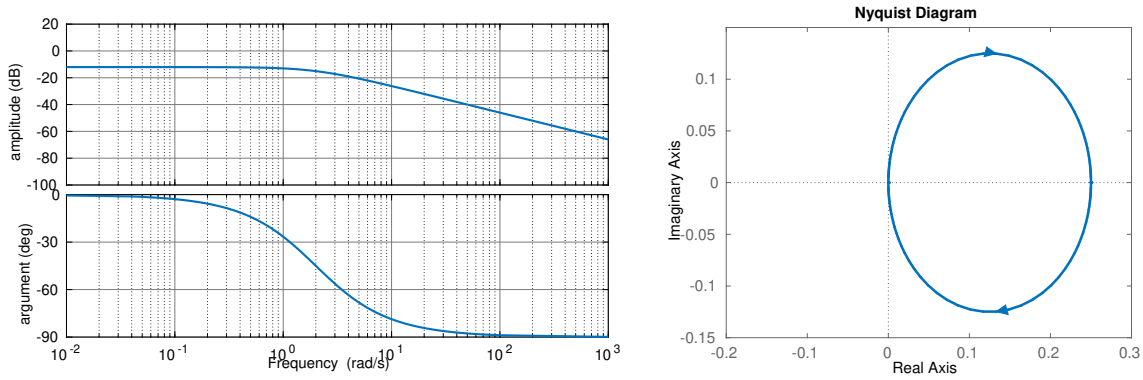


Figure 1.4: Bode (left) and Nyquist (right) diagrams for system (1.10) with $\lambda = 2, b = 0.5$.

Once a system is described in a discrete-time setup, the next question is about stability properties. This will be discussed in continuation, followed by a short recall of the z-transform, the discrete-time counterpart to the Laplace transform, its properties and application to the analysis of discrete-time systems.

1.2.1 From continuous- to discrete-time models

Consider the linear time-invariant system

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t), \quad t \geq 0, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{1.11}$$

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) + D\boldsymbol{u}(t) \tag{1.12}$$

with the solution

$$\boldsymbol{x}(t) = \exp(At)\boldsymbol{x}_0 + \int_0^t \exp(A(t-\tau))B\boldsymbol{u}(\tau)\mathrm{d}\tau, \quad t \geq 0.$$

After a given time dt the solution starting at $\boldsymbol{x}(t)$ at time $t \geq 0$ is thus given by

$$\boldsymbol{x}(t+\mathrm{d}t) = \exp(A\,dt)\boldsymbol{x}(t) + \int_t^{t+dt} \exp(A(t+\mathrm{d}t-\tau))B\boldsymbol{u}(\tau)\mathrm{d}\tau.$$

Accordingly, denoting $\boldsymbol{x}[k] = \boldsymbol{x}(k\,dt)$ und considering a constant input $\boldsymbol{u}(t) = \boldsymbol{u}[k]$ for $t \in [(k-1)dt, k\,dt) = [t, t+dt)$ it follows that

$$\boldsymbol{x}[k+1] = A_d\boldsymbol{x}[k] + B_d\boldsymbol{u}[k], \quad k \in \mathbb{N}, \quad \boldsymbol{x}[0] = \boldsymbol{x}_0 \tag{1.13a}$$

$$\boldsymbol{y}[k] = C\boldsymbol{x}[k] + D\boldsymbol{u}[k] \tag{1.13b}$$

with

$$A_d = \exp(A\,dt), \quad B_d = \int_0^{dt} \exp(A(dt-\tau))B\mathrm{d}\tau \tag{1.13c}$$

Exercise 1.1 (*).

(a) Show that $\int_t^{t+dt} e^{A(t+dt-\tau)}\mathrm{d}\tau = \int_0^{dt} e^{A(dt-\tau)}\mathrm{d}\tau$ *holds for all* $dt > 0$.

(b) Derive the relation (1.13) *and in particular* (1.13c).

Example 1.6. *Consider again the first order continuous-time system (see also Exercise 1.1 and 1.3)*

$$\dot{x}(t) = \lambda x(t) + bu(t), \quad x(0) = x_0 \tag{1.14a}$$
$$y(t) = x(t). \tag{1.14b}$$

Its discrete-time equivalent for the case that $u(t) = u[k]$ *for* $t \in [(k-1)\mathrm{d}t, k\mathrm{d}t)$ *is obtained as follows:*

$$\begin{aligned} x[k+1] = x(t+dt) &= e^{\lambda dt}x(t) + \int_t^{t+dt} e^{\lambda(t+dt-\tau)}bu(t+\tau)\mathrm{d}\tau \\ &= e^{\lambda dt}x[k] + bu[k]\int_t^{t+dt} e^{\lambda(t+dt-\tau)}\mathrm{d}\tau \\ &= e^{\lambda dt}x[k] + bu[k]\frac{1}{-\lambda}e^{\lambda(t+dt)}\left(e^{-\lambda(t+dt)} - e^{-\lambda t}\right) \\ &= \underbrace{e^{\lambda dt}}_{=:a_d}x[k] \underbrace{-\frac{b}{\lambda}\left(1-e^{\lambda dt}\right)}_{=:b_d}u[k] \end{aligned}$$

Summarizing we have

$$x[k+1] = x(t+dt) = a_d x[k] + b_d u[k], \quad k \in \mathbb{N}, \quad x[0] = x_0 \tag{1.15a}$$
$$y[k] = x[k] \tag{1.15b}$$

with $a_d = e^{\lambda dt}$, $\quad b_d = -\frac{b}{\lambda}\left(1-e^{\lambda dt}\right)$.

For an illustration consider $\lambda = -2$, $\quad b = 3$, $\quad dt = 1$, $\quad T_u = 10$, $\quad u[k] = e^{-t[k]/8}\sin\left(\frac{2\pi}{T_u}t[k]\right)$. *The associated solution of the continuous-time and discrete-time systems are shown together with a the input* $u(t)$ *in Fig. 1.5. It can be seen that the solution of the discrete-time model* (1.15) *exactly corresponds to the solution of the continuous-time model* (1.14) *at the discrete time instances* $t_k = kdt$, $k \in \mathbb{N}$.

Remark 1.1

(a) Using the analytic (i.e., exact) solution for the differential equation and evaluating the solution at sampled (i.e., discrete time) steps, one obtains an exact solution at the sampling instants $t[k] = kdt$.

(b) This can only be applied to such models, for which the analytic solution is known, i.e. for linear models and maybe for simple enough nonlinear models. For complex nonlinear models this approach is no longer feasible.

An alternative approach to using the exact analytic solution is using explicit numerical integration schemes. Consider a forward difference scheme (*upwind* or explicit Euler)

$$\frac{\boldsymbol{x}(t+dt) - \boldsymbol{x}(t)}{dt} \approx \dot{\boldsymbol{x}}(t) = A\boldsymbol{x} + B\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

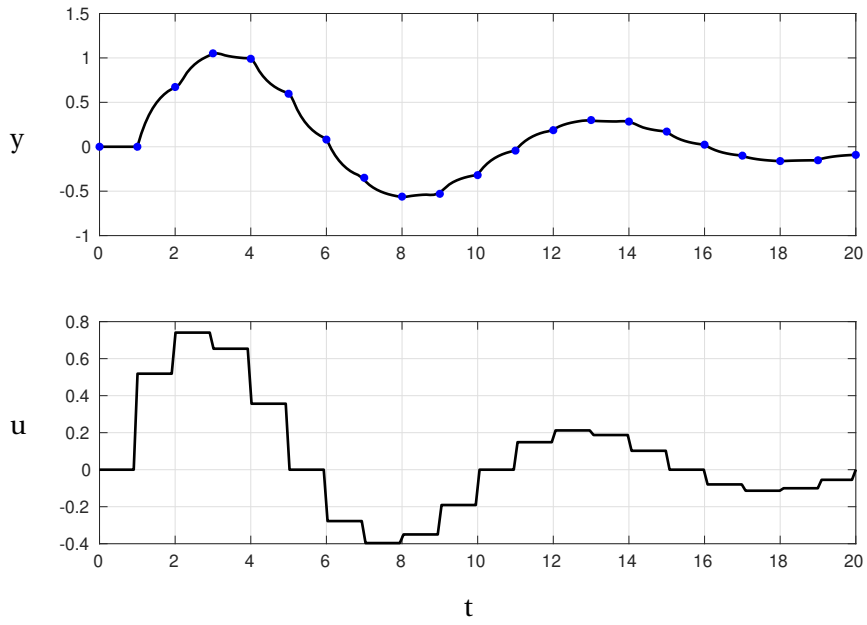


Figure 1.5: Comparison of the solutions (upper diagram) for the continuous- and discrete-time models (1.14) (black curve) and (1.15) (blue dots) with the associated input signal (lower diagram).

so that one directly obtains

$$\boldsymbol{x}(t+dt) \approx \boldsymbol{x}(t) + dt\,(A\boldsymbol{x}(t) + B\boldsymbol{u}(t))$$

and thus the explicit discrete-time scheme

$$\boldsymbol{x}[k+1] = (I + dtA)\,\boldsymbol{x}[k] + dtB\boldsymbol{u}[k], \quad \boldsymbol{u}[k] = \boldsymbol{u}(t[k]), \quad \boldsymbol{x}[0] = \boldsymbol{x}_0.$$

One again ends up with a discrete-time (sampled) model equation of the form

$$\begin{aligned}\boldsymbol{x}[k+1] &= A_d\boldsymbol{x}[k] + B_d\boldsymbol{u}_d[k], \quad \boldsymbol{x}[0] = \boldsymbol{x}_0, \quad \boldsymbol{u}_d[k] = \boldsymbol{u}(t[k])\\ \boldsymbol{y}[k+1] &= C\boldsymbol{x}[k] + D\boldsymbol{u}_d[k].\end{aligned}$$

Note that this obviously has some limitations on the sampling time dt:

- dt must be small enough in relation to the rate of change of the input signal $\boldsymbol{u}$. For example for $u(t) = \sin(\omega_0 t)$ it follows by the Nyquist-Shannon theorem that the essential information of the input signal is only included in the samples if $dt \le \frac{1}{2\omega}$.
- Numerically, even for a Hurwitz matrix, if dt is too large, instability can occur.

Example 1.7. *Considering the dynamics* (1.14) *from the previous example with* $\lambda = -2, \quad b = 3$ *and* $dt = 1$ *it follows that*

$$x[k+1] = (1-2dt)x[k] + 3dtu[k] = -x[k] + 3u[k].$$

In consequence, even for $u[k] = 0, \quad x[0] \neq 0$ *the solution* $x[k]$ *will oscillate between* $x[0]$ *and* $-x[0]$ *for all time, showing that no good approximation of the real behavior (a stable step response) is obtained.*

Considering now the input signal $u[k] = e^{-t[k]/8} \sin\left(\frac{2\pi}{T_u} t[k]\right)$ *with* $T_u = 10$ *and* $dt = 1$ *the behavior shown in Figure 1.6 is obtained.*

The situation becomes worse when $dt = 1.2$ *is considered, as shown in Figure 1.7.*

If the step-size is chosen small enough, then a good approximation is obtained. This can be seen in Figure 1.8 for the choice $dt = 0.5$.

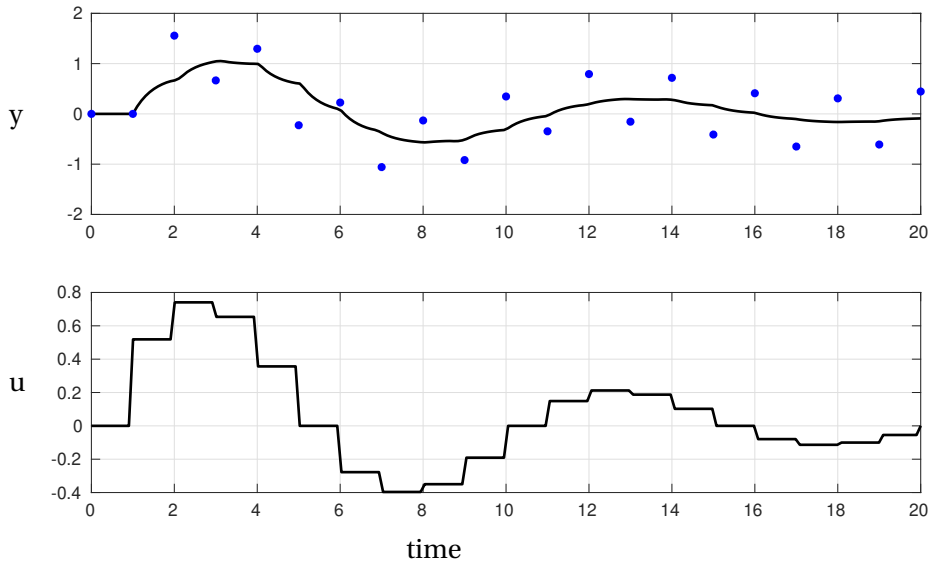


Figure 1.6: Comparison of the solutions of the continuous-time model and the explicitely discretized models with $dt = 1.0$.

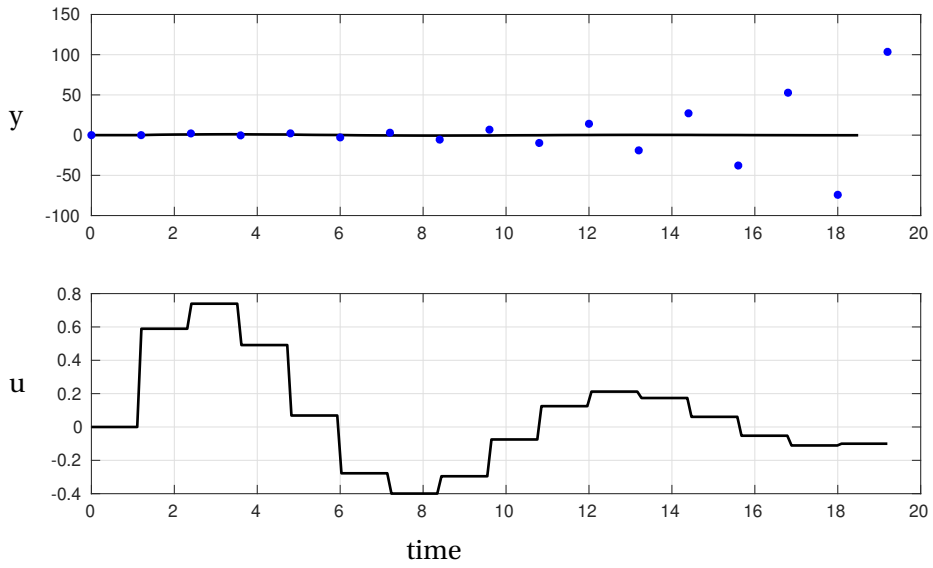


Figure 1.7: Comparison of the solutions of the continuous-time model and the explicitely discretized models with $dt = 1.2$.

The advantage of this direct discretization approach is that it can be used for nonlinear systems as well:

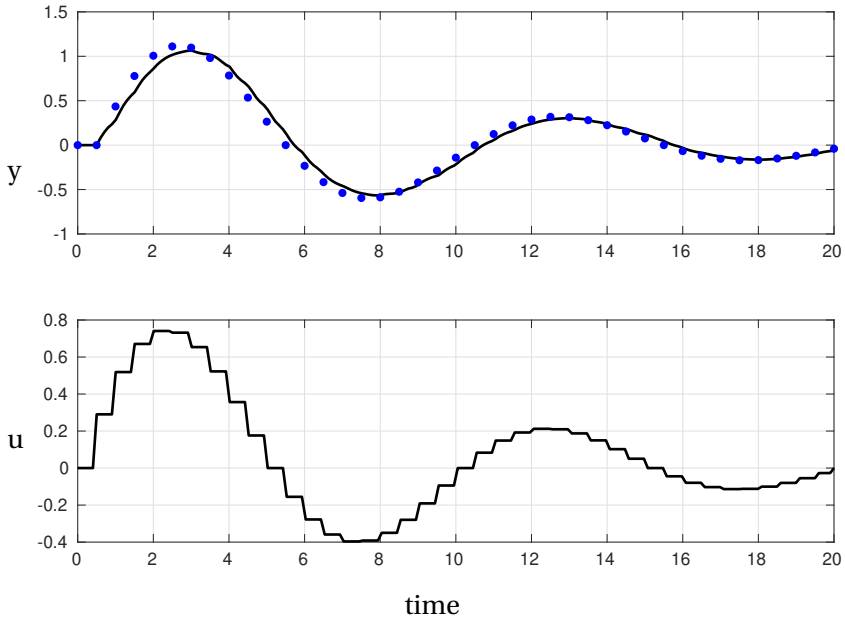


Figure 1.8: Comparison of the solutions of the continuous-time model and the explicitely discretized models with $dt = 0.5$.

Consider a system modeled by the the general nonlinear ode

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{u}) \end{aligned}$$

Using the explicit Euler scheme

$$\frac{\boldsymbol{x}(t+dt) - \boldsymbol{x}(t)}{dt} \approx \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

one obtains

$$\begin{aligned} \boldsymbol{x}[k+1] = \boldsymbol{x}(t+dt) &\approx \boldsymbol{x}[k] + dt\,\boldsymbol{f}(\boldsymbol{x}[k], \boldsymbol{u}_d[k]), \quad \boldsymbol{x}[0] = \boldsymbol{x}_0 \\ \boldsymbol{y}[k] &= \boldsymbol{h}(\boldsymbol{x}[k], \boldsymbol{u}_d[k]). \end{aligned}$$

If $\boldsymbol{f}(\boldsymbol{x})$ does not change to much with respect to $\boldsymbol{x}$ than for small enough step sizes dt a good approximation can be obtained using this method. It can be shown that the error of the approximation of the explicit Euler method is of order $\mathcal{O}(dt)$, meaning that it scales linearly in the step size [GH10].

Beyond this direct Euler discretization more complex numerical discretization methods can be used that achieve a better approximation behavior even for moderate step sizes. An examples is the Heun method (here presented for a system without input), which is kind of a prediction (with an explicit Euler method) and correction scheme yielding

$$\begin{aligned} \boldsymbol{x}[0] &= \boldsymbol{x}_0 \\ \boldsymbol{x}_e[k+1] &= \boldsymbol{x}[k] + dt\,\boldsymbol{f}(\boldsymbol{x}[k]) \\ \boldsymbol{x}[k+1] &= \boldsymbol{x}[k] + \frac{dt}{2}\left(\boldsymbol{f}(\boldsymbol{x}[k]) + \boldsymbol{f}(\boldsymbol{x}_e[k+1])\right) \\ \boldsymbol{y}[k] &= \boldsymbol{h}(\boldsymbol{x}[k]), k \in \mathbb{N}. \end{aligned}$$

In the correction scheme the mean value of $\boldsymbol{f}$ evaluated at the points $\boldsymbol{x}[k]$ and $\boldsymbol{x}_e[k+1]$ is used to obtain a better prediction of the next value $\boldsymbol{x}[k+1]$. The approximation error of the Heun method is of order $\mathcal{O}^2(dt)$ [GH10], meaning that it depends on the square of dt and thus with decreasing dt quickly becomes smaller. Accordingly, the Heun method achieves a better performance than the explicit Euler approximation at the cost of a more complex implementation and additional storage.

More advanced methods with better approximation performance are the different Runge-Kutta methods [GH10] which are e.g. implemented in standard distributions of MATLAB or PYTHON/SCIPY.

1.2.2 Stability of discrete-time systems

It is important to *a priori* know the maximum sample time that can be used to capture the dynamic behavior of a given system and to ensure stability at the same time. For this purpose, in the sequel the basics of discrete-time stability theory for linear time-invariant systems are recalled [Oga95; DB05].

Definition 1.3

A point $\boldsymbol{x}^* \in \mathbb{R}^n$ is called a fixed point of the discrete-time system $\boldsymbol{x}[k+1] = \boldsymbol{f}(\boldsymbol{x}[k])$ if it holds that $\boldsymbol{x}^* = \boldsymbol{f}(\boldsymbol{x}^*)$.

Definition 1.4

A fixed point $\boldsymbol{x}^*$ is stable (in the sense of Lyapunov) if for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $\boldsymbol{x}_0$ satisfying $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \delta$ it holds that $\|\boldsymbol{x}[k] - \boldsymbol{x}^*\| \leq \epsilon$ for all $k \geq 0$. If it is not stable it is called unstable.

Definition 1.5

A fixed point $\boldsymbol{x}^*$ is asymptotically stable in a set $D \subseteq \mathbb{R}^n$ if it is stable and attractive in D, i.e. for every $\boldsymbol{x}_0 \in D$ it holds that $\lim_{k\to\infty} \|\boldsymbol{x}[k] - \boldsymbol{x}^*\| = 0$.

Recall that the (closed) unit circle $\bar{U}_1$ is defined as the set of all complex numbers with modulus less or equal to 1, i.e.

$$\bar{U}_1 = \{\lambda \in \mathbb{C} \,|\, |\lambda| \leq 1\}. \tag{1.16}$$

Theorem 1.2

The fixed point $\boldsymbol{x}^* = \boldsymbol{0}$ of the linear system $\boldsymbol{x}[k+1] = A\boldsymbol{x}[k]$ is stable if and only if the eigenvalues of A are contained in the (closed) unit circle $\bar{U}_1 \subset \mathbb{C}$, and asymptotically stable iff they are contained in the open unit circle U_1.

To illustrate this result, recall example 1.7 with the explicit Euler numerical integration scheme and let $u = 0$:

- For $dt = 1$ it follows that $x[k+1] = -x[k]$, meaning that $x = 0$ is stable but not asymptotically stable
- For $dt = 1.2$ it follows that $x[k+1] = -1.4x[k]$, meaning that $x = 0$ is unstable
- For $dt = 0.5$ it follows that $x[k+1] = 0x[k]$, meaning that $x = 0$ is asymptotically stable

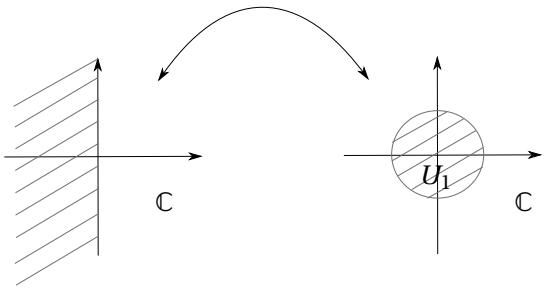


Figure 1.9: Mapping of the stability regions of continuous- and discret-time systems. The left-half complex plane is mapped by $e^{\lambda t}$ with $t \geq 0$ into the unit circle.

To see the relation between continuou- and discrete-time stability, recall that for continuous-time systems it must hold that all eigenvalues λ_i are contained in the left half complex plane $\mathbb{C}_-$. This is maped by $\exp(\lambda_i t)$ exactly into the unit circle U_1 as depicted in Figure 1.9.

1.2.3 The z-transform

Consider the sampled data signal $y[k]$. The (unilateral) z-transform is defined as [OS14; DB05]

$$\hat{y}(z) = \mathcal{Z}\{y[k]\}(z) = \sum_{k=0}^{\infty} y[k] z^{-k}, \qquad z = a + jb \in \mathbb{C}.$$

Given that the z-transform is given by an infinite series it has to be clarified for which values of z it converges. This leads to the concept of the region of convergence $\mathcal{R}(y) \subset \mathbb{C}$ for a given signal y defined as

$$\mathcal{R}(y) = \left\{ z \in \mathbb{C} \,\middle|\, \left| \sum_{k=0}^{\infty} y[k] z^{-k} \right| < \infty \right\}.$$

Example 1.8. *For $y[k] = q^k, |q| < 1$, recalling the geometric series it follows that*

$$\mathcal{Z}\{y[k]\}(z) = \sum_{k=0}^{\infty} q^k z^{-k} = \sum_{k=0}^{\infty} \left(\frac{q}{z}\right)^k = \frac{1}{1 - \frac{q}{z}} = \frac{z}{z-q} \quad \forall \left|\frac{q}{z}\right| < 1$$

- *Note that there is a pole at $z = q$, thus the condition $|z| > |q|$ means, that $\mathcal{R}(y)$ encloses the region around the pole.*
- *In accordance with the stability theory, considering the relation of $y[k+1]$ and $y[k]$, it is clear that in the present case $y[k+1] < y[k]$, meaning that the eigenvalue of the dynamics is less than 1 and the associated discrete-time system is asymptotically stable.*

Some important properties of the z-transform are recalled next:

- **Linearity:**
$$\mathcal{Z}\{a_1 y_1[k] + a_2 y_2[k]\}(z) = a_1 \hat{y}_1(z) + a_2 \hat{y}_2(z)$$
- **Time delay** $(0 < \Delta \in \mathbb{N})$**:**
$$\mathcal{Z}\{y[k-\Delta]\}(z) = z^{-\Delta} \hat{y}(z)$$

- **Positive time shift:**
$$\mathcal{Z}\{y[k+1]\}(z) = z\hat{y}(z) - zy[0]$$

- **Positive time shift** $(0 < \Delta \in \mathbb{N})$**:**
$$\mathcal{Z}\{y[k+\Delta]\}(z) = z^{\Delta}\hat{y}(z) - \sum_{n=0}^{\Delta-1} y[n]z^{-(n-\Delta)}$$

- **Backward difference:**
$$\mathcal{Z}\{y[k] - y[k-1]\}(z) = \left(1 - z^{-1}\right)\hat{y}(z), \quad y[k] = 0 \text{ for } k < 0$$

- **Forward difference:**
$$\mathcal{Z}\{y[k+1] - y[k]\}(z) = (z-1)\hat{y}(z) - zy[0]$$

- **Scaling:**
$$\mathcal{Z}\{a^k y[k]\}(z) = \hat{y}\left(\frac{z}{a}\right)$$

- **Convolution:**
$$\mathcal{Z}\{y_1[k] * y_2[k]\}(z) = \hat{y}_1(z)\hat{y}_2(z)$$

Having these properties at hand it is e.g. straight-forward to calculate the discrete-time transfer function of the system (1.13) as

$$\hat{g}(z) = C(zI - A_d)^{-1}B_d + D. \tag{1.17}$$

Remark 1.2

- In general, a discrete-time transfer function $\hat{g}(z) = \frac{\hat{y}(z)}{\hat{u}(z)}$ is associated to a stable dynamics if all poles are contained in the (closed) unit circle.
- If the poles of $\hat{g}(z)$ are contained in the open unit circle the associated system is asymptotically stable.
- In the case of a stable process the inverse z-transform is given by
$$y[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{y}\left(e^{j\omega}\right)e^{j\omega k}d\omega$$
(i.e. the countour integral along the unit circle).

Example 1.9. *For the following signals one obtains*

- *Unit step*
$$u[k] = h[k] = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0 \end{cases}, \quad \Rightarrow \quad \mathcal{Z}\{u[k]\}(z) = \frac{1}{1 - z^{-1}}, \quad |z| > 1$$

- *Dirac (unit) impulse*

$$u[k] = \delta[k] = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}, \qquad \Rightarrow \quad \mathcal{Z}\{u[k]\}(z) = 1$$

- *First order step response (follows from the scaling property above)*

$$y[k] = a^k h[k], \qquad \Rightarrow \quad \mathcal{Z}\{y[k]\}(z) = \frac{1}{1 - az^{-1}}, \quad |z| > |a|$$

To obtain the z-transform of a given signal from the knowledge of Laplace transform some of the following approaches are typically used:

- The bilinear transformation (*Tustin* method)

$$\hat{u}_Z(z) = \hat{u}_L(s)|_{s=\frac{2}{dt}\frac{z-1}{z+1}} \tag{1.18}$$

- The starred transformation

$$\hat{u}_Z(z) = \hat{u}_L(s)|_{s=\frac{1}{dt}\ln(z)}. \tag{1.19}$$

References

[Ari78] R. Aris. *Mathematical Modeling Techniques*. Dover, 1978 (cit. on p. 1).

[DB05] R. C. Dorf and R. H. Bishop. *Modern Control Systems*. Pearson Prentice Hall, 2005 (cit. on pp. 14, 15).

[Det88] J. W. Dettman. *Mathematical Methods in Physics and Engineering*. Dover Publications, McGraw-Hill, 1988 (cit. on p. 6).

[Dym07] H. Dym. *Linear algebra in action*. American Mathematical Society, 2007 (cit. on p. 2).

[Fis09] G. Fischer. *Lineare Algebra*. Vieweg, 2009 (cit. on p. 2).

[GH10] D. Griffiths and D. J. Higham. *Numerical Methods for Ordinary Differential Equations*. Springer-Verlag, London, 2010 (cit. on pp. 13, 14).

[Har64] P. Hartman. *Ordinary Differential Equations*. John Wiley and Sons, New York, 1964 (cit. on p. 2).

[Kai80] T. Kailath. *Linear Systems*. Prentice Hall, Inc., 1980 (cit. on pp. 2, 6).

[Oga95] K. Ogata. *Discrete-time control systems*. Pearson Prentice Hall, 1995 (cit. on p. 14).

[OS14] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Pearson Prentice Hall, 2014 (cit. on p. 15).

[Tes12] G. Teschl. *Ordinary Differential Equations and Dynamical Systems*. AMS, 2012 (cit. on p. 2).

Introduction to system identification

2

2.1 General notions and concepts

In this section basic notions and important concepts in the identification of linear time-invariant systems are summarized. The intention is to give a general idea of the underlying problem and some solution approaches. It is not possible to include all aspects of system identification here. For additional information and to further deepen the discussion the reader is referred to the literature, in particular [Ari94; ÅM08; Dym04].

2.1.1 Mathematical modeling

First of all consider the general questions, which have to be clarified before starting over prior to any system identification.

Remark 2.1

What is a *mathematical* model?

A mathematical representation of certain aspects and relations in a given system in form of algebraic, differential or difference equations.

For which purposes are mathematical models employed?

- Analysis for understanding phenomena ⟶ physics, biology, economics, etc.
- Prediction of future behavior on the basis of present (and evtl. past) information ⟶ weather forecast, stocks market, etc.
- Design of a prescribed system behavior by systematic construction approaches ⟶ Engineering applications (electrical, chemical, etc.)
- System monitoring ⟶ energy grid, water distribution systems, chemical process plants, etc.
- Control design ⟶ stabilization, trajectory tracking, disturbance rejection, etc.

2.1.2 Historical perspective and isomorphisms of concepts

There are two main branches of science and engineering that led to substantially new concepts over the past centuries that are nowadays standard notions in systems theory. Some main developments are summarized in the sequel. For more detailed descriptions the reader is referred e.g. to the book [ÅM08].

2.1.2.1 The heritage of mechanics

Mechanics, as a discipline of physics has introduced important basic concepts and approaches that are nowadays used as a standard in system modeling. Some of the many precedents which have to be named here are the following :

- Experimental studies of plenatary motions (e.g. by Tycho Brahe and Johannes Kepler) ⟶ planet motions follow elliptic paths
- Theoretical explanations by Newton ⟶ explanation of elliptic movement by law of gravity and $F = ma$

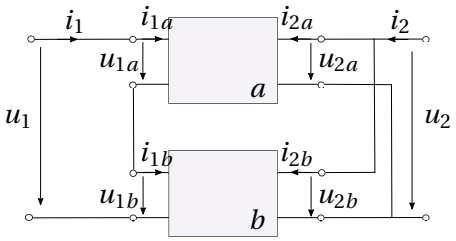
This important step led to the finding that the motion of planets can be well predicted if the position and velocities of all planets in the solar systems are known for a given time t.

With this idea in mind the basic notion of dynamical systems theory was established. In particular the notion of the state of a system as the set of variables that must be known at a given time t to predict the future behavior of the system (*in the case of the planetary system, the position and velocities of all planets*)

2.1.2.2 The heritage of electrical engineering

Some of the main conceptual breakthroughs that came along with the development of electrical engineering are summarized next:

- The design of electrical amplifiers led to the analysis of input-output behavior
- Complex systems are viewed as interconnections of different *simple* systems with respective inputs and outputs.



- Analysis methods using step response ($u(t) = h(t)$) and frequency responses ($u(t) = a\sin(\omega t)$).

These notions are widely used in system modeling and identification.

2.1.3 Limitations of a model

With all these important precedents and the highly developed state of knowledge and understanding in modern science and with all the technological breakthroughs that have been achieved it is important to remember that a mathematical model is always a trade-off between accuracy and complexity. This important fact can be illustrated e.g. for a simple mass-spring-damping system.

> **Example 2.1.** ***Mass-spring-damping system:***
>
> - *A precise model would require to describe the mass and stress distributions along the spring, eventual rotational movements of the mass with inhomogenous mass distribution and moments of inertia, different types of friction, etc.*

- *A simple model, considering a homogenous, linear force of the spring and a point mass only accounts for some basic characteristics such as frequency of oscillation, damping, etc.*
- *The simple model can anyway be used for design and control purposes.*

This simple discussion should emphatize the fact, that the *quality of a model* depends on its later use. In accordance, when considering system identification it is important to keep in mind for which purpose the model will be used that is about to be identified.

2.2 Different types of model identification

A model for a dynamical system always has two components, namely structure and parameters.

Example 2.2. *For the first order model* $\dot{x}(t) = \lambda x(t), \quad x(0) = x_0, \quad y(t) = x(t)$ *one has the following:*

- *structure: first order linear time-invariant autonomous differential equation*
- *parameter:* λ.

According to this fact the model identification methods can be characterized in three groups:

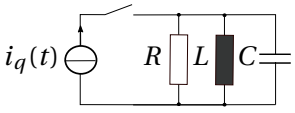
- White-box modeling: First-principles modeling and experiment-based parameter identification (model structure known, parameters known or experimentally determined)
- Grey-box modeling: First-principles modeling and combined experimental and data-driven parameter identification (model structure known, parameters (partially) unknown)
- Black-box modeling: Model and parameter identification using standard models and data-driven parameter identification (model structure and parameters unknown)

In the sequel these approaches will be discussed with more detail.

2.2.1 White-box identification

In the white-box identification approach the model structure is known from (e.g. physics-based) first principles modeling and the parameters are either known or can be determined by simple experiments, like weighing a mass, measuring a length, or are known from data sheets of electrial elements, etc.

Example 2.3. *Consider the following RLC circuit*



By Kirchhoff's laws it holds that

$$i_q(t) - i_R(t) - i_L(t) - i_C(t) = 0, \quad u_R(t) = u_C(t) = u_L(t).$$

Furthermore, the contained electrical elements satisfy

$$i_R(t) = \frac{u_R(t)}{R}, \quad i_C(t) = C\frac{du_C(t)}{dt}, \quad u_L(t) = L\frac{di_L(t)}{dt}.$$

By substitution it follows that

$$i_q(t) - \frac{L}{R}\frac{di_L(t)}{dt} - i_L(t) - LC\frac{d^2 i_L(t)}{dt^2} = 0.$$

Finally, the model is given by

$$\frac{d^2 i_L(t)}{dt^2} + \frac{1}{RC}\frac{di_L(t)}{dt} + \frac{1}{LC} i_L(t) = \frac{1}{LC} i_q(t)$$
$$i_L(0) = i_{L0}, \quad \frac{di_L(0)}{dt} = \frac{1}{L} u_{C0}.$$

Accordingly, the following can be said about the structure and parameters of the model:

- *structure: inhomogenous (because of $i_q(t)$) second-order linear time-invariant differential equation*
- *parameters: R, L, C, which are previously known from data sheets (at least approximately...).*

Based on the second-order model a set of first-order differential equations can be used following the standard state-space representation with $x_1 = i_L, x_2 = \frac{di_L}{dt}$ and $u = i_q$, leading to

$$\dot{x}_1 = x_2, \qquad x_1(0) = x_{10}$$
$$\dot{x}_2 = -\frac{1}{LC} x_1 - \frac{1}{RC} x_2 + \frac{1}{LC} u, \qquad x_2(0) = x_{20}$$

Using the model we can predict e.g. the behavior for a step-wise input $u(t) = i_q(t) = \hat{i}_q h(t)$ with $h(t)$ being the Heaviside unit step:

- *Consider $y = x_1 = i_L$*
- *If $R > \frac{L}{C}$ a damped oscillation will be observed with frequency*

$$\omega = \sqrt{\frac{4R^2C - L}{4L(RC)^2}}$$

and damping rate constant

$$d = \frac{1}{2RC}$$

converging to the stationary value

$$y_\infty = \hat{i}_q$$

as shown in Figure 2.1.

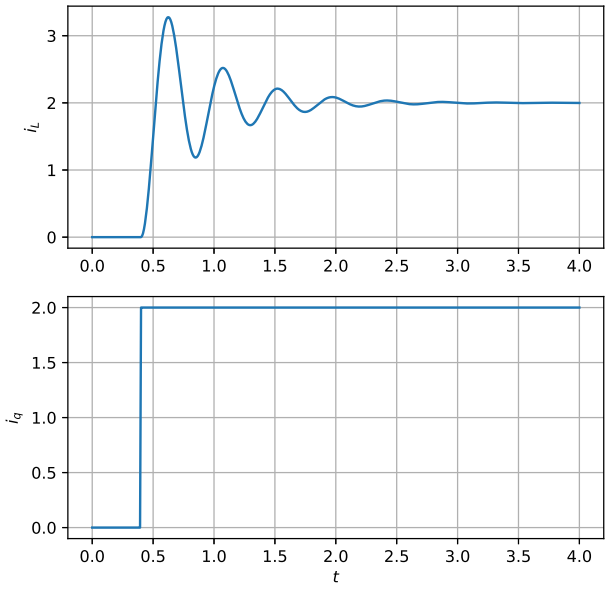
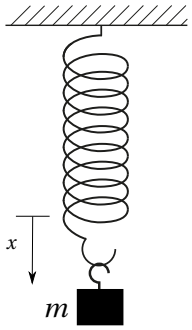


Figure 2.1: Typical step response for the *RLC* circuit in example 2.3

Example 2.4. *Consider the mass-spring system shown in conitnuation:*



The application of Newton's law and Hook's law yield that for neglectable damping and friction a model is given by

$$m\ddot{x} = -kx + mg$$

with initial conditions $x(0) = x_0, \dot{x}(0) = v_0$. *The model structure is thus given by a second order autonomous differential equation. The model parameters are given by* m, g, k.

In the equilibrium point it holds that $\ddot{x} = 0, \dot{x} = 0$ *implying* $kx^* = mg$ *or*

$$x^* = \frac{mg}{k}.$$

A suitable state-space representation is determined with the coordinate shift

$$x_1 = x - x^*, \quad x_2 = \dot{x}_1$$
$$\dot{x}_1 = x_2, \qquad\qquad x_1(0) = x_{10}$$

$$\dot{x}_2 = -\frac{k}{m} x_1, \qquad x_2(0) = x_{20}$$

In this new form the model structure is given by a set of two autonomous ordinary differential equations of first order. The model parameters are given by m, k, meaning that there is one parameter less than in the original model (which is hidden in the coordinate shift).

The next question is how to identify these parameters?

- *The mass can simply be weighed.*
- *The stiffness k of the spring can eventually be read out of a data sheet or determined by experiment. An experimental determination of k could be carried out using the following steps:*

  As stated above, in equilibrium it holds that $x^* = \frac{mg}{k}$ or equivalently

  $$k = \frac{mg}{x^*}.$$

  Accordingly, for q different masses m_i $i = 1, \ldots, q$ one obtaines q different equilibrium points x_i^* and hence q values k_i, $i = 1, \ldots, q$. A suitable parameter k for the model can then be calculated e.g. using the arithmetic mean value k_a or the geometric mean value k_g according to

  $$k_a = \frac{1}{q} \sum_{i=1}^{q} k_i, \qquad k_g = \left(\prod_{i=1}^{q} k_i \right)^{\frac{1}{q}}$$

  Data from such an experiment is shown in the following table:

Mass [g]	Height [cm]	Extension	X-value	Stiffness k
0,00	52,00	0,00	-25,00	--
62,80	41,50	10,50	-14,50	58,67
81,10	37,00	15,00	-10,00	53,04
132,80	***27,00***	***25,00***	***0,00***	***52,11***
146,50	22,50	29,50	4,50	48,72
151,10	21,50	30,50	5,50	48,60
172,26	17,00	35,00	10,00	48,28
185,98	14,00	38,00	13,00	48,01
195,60	11,00	41,00	16,00	46,80

zero position	Gravity	Arith. mean	Geom. mean	Exp. Weight
25	9,81	50,53	50,40	132,80

- *For a model validation a step response can be used. The predicted frequency of the oscillation is given by*

  $$\omega = \sqrt{\frac{k}{m}}$$

  The measured frequencies are shown in the following table:

Frequencies			
Calc. (arit)	Calc. (geom)	Direct k(132,8]	Experiment
0,62	0,62	0,63	0,63

- *If the result is not accurate enough additional experiments can be carried out to improve the parameter value.*

2.2.2 General considerations about the identification process

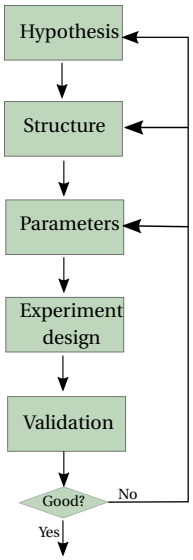
The White-box identification employs first-principle modelling techniques together with simple determination of system parameters using experiments or data sheets.

Actually, any kind of identification process depends on and is limited by the measurement capabilities following the rule

> *Any model can just be as good as the sensor that is used for its validation.*

Typically, several iterations are needed to end with a model that is useful for a given purpose and fits all requirements (like real-time capabilities, accuracy, observability, controllability, etc.).

The consideration above yield the following scheme that is representative for the white-box identification in particular but also for any identification process in general.



1. Clarify hypothesis including system boundaries
2. Choose a system structure (first principles, standard transfer function or state-space model)
3. Identify parameters (data sheets, experiments, parameter optimization)
4. Design a validation experiment and obtain (new) measurements to test the model
5. Inspect the correspondence with measurement (goodness of fit)
6. Iterate the above until a good fit is obtained

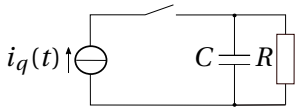
2.2.3 Grey-box identification

When the model is known from first principles modeling but some of the parameters or all paramters are unknown the input-output data must be used to determine the unknown parameters. The approach is known as grey-box identification (model structure known but parameters (*partially*) unknown) [Tul93; Soh98; Boh91; Boh06; Ste84] The determination of the unknown parameters can be carried out using different types of data sets, e.g.

1. Step or impulse response (continuous or sampled)
2. Frequency response (Bode or Nyquist diagram)

After a suitable set of parameters is obtained using one data set, other, independent data sets (if at hand) should be used to verify the identified model.

Example 2.5. *Consider the RC-circuit with known resistence R but unknown capacitance C.*



According to basic laws of electrical circuit modeling, a continuous-time model for the voltage at the capacitance is given by

$$\dot{u}_C = -\frac{1}{RC}u_C + \frac{1}{C}i_q, \quad u_C(0) = u_{C0}$$

Suppose that a discrete-time (sampled) input-output data set is at hand. Following the discussion in Section 1.2.1, the exact discrete-time model for sampling time Δt is given by

$$u_C[k+1] = a_d u_C[k] + b_d i_q[k], \quad u_C[0] = u_{C0}, \quad a_d = e^{-\frac{\Delta t}{RC}}, \quad b_d = R(1-a_d).$$

Considering the input-output data set is given by $\{i_q[0],\ldots,i_q[m]\},\{u_C[0],\ldots,u_C[m]\}$

Using the discrete-time model it follows that

$$\begin{aligned} u_C[1] &= a_d u_C[0] + b_d i_q[0] \\ u_C[2] &= a_d u_C[1] + b_d i_q[1] \\ &\cdots \\ u_C[m] &= a_d u_C[m-1] + b_d i_q[m-1] \end{aligned}$$

Given that C is unknown, both a_d and b_d are unknown.

Considering three subsequent data points, i.e. for $n-1, n$ and $n+1$, it follows that

$$\begin{aligned} u_C[n] &= a_d u_C[n-1] + b_d i_q[n-1] \\ u_C[n+1] &= a_d u_C[n] + b_d i_q[n] \end{aligned}$$

In Matrix notation this reads

$$\begin{bmatrix} u_C[n] \\ u_C[n+1] \end{bmatrix} = \begin{bmatrix} u_C[n-1] & i_q[n-1] \\ u_C[n] & i_q[n] \end{bmatrix} \begin{bmatrix} a_d \\ b_d \end{bmatrix}$$

As long as the rows (or columns) of the matrix

$$M = \begin{bmatrix} u_C[n-1] & i_q[n-1] \\ u_C[n] & i_q[n] \end{bmatrix}$$

are linearly independend (what in non-equilibrium conditions is expected to hold...) M can be inverted and the parameters a_d and b_d can be determined according to

$$\begin{bmatrix} a_d \\ b_d \end{bmatrix} = M^{-1} \begin{bmatrix} u_C[n] \\ u_C[n+1] \end{bmatrix} = \begin{bmatrix} \frac{i_q[n]u_C[n]-i_q[n-1]u_C[n+1]}{u_C[n-1]i_q[n]-u_C[n]i_q[n-1]} \\ \frac{-u_C[n]^2+u_C[n-1]u_C[n+1]}{u_C[n-1]i_q[n]-u_C[n]i_q[n-1]} \end{bmatrix}$$

The capacitance C is then determined by the relation

$$C = -\frac{\Delta t}{R\ln(a_d)}.$$

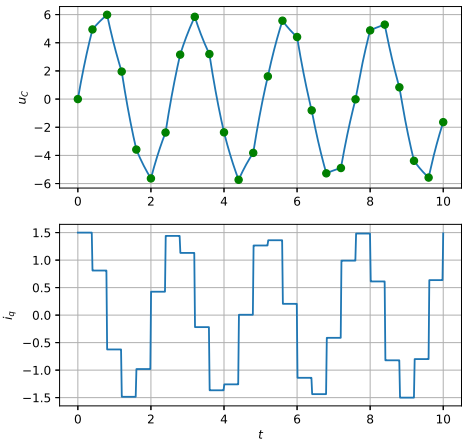
Note that the approach used in Example 2.5 also works if both R and C are unkown because $b_d = R(1-a_d)$ is also contained in the solution vector. In this case it holds that

$$R = \frac{b_d}{1-a_d}, \quad C = \frac{-\Delta t}{R\ln(a_d)} = \frac{-\Delta t(1-a_d)}{b_d \ln(a_d)}$$

Depending on the particular values of $i_q[n-1], i_q[n], u_C[n-1], u_C[n], u_C[n+1]$ (i.e. for a given data set the value of n) the equations to be solved can be more or less well conditioned (in the sense of the numerical solver used for the matrix inversion).

Example 2.6. *Comparison of exact values $R = 10\Omega, C = 0.1$F and data generated using the numerical solution in* PYTHON *for $u(t) = 1.5\cos(2.5t)$ and sampling time $\Delta t = 0.4$*

n	R	C
1.0	10.0	0.10000000000000002
2.0	9.9999999999999964	0.10000000000000005
3.0	10.000000000000004	0.10000000000000002
4.0	10.0	0.10000000000000002
5.0	10.000000000000002	0.10000000000000001
6.0	9.9999999999999982	0.10000000000000003
7.0	10.000000000000002	0.10000000000000001
8.0	9.9999999999999982	0.10000000000000003
9.0	10.000000000000005	0.10000000000000001
10.0	10.000000000000002	0.10000000000000001
11.0	10.000000000000002	0.10000000000000001
12.0	10.000000000000002	0.10000000000000001
13.0	9.9999999999999964	0.10000000000000001
14.0	10.0	0.10000000000000002
15.0	9.9999999999999982	0.10000000000000003
16.0	9.9999999999999964	0.10000000000000001
17.0	10.000000000000005	0.10000000000000003
18.0	10.000000000000005	0.10000000000000003
19.0	10.000000000000005	0.10000000000000001
20.0	9.9999999999999929	0.10000000000000001
21.0	10.000000000000002	0.10000000000000001
22.0	9.9999999999999964	0.10000000000000001
23.0	10.0	0.10000000000000002
24.0	9.9999999999999964	0.10000000000000001



Green dots: data points
Blue line: numerical solution of ode.

It can be seen that good results are obtained that are close to the actual values of R, C.

Note that in Example 2.6 an ideal situation is simulated in which the data does not contain any kind of uncertainty. Thus the obtained parameter values pretty much correspond to the real ones. In real measurement data always measurement uncertainty and noise are contained. Taking into account all the data points will then statistically (*to be clarified during the subsequent discussion*) yield better results.

A common measure of goodness of fit is the mean squared error (mse). Let $\boldsymbol{x}(t;\boldsymbol{x}_0,\boldsymbol{p},\boldsymbol{u}(\cdot))$ be the solution of the continuous-time model equations for a given parameter vector $\boldsymbol{p}$ and input $\boldsymbol{u}(t)$, $t \geq 0$. Let $y_s[k]$, $k = 1,\ldots,m$ be the measured sensor data values at the time instances $t_k = k\Delta t$ and $y[k] = \boldsymbol{c}^T\boldsymbol{x}(k\Delta t)$ the calculated ones. Then the mse is defined as

$$e_{\boldsymbol{p}} = \frac{1}{m}\sum_{k=1}^{m}\big(y[k]-y_s[k]\big)^2. \tag{2.1}$$

Consider the general first order discrete-time model

$$y[k+1] = ay[k] + bu[k]$$

with the parameter vector $\boldsymbol{p} = [a,b]^T$. Suppose that m input and sensor data points $(u[k], y_s[k])$ are given. The input-output data can be used to obtain $m-1$ linear equations

$$\begin{aligned} y_s[1] &= ay_s[0] + bu[0] \\ &\vdots \\ y_s[m] &= ay_s[m-1] + bu[m-1] \end{aligned}$$

or equivalently in matrix notation

$$\begin{bmatrix} y_s[1] \\ \vdots \\ y_s[m-1] \end{bmatrix} = \begin{bmatrix} y_s[0] & u[0] \\ \vdots & \vdots \\ y_s[m-1] & u[m-1] \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

This representation can be written as

$$\boldsymbol{y}_s = M\boldsymbol{p}, \quad M \in \mathbb{R}^{(m-1)\times 2}.$$

Given the dimensions of M, this matrix cannot be directly inverted. A possibility to obtain a solution $\boldsymbol{p}$ is by rewriting the linear equation as a quadratic minimization problem:

> Find $\boldsymbol{p}$ to minimize the mean squared error (mse), i.e.
>
> $$\min_{\boldsymbol{p}} \frac{1}{m}\big(\boldsymbol{y}_s - M\boldsymbol{p}\big)^T\big(\boldsymbol{y}_s - M\boldsymbol{p}\big).$$

The mse can be written as

$$e_{\boldsymbol{p}} = \boldsymbol{y}_s^T\boldsymbol{y}_s - 2\boldsymbol{p}_s^T M^T\boldsymbol{y}_s + \boldsymbol{p}^T M^T M\boldsymbol{p}.$$

This expression attains a minimum if $\boldsymbol{p}$ is chosen so that

$$\frac{\partial e_{\boldsymbol{p}}}{\partial \boldsymbol{p}} = -2M^T\boldsymbol{y}_s + 2M^T M\boldsymbol{p} = \boldsymbol{0}$$

what is achieved for

$$\boldsymbol{p} = (M^T M)^{-1}M^T\boldsymbol{y}_s. \tag{2.2}$$

Note that $M^T M \in \mathbb{R}^{2\times 2}$ is given by

$$M^T M = \begin{bmatrix} \sum_{k=0}^{m-1} y_s^2[k] & \sum_{k=0}^{m-1} y_s[k]u[k] \\ \sum_{k=0}^{m-1} y_s[k]u[k] & \sum_{k=0}^{m-1} u^2[k] \end{bmatrix}$$

Thus, if the columns of M are linearly independent, i.e. if the input and output sequences are not just related by $y[k] = d\,u[u]$ for some $d \in \mathbb{R}$ [1] the quadratic matrix $M^T M$ can be inverted and thus the matrix

$$M^\dagger = \left(M^T M\right)^{-1} M^T \tag{2.3}$$

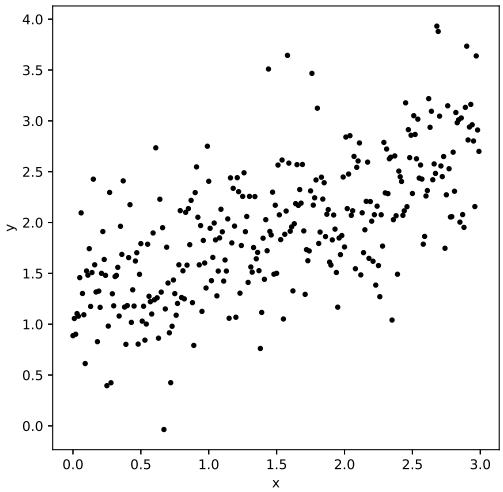
can be calculated. The matrix $M^\dagger$ is called the (Moore-Penrose) pseudo-inverse of M [Dym07; CM91] and is frequenctly used in very different problems whenever a non-quadratic linear equation system has to be solved. From the above considerations it can be seen that the pseudo-inverse exists whenever the columns of the matrix M are linearly independent.

[1] This is the case in non-stationary systems and can allways be ensured for the regarded class of discrete-time models if the input is non-stationary.

2.2.4 Black-box identification

The term black-box identification refers to the situation when no *a priori* information is given about the structure of the model [Lju87]. In this case, both a possible structure and a suitable parameter set have to be identified. This is typically carried out using some normal form (*experience based*). Similar to the grey-box approach, the input (excitation) signal is then chosen so that the parameters become identifiable using e.g. a minimum squarred error identification. After a suitable set of parameters is obtained using one data set, other, independent data sets (if at hand) should be used to verify the identified model. Special care has to be taken when measurement noise is present.

The most simple example of such a situation is the problem of identifying a straight line from noisy measurements.



For the straight line the model structure is given by

$$y = ax + b$$

with the two parameters a, b. Suppose that m samples with a stochastic uncertainty (frequently called noise) are given in the form

$$y_i = ax_i + b + w_i$$

where w_i represents the noise, i.e. the values of a random variable. In absence of noise, on the one hand, the approach from the preceding subsection, based on the minimization of the mean squared error (mse) yields a good fit of the model. In presence of noise, on the other hand, the question of whether this mse approach is useful or not will depend on the specific nature and properties of the noise. Thus, treating with noisy data requires first to set up a noise model, i.e. a specific characterization that can be used to analyze the goodness of fit using a specific parameter identification approach like e.g. mse minimization.

2.2.4.1 Noise model

To identify a noise model basically means to determine the associated probability distribution function (pdf). Typically the white noise model associated to a (Gaussian) normal distribution is considered (motivated by central limit theorems [Dur10]) with mean value μ and standard deviation σ, denoted

by $w_i \sim \mathcal{N}(\mu, \sigma^2)$. The associated probability that w_i attains a value less or equal to x is then given by

$$P(w_i \leq x) = \int_{-\infty}^{x} \varphi(\zeta) \mathrm{d}\zeta,$$

where φ is the Gaussian probability density function

$$\varphi(\zeta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\zeta-\mu)^2}{2\sigma^2}}. \tag{2.4}$$

The typical bell-shape of this function is shown in Figure 2.2

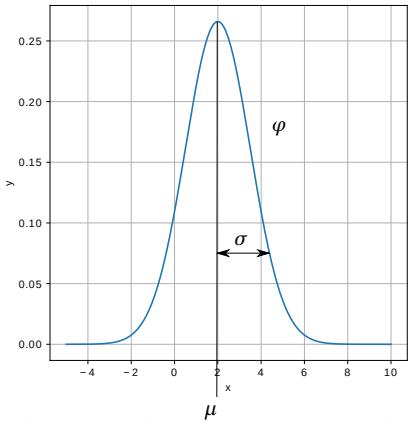


Figure 2.2: Gaussian probability distribution $\varphi(x)$ associated to the white noise process.

In accordance, the measurement noise $w[k]$ is considered as outcome of a stochastic process

$$w[k] = \mathcal{W}_k : \{W(k) \mid k \in \mathbb{N}\}$$

and $\mathcal{W}(k) : (\mathbb{R}, \mathcal{B}, P) \rightarrow (\mathbb{R}, \Sigma)$ being a stochastic variable. The mean value is given by

$$E\{w[n]\} = \int_{-\infty}^{\infty} \zeta \varphi(\zeta) \mathrm{d}\zeta = \mu,$$

and for the covariance between different time instances it holds that

$$\mathrm{Cov}\{w[n]\, w[n+m]\} = \sigma^2 \delta_{n,m}$$

with $\delta_{n,m}$ being the Kronecker-δ.

2.2.4.2 Maximum-likelihood parameter estimation

The likelihood $L(\boldsymbol{p})$ that a given parameter vector $\boldsymbol{p} = [a, b]^T$ corresponds to the given data set (x_i, y_i) with $i = 1, \ldots, m$ can be determined using the related joint probability distribution function (pdf), say

$$L(\boldsymbol{p}) = \phi(\{y[0] = y_0\} \cap \{y[1] = y_1\} \cap \cdots \cap \{y[m] = y_m\}; \boldsymbol{p}). \tag{2.5}$$

Considering that all samples are statistically indepent and identically distributed (i.i.d.) with pdf $\varphi(y[k];\boldsymbol{p})$ it turns out that

$$L(\boldsymbol{p}) = \prod_{k=0}^{m} \varphi(y[k];\boldsymbol{p}).$$

The problem of determining the best parameter vector $\boldsymbol{p}^*$ can now be interpreted as the maximization of the likelihood, i.e.

$$\boldsymbol{p}^* = \text{argmax}_{\boldsymbol{p}\in\mathbb{R}^2} L(\boldsymbol{p}). \tag{2.6}$$

Given that the logarithm is monotonically increasing, the maximum of L and the maximum of $\ln(L)$ are located at the same position $\boldsymbol{p}^*$.

Consider that the individual distributions $\varphi(y[k];\boldsymbol{p})$ correspond to a normal distribution, centered at the paramter vector $\boldsymbol{p}^*$ with covariance matrix C =diag(σ_1,σ_2), i.e.

$$\varphi(y[k];\boldsymbol{p}) = \frac{1}{\sqrt{2\pi\sigma_1\sigma_2}} e^{-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{p}^*)^T C^{-1}(\boldsymbol{p}-\boldsymbol{p}^*)}$$

It holds that

$$\begin{aligned}\ln(L(\boldsymbol{p})) &= \ln\left(\prod_{k=0}^{m} \varphi(y[k];\boldsymbol{p})\right) = \sum_{k=0}^{m} \ln\left(\varphi(y[k];\boldsymbol{p})\right) \\ &= -m\ln\left(\sqrt{2\pi\sigma_1\sigma_2}\right) - \frac{m}{2}\sum_{k=0}^{m} (\boldsymbol{p}-\boldsymbol{p}^*)^T C^{-1}(\boldsymbol{p}-\boldsymbol{p}^*)\end{aligned}$$

This corresponds to the problem of minimizing the mean squarred error w.r.t. $\boldsymbol{p}$ using the weight matrix C^{-1}. To find the maximum derive w.r.t. $\boldsymbol{p}$ to obtain the condition

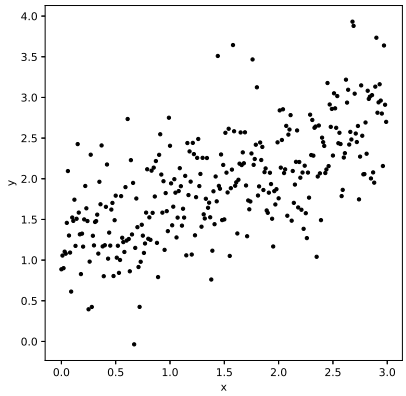
$$2C^{-1}\boldsymbol{p} - 2C^{-1}\boldsymbol{p}^* = 0 \quad \Leftrightarrow \quad \boldsymbol{p} = \boldsymbol{p}^*.$$

This shows the following:

> For independent and identically normal distributed samples the minimum mean squarred error (or minimum sum of squares) identification is equivalent to the maximum likelihood estimation.

2.2.4.3 Back to the straight-line

Lets come back to the initial problem of finding the parameters for a straight line that best fit the given data set



Recall the model $y_i = ax_i + b + w_i$, $i = 1,\dots,m$ and suppose that the noise is a white noise, i.i.d., here normal distributed with stochastically independent sampling distributions. Accordingly, for all $i = 1,\dots,m$ it holds that

$$E\{y_i - ax_i + b\} = E\{w_i\} = 0.$$

Summing up the squarred errors for each $i = 1,\dots,m$ yields (assuming $\sigma_1 = \sigma_2$)

$$E\left\{\sum_{i=0}^{m}(y_i - ax_i + b)^2\right\} = \sigma^2.$$

For a given parameter vector $\boldsymbol{p} = [a', b']^T \neq \boldsymbol{p}^*$ it holds on the other side that

$$E\left\{\sum_{i=0}^{m}(y_i - a'x_i + b')^2\right\} = l(\boldsymbol{p})$$

where $l(\boldsymbol{p})$ is not necessarily zero. Accordingly, the parameter identification problem can be addressed as the problem of minimizing w.r.t. $\boldsymbol{p}$ the sum of squarred errors $l(\boldsymbol{p})$.

Again, the sum of squares can be written in matrix notation

$$l(\boldsymbol{p}) = \left(\boldsymbol{y} - M\boldsymbol{p}\right)^T\left(\boldsymbol{y} - M\boldsymbol{p}\right) \geq 0, \quad \boldsymbol{y} = \begin{bmatrix} y[0] \\ \vdots \\ y[m] \end{bmatrix}, \quad M = \begin{bmatrix} x[0] & 1 \\ \vdots & \vdots \\ x[m] & 1 \end{bmatrix}$$

according to which the minimum squarred error is calculated using the Moore-Penrose pseudo-inverse[2] $M^\dagger = (M^T M)^{-1} M^T$ and given by

$$\boldsymbol{p}^* = M^\dagger \boldsymbol{y} = \begin{bmatrix} \frac{\sum_{i=0}^{m} x_i y_i - (\sum_{i=0}^{m} x_i)(\sum_{i=0}^{m} y_i)}{\sum_{i=0}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2} \\ \frac{(\sum_{i=0}^{m} x_i^2)(\sum_{i=0}^{m} y_i) - (\sum_{i=0}^{m} x_i)(\sum_{i=0}^{m} x_i y_i)}{\sum_{i=0}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2} \end{bmatrix}.$$

This solution corresponds to the best fit in mean considering white noise measurement uncertainty.

For the line, this process is equivalent to the linear regression approach from statistics. For the particular data considered here the following fit is obtained

- Identified model: $y = 0.52065884x + 1.18850078$.

[2] This exists given that the vector of x-data is increasing and the 1-vector is constant.

- Model used for data generation: $y = 0.5x + 1.2$.

Accordingly, a rather parameter good fit is obtained. The resulting straight line is shown in Figure 2.3. Note that an even better approximation would be possible when more data points would be at hand.

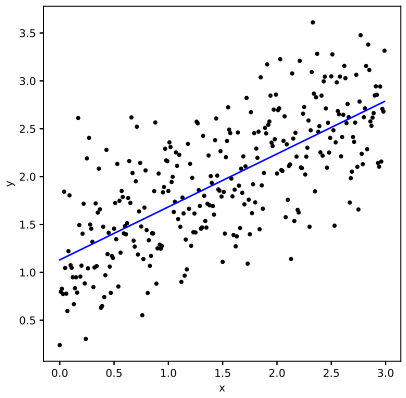


Figure 2.3: Data set with identified straight line following the minimum squared error approach.

2.2.4.4 Standard models for black-box identification

When no structure about the system dynamics is known *a priori* standard models from statistics are considered. Typically these are classified in terms of auto-regression (AR), moving avarege (MA) and the presence of exogenous inputs (X) [Lju87; HD12].

Name	Expression
AR(p)	$y[k+1] = \sum_{i=0}^{p} a_i\, y[k-i] + b + w[k]$
MA(q)	$y[k+1] = \mu + \sum_{l=0}^{q} d_l\, w[k-l]$
ARMA(p,q)	$y[k+1] = \mu + \sum_{i=1}^{p} a_i\, y[k-i] + \sum_{l=0}^{q} d_l\, w[k-l]$
ARMAX(p,q,r)	$y[k+1] = \mu + \sum_{i=1}^{p} a_i\, y[k-i] + \sum_{l=0}^{q} d_l\, w[k-l] + \sum_{n=1}^{r} b_n\, u[k-n]$

Table 2.1: Standard models in black-box identification.

The parameters are normally determined using the associated input-output and disturbance data using a minimum squarred error approach for determining the mean value of the parameters.

As seen above, assuming a normal distribution of the probability over the parameter space this approach is equivalent to the maximum-likelihood estimation.

In case that the noise is not white but colored, form filters can be employed to construct the colored noise from white noise.

Further statistical properties of the mean squarred approach can be shown, but this goes beyond the scope of this introductory notes (see the reference list for further information).

References

[ÅM08] K. J. Åström and R. M. Murray. *Feedback systems : an introduction for scientists and engineers*. Princeton University Press, New Jersey, 2008 (cit. on p. 19).

[Ari94] R. Aris. *Mathematical modeling techniques*. Dover Publications Inc., New York, 1994 (cit. on p. 19).

[Boh06] T. P. Bohlin. *Practical Grey-box Process Identification: Theory and Applications*. Springer Verlag, London, 2006 (cit. on p. 26).

[Boh91] T. P. Bohlin. *Interactive System Identification: Prospects and Pitfalls*. Springer Verlag, Berlin, 1991 (cit. on p. 26).

[CM91] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Dover, 1991 (cit. on p. 29).

[Dur10] R. Durrett. *Probability: theory and examples*. Cambridge University Press, 2010 (cit. on p. 30).

[Dym04] C. L. Dym. *The principles of mathematical modeling*. Elsevier, 2004 (cit. on p. 19).

[Dym07] H. Dym. *Linear algebra in action*. American Mathematical Society, 2007 (cit. on p. 29).

[HD12] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2012 (cit. on p. 34).

[Lju87] L. Ljung. *System identification: Theory for the user*. Prentice Hall, New Jersey, 1987 (cit. on pp. 30, 34).

[Soh98] B. Sohlberg. *Supervision and Control for Industrial Processes. Advances in Industrial Control*. Springer Verlag, London, 1998 (cit. on p. 26).

[Ste84] G. Stephanopoulos. *Chemical Process Control. An Introduction to Theory and practice*. Prentice Hall, New Jersey, 1984 (cit. on p. 26).

[Tul93] H. Tulleken. „Grey-box modelling and identification using physical knowledge and bayesian techniques“. In: *Automatica* 29 (2) (1993), pp. 285–308 (cit. on p. 26).

Observer design for linear systems

3

3.1 Observability and detectability

Consider the system

$$\begin{aligned} \dot{\boldsymbol{x}} &= A\boldsymbol{x} + B\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y} &= C\boldsymbol{x} \end{aligned} \tag{3.1}$$

with $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{m\times n}$.

Recall that the solution of (3.1) is given by

$$\boldsymbol{x}(t) = S(t)\boldsymbol{x}_0 + \int_0^t S(t-\tau)B\boldsymbol{u}(\tau)d\tau = \boldsymbol{\Phi}(t;\boldsymbol{x}_0,\boldsymbol{u}(\cdot)) \tag{3.2}$$

where $\Phi(T;x_0)$ denotes the flow of the system at time T starting at x_0 and being steered by the control input $u(\cdot)$. The fundamental solution $S(t)$ is given by the matrix exponential

$$S(t) = \exp(At).$$

Furthermore, the inverse of $S(t)$ is given by $S^{-1}(t) = S(-t)$.

3.1.1 Observability

We define observability as follows:

Definition 3.1

System (3.1) is called completely observable in time $T > 0$ if any initial condition $\boldsymbol{x}_0 \in \mathbb{R}^n$ can be uniquely determined by the output $\boldsymbol{y}(t)$, $t \in [0, T]$ and the input $\boldsymbol{u}(\cdot) \in L([0, T])$.

In this framework, two trajectories which start from $\boldsymbol{x}_0^1$ and $\boldsymbol{x}_0^2$, respectively, and produce the same output signal $\boldsymbol{y}(t)$ are called *indistinguishable*. Indistinguishability will play a major role when discussing the observability of nonlinear systems, and thus we define it here.

Definition 3.2

The set of states $\boldsymbol{x}_0$ producing the same output signal $\boldsymbol{y}(t)$ for a given input $\boldsymbol{u}(\cdot)$ is called the indistiguishable set

$$\mathscr{I}_u(\boldsymbol{x}_0) := \{\boldsymbol{x} \in \mathbb{R}^n \,|\, C\Phi(t;\boldsymbol{x},\boldsymbol{u}(\cdot)) \equiv C\Phi(t;\boldsymbol{x}_0,\boldsymbol{u}(\cdot))\}.$$

Clearly, the following result holds.

Theorem 3.1

A dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \quad \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ is (uniformly) observable at $\boldsymbol{x}_0$, if for all $\boldsymbol{x}_0 \in \mathbb{R}^n$ (and all $\boldsymbol{u}(\cdot)$) the set

$$\mathscr{I}_u(\boldsymbol{x}_0) = \{\boldsymbol{x}_0\}.$$

Observability can thus be viewed as a kind of invertibility property which can be analyzed on the basis of the analytic solution (3.2), by trying to solve for $\boldsymbol{x}_0$. Actually, in the trivial case that $C = I$ (or if C has $rank(C) = n$) and thus ($\boldsymbol{y}(T) = \boldsymbol{x}(T)$) it would follow that

$$\boldsymbol{x}_0 = S(-T)\left(\boldsymbol{y}(T) - \int_0^T S(T-\tau)B\boldsymbol{u}(\tau)d\tau\right).$$

A similar result is obtained for the case that C is invertible, given that this implies that

$$\boldsymbol{x}_0 = S(-T)\left(C^{-1}\boldsymbol{y}(T) - \int_0^T C^{-1}S(T-\tau)B\boldsymbol{u}(\tau)d\tau\right).$$

3.1.1.1 The Kalmann Observability criterion

Now in the case that rank$(C) = m < n$, we have m equations

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t)$$

so that for determining $\boldsymbol{x}(t)$ (and thus $\boldsymbol{x}_0$) we need $n - m$ additional equations. Knowing the signal $\boldsymbol{y}(t)$ over the time interval $[0, T]$, we know, in principle, its derivatives and can establish the set of equqations

$$\begin{aligned}
&\boldsymbol{y}(t) = C\boldsymbol{x}(t)\\
&\dot{\boldsymbol{y}}(t) = C\dot{\boldsymbol{x}}(t) = CA\boldsymbol{x}(t) + CB\boldsymbol{u}(t)\\
&\ddot{\boldsymbol{y}}(t) = CA^2\boldsymbol{x}(t) + CAB\boldsymbol{u}(t) + CB\dot{\boldsymbol{u}}(t)\\
&\vdots\\
&\boldsymbol{y}^{(n-1)}(t) = CA^{n-1}\boldsymbol{x}(t) + \textstyle\sum_{i=0}^{n-2} CA^iB\boldsymbol{u}^{(n-2-i)}(t)
\end{aligned} \quad . \tag{3.3}$$

The coefficient

$$m_k = CA^kB \tag{3.4}$$

is also known as k-th Markov parameter, and weights the influence of the m-th derivative of u in the $(m+k+1)$-th derivative of the output $\boldsymbol{y}$. Rearranging equation (3.3) so that on the left known terms and on the right the unknown ones appear, one obtains the equation

$$\mathscr{Y}(t) - \mathscr{U}(t) = \mathscr{K}_o\boldsymbol{x}(t) \tag{3.5}$$

with the known vectors

$$\mathscr{Y}(t) = \begin{bmatrix} \boldsymbol{y}(t) \\ \dot{\boldsymbol{y}}(t) \\ \vdots \\ \boldsymbol{y}^{(n-1)}(t) \end{bmatrix}, \quad \mathscr{U}(t) = \begin{bmatrix} 0 \\ m_0\boldsymbol{u}(t) \\ \vdots \\ \sum_{i=0}^{n-2} m_i\boldsymbol{u}^{(n-2-i)}(t) \end{bmatrix} \tag{3.6}$$

and the known matrix

$$\mathcal{K}_o(A,C) = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{mn\times n}. \tag{3.7}$$

This matrix is called **Kalman observability matrix** for Kalman having been the first in writing it down. The system of mn equations (3.5) has a unique solution $\boldsymbol{x}(t)$ (and thus a unique solution for $\boldsymbol{x}_0$) if and only if $\mathcal{K}_o$ has rank n, i.e.

$$\operatorname{rank}(\mathcal{K}_o) = n. \tag{3.8}$$

Given that the matrix $\mathcal{K}_o$ is independent of time, it follows that the observability property for linear time invariant systems of the form (3.1) is independent of the final time T. These fundamental results are summarized in the following theorem.

Theorem 3.2

The system (3.1) is completely observable in time $T > 0$ if and only if the Kalman observability matrix $\mathcal{K}_o$ (3.7) has $\operatorname{rank}(\mathcal{K}_o) = n$. Moreover, in this case it is observable for any $T > 0$.

To be more specific, let the system be completely observable and consider the quadratic matrix $\mathcal{K}_{O,m}$ consisting of (arbitrary) n linearly independent rows of the $nm \times n$ matrix $\mathcal{K}_o$. Then

$$\boldsymbol{x}(t) = \mathcal{K}_{O,m}^{-1}(\mathscr{Y}(t) - \mathscr{U}(t))$$

and thus from (3.2)

$$\boldsymbol{x}_0 = S(-t)\left(\mathcal{K}_{O,m}^{-1}(\mathscr{Y}(t) - \mathscr{U}(t)) - \int_0^t S(t-\tau)Bu(\tau)d\tau\right)$$

Finally, the reader should convince himself that the rank of the Kalman observability matrix is independent of the choice of the basis of the state space, i.e. for any regular transformation T

$$\tilde{A} = TAT^{-1},\ \tilde{C} = CT^{-1} \quad \Rightarrow \quad \operatorname{rank}(\mathcal{K}_o(A,C)) = \operatorname{rank}(\mathcal{K}_O(\tilde{A},\tilde{C})). \tag{3.9}$$

This feature will be important later on.

3.1.1.2 The Observability Gramian

Recall that the output signal can be written in terms of the solution of the differential equation in (3.1). For the case that $\boldsymbol{u} = 0$ this yields

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) = CS(t)\boldsymbol{x}_0.$$

If for any $t \geq 0$ the matrix $CS(t)$ would have full rank it would be invertible and the initial state $\boldsymbol{x}_0$ would be uniquely determined through $\boldsymbol{x}_0 = [CS(t)]^{-1}\boldsymbol{y}(t)$. As this is in general not possible, because C does not have rank n, one can instead try to use a pseudo-inverse-based construction by considering the signal

$$\bar{\boldsymbol{y}}(t) = S^{\mathsf{T}}(t)C^{\mathsf{T}}\boldsymbol{y}(t) = S^{\mathsf{T}}(t)C^{\mathsf{T}}CS(t)\boldsymbol{x}_0.$$

The matrix $S^{\intercal}(t)C^{\intercal}CS(t)$ is quadratic at any time $t \geq 0$. On the other hand, whether this matrix is invertible or not will depend on the particular structure of $S(t)$ and thus on the time t. This motivates to go a step further, given that it is assumed that $\boldsymbol{y}(t)$ is known over a complete interval of time, and consider the weighted integral of the measurement given by

$$\boldsymbol{\psi}(t) = \int_0^t S(\tau)^T C^T \boldsymbol{y}(\tau) d\tau. \tag{3.10}$$

This can be easily computed on the knowledge of $\boldsymbol{y}$ and S. Again, for $\boldsymbol{u} = 0$ this can be written as

$$\boldsymbol{\psi}(t) = \int_0^t S(\tau)^T C^T CS(\tau) d\tau \boldsymbol{x}_0 = \mathscr{G}_O(t)\boldsymbol{x}_0$$

where

$$\mathscr{G}_O(t) = \int_0^t S(\tau)^T C^T CS(\tau) d\tau \tag{3.11}$$

is called the Gramian observability matrix. If $\mathscr{G}_O(t)$ is invertible (i.e. it has rank n), we have

$$\boldsymbol{x}_0 = \mathscr{G}_O(t)^{-1}\boldsymbol{\psi}(t) \tag{3.12}$$

implying that the system is completely observable in time t. In the general case, with $\boldsymbol{u} \neq 0$ this becomes

$$\boldsymbol{\psi}_u(t) = \int_0^t S(\tau)^T C^T \boldsymbol{y}(\tau) d\tau = \int_0^t S(\tau)^T C^T C \left(S(\tau)\boldsymbol{x}_0 + \int_0^\tau S(\tau - s) B\boldsymbol{u}(s) ds d\tau \right)$$

which can be computed on the basis of $\boldsymbol{y}$ and $\boldsymbol{u}$ over the time interval $[0, t]$, and if $\text{rank}(\mathscr{G}_O(t)) = n$ we have

$$\boldsymbol{x}_0 = \mathscr{G}_O(t)^{-1} \left[\boldsymbol{\psi}_u(t) - \int_0^t S(\tau)^T C^T CS(\tau) \left(\int_0^\tau S(-s) B\boldsymbol{u}(s) ds \right) d\tau \right],$$

implying the complete observability for $t > 0$.

As shown above, observability of a linear system is independent of the particular time $t > 0$, so it results that if the observability Gramian has full rank for a particular $t > 0$, it has full rank for any $t > 0$. By taking into account the Theorem of Cayleyigh-Hamilton $\mathscr{G}_O$ can further be analytically related to the Kalman observability matrix $\mathscr{K}_o$. The preceding results are summarized in the following theorem.

Theorem 3.3

For the linear time-invariant system $\Sigma(A, B, C)$ (3.1) the following properties are equivalent:

1. The system is completely observable in time $T > 0$.
2. The system is completely observable for any time $T > 0$.
3. The Kalman observability matrix $\mathscr{K}_o(A, C)$ (3.7) has rank n.
4. The Gramian observability matrix $\mathscr{G}_O(t)$ (3.11) has rank $n \; \forall \; t > 0$.

3.1.2 Detectability

Clearly, observability is a property which depends on the inherent interaction mechanisms between the system states x_i and the measurement (or output) $\boldsymbol{y}$, and thus depends as well as on the matrix C as on the structure of A. For this reason, observability (as well as controllability) are sometimes called *structural properties*. Given that observability represents an important issue in system monitoring and feedback control problems, the choice of an adequate sensor for a given system has to be carefully analyzed.

In many cases the system is not observable, and there do not exist sensors which whould enhance the observability properties of the system. For these cases, the associated weaker concept of **detectability** is important.

Definition 3.3

The pair (A, C) (3.1) is called *detectable* if (with $\boldsymbol{u} = 0$)

$$\boldsymbol{y} \equiv \boldsymbol{0} \quad \Rightarrow \quad \lim_{t\to\infty} ||\boldsymbol{x}(t)|| = 0. \tag{3.13}$$

Note that the first condition ($\boldsymbol{y} \equiv 0$) implies that $\boldsymbol{y}$ and its derivatives are zero, i.e.

$$\mathscr{Y}(t) = \boldsymbol{0}$$

with $\mathscr{Y}$ defined in (3.6). In the case that the system is completely observable this implies that $\boldsymbol{x}(t) \equiv \boldsymbol{0}$, or equivalently $\boldsymbol{x}_0 = \boldsymbol{0}$. Thus, we have the following result.

Theorem 3.4

If the system $\Sigma(A, B, C)$ is completely observable, it is detectable.

This relation clearly shows that observability is a stronger property than detectability. The reader should convince himselve that the inverse implication does not hold.

In the case that the system is not completely observable the observability matrix $rank(\mathscr{K}_o) = r < n$ and it follows from (3.5) (with $\boldsymbol{u} \equiv 0$) that $\boldsymbol{y} \equiv \boldsymbol{0}$ implies that $\boldsymbol{x}(t)$ lies within the nullspace $\mathscr{N}_O$ of the observability matrix $\mathscr{K}_o$, or equivalently of $\mathscr{K}_{o,m}$. The complement of $\mathscr{N}_O$ in $\mathbb{R}^n$ is called the observable subspace, and we will denote it by $\mathscr{O} \subseteq \mathbb{R}^n$. Clearly, if the system is completely observable, the nullspace $\mathscr{N}_O = \{\boldsymbol{0}\}$ and $\mathscr{O} = \mathbb{R}^n$.

Now, consider the map defined by the transposed of the observability matrix $\mathscr{K}_o$, i.e.

$$\mathscr{K}_o^T \in \mathbb{R}^{n\times mn} : \mathfrak{Y} \to \mathbb{R}^n$$

where $\mathfrak{Y}$ denotes a generalized space of output functions $y : \mathbb{R} \to \mathbb{R}^m$ with their first $n-1$ successive time derivatives, in the sense of the vector $\mathscr{Y}$ in (3.5). Clearly, $\mathscr{Y} \in \mathfrak{Y}$. It follows that $\mathscr{O}$ is the image of the map $\mathscr{K}_o^T$, i.e.

$$\mathscr{O} = \mathscr{R}(\mathscr{K}_o^T) \tag{3.14}$$

with dimension equal to the number of independent rows of $\mathscr{K}_o$. By the same reasoning it follows that $\mathscr{N}_o = \ker(\mathscr{K}_o)$. Note that

$$A^T\mathscr{K}_o^T = \begin{bmatrix} A^TC^T & A^T(CA)^T & \cdots & A^T(CA^{n-1})^T \end{bmatrix}$$

$$= \begin{bmatrix} (CA)^T & (CA^2)^T & \cdots & (CA^n)^T \end{bmatrix}$$

and, in virtue of the Theorem of Cayleigh-Hamilton, it follows that

$$\mathscr{R}(A^T \mathscr{K}_o) \subseteq \mathscr{R}(\mathscr{K}_o) = \mathscr{O}, \tag{3.15}$$

meaning that the observable subspace $\mathscr{O}$ is A^T-invariant, i.e. for all $\boldsymbol{x}^* \in \mathscr{O}$ it follows that $A^T \boldsymbol{x}^* \in \mathscr{O}$.

The A^T-invariance of $\mathscr{O}$ is important in the following reasoning. Given that $\dim(\mathscr{O}) = \text{rank}(\mathscr{K}_o^T) = r$, there are r linearly independent column vectors $\boldsymbol{\rho}_i$, $i = 1, \ldots, r$ of the matrix $\mathscr{K}_o^T$ that form a basis of the r-dimensional subspace $\mathscr{O} \subset \mathbb{R}^n$, with the orthogonal complement $\mathscr{N}_O = \mathbb{R}^n \setminus \mathscr{O}$. Now, let $\boldsymbol{\kappa}_i$, $i = 1, \ldots, n-r$ be a basis of this complement $\mathscr{N}_O$. Define the transformation

$$T = \begin{bmatrix} \boldsymbol{\rho}_1^T & \cdots & \boldsymbol{\rho}_r^T & \boldsymbol{\kappa}_1 & \cdots & \boldsymbol{\kappa}_{n-r} \end{bmatrix}, \quad \boldsymbol{\xi} = T^{-1} \boldsymbol{x} \tag{3.16}$$

and the associated matrices

$$\tilde{A} = T^{-1} A T, \quad \tilde{C} = C T.$$

Denote by $\tilde{\mathscr{O}} = T^{-1}\mathscr{O}$, $\tilde{\mathscr{N}_O} = T^{-1}\mathscr{N}_O$, and note that this transformation implies that the first r components of the vector $\boldsymbol{\xi} = T^{-1}\boldsymbol{x}$ correspond to components in the r-dimensional observable subspace, while the remaining $n-r$ components to the non-observable subspace, i.e.

$$\boldsymbol{\xi}_o = \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{0} \end{bmatrix} \in \tilde{\mathscr{O}}, \quad \boldsymbol{\xi}_{no} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\xi}_2 \end{bmatrix} \in \tilde{\mathscr{N}_O} \tag{3.17}$$

with $\boldsymbol{\xi}_1 \in \mathbb{R}^r, \boldsymbol{\xi}_2 \in \mathbb{R}^{n-r}$, or equivalently $\boldsymbol{x}_i = T^{-1}\boldsymbol{\xi}_i \in \mathscr{O}$, $i = 1,2$. Thus, for any $\boldsymbol{\xi} \in T^{-1}\mathscr{O}$ it holds that

$$\tilde{A}^T \boldsymbol{\xi} \in \tilde{\mathscr{O}} \quad \Leftrightarrow \quad \tilde{A}^T \boldsymbol{\xi}_o = \begin{bmatrix} \tilde{A}_{1,1}^T & \tilde{A}_{2,1}^T \\ \tilde{A}_{1,2}^T & \tilde{A}_{2,2}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} A_{1,1}^T \boldsymbol{\xi}_1 \\ A_{1,2}^T \boldsymbol{\xi}_1 \end{bmatrix}$$

but $\tilde{A}^T \boldsymbol{\xi}_o \in \tilde{\mathscr{O}}$ and thus $A_{1,2}^T = 0$. This implies that the matrix $\tilde{A}$ is of the form

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{1,1} & 0 \\ \tilde{A}_{2,1} & \tilde{A}_{2,2} \end{bmatrix}, \quad \tilde{A}^T = \begin{bmatrix} \tilde{A}_{1,1}^T & \tilde{A}_{2,1}^T \\ 0 & \tilde{A}_{2,2}^T \end{bmatrix} \tag{3.18}$$

Further, as the image of C^T (or $\tilde{C}^T$) lies in $\mathscr{O}$ (or $\tilde{\mathscr{O}}$), it follows that

$$\tilde{C} = \begin{bmatrix} \tilde{c}_1 & \cdots & \tilde{c}_r & 0 & \cdots & 0 \end{bmatrix}. \tag{3.19}$$

The matrix $\tilde{B}$ does not have a particular structure in the general case.

This shows, that any linear time-invariant system of the form (3.1) can be brought by a regular state transformation into a form in which the observable part is decoupled from the unobservable one. In terms of this decomposition, the detectability property can now be interpreted easily.

Consider $\boldsymbol{y} = C\boldsymbol{x} \equiv \boldsymbol{0}$, and thus $\tilde{\boldsymbol{y}} = \tilde{C}\boldsymbol{\xi} \equiv \boldsymbol{0}$. Having in mind the form of the matrices $\tilde{A}$ and $\tilde{C}$, it follows that $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r \equiv 0$. Thus, the remaining (possibly non-zero) state is of the form $\boldsymbol{\xi}_{no}$ in (3.17), and can be written as

$$\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\xi}_2 \end{bmatrix}$$

with $\boldsymbol{\xi}_2$ having the dynamics

$$\dot{\boldsymbol{\xi}}_2 = \tilde{A}_{2,2}\boldsymbol{\xi}_2.$$

Accordingly, the system is detectable in the sense of Definition 3.3 if and only if the dynamics of the unobservable part are asymptotically stable, i.e. $\lim_{t\to\infty} ||\boldsymbol{\xi}_2(t)|| = 0$. These results are summarized in the following theorem.

Theorem 3.5

Any system of the form (3.1) with $\operatorname{rank}(\mathcal{K}_o) = r \leq n$ can be decomposed by a regular transformation into the form

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\xi} &= \begin{bmatrix} \tilde{A}_{1,1} & 0 \\ \tilde{A}_{2,1} & \tilde{A}_{2,2} \end{bmatrix} \boldsymbol{\xi} + \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix} \boldsymbol{u}, \quad \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0 \\ \tilde{\boldsymbol{y}} &= \begin{bmatrix} \tilde{C}_1 & 0 \end{bmatrix} \boldsymbol{\xi} \end{aligned} \tag{3.20}$$

with $\tilde{A}_{1,1} \in \mathbb{R}^{r\times r}$, $\tilde{A}_{2,1} \in \mathbb{R}^{(n-r)\times r}$, $\tilde{A}_{2,2} \in \mathbb{R}^{(n-r)\times(n-r)}$, $\tilde{B}_1 \in \mathbb{R}^{r\times p}$, $\tilde{B}_2 \in \mathbb{R}^{(n-r)\times p}$, $\tilde{C}_1 \in \mathbb{R}^{r\times m}$ and with the pair $(\tilde{A}_{1,1}, \tilde{C}_1)$ being completely observable. The system is detectable if and only if $\tilde{A}_{2,2}$ is Hurwitz.

In virtue of Definition the indistinguishable subset of $\mathbb{R}^n$ associated to the representation (3.20) is independent of the input and given by

$$\mathcal{I}(\boldsymbol{x}_0) = \{\boldsymbol{x} \in \mathbb{R}^n \,|\, \boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{v}, \quad \boldsymbol{v} \in \operatorname{span}\{\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_{n-r}\}\} \tag{3.21}$$

and defines a separation of the state space into parallel slides passing through the initial state $\boldsymbol{x}_0$.

3.1.3 The (single-output) observability normal form

Besides the descomposition normal form (3.20) which exists for any system, there is an important particular intrinsic structure of any completely observable system that will be analyzed next. Suppose that we want to transform the system into such coordinates that the measured output corresponds to the last state $\boldsymbol{z}_n$, i.e.

$$\boldsymbol{y} = \tilde{\boldsymbol{c}}^T \boldsymbol{z} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \boldsymbol{z} \tag{3.22}$$

If the system is completely observable, then $\operatorname{rang}(\mathscr{K}_o) = n$, and there exists a unique solution to the equation

$$\mathscr{K}_O \hat{\boldsymbol{w}} = \tilde{\boldsymbol{c}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{3.23}$$

given by

$$\hat{\boldsymbol{w}} = \mathscr{K}_o^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

On the other hand, equation (3.23) can be written as

$$\begin{bmatrix} \boldsymbol{c}^T \hat{\boldsymbol{w}} \\ \vdots \\ \boldsymbol{c}^T A^{n-1} \hat{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{w}}^T \boldsymbol{c} \\ \vdots \\ \hat{\boldsymbol{w}}^T (A^{n-1})^T \boldsymbol{c} \end{bmatrix} = T^T \boldsymbol{c} = \tilde{\boldsymbol{c}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

with the matrix

$$T = \begin{bmatrix} \hat{\boldsymbol{w}} & A\hat{\boldsymbol{w}} & \cdots & A^{n-1}\hat{\boldsymbol{w}} \end{bmatrix}, \quad \hat{\boldsymbol{w}} = \mathscr{K}_o^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{3.24}$$

Denote by $\boldsymbol{s}_i^T$, $i = 0, \ldots, n-1$ the rows of the inverse matrix T^{-1}. The following two identities hold:

$$T^{-1}AT = \begin{bmatrix} \boldsymbol{s}_0^T A\hat{\boldsymbol{w}} & \cdots & \boldsymbol{s}_0^T A^n \hat{\boldsymbol{w}} \\ \vdots & & \vdots \\ \boldsymbol{s}_{n-1}^T A\hat{\boldsymbol{w}} & \cdots & \boldsymbol{s}_{n-1}^T A^n \hat{\boldsymbol{w}} \end{bmatrix}, \quad T^{-1}T = \begin{bmatrix} \boldsymbol{s}_0^T \hat{\boldsymbol{w}} & \cdots & \boldsymbol{s}_0^T A^{n-1} \hat{\boldsymbol{w}} \\ \vdots & & \vdots \\ \boldsymbol{s}_{n-1}^T \hat{\boldsymbol{w}} & \cdots & \boldsymbol{s}_{n-1}^T A^{n-1} \hat{\boldsymbol{w}} \end{bmatrix} = I.$$

Comparing both products it results that for $i = 1, \ldots, n-1$ the i-th column of $T^{-1}AT$ is identical to the $(i+1)$-th column of the identity matrix. This in turn implies that the matrix $T^{-1}AT$ has the a sub-diagonal with unit entries so that the elements of the last column are exactly the negative coefficients of the characteristic polynomial, i.e. for the characteristic polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_o$ we have $(T^{-1}AT)_{in} = -a_{i-1}$. Given that the characteristic polynomial is invariant with respect to regular

transformations, these coefficients are exactly the ones of the characteristic polynomial associated to the original matrix A. Thus the transformed dynamics are given by

$$\begin{aligned} \dot{\boldsymbol{z}} &= A_O \boldsymbol{z} + B_O \boldsymbol{u}, \quad \boldsymbol{z}(0) = \boldsymbol{z}_0 \\ y &= \boldsymbol{c}_O^T \boldsymbol{z} \end{aligned} \tag{3.25}$$

with

$$A_O = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 & -a_0 \\ 1 & 0 & & & 0 & -a_1 \\ 0 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & 1 & 0 & -a_{n-2} \\ 0 & \cdots & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix}, \quad B_O = T^{-1}B, \quad \boldsymbol{c}_O^T = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$$

This particular structure is called ***observability normal form***. It can be easily verified that

$$y = \boldsymbol{c}_O^T \boldsymbol{z} = \boldsymbol{c}^T T \boldsymbol{z} = \boldsymbol{c}^T T T^{-1} \boldsymbol{x} = \boldsymbol{c}^T \boldsymbol{x}.$$

Given that the only condition for the transformability is that the Kalman observability matrix $\mathcal{K}_o$ has rank n, the following result is obtained.

Theorem 3.6

The LTI system (3.1) can be transformed into observability normal form (3.25) if and only if it is completely observable.

Proof The sufficiency has allready been shown above. For the necessity, note that if the system is not completely observable, then the transformation matrix T (3.24) does not exist, because $\mathcal{K}_o$ is not invertible. Q.E.D.

The observability normal form is particularly useful in the design of a state observer, as will be analyzed in the next section.

3.2 Observer Design for LTI Systems

An observer is a system for the reconstruction of the system state based on the knowledge of the measurement and the input over a certain time-interval. The notion of observability underlies this enterprise, and the relations will be studied in this section with detail. First of all, let us introduce formally the concept of an observer.

Definition 3.4

A dynamical system $\hat{\Sigma}$ with state $\hat{\boldsymbol{x}}$ is called an observer for the system Σ with state $\boldsymbol{x}$, if it holds that

$$\lim_{t \to \infty} ||\hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)|| = 0. \tag{3.26}$$

David Luenberger [Lue71] proposed the following simple observer scheme (called the *Luenberger observer*) for the system (3.1)

$$\dot{\hat{\boldsymbol{x}}} = A\hat{\boldsymbol{x}} + B\boldsymbol{u} - L\big(C\hat{\boldsymbol{x}} - \boldsymbol{y}\big), \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0. \tag{3.27}$$

The observer can be viewed as consisting of two parts:

- A copy of the system model (3.1) $\Sigma(A,B,C)$ for the state prediction on the basis of the considered initial value $\hat{\boldsymbol{x}}_0$.
- An innovation scheme for model adaptation in dependence of the measured output $\boldsymbol{y}(t)$.

The associated block diagram is presented in Figure 3.1. Clearly, if the initial value $\hat{\boldsymbol{x}}_0$ and the

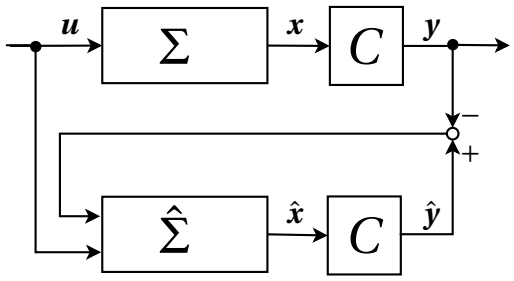


Figure 3.1: Structure of the Luenberger observer.

simulation model $\hat{\Sigma}$ coincide with the actual ones $\boldsymbol{x}_0, \Sigma$, the simulation part of the observer would predict the correct state $\hat{\boldsymbol{x}}$ at any time. In most cases, nevertheless, the initial value is not exactly known, so that the simulation part will not be able to correctly predict the state evolution. In this case, the adaptation mechanism tries to steer the predicted state towards the actual one, through the information contained in the measurement. Thus, it becomes clear that the underlying notion of observability and detectability are crucial for the performance of the observer. To analyze how observability properties determine the observer functioning consider the observation error

$$\boldsymbol{e} = \hat{\boldsymbol{x}} - \boldsymbol{x}, \tag{3.28}$$

with dynamics

$$\dot{\boldsymbol{e}} = \dot{\hat{\boldsymbol{x}}} - \dot{\boldsymbol{x}} = A\hat{\boldsymbol{x}} + B\boldsymbol{u} - LC(\hat{\boldsymbol{x}} - \boldsymbol{x}) - A\boldsymbol{x} - B\boldsymbol{u} = (A - LC)\boldsymbol{e}. \tag{3.29}$$

Thus, to ensure (3.26), the spectrum of the matrix $A - LC$ has to be contained in the open left-half plane of the complex numbers, i.e.

$$\sigma(A - LC) \in \mathbb{C}_-$$

Let us analyze this matrix with more attention, and first consider the SISO case with a completely observable system. In this case the dynamics can be brought into *observability normal form* (3.25) by

the regular transformation T (3.24). When applying the Luenberger observer structure (3.27) in these coordinates, the following particular form is obtained

$$\tilde{A}-L\tilde{C}=\begin{bmatrix}0 & \cdots & \cdots & 0 & -a_0\\ 1 & \ddots & & \vdots & \vdots\\ 0 & \ddots & \ddots & \vdots & \vdots\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & 1 & -a_{n-1}\end{bmatrix}-\begin{bmatrix}0 & \cdots & l_1\\ & \vdots & \\ & \vdots & \\ & \vdots & \\ 0 & \cdots & l_n\end{bmatrix}=\begin{bmatrix}0 & \cdots & \cdots & 0 & -a_0-l_1\\ 1 & \ddots & & \vdots & \vdots\\ 0 & \ddots & \ddots & \vdots & \vdots\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & 1 & -a_{n-1}-l_n\end{bmatrix} \tag{3.30}$$

and the characteristic polynomial is given by

$$\mathscr{P}[\tilde{A}-L\tilde{C}](\lambda)=\lambda^n+(a_{n-1}+l_n)\lambda^{n-1}+\ldots+(a_1+l_2)\lambda+(a_0+l_1). \tag{3.31}$$

This shows that an arbitrary characteristic polynomial can be assigned to the observer dynamics by choosing the coefficients l_i. In particular, this can be done using the Ackerman formula. In MATLAB or OCTAVE this can be implemented using the `place` and `acker` commands for the dual system (**??**). This subtle fact implies that in the case of a completely observable system the observer convergence can be ensured by adequately choosing the observer gains l_i. Moreover, this does not rely on the system stability, i.e. *an exponentially convergent observer can be designed for unstable and stable system in the same way* (clearly, for unstable ones the observer gains will need to be larger). Finally, it should be noted that the fact that an arbitrary characteristic polynomial can be imposed implies that the observer can be made, in principle[1], *arbitrarily fast.*

Now, consider the case that the system is not completely observable. Then, according to Theorem 3.5, by a regular transformation the system can be brought into the form (3.20) with the pair $(\tilde{A}_{1,1},\tilde{C}_1)$ being completely observable. The application of the observer structure (3.27) in this case yields

$$\tilde{A}-\tilde{L}\tilde{C}=\begin{bmatrix}\tilde{A}_{1,1} & 0\\ \tilde{A}_{2,1} & \tilde{A}_{2,2}\end{bmatrix}-\begin{bmatrix}\tilde{L}_1\tilde{C}_1 & 0\\ \tilde{L}_2\tilde{C}_2 & 0\end{bmatrix}=\begin{bmatrix}\tilde{A}_{1,1}-\tilde{L}_1\tilde{C}_1 & 0\\ \tilde{A}_{2,1}-\tilde{L}_2\tilde{C}_1 & \tilde{A}_{2,2}\end{bmatrix} \tag{3.32}$$

showing that the spectrum of the matrix $\tilde{A}-\tilde{L}\tilde{C}$ is given by

$$\sigma(\tilde{A}-\tilde{L}\tilde{C})=\sigma(\tilde{A}_{1,1}-\tilde{L}_1\tilde{C}_1)\cup\sigma(\tilde{A}_{2,2}). \tag{3.33}$$

This means that, besides the fact that the completely observable part $(\tilde{A}_{1,1},\tilde{C}_1)$ can, in principle, be made to converge arbitrarily fast as shown before, the observer convergence completely relies on the spectrum of the unobservable part associated to $\tilde{A}_{2,2}$. Thus, it is necessary for the existence of an asymptotically convergent observer that $\sigma(\tilde{A}_{2,2})\in\mathbb{C}_-$, i.e. that the system is detectable. These results are summarized in the following theorem.

Theorem 3.7

There exists a matrix $L\in\mathbb{R}^{n\times m}$ such that the dynamics (3.27) is an observer for (3.1) if and only if the system (3.1) is detectable. If (3.1) is completely observable the convergence rate can be assigned arbitrarily.

[1] The restriction comes from the fact that in presence of measurement noise this will be amplified proportional to the observer gains.

3.3 Reduced order and unknown-input observers

In this section we want to address two scenarios: (i) the system dynamics are very large with several measurements given, and (ii) the system dynamics are influenced by unknown exogeneous inputs or perturbations. In the first case the question arises on how to construct an observer without the necessity of solving an unnecessary large set of differential equations, having several state measurements at hand. In the second case we are interested in reconstructing exactly the state vector in spite of the unknown perturbations.

3.3.1 The reduced order observer

For a full rank matrix C, consider the state transformation

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_m \\ \boldsymbol{z}_o \end{bmatrix} = \underbrace{\begin{bmatrix} C \\ T_o \end{bmatrix}}_{=:V} \boldsymbol{x} \tag{3.34}$$

where T_o is such that V is regular. The dynamics in the new coordinates reads

$$\begin{aligned} \dot{\boldsymbol{z}} &= \tilde{A}\boldsymbol{z} + \tilde{B}\boldsymbol{u} \\ \boldsymbol{y} &= \boldsymbol{z}_m \end{aligned}$$

where

$$\tilde{A} = VAV^{-1} = \begin{bmatrix} \tilde{A}_m & \tilde{A}_{mo} \\ \tilde{A}_{om} & \tilde{A}_o \end{bmatrix}, \quad \tilde{B} = VB = \begin{bmatrix} \tilde{B}_m \\ \tilde{B}_o \end{bmatrix}. \tag{3.35}$$

Given that the first component vector $\boldsymbol{z}_m = \boldsymbol{y}$ is completely measured, it is unnecessary to include it in the observer dynamics and an observer of reduced $(n-m)$-dimensional state vector can be designed. For this purpose consider for the moment that $\dot{y}$ would be measurable, and define the ficticious measurement

$$\boldsymbol{\psi} = \dot{\boldsymbol{z}}_m - A_m\boldsymbol{z}_m - B_m\boldsymbol{u} = A_{mo}\boldsymbol{z}_o$$

and construct the following Luenberger observer

$$\begin{aligned} \dot{\hat{\boldsymbol{z}}}_o &= A_o\hat{\boldsymbol{z}}_o + A_{om}\boldsymbol{z}_m + B_o\boldsymbol{u} - L(A_{mo}\hat{\boldsymbol{z}}_o - \boldsymbol{\psi}) \\ &= A_o\hat{\boldsymbol{z}}_o + A_{om}\boldsymbol{z}_m + B_o\boldsymbol{u} - L(A_{mo}\hat{\boldsymbol{z}}_o - \dot{\boldsymbol{z}}_m + A_m\boldsymbol{z}_m + B_m\boldsymbol{u}) \end{aligned}$$

Now, introduce the new state vector

$$\boldsymbol{\chi}_o = \hat{\boldsymbol{z}}_o - L\boldsymbol{y} = \hat{\boldsymbol{z}}_o - L\boldsymbol{z}_m. \tag{3.36}$$

The dynamics of $\boldsymbol{\chi}_o$ are given by

$$\begin{aligned} \dot{\boldsymbol{\chi}}_o &= A_o\hat{\boldsymbol{z}}_o + A_{om}\boldsymbol{z}_m + B_o\boldsymbol{u} - L(A_{mo}\hat{\boldsymbol{z}}_o - \dot{\boldsymbol{z}}_m + A_m\boldsymbol{z}_m + B_m\boldsymbol{u}) - L\dot{\boldsymbol{y}} \\ &= A_o(\boldsymbol{\chi}_o + L\boldsymbol{y}) + A_{om}\boldsymbol{y} + B_o\boldsymbol{u} - L(A_{mo}(\boldsymbol{\chi}_o + L\boldsymbol{y}) + A_m\boldsymbol{y} + B_m\boldsymbol{u}) \end{aligned}$$

and do no more depend on the knowledge of $\dot{\boldsymbol{z}}_m = \dot{\boldsymbol{y}}$. Rearranging the terms in the preceeding equation, the dynamics of the ***reduced order observer*** are obtained:

$$\dot{\boldsymbol{\chi}}_o = (A_o - LA_{mo})\boldsymbol{\chi}_o + (B_o - LB_m)\boldsymbol{u} + (A_oL + A_{om} - L(A_{mo}L + A_m))\boldsymbol{y}, \quad \boldsymbol{\chi}(0) = \boldsymbol{\chi}_0 \tag{3.37}$$

In the same way as in the existence result presented in Theorem 3.7, the following theorem is obtained.

Theorem 3.8

A reduced order observer (3.37) exists for the system (3.1) if and only if there exists a matrix T_o such that the pair $(\tilde{A}_o, \tilde{A}_{mo})$ (3.35) is detectable. If this pair is furthermore completely observable, then the convergence speed can be assigned arbitrarily.

Proof: Consider the observation error

$$\boldsymbol{\epsilon}_o := \hat{\boldsymbol{z}}_o - \boldsymbol{z}_o = \boldsymbol{\chi}_o + L\boldsymbol{y} - \boldsymbol{z}_o.$$

Expanding terms in the associated observation error dynamics yields

$$\begin{aligned}\dot{\boldsymbol{\epsilon}}_o =& (A_o - LA_{mo})(\boldsymbol{z}_o + \boldsymbol{\epsilon}_o - L\boldsymbol{y}) + (B_o - LB_m)\boldsymbol{u} + (A_oL + A_{om} - L(A_{mo}L + A_m))\boldsymbol{y} \\ &+ L(A_m\boldsymbol{y} + A_{mo}\boldsymbol{z}_o + B_m\boldsymbol{u}) - (A_{om}\boldsymbol{y} + A_o\boldsymbol{z}_o + B_o\boldsymbol{u}) \\ =& (A_o - LA_{mo})\boldsymbol{\epsilon}_o.\end{aligned}$$

This shows that $\lim_{t\to\infty} ||\boldsymbol{\epsilon}_o|| = 0$ if and only if $\sigma(A_o - LA_{mo}) \subset \mathbb{C}_-$. According to the above considerations on the observability and the complete order Luenberger observer this can be achieved if and only if the pair (A_o, A_{mo}) is detectable. If it is completely observable, then all eigenvalues of $(A_o - LA_{mo})$ can be assigned arbitrarily. □

3.3.2 Unknown-input observers

The reduced order observer provides an interesting means to reconstruct the unmeasured state without necessity of estimating the measured one. This can be particularly usefull if the measured state dynamics are not completely known, e.g. due to some perturbation or exogeneous input acting on the system. To analyze this situation further, consider the following dynamics

$$\begin{aligned}\dot{\boldsymbol{x}} &= A\boldsymbol{x} + B\boldsymbol{u} + E\boldsymbol{w}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y} &= C\boldsymbol{x}.\end{aligned} \tag{3.38}$$

Given that the presence of the unknown input $\boldsymbol{w}(\cdot) \in \mathbb{R}^q$ implies an additional degree of freedom, the properties of observability and detectability are usually called ***strong observability*** and ***strong detectability***, but are defined in the same way as above. It also holds that strong observability implies strong detectability. In the unknown input case an additional concept appears to be necessary:

Definition 3.5

The system (3.38) is called ***strong*** **detectable*** if $\boldsymbol{y} \to 0$ implies $\boldsymbol{x} \to 0$.

To see the difference between strong and strong* detectability, consider the system [Hau83]

$$\ddot{y} + \dot{y} + y = w \tag{3.39}$$

with state $\boldsymbol{x} = \begin{bmatrix} y & \dot{y} \end{bmatrix}^T$. Clearly, for $y \equiv 0$ it is necessary that $\dot{y} = 0$, $\ddot{y} = 0$ and $w = 0$, so $\boldsymbol{x} = 0$, implying the strong observability of the system. Nevertheless for $y = t^{-1}\sin(t^2)$ it follows that $y \to 0$ but $\dot{y} = -t^{-2}\sin(t^2) + 2\sin(t^2) \to 2\sin(t^2) \neq 0$. Thus with w obtained from substituting $y = t^{-1}\sin(t^2)$ into

(3.39) it follows that $y \to 0$ but $\boldsymbol{x} \not\to \boldsymbol{0}$. This illustrates the difference between both concepts, and shows that, actually, strong* detectability is a stronger property than strong observability.

From the structure of the general Luenberger observer (3.27) the strong* detectability can also be motivated in terms of the observation error as follows: when the correction term yields output convergence, i.e. $C\boldsymbol{e} \to 0$, the observability or detectability property ensures that $\boldsymbol{e} \to 0$. If this property would not be given, it would be impossible to construct a convergent Luenberger observer.

Hautus [Hau83] provided a characterization of the conditions for strong detectability, strong observability and strong* detectability. To state these results, consider first the Laplace transform of (3.38) with $u = 0$ in the following form

$$(sI - A)\hat{\boldsymbol{X}}(s) - E\hat{\boldsymbol{W}}(s) = \boldsymbol{x}_0, \quad C\hat{\boldsymbol{X}}(s) = \hat{\boldsymbol{Y}}(s)$$

or equivalently

$$\begin{bmatrix} (sI - A) & -E \\ C & 0 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{X}}(s) \\ \hat{\boldsymbol{W}}(s) \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_0 \\ \hat{\boldsymbol{Y}}(s) \end{bmatrix}$$

Note that the complex numbers s for which the preceeding matrix presents a loss of rank are exactly the invariant zeros of the system [MK76]. In terms of these zeros following theorem can be stated.

Theorem 3.9

The system (3.38) is stongly detectable if and only if all its invariant zeros s satisfy $\Re(s) < 0$, i.e.

$$\operatorname{rank} \begin{bmatrix} (sI - A) & -E \\ C & 0 \end{bmatrix} < n + \operatorname{rank}(E), \quad \Rightarrow \quad \Re(s) < 0. \tag{3.40}$$

It is strongly observable if it has no invariant zeros, and strongly* detectable if it is strongly detectable and in addition

$$\operatorname{rank}(CE) = \operatorname{rank}(E). \tag{3.41}$$

The rank condition (3.41) implies that the unkown input enters the measured channels only, i.e. has relative (vector) degree 1.

The basic idea in constructing an unknown-input observer consists in determining a state transformation

$$\boldsymbol{z} = T\boldsymbol{x} = \begin{bmatrix} SC \\ M \end{bmatrix} \boldsymbol{x} \tag{3.42}$$

with

$$ME = 0, \quad S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \quad S_1 CE = I_q, \quad S_2 CE = 0_{(m-q)\times q}.$$

Note that the matrix S can be determined if and only if the rank condition (3.41) is satisfied. The new state $\boldsymbol{z}$ is thus given by

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix}, \quad \boldsymbol{z}_1 = S_1 C\boldsymbol{x}, \quad \boldsymbol{z}_2 = \begin{bmatrix} S_2 C \\ M \end{bmatrix} \boldsymbol{x},$$

and its dynamics by

$$\begin{aligned}\dot{\boldsymbol{z}}_1 &= \tilde{A}_{11}\boldsymbol{z}_1 + \tilde{A}_{12}\boldsymbol{z}_2 + B_1\boldsymbol{u} + \boldsymbol{w}\\ \dot{\boldsymbol{z}}_2 &= \tilde{A}_{21}\boldsymbol{z}_1 + \tilde{A}_{22}\boldsymbol{z}_2 + B_2\boldsymbol{u}\\ \tilde{y} &= \begin{bmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{y}_2\end{bmatrix} = \tilde{C}\boldsymbol{z}\end{aligned}$$

with

$$\begin{aligned}\tilde{A} &= TAT^{-1} = \begin{bmatrix}\tilde{A}_{11} & \tilde{A}_{12}\\ \tilde{A}_{21} & \tilde{A}_{22}\end{bmatrix}, \quad \tilde{B} = TB = \begin{bmatrix}\tilde{B}_1\\ \tilde{B}_2\end{bmatrix},\\ \tilde{C} &= SCT^{-1} = \begin{bmatrix}\tilde{C}_1 & 0_{q\times(n-q)}\\ 0_{(m-q)\times q} & \tilde{C}_2\end{bmatrix} = \begin{bmatrix}I_q & 0_{q\times(m-q)} & 0_{q\times(n-m)}\\ 0_{(m-q)\times q} & I_{m-q} & 0_{(m-q)\times(n-m)}\end{bmatrix}\end{aligned}$$

An obvious choice for a reduced order observer which does not depend on the unknown input is given by

$$\begin{aligned}\dot{\hat{\boldsymbol{z}}}_2 &= \tilde{A}_{21}\boldsymbol{y}_1 + \tilde{A}_{22}\hat{\boldsymbol{z}}_2 + B_2\boldsymbol{u} - L(\tilde{C}_2\hat{\boldsymbol{z}}_2 - \boldsymbol{y}_2)\\ \hat{\boldsymbol{x}} &= T^{-1}\begin{bmatrix}\boldsymbol{y}_1\\ \hat{\boldsymbol{z}}_2\end{bmatrix}\end{aligned} \tag{3.43}$$

Considering the observation error

$$\boldsymbol{\epsilon} = \hat{\boldsymbol{x}} - \boldsymbol{x} = T^{-1}(\hat{\boldsymbol{z}} - \boldsymbol{z})$$

with dynamics

$$\begin{aligned}\dot{\boldsymbol{\epsilon}} &= T^{-1}\left(\begin{bmatrix}\dot{\boldsymbol{z}}_1\\ \tilde{A}_{21}\boldsymbol{z}_1 + \tilde{A}_{22}\hat{\boldsymbol{z}}_2 + B_2\boldsymbol{u} - L(\tilde{C}_2\hat{\boldsymbol{z}}_2 - \boldsymbol{y}_2)\end{bmatrix} - \begin{bmatrix}\dot{\boldsymbol{z}}_1\\ \tilde{A}_{21}\boldsymbol{z}_1 + \tilde{A}_{22}\boldsymbol{z}_2 + B_2\boldsymbol{u}\end{bmatrix}\right)\\ &= T^{-1}\left(\begin{bmatrix}\boldsymbol{0}\\ (\tilde{A}_{22} - L\tilde{C}_2)(\hat{\boldsymbol{z}}_2 - \boldsymbol{z}_2)\end{bmatrix}\right)\end{aligned}$$

it follows that

$$||\boldsymbol{\epsilon}(t)|| \le ||T^{-1}||\,||\hat{\boldsymbol{z}}_2(t) - \boldsymbol{z}_2(t)|| \le ||T^{-1}||\,e^{\lambda_O t}||\hat{\boldsymbol{z}}_{20} - \boldsymbol{z}_{20}||$$

where λ_O is the maximum eigenvalue of the matrix $(\tilde{A}_{22} - L\tilde{C}_2)$. This, in turn, can be assigned arbitrarily if the pair $(\tilde{A}_{22}, \tilde{C}_2)$ is completely observable. If it is not completely observable, but detectable, then L can be chosen so that $\Re(\lambda_O) < 0$. Note that the detectability of the pair $(\tilde{A}_{22}, \tilde{C}_2)$ is equivalent to the strong detectability of the complete system. Given that strong* detectability corresponds to strong detectability plus the relative degree condition (3.41) which underlies the existence of the state transformation, the existence of the designed unknown input observer requires that the system is strong* detectable. These results are summarized in the following theorem.

Theorem 3.10

The unknown-input observer (3.43) exists if and only if the system is stong* detectable.

3.4 Discrete-time observability and observer design

The observability analysis and observer design for discrete-time systems basically follows the same steps as outline above for the continuous-time setup with the difference that instead of time derivatives discrete time steps are considered [Kai80]. To clarify this, consider the discrete-time linear system

$$\boldsymbol{x}[k+1] = A_d\boldsymbol{x}[k] + B_d\boldsymbol{u}[k], \quad \boldsymbol{x}[0] = \boldsymbol{x}_0 \tag{3.44a}$$

$$\boldsymbol{y}[k] = C\boldsymbol{x}[k]. \tag{3.44b}$$

The basic definition of observability is similar to the continuous-time case.

Definition 3.6

System (3.44) is called completely observable within K steps, if any initial condition $\boldsymbol{x}_0 \in \mathbb{R}^n$ can be uniquely determined by knowledge of the output $\boldsymbol{y}[k]$ and input $\boldsymbol{u}[k], k \in [0, K]$.

To derive a method for accessing the discrete-time observability conside a sequence of measurement samples

$$\begin{aligned}
\boldsymbol{y}[0] &= C\boldsymbol{x}[0] \\
\boldsymbol{y}[1] &= C\boldsymbol{x}[1] = CA_d\boldsymbol{x}[0] + CB_d\boldsymbol{u}[0] \\
\boldsymbol{y}[2] &= C\boldsymbol{x}[2] = CA_d\boldsymbol{x}[1] + CB_d\boldsymbol{u}[1] = CA_d^2\boldsymbol{x}[0] + CA_dB_d\boldsymbol{u}[0] + CB_d\boldsymbol{u}[1] \\
&\vdots \\
\boldsymbol{y}[n-1] &= CA_d^{n-1}\boldsymbol{x}[0] + \sum_{i=0}^{n-1} CA_d^{n-1-i}B_d\boldsymbol{u}[i].
\end{aligned}$$

Joining on the left-hand side all the known terms and the unknown ones on the right-hand side yields the equation

$$\mathscr{Y} - \mathscr{U} = \mathscr{K}_{od}\boldsymbol{x}_0 \tag{3.45}$$

with $\mathscr{Y}, \mathscr{U} \in \mathbb{R}^{nm}$ and $\mathscr{K}_{od} \in \mathbb{R}^{nm\times n}$ given by

$$\mathscr{Y} = \begin{bmatrix} \boldsymbol{y}[0] \\ \vdots \\ \boldsymbol{y}[n-1] \end{bmatrix}, \quad \mathscr{U} = \begin{bmatrix} \boldsymbol{0} \\ CB_d\boldsymbol{u}[0] \\ \vdots \\ \sum_{i=0}^{n-1} CA_d^{n-1-i}B_d\boldsymbol{u}[i] \end{bmatrix}, \quad \mathscr{K}_{od} = \begin{bmatrix} C \\ CA_d \\ \vdots \\ CA_d^{n-1} \end{bmatrix}.$$

The matrix $\mathscr{K}_{od}$ is the discrete-time Kalman observability matrix. If this matrix has rank n then equation (3.45) can be uniquely solved for $\boldsymbol{x}_0$, either by considering an invertible $n \times n$ submatrix $\mathscr{K}_{od}^*$ of $\mathscr{K}_{od}$ and considering

$$\boldsymbol{x}_0 = \left(\mathscr{K}_{od}^*\right)^{-1}(\mathscr{Y} - \mathscr{U}),$$

or the (left) pseudo-inverse of $\mathscr{K}_{od}$, i.e.

$$\boldsymbol{x}_0 = \mathscr{K}_{od}^{\dagger}(\mathscr{Y} - \mathscr{U}), \quad \mathscr{K}_{od}^{\dagger} = \left(\mathscr{K}_{od}^{\mathsf{T}}\mathscr{K}_{od}\right)^{-1}\mathscr{K}_{od}.$$

In the case that the rank of $\mathscr{K}_{od}$ is $r < n$ one can follow the reasoning of the continuous-time setup and split the state space into an observable and unobservable subspace, where the observable one

is spanned by the r linear independent row vectors of the matrix $\mathscr{K}_{od}$. This yields to the Kalman decomposition

$$\begin{bmatrix} \boldsymbol{x}_o[k+1] \\ \boldsymbol{x}_n[k+1] \end{bmatrix} = \begin{bmatrix} \tilde{A}_{d,o} & 0 \\ \tilde{A}_{d,no} & \tilde{A}_{d,n} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_o[k] \\ \boldsymbol{x}_n[k] \end{bmatrix} + \tilde{B}_d \boldsymbol{u} \tag{3.46a}$$

$$\boldsymbol{y}[k] = \begin{bmatrix} \tilde{C}_o & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_o[k] \\ \boldsymbol{x}_n[k] \end{bmatrix} \tag{3.46b}$$

As pointed out in the discussion of the observability of continuous-time systems, such a separation is useful to analyze the detectability of the system, which is ensured if the eigenvalues of the matrix $\tilde{A}_{d,n}$ are contained in the open unit circle U_1.

Note that in case of observability of a linear discrete-time systems it is in principle possible to determine the initial state after a maximum of n steps. This holds true because the vectors $\mathscr{Y}$ and $\mathscr{U}$ can actually be simply calculated and equation (3.45) solved. In continuous-time systems this is practically not possible, because these vectors depend on time derivatives of $\boldsymbol{y}$ and $\boldsymbol{u}$, which can only be approximated. This approximation takes time, and is basically the reason why the Luenberger observer only converges asymptotically and not in finite time. For a discrete-time system thus an observer can be constructed by gathering the output and input samples and then inverting equation (3.45). This seems interesting overall for small-scale system, i.e. with a small number of states. For large-scale systems, i.e. with a huge number of states this quickly becomes slow and implies the inversion of a large matrix, what in turn requires typically again a lot of time and numerically ressources. Besides this reason, in the case that the sampling rate is fast in comparison to the system dynamics the differences between the rows of $\mathscr{K}_{od}$ will be very small because the associated differences between $\boldsymbol{y}[k]$ and $\boldsymbol{y}[k+1]$ will be small. In consequence the numerical invertibility of $\mathscr{K}_{od}$ can be ill conditioned and cause some trouble.

As a simple alternative to the direct inversion of (3.45), consider the discrete-time Luenberger observer

$$\hat{\boldsymbol{x}}[k+1] = A_d \hat{\boldsymbol{x}}[k] + B_d \boldsymbol{u}[k] - L\left(C\hat{\boldsymbol{x}}[k] - \boldsymbol{y}[k]\right), \quad \hat{\boldsymbol{x}}[0] = \hat{\boldsymbol{x}}_0 \tag{3.47}$$

that consists of a predictor-corrector scheme, just like the continuous-time counterpart. Considering the associated observation error $\boldsymbol{e} = \hat{\boldsymbol{x}} - \boldsymbol{x}$ with the observation error dynamics

$$\boldsymbol{e}[k+1] = (A_d - LC)\boldsymbol{e}[k], \quad \boldsymbol{e}[0] = \boldsymbol{e}_0. \tag{3.48}$$

Obviously, the eigenvalues of the matrix $A_d - LC$ must be contained in the open unit circle in order to ensure the asymptotic stability of $\boldsymbol{e} = \boldsymbol{0}$ and thus the convergence of the observer, i.e. $\lim_{k\to\infty} \|\hat{\boldsymbol{x}}[k] - \boldsymbol{x}[k]\| = 0$.

To analyze the relation between the continuous-time and discrete-time Luenberger observer, consider the former one with sampled measurement, yielding the consideration of $\boldsymbol{y}(t) = \boldsymbol{y}[k]$ and $\boldsymbol{u}(t) = \boldsymbol{u}[k]$ forall $t \in [k\Delta t, (k+1)\Delta t)$ with Δt being the sampling interval. Of course this assumption will only be approximately valid, in particular for the output signal, and will strongly depend on the sampling rate. In particular, for the case that the system dynamics is slow in comparison to the sampling rate the assumption that $C\boldsymbol{x}(t)$ remains constant over a sampling interval is reasonable. The Luenberger observer can be written as

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}} &= A\hat{\boldsymbol{x}} + B\boldsymbol{u} - L(C\hat{\boldsymbol{x}} - \boldsymbol{y}) \\ &= (A - LC)\hat{\boldsymbol{x}} + B\boldsymbol{u} + L\boldsymbol{y}. \end{aligned}$$

Under the above assumptions one has

$$\begin{aligned}\hat{\boldsymbol{x}}((k+1)\Delta t) &= e^{(A-LC)\Delta t}\hat{\boldsymbol{x}}(k\Delta t) + \int_0^{\Delta t} e^{(A-LC)(\Delta t-\tau)}\left(B\boldsymbol{u}[k] + L\boldsymbol{y}[k]\right)\mathrm{d}\tau \\ &= e^{(A-LC)\Delta t}\hat{\boldsymbol{x}}(k\Delta t) + \int_0^{\Delta t} e^{(A-LC)(\Delta t-\tau)}\mathrm{d}\tau\left(B\boldsymbol{u}[k] + L\boldsymbol{y}[k]\right) \\ &= A_{Ld}\hat{\boldsymbol{x}}(k\Delta t) + B_{Ld}\boldsymbol{u}[k] + L_d\boldsymbol{y}[k],\end{aligned}$$

with

$$A_{Ld} = e^{(A-LC)\Delta t}, \quad B_{Ld} = \int_0^{\Delta t} e^{(A-LC)(\Delta t-\tau)}\mathrm{d}\tau B, \quad L_d = \int_0^{\Delta t} e^{(A-LC)(\Delta t-\tau)}\mathrm{d}\tau L.$$

On the other hand the discrete-time Luenberger observer (3.47) can be written as

$$\hat{\boldsymbol{x}}[k+1] = (A_d - LC)\hat{\boldsymbol{x}}[k] + B_d\boldsymbol{u}[k] + L\boldsymbol{y}[k].$$

Accordingly, both observers have the same general structure and will yield the same convergence behavior in discrete-time if

$$(A_d - LC) = A_{Ld}, \quad B_d = B_{Ld}, \quad L = L_d.$$

This shows that both approaches will yield similar behavior and the decision whether a continuous or discrete-time design is used can thus basically be taken on the basis of the sampling rate in relation to the system dynamics. For fast dynamics rather fast sampling rates will be necessary, while for very slow dynamics lower sampling rates can still be considered as pseudo-continuous measurements. As discussed above in the context of the discrete-time observability analysis, a fast dynamics in combination with a slow sampling rate can yield sometimes a fast, finite-time convergence by direclty solving the equation set (3.45).

The possibility to find an appropriate correction gain matrix L for which it is ensured that the eigenvalues of the matrix $A_d - LC$ are contained in the open unit circle U_1 can again be analyzed using the Kalman decomposition (3.46). A straightforward calculation shows that

$$\tilde{A}_d - L\tilde{C} = \begin{bmatrix}\tilde{A}_{d,o} & 0 \\ \tilde{A}_{d,no} & \tilde{A}_{d,n}\end{bmatrix} - \begin{bmatrix}L_o \\ L_n\end{bmatrix}\begin{bmatrix}\tilde{C}_o & 0\end{bmatrix} = \begin{bmatrix}\tilde{A}_{d,o} - L_o\tilde{C}_o & 0 \\ \tilde{A}_{d,no} - L_n\tilde{C}_o & \tilde{A}_{d,n}\end{bmatrix}.$$

Due to the triangular structure of the matrix, the eigenvalues are given by the union of the eigenvalues of the submatrices $\tilde{A}_{d,o} - L_o\tilde{C}_o$ and $\tilde{A}_{d,n}$. The eigenvalues of $\tilde{A}_{d,o} - L_o\tilde{C}_o$ can be assigned arbitrarily, because by assumption the pair $(\tilde{A}_{d,o}, \tilde{C}_o)$ is observable. The eigenvalues of $\tilde{A}_{d,n}$ can not be moved by means of the correction term. Thus, the necessary and sufficient condition for the existence of a correction gain matrix L is that the eigenvalues of $\tilde{A}_{d,n}$ are contained (already) in the open unit circle U_1.

References

[Hau83] M. L. J. Hautus. „Strong detectability and observers“. In: *Linear Algebra Appl.* 50 (1983), pp. 353–368 (cit. on pp. 49, 50).

[Kai80] T. Kailath. *Linear Systems*. Prentice Hall, Inc., 1980 (cit. on p. 52).

[Lue71] D. G. Luenberger. „An introduction to observers“. In: *IEEE Trans. Autom. Control.* 16 (6) (1971), pp. 596–602 (cit. on p. 46).

[MK76] A. G. J. MacFarlane and N. Karcanias. „Poles and zeros of linear multivariable systems : a survey of the algebraic, geometric and complex- variable theory“. In: *International Journal of Control* 24(1) (1976), pp. 33–74 (cit. on p. 50).

Stochastic optimal state estimation

4

Stochasticity is something that every real system is subject to. Nevertheless, as far as stochastic effects can be neglected without loss of essential information, i.e. when stochastic fluctuations are small enough, the deterministic approaches discussed above yield sufficiently good results. When the stochastic effects become larger, e.g. when the uncertainties in a measurement become considerable large in comparison to the measurement values, approaches are required that achieve an ensured functioning. To be more precise, the effect of stochastic uncertainty on the prediction has to be quantitatively delimited. In the theory of linear observer design for state estimation the mostly recognized and very efficient technique for this task is the Kalman-Bucy filter for continuous-time systems, and the Kalman filter for discrete-time systems. These techniques will be discussed in the sequel and later extended for joint state and parameter estimation in the continuous-time setup.

4.1 A primer on linear stochastic systems

Consider a linear time invariant system with stochastic excitation which is equipped with a sensor system with associated measurement noise

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + G\boldsymbol{w}(t) \tag{4.1a}$$

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) + \boldsymbol{v}(t) \tag{4.1b}$$

where $\boldsymbol{x}(t)$ is the state and $\boldsymbol{u}(t)$ the deterministic (known) input, $\boldsymbol{w}(t), \boldsymbol{v}(t)$ are the process and measurement noise, respectively, considered as generated by (stationary) Gaussian processes satisfying

$$E[\boldsymbol{w}(t)] = \mathbf{0}, \quad E\left[\boldsymbol{w}(t)\boldsymbol{w}^{\mathsf{T}}(\tau)\right] = Q\delta(t-\tau)$$
$$E[\boldsymbol{v}(t)] = \mathbf{0}, \quad E\left[\boldsymbol{v}(t)\boldsymbol{v}^{\mathsf{T}}(\tau)\right] = R\delta(t-\tau)$$

and Q, R being the associated (constant) covariance matrices, respectively, and $\boldsymbol{v}(t)$ and $\boldsymbol{w}(t)$ are supposed to be uncorrelated (for all times), i.e.

$$E[\boldsymbol{w}(t)\boldsymbol{v}^{\mathsf{T}}(\tau)] = 0 \quad \forall\, t, \tau \geq 0.$$

These noise sources are typically represented using the notation $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, Q)$ and $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, R)$, meaning that they are produced by a normally distributed variable with mean $\mathbf{0}$ and covariance Q, or R, respectively. See also Section 2.2.4.1 for further discussion on the noise model.

Given the stochastic excitation, the state $\boldsymbol{x}(t)$ is also a stochastic process. Thus the initial condition $\boldsymbol{x}(0)$ is considered as a stochastic variable with a Gaussian distribution with mean value $E[\boldsymbol{x}(0)] = \bar{\boldsymbol{x}}_0$ and covariance[1] $E[(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)^{\mathsf{T}}] = \Sigma_0$. It is assumed that $\boldsymbol{x}(0)$ is statistically independent of $\boldsymbol{w}$, i.e. $E[\boldsymbol{x}(0)G_d\boldsymbol{w}^{\mathsf{T}}(t)] = 0$ for all $t \geq 0$.

[1] Note that in the scalar case $\Sigma_0 = \sigma^2$ with sigma being the standard deviation of the associated Gaussian distribution of the initial condition.

Considering only white process noise $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, Q)$ the stochastic terms can be completely characterized by the expectaction (or mean) and the covariance, because these two parameters completely determine the Gaussian probability density function $\varphi(x)$ (2.4). Accordingly, in the sequel a characterization of the time evolution of the expectation and covariance will be derived.

The time evolution of the expectation $\bar{\boldsymbol{x}}(t) = E[\boldsymbol{x}(t)]$ is determined according to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{x}} &= \frac{\mathrm{d}}{\mathrm{d}t}E[\boldsymbol{x}(t)] = \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\boldsymbol{x}(t)\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\dot{\boldsymbol{x}}(t)\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= E[\dot{\boldsymbol{x}}(t)] = E[A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + G\boldsymbol{w}(t)].\end{aligned}$$

Recalling that $E[\boldsymbol{w}(t)] = \mathbf{0}$ this yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\boldsymbol{x}}(t) = A\bar{\boldsymbol{x}}(t) + B\boldsymbol{u}(t), \quad \bar{\boldsymbol{x}}(0) = \bar{\boldsymbol{x}}_0. \tag{4.2}$$

It should be noted that this means, that the time evolution of the mean value completely corresponds to the solution of the deterministic differential equation that has been considered in the preceding chapters. In particular, the solution is given by the variation of constants formula

$$\bar{\boldsymbol{x}}(t) = S(t)\bar{\boldsymbol{x}}_0 + \int_0^t S(t-\tau)B\boldsymbol{u}(\tau)\mathrm{d}\tau, \tag{4.3}$$

with the state transition matrix $S(t) = e^{At}$.

The solution of the differential equation (4.1) can be written using the variation of constants formula as

$$\boldsymbol{x}(t) = S(t)\boldsymbol{x}_0 + \int_0^t S(t-\tau)B\boldsymbol{u}(\tau)\mathrm{d}\tau + \int_0^t S(t-\tau)G\boldsymbol{w}(\tau)\mathrm{d}\tau.$$

Thus it follows that

$$\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t) = S(t)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0) + \int_0^t S(t-\tau)G\boldsymbol{w}(\tau)\mathrm{d}\tau.$$

In consequence, the covariance is then given by

$$\begin{aligned}\Sigma(t) &= E[(\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t))(\boldsymbol{x}(t) - \bar{\boldsymbol{x}}(t))^{\mathsf{T}}]\\ &= E\left[\left(S(t)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0) + \int_0^t S(t-\tau)G\boldsymbol{w}(\tau)\mathrm{d}\tau\right)\left(S(t)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0) + \int_0^t S(t-\tau)G\boldsymbol{w}(\tau)\mathrm{d}\tau\right)^{\mathsf{T}}\right]\\ &= S(t)E[(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)^{\mathsf{T}}]S^{\mathsf{T}}(t) + S(t)\int_0^t E\left[(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)\boldsymbol{w}^{\mathsf{T}}(\tau)\right]G^{\mathsf{T}}S^{\mathsf{T}}(t-\tau)\mathrm{d}\tau +\\ &\quad + \int_0^t S(t-\tau)GE\left[\boldsymbol{w}(\tau)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)^{\mathsf{T}}\right]\mathrm{d}\tau S^{\mathsf{T}}(t) + \int_0^t\int_0^t S(t-\tau)GE[\boldsymbol{w}(\tau)\boldsymbol{w}^{\mathsf{T}}(s)]G^{\mathsf{T}}S^{\mathsf{T}}(t-s)\mathrm{d}\tau\mathrm{d}s.\end{aligned}$$

The first term contains the dependence on the initial covariance $\Sigma(0) = \Sigma_0$. From the statistical independence of $\boldsymbol{x}(0)$ and $\boldsymbol{w}(t)$ (for all $t \geq 0$) it follows that

$$E[(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)\boldsymbol{w}^{\mathsf{T}}(t)] = E[\boldsymbol{w}(t)(\boldsymbol{x}(0) - \bar{\boldsymbol{x}}_0)^{\mathsf{T}}] = 0.$$

Given that $\boldsymbol{w}(t)$ is a Gaussian (white noise) process it holds that

$$E[\boldsymbol{w}(\tau)\boldsymbol{w}^{\mathsf{T}}(s)] = Q\delta(\tau - s).$$

Furthermore, recall that $S(t-\tau) = S(t)S(-\tau) = S(-\tau)S(t)$. With this, the double integral term results as

$$\begin{aligned}
&\int_0^t \int_0^t S(t-\tau)GE[\boldsymbol{w}(\tau)\boldsymbol{w}^\mathsf{T}(s)]G^\mathsf{T}S^\mathsf{T}(t-s)\mathrm{d}\tau\mathrm{d}s \\
&= \int_0^t \int_0^t S(t-\tau)GQ\delta(\tau-s)G^\mathsf{T}S^\mathsf{T}(t-s)\mathrm{d}\tau\mathrm{d}s \\
&= \int_0^t S(t)S(-\tau)GQG^\mathsf{T}\big(S(t)S(-\tau)\big)^\mathsf{T}\mathrm{d}\tau \\
&= S(t)\int_0^t S(-\tau)GQG^\mathsf{T}S^\mathsf{T}(-\tau)\mathrm{d}\tau S^\mathsf{T}(t)
\end{aligned}$$

Accordingly, the above expression for the covariance $\Sigma(t)$ simplifies to

$$\Sigma(t) = S(t)\left\{\Sigma_0 + \int_0^t S(-\tau)GQG^\mathsf{T}S^\mathsf{T}(-\tau)\mathrm{d}\tau\right\}S^\mathsf{T}(t). \tag{4.4}$$

As a result of this dependence it follows that for $\boldsymbol{u} = \mathbf{0}$ the covariance matrix Σ satisfies the differential equation

$$\dot{\Sigma}(t) = A\Sigma(t) + \Sigma(t)A^\mathsf{T} + GQG^\mathsf{T}, \quad \Sigma(0) = \Sigma_0. \tag{4.5}$$

Having established this, it directly follows that the state will remain normally distributed for all times when the initial distribution is normal (i.e. Gaussian) and the probability density function of the state distribution can be established for all time instances by the solutions of the deterministic differential equations for the expectation (4.2) and covariance (4.5).

Example 4.1. *Consider the first order stochastic system* [a]

$$\dot{x} = -\lambda x + w, \quad w \sim \mathcal{N}(0, q).$$

Consider that the expectation for the initial value satisfies $E[x(0)] = \bar{x}_0$ *and the variance* $E[x^2(0)] = \varsigma_0 = \sigma_0^2$ *with the standard deviation* σ *of the associated Gaussian probability density function*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x}_0)^2}{2\sigma^2}}$$

shown in Figure 4.1 (left) for $\bar{x}_0 = 0.5$ *and* $\sigma_0 = 0.3$.

According to the above considerations the mean value $e(t) = E[x(t)]$ *is the solution of the deterministic differential equation*

$$\dot{e} = -\lambda e, \quad e(0) = \bar{x}_0 \quad \Rightarrow \quad e(t) = e^{-\lambda t}\bar{x}_0$$

and converges exponentially to zero with a characteristic time $t_c = \frac{1}{\lambda}$. *The covariance* $\varsigma = E[x^2(t)]$ *satisfies*

$$\dot{\varsigma} = -2\lambda\varsigma + q, \quad \varsigma(0) = \sigma_0^2 \quad \Rightarrow \quad \varsigma(t) = e^{-2\lambda t}\sigma_0^2 + \frac{q}{2\lambda}\left(1 - e^{-2\lambda t}\right).$$

This implies that at any time the state probability distribution is given by a Gaussian with mean at $e(t)$ *and variance* $\sigma(t)$ *given by*

$$e(t) = e^{-\lambda t}\bar{x}_0, \quad \sigma(t) = \sqrt{\varsigma(t)}.$$

The $\pm\sigma$ confidence interval $[e(t)-\sigma, e(t)+\sigma]$ is shown in Figure 4.1 (right) together with a simulation example run with numerically generated white noise for the values $\lambda = 1, q = 0.5$. It can be seen that the solution is contained in this interval which grows over time, implying that the predictive certainty becomes smaller while the time increases. The stationary distribution if shown in Figure 4.1 (left) with mean value 0 and standard deviation σ_∞.

[a] This system basically resembles the Ornstein-Uhlenbeck process that is widely considered in the study of stochastic processes [RW00].

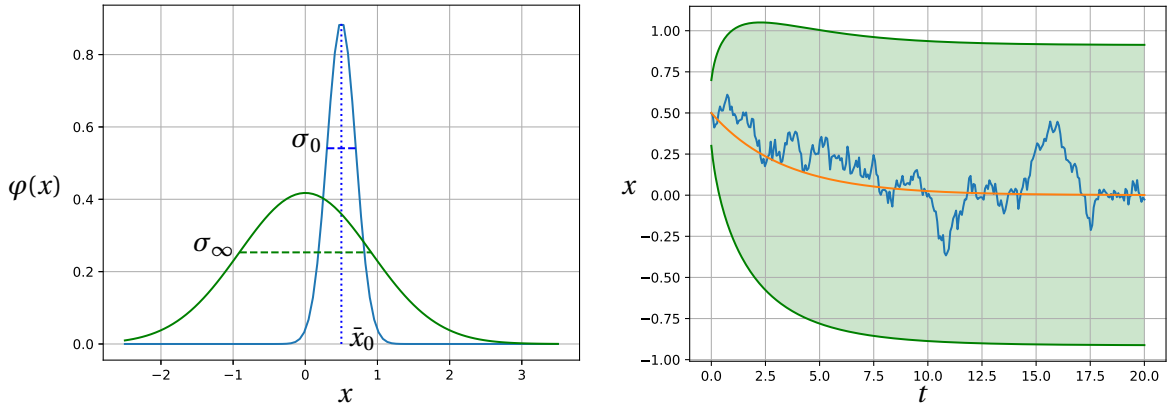


Figure 4.1: Left: Initial and stationary probability density functions. Right: Numerical simulation (blue line) with expectation (orange line) and $\pm\sigma(t)$ confidence interval (shaded region).

4.2 The Kalman-Bucy filter

Having the considerations from the preliminary section as point of departure the stochastic state estimation problem is formaly stated as follows:

Problem formulation: Design a stochastic observer (filter) which provides a state estimate[a] $\hat{\boldsymbol{x}}(t)$ that is unbiased, i.e. the associated estimation error $\boldsymbol{e}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ satisfies

$$E[\boldsymbol{e}(0)] = \boldsymbol{0} \quad \Rightarrow \quad E[\boldsymbol{e}(t)] = \boldsymbol{0}$$

and has minimum error covariance, i.e. $E[\boldsymbol{e}(t)\boldsymbol{e}^\mathsf{T}(t)]$ is minimum.

[a] In the stochastic framework, instead of observation the term estimation is used. Nevertheless, in the common use of terminology these terms are mixed in the deterministic set-up.

This problem has been solved by Rudolph Kalman and Richard Bucy in the 1960's [KB61] and later been derived using alternative methods (see e.g. [AT67; Gel78]). In the following a simple approach by directly using the Luenberger observer structure is derived using a simple optimization approach.

Consider a Luenberger observer with a time-varying correction gain matrix $L(t)$

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - L(t)\big(C\hat{\boldsymbol{x}}(t) - \boldsymbol{y}(t)\big), \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0$$

The associated state estimation error $\boldsymbol{e}(t) = \hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$ thus satisfies the differential equation

$$\dot{\boldsymbol{e}}(t) = (A - L(t)C)\boldsymbol{e}(t) + L(t)\boldsymbol{v}(t) - G\boldsymbol{w}(t), \quad \boldsymbol{e}(0) = \boldsymbol{e}_0. \tag{4.6}$$

This observer is unbiased, given that

$$E[\boldsymbol{e}_0] = 0, \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}t} E[\boldsymbol{e}(0)] = (A - L(t)C)E[\boldsymbol{e}(0)] = \mathbf{0}$$

implying that $E[\boldsymbol{e}] = \mathbf{0}$ is an equilibrium point for the expectation.

The error covariance is given by $P(t) = E[\boldsymbol{e}(t)\boldsymbol{e}^\mathsf{T}(t)] = P^\mathsf{T}(t) > 0$ with $P(0) = E[\boldsymbol{e}_0\boldsymbol{e}_0^\mathsf{T}]$ and satisfies

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} P(t) = E[\dot{\boldsymbol{e}}(t)\boldsymbol{e}^\mathsf{T}(t)] + E[\boldsymbol{e}(t)\dot{\boldsymbol{e}}^\mathsf{T}(t)] \\
&= (A - L(t)C)P(t) + L(t)E[\boldsymbol{v}(t)\boldsymbol{e}^\mathsf{T}(t)] - GE[\boldsymbol{w}(t)\boldsymbol{e}^\mathsf{T}(t)] \\
&\qquad + P(t)(A - L(t)C)^\mathsf{T} + E[\boldsymbol{e}(t)\boldsymbol{v}^\mathsf{T}(t)]L^\mathsf{T}(t) - E[\boldsymbol{e}(t)\boldsymbol{w}^\mathsf{T}(t)]G^\mathsf{T}
\end{aligned}$$

From the differential equation (4.6) for $\boldsymbol{e}(t)$ it follows on the other side that

$$\boldsymbol{e}(t) = S_L(t)\boldsymbol{e}_0 + \int_0^t S_L(t-\tau)\left(L(\tau)\boldsymbol{v}(\tau) - G\boldsymbol{w}(\tau)\right)\mathrm{d}\tau$$

with $S_L(t)$ being the state transition matrix (fundamental solution) of the time-varying ode with $\boldsymbol{v} = \mathbf{0}$ and $\boldsymbol{w} = \mathbf{0}$. The fundamental solution $S(t)$ satisfies furthermore $S(t-\tau) = S(t)S(-\tau)$.

Substituting the solution for $\boldsymbol{e}(t)$ into the expressions $E[\boldsymbol{v}(t)\boldsymbol{e}^\mathsf{T}(t)]$, $E[\boldsymbol{w}(t)\boldsymbol{e}^\mathsf{T}(t)]$,$E[\boldsymbol{e}(t)\boldsymbol{v}^\mathsf{T}(t)]$, $E[\boldsymbol{e}(t)\boldsymbol{w}^\mathsf{T}(t)]$ and recalling that for all $t \geq 0$ it holds that $E[\boldsymbol{v}(t)\boldsymbol{e}_0^\mathsf{T}] = E[\boldsymbol{e}_0\boldsymbol{v}^\mathsf{T}(t)] = 0$, $E[\boldsymbol{e}_0, \boldsymbol{w}^\mathsf{T}(t)] = E[\boldsymbol{w}(t)\boldsymbol{e}_0^\mathsf{T}] = 0$ and $E[\boldsymbol{v}(t), \boldsymbol{w}^\mathsf{T}(t)] = E[\boldsymbol{w}(t), \boldsymbol{v}^\mathsf{T}(t)] = 0$ it follows that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} P(t) &= (A - L(t)C)P(t) + P(t)(A - L(t)C)^\mathsf{T} + L(t)\int_0^t E[\boldsymbol{v}(t)\boldsymbol{v}^\mathsf{T}(\tau)]L^\mathsf{T}(\tau)S_L^\mathsf{T}(t-\tau)\mathrm{d}\tau \\
&\quad + \int_0^t S_L(t-\tau)L(\tau)E[\boldsymbol{v}(\tau)\boldsymbol{v}^\mathsf{T}(t)]\mathrm{d}\tau L^\mathsf{T}(t) + G\int_0^t E[\boldsymbol{w}(t)\boldsymbol{w}^\mathsf{T}(\tau)]G^T S_L^\mathsf{T}(t-\tau)\mathrm{d}\tau \\
&\quad + \int_0^t S_L(t-\tau)GE[\boldsymbol{w}(\tau)\boldsymbol{w}^\mathsf{T}(t)]G^T\mathrm{d}\tau
\end{aligned}$$

Note that for the Dirac δ-Function the following holds

$$\int_0^a \delta(t-a)f(t)\mathrm{d}t = \frac{1}{2}f(a)$$

given that the impulse is evaluated at the upper limit of the integral (*and thus loosely speaking, only half the value of the impulse is obtained* [AT67]). Accordingly, as $\boldsymbol{v}$ and $\boldsymbol{w}$ are Gaussian (white-noise) processes it holds that

$$\begin{aligned}
&L(t)\int_0^t E[\boldsymbol{v}(t)\boldsymbol{v}^\mathsf{T}(\tau)]L^\mathsf{T}(\tau)S_L^\mathsf{T}(t-\tau)\mathrm{d}\tau + \int_0^t S_L(t-\tau)L(\tau)E[\boldsymbol{v}(\tau)\boldsymbol{v}^\mathsf{T}(t)]\mathrm{d}\tau L^\mathsf{T}(t) = L(t)RL^\mathsf{T}(t) \\
&G\int_0^t E[\boldsymbol{w}(t)\boldsymbol{w}^\mathsf{T}(\tau)]G^T S_L^\mathsf{T}(t-\tau)\mathrm{d}\tau + \int_0^t S_L(t-\tau)GE[\boldsymbol{w}(\tau)\boldsymbol{w}^\mathsf{T}(t)]G^T\mathrm{d}\tau = GQG^T
\end{aligned}$$

Summarizing the estimation error covariance matrix satisfies the Differential Equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = AP + PA^\mathsf{T} - L(t)CP - PC^\mathsf{T}L^\mathsf{T}(t) + L(t)RL^\mathsf{T}(t) + GQG^T$$

with $P(0) = P_0$. Completing squares on the right-hand side one obtains

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(t) &= AP + PA^\intercal + GQG^T - P(t)C^\intercal R^{-1}CP(t) + \left(L(t) - P(t)C^\intercal R^{-1}\right)R\left(L(t) - P(t)C^\intercal R^{-1}\right)^\intercal \\ &\geq AP + PA^\intercal + GQG^T - P(t)C^\intercal R^{-1}CP(t).\end{aligned}$$

As $P(t) \geq 0$ for all $t \geq 0$ it is clear that $P(t)$ is minimized if $\frac{\mathrm{d}}{\mathrm{d}t}P(t)$ is minimized (i.e. becomes more negative). Thus, the optimal choice of $L(t)$ is given by

$$L(t) = P(t)C^\intercal R^{-1}. \tag{4.7}$$

The resulting state estimation scheme is summarized next.

Kalman-Bucy-Filter: The unbiased minimum covariance estimator for the stochastic state estimation problem is given by

$$\dot{\hat{\boldsymbol{x}}}(t) = A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) - L(t)\left(C\hat{\boldsymbol{x}}(t) - \boldsymbol{y}(t)\right), \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0 \tag{4.8a}$$

$$L(t) = P(t)C^\intercal R^{-1} \tag{4.8b}$$

where $P(t)$ is the solution of the Ricatti Differential Equation

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t) = AP(t) + P(t)A^\intercal + GQG^T - P(t)C^\intercal R^{-1}CP(t), \quad P(0) = P_0. \tag{4.8c}$$

As stated before, for $E[\boldsymbol{e}_0] = \boldsymbol{0}$ the Kalman-Bucy Filter yields an unbiased estimation. If statistical information is at hand P_0 can be chosen as $E[\boldsymbol{e}_0\boldsymbol{e}_0^\intercal]$ and $\hat{\boldsymbol{x}}_0$ as $E[\boldsymbol{x}_0] = \bar{\boldsymbol{x}}_0$.

Note that for $\boldsymbol{x} \in \mathbb{R}^n$ the $n \times n$ covariance matrix P has n^2 entries. Thus, the solution of the Ricatti Differential Equation may cause a siginificant computational load in the implementation. Due to symmetry (i.e., $P^\intercal(t) = P(t)$) the number of independent entries is reduced to $\frac{1}{2}n(n+1)$.

Given that A, B, G, C are constant matrices one can use the associated stationary version of the differential equation, namely the Algebraic Ricatti Equation (ARE)

$$0 = AP + PA^\intercal + GQG^T - PC^\intercal R^{-1}CP, \quad L = PC^\intercal R^{-1} \tag{4.9}$$

The solution of the associated equation set for the Kalman-Bucy Filter based on the solution of the ARE (4.9) (i.e., for given system matrices obtain the gain L and stationary value of P) is implemented in MATLAB in the function `kalman` (see e.g. `https://de.mathworks.com/help/control/ref/kalman.html`).

Example 4.2. *Consider again the first order stochastic system*

$$\begin{aligned}\dot{x} &= -\lambda x + w, & w &\sim \mathcal{N}(0, q), \\ y &= x + v, & v &\sim \mathcal{N}(0, r).\end{aligned}$$

As before let $E[x(0)] = \bar{x}_0$ *and* $E[x^2(0)] = \varsigma_0 = \sigma_0^2$.

The Kalman-Bucy Filter is given by

$$\dot{\hat{\boldsymbol{x}}} = -\lambda\hat{\boldsymbol{x}} - l(x - y), \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0, \quad l = \frac{p}{r},$$

with the error covariance p being the solution of the Ricatti-Differential Equation

$$\dot{p} = -2\lambda p + q - \frac{p^2}{r}, \quad p(0) = \sigma_0^2.$$

The numerical solution is shown in Figure 4.2 for the case $\lambda = 1, q = 0.5, r = 0.2, \sigma_0 = 0.2$. It can be seen that $\hat{\boldsymbol{x}}$ converges into the vecinity of x in such a way that the real state $x(t)$ is included in the $\pm\sigma(t)$ confidence interval where

$$\sigma(t) = \sqrt{p(t)}.$$

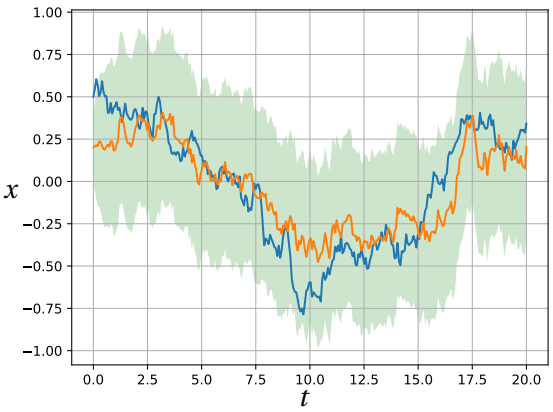


Figure 4.2: Numerical simulation (blue line) and estimated value (orange line) with $\pm\sigma(t)$ confidence interval (green shaded region).

4.3 Sampled data stochastic systems

In this section the discrete-time counterpart of the stochastic optimal Kalman-Bucy Filter is addressed. For this purpose consider the discrete-time system

$$\boldsymbol{x}[k+1] = A_d\boldsymbol{x}[k] + B_d\boldsymbol{u}[k] + G_d\boldsymbol{w}[k], \qquad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, Q) \tag{4.10a}$$

$$\boldsymbol{y}[k] = C\boldsymbol{x}[k] + \boldsymbol{v}[k], \qquad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, R) \tag{4.10b}$$

with initial condition according to normal distributed random variable $\boldsymbol{x}(0) \sim \mathcal{N}(\bar{\boldsymbol{x}}_0, \sigma^2)$.

As discussed in the first chapters, such a system representation can be obtained e.g. using the exact discretization of a continuous time system assuming that A is nonsingular. In this case

$$A_d = \exp(A\Delta t), \quad B_d = -A^{-1}(I - A_d)B, \quad G_d = -A^{-1}(I - A_d)G.$$

It should be noted that the noise covariances will probably vary when a sampled version is considered [Gel78].

The mean value $\bar{\boldsymbol{x}}[k] = E[\boldsymbol{x}[k]]$ of the stochastic variable $\boldsymbol{x}[k]$ is given by

$$\bar{\boldsymbol{x}}[k] = E[\boldsymbol{x}[k]] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \boldsymbol{x}[k] \mathrm{d}x \tag{4.11}$$

and satisfies

$$\bar{\boldsymbol{x}}[k+1] = E[A_d\boldsymbol{x}[k] + B_d\boldsymbol{u}[k] + G_d\boldsymbol{w}[k]] = A_d\bar{\boldsymbol{x}}[k] + B_d\boldsymbol{u}[k], \quad \bar{\boldsymbol{x}}[0] = \bar{\boldsymbol{x}}_0. \tag{4.12}$$

Thus it holds that

$$\boldsymbol{x}[k+1] - \bar{\boldsymbol{x}}[k+1] = A_d(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k]) + G_d\boldsymbol{w}[k], \quad (\boldsymbol{x}[0] - \bar{\boldsymbol{x}}[0]) = \boldsymbol{x}[0] - \bar{\boldsymbol{x}}_0.$$

The covariance $\Sigma[k] = E[(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}]$ then satisfies

$$\begin{aligned}
\Sigma[k+1] &= E[(A_d(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k]) + G_d\boldsymbol{w}[k])(A_d(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k]) + G_d\boldsymbol{w}[k])^\mathsf{T}] \\
&= E[A_d(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}A_d^\mathsf{T}] + E[A_d(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])\boldsymbol{w}^\mathsf{T}[k]G_d^\mathsf{T}] \\
&\qquad + E[G_d\boldsymbol{w}[k](\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}A_d^\mathsf{T}] + E[G_d\boldsymbol{w}[k]\boldsymbol{w}^\mathsf{T}[k]G_d^\mathsf{T}] \\
&= A_dE[(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}]A_d^\mathsf{T} + A_dE[(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])\boldsymbol{w}^\mathsf{T}[k]]G_d^\mathsf{T} \\
&\qquad + G_dE[\boldsymbol{w}[k](\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}]A_d^\mathsf{T} + G_dE[\boldsymbol{w}[k]\boldsymbol{w}^\mathsf{T}[k]]G_d^\mathsf{T}.
\end{aligned}$$

Given that $E[(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}] = \Sigma[k]$, $E[(\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])\boldsymbol{w}^\mathsf{T}[k]] = E[\boldsymbol{w}[k](\boldsymbol{x}[k] - \bar{\boldsymbol{x}}[k])^\mathsf{T}] = 0$ and $E[\boldsymbol{w}[k]\boldsymbol{w}^\mathsf{T}[k]] = Q$ it follows that the covariance satisfies

$$\Sigma[k+1] = A_d\Sigma[k]A_d^\mathsf{T} + G_dQG_d^\mathsf{T}, \quad \Sigma[0] = \Sigma_0. \tag{4.13}$$

Example 4.3. *Consider the scalar system*

$$x[k+1] = 0.3x[k] + \sin[k] + 0.5w[k], w[k] \sim \mathcal{N}(0, q), \quad x[0] \sim \mathcal{N}(\bar{x}_0, \sigma_0^2),$$

with $q = 0.8, \sigma_0 = 0.2$. *According to the above considerations, in particular* (4.12) *and* (4.13) *the mean value* $\bar{\boldsymbol{x}}[k]$ *and the covariance* $\varsigma[k]$ *satisfy*

$$\bar{x}[k+1] = 0.3\bar{x}_0 + \sin[k], \qquad \bar{x}[0] = \bar{x}_0$$

$$\varsigma[k+1] = 0.3^2\varsigma[k] + 0.25q, \qquad \varsigma[0] = \sigma^2.$$

According the preceding discussions, the solution $x[k]$ is standard distributed with mean $\bar{x}[k]$ and standard deviation $\sigma[k] = \sqrt{\varsigma[k]}$. The associated time evolution of the state $x[k]$, the mean value $\bar{x}$ and the associated $\pm\sigma[k]$ interval are shown in Figure 4.3. It can be seen that the mean value $\bar{x}$ converges to the oscillating solution induced by the periodic system input, and the state is contained almost all the time in the $\bar{x} \pm 3\sigma$ confidence interval.

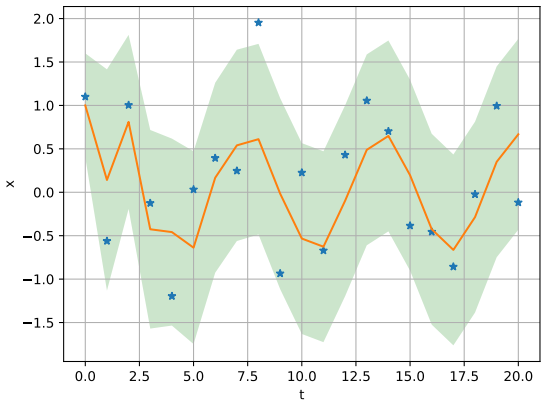


Figure 4.3: Illustration of the state evolution $x[k]$ (blue) with associated mean value $\bar{x}[k]$ (orange) and $\bar{x}[k] \pm 3\sigma[k]$ interval (green-shaded region).

4.4 The Kalman Filter

The following derivations are similar to the ones in the standard literature on sampled data, i.e. discrete-time state estimation (see e.g. [Gel78; Kal60; Lof90]). Consider a state estimate $\hat{\boldsymbol{x}}[k-1]$ is given. This estimate yields the prediction

$$\hat{\boldsymbol{x}}_p[k] = A_d\hat{\boldsymbol{x}}[k-1] + B_d\boldsymbol{u}[k-1]$$

with a prediction error $\boldsymbol{e}_p[k] = \hat{\boldsymbol{x}}_p[k] - \boldsymbol{x}[k]$. Incorporating the measurement information the prediction $\hat{\boldsymbol{x}}_p[k]$ is improved according to

$$\hat{\boldsymbol{x}}[k] = \hat{\boldsymbol{x}}_p[k] + L[k]\left(\boldsymbol{y}[k] - C\hat{\boldsymbol{x}}_p[k]\right)$$

In this context the term

$$\boldsymbol{i}[k] = \boldsymbol{y}[k] - C\hat{\boldsymbol{x}}_p[k]$$

is called innovation. The state estimate can then be expressed as

$$\hat{\boldsymbol{x}}[k] = (I - L[k]C)\,\hat{\boldsymbol{x}}_p[k] + L[k]C\boldsymbol{x}[k] + L[k]\boldsymbol{v}[k]$$

Thus the estimation error at $t = t_k$ is given by

$$\begin{aligned}\boldsymbol{e}[k] &= \hat{\boldsymbol{x}}[k] - \boldsymbol{x}[k] \\ &= (I - L[k]C)\,\hat{\boldsymbol{x}}_p[k] + L[k]C\boldsymbol{x}[k] + L[k]\boldsymbol{v}[k] - \boldsymbol{x}[k] \\ &= (I - L[k]C)\left(\hat{\boldsymbol{x}}_p[k] - \boldsymbol{x}[k]\right) + L[k]\boldsymbol{v}[k]\end{aligned}$$

yielding

$$\boldsymbol{e}[k] = (I - L[k]C)\,\boldsymbol{e}_p[k] + L[k]\boldsymbol{v}[k] \tag{4.14}$$

with the prediction error $\boldsymbol{e}_p[k]$ given by

$$\boldsymbol{e}_p[k] = \hat{\boldsymbol{x}}_p[k] - \boldsymbol{x}[k] = A_d \boldsymbol{e}[k-1] - G_d \boldsymbol{w}[k-1]. \tag{4.15}$$

The associated covariance of the prediction error is given by

$$\begin{aligned} P_p[k] &= E\left[\boldsymbol{e}_p[k]\boldsymbol{e}_p^\mathsf{T}[k]\right] \\ &= E\left[(A_d\boldsymbol{e}[k-1] - G_d\boldsymbol{w}[k-1])(A_d\boldsymbol{e}[k-1] - G_d\boldsymbol{w}[k-1])^\mathsf{T}\right] \\ &= E\left[A_d\boldsymbol{e}[k-1]\boldsymbol{e}^\mathsf{T}[k-1]A_d^\mathsf{T}\right] - E\left[A_d\boldsymbol{e}[k-1]\boldsymbol{w}^\mathsf{T}[k-1]G_d^\mathsf{T}\right] \\ &\qquad - E\left[G_d\boldsymbol{w}[k-1]\boldsymbol{e}^\mathsf{T}[k-1]A_d^\mathsf{T}\right] + E\left[G_d\boldsymbol{w}[k-1]\boldsymbol{w}^\mathsf{T}[k-1]G_d^\mathsf{T}\right]. \end{aligned}$$

Note that the estimation error $\boldsymbol{e}[k-1]$ does not depend on $\boldsymbol{w}[k-1]$ according to (4.14)-(4.15) but on $\boldsymbol{w}[k-2]$. As $E[\boldsymbol{w}[k-1]\boldsymbol{w}^\mathsf{T}[k-2]] = 0$ it results that $E[\boldsymbol{e}[k-1]\boldsymbol{w}[k-1]] = 0$ and the expression for the prediction error covariance simplifies to

$$P_p[k] = A_d P[k-1] A_d^\mathsf{T} + G_d Q G_d^\mathsf{T}. \tag{4.16}$$

The covariance $P[k]$ of the estimation error $\boldsymbol{e}[k]$ is given by

$$\begin{aligned} P[k] &= E\left[\boldsymbol{e}[k]\boldsymbol{e}^\mathsf{T}[k]\right] \\ &= E\left[\left((I - L[k]C)\,\boldsymbol{e}_p[k] + L[k]\boldsymbol{v}[k]\right)\left((I - L[k]C)\,\boldsymbol{e}_p[k] + L[k]\boldsymbol{v}[k]\right)^\mathsf{T}\right] \end{aligned}$$

Given that $E[\boldsymbol{v}[k-1]\boldsymbol{v}^\mathsf{T}[k-2]] = 0$ it follows from (4.14) that $\boldsymbol{v}$ and $\boldsymbol{e}_p$ are statistically independent (i.e. uncorrelated) and thus $E\left[\boldsymbol{e}_p[k]\boldsymbol{v}^\mathsf{T}[k]\right] = 0$. Thus the expression for $P[k]$ simplifies to

$$\begin{aligned} P[k] &= (I - L[k]C)\,E\left[\boldsymbol{e}_p[k]\boldsymbol{e}_p^\mathsf{T}[k]\right](I - L[k]C)^\mathsf{T} + L[k]E[\boldsymbol{v}[k]\boldsymbol{v}^\mathsf{T}[k]]L^\mathsf{T}[k] \\ &= (I - L[k]C)\,P_p[k]\,(I - L[k]C)^\mathsf{T} + L[k]RL^\mathsf{T}[k] \\ &= P_p[k] - L[k]CP_p[k] - P_p[k]C^\mathsf{T}L^\mathsf{T}[k] + L[k]CP_p[k]C^\mathsf{T}L^\mathsf{T}[k] + L[k]RL^\mathsf{T}[k] \end{aligned}$$

Completing squares on the right-hand side yields

$$\begin{aligned} P[k] &= P_p[k] - P_p[k]C^\mathsf{T}\left(CP_p[k]C^\mathsf{T} + R\right)^{-1}CP_p[k] + \\ &+ \left(L[k] - P_p[k]C^\mathsf{T}(CP_p[k]C^\mathsf{T} + R)^{-1}\right)\left(CP_p[k]C^\mathsf{T} + R\right)\left(L[k] - P_p[k]C^\mathsf{T}(CP_p[k]C^\mathsf{T} + R)^{-1}\right)^\mathsf{T} \\ &\geq P_p[k] - P_p[k]C^\mathsf{T}\left(CP_p[k]C^\mathsf{T} + R\right)^{-1}CP_p[k] \end{aligned}$$

Thus, the minimum estimation error covariance $P[k]$ is obtained by choosing

$$L[k] = P_p[k]C^\mathsf{T}(CP_p[k]C^\mathsf{T} + R)^{-1}.$$

In this case $L[k]$ is called the Kalman gain. With $L[k]$ chosen this way the covariance of the estimation error can be compactly written in function of the covariance of the prediction error as

$$P[k] = (I - L[k]C)\,P_p[k].$$

Summarizing the above one obtains the Kalman Filter equations [Kal60; Gel78].

Kalman Filter equations

- Prediction:

$$\hat{\boldsymbol{x}}_p[k] = A_d\hat{\boldsymbol{x}}[k-1] + B_d\boldsymbol{u}[k-1], \qquad \hat{\boldsymbol{x}}[0] = \hat{\boldsymbol{x}}_0 \tag{4.17a}$$
$$P_p[k] = A_d P[k-1] A_d^\intercal + G_d Q G_d^\intercal, \qquad P[0] = P_0 \tag{4.17b}$$

- Kalman gain

$$L[k] = P_p[k]C^\intercal \left(CP_p[k]C^\intercal + R\right)^{-1} \tag{4.17c}$$

- Correction (innovation)

$$\hat{\boldsymbol{x}}[k] = (I - L[k]C)\,\hat{\boldsymbol{x}}_p[k] + L[k]\boldsymbol{y}[k] \tag{4.17d}$$
$$P[k] = (I - L[k]C)\,P_p[k] \tag{4.17e}$$

This approach is illustrated next with a simple example.

Example 4.4. *Consider again the scalar system*

$$x[k+1] = 0.7x[k] + \sin[k] + 0.5w[k], \qquad w[k] \sim \mathcal{N}(0, q)$$
$$y[k] = x[k] + v[k], \qquad v[k] \sim \mathcal{N}(0, r),$$

with initial condition $\boldsymbol{x}[0] \sim \mathcal{N}(\boldsymbol{x}_0, \sigma^2)$, *and* $q = 0.8, r = 0.1, \sigma = 0.2$.

The Kalman Filter is implemented according to (4.17), *i.e.*

$$\hat{x}_p[k] = 0.7\hat{x}[k-1] + \sin[k-1], \qquad \hat{x}[0] = \hat{x}_0$$
$$p_p[k] = 0.7^2 p[k-1] + 0.5^2 q, \qquad p[0] = p_0$$
$$l[k] = \frac{p_p[k]}{p_p[k] + r}$$
$$\hat{x}[k] = (1 - l[k])\,\hat{x}_p[k] + l[k]y[k]$$
$$p[k] = (1 - l[k])\,p_p[k].$$

At each time instant the associated standard deviation is given by

$$\sigma_{kf}[k] = \sqrt{p[k]}.$$

The results are depicted in Figure 4.4 comparing the state $x[k]$ *(blue stars), its Kalman-Filter-based estimate* $\hat{x}[k]$ *(orange line), the mean value* $\bar{x}[k]$ *(green line) calculated as in Example 4.3 with the associated system covariance* $\sigma[k] = \sqrt{\varsigma[k]}$ *based* $\bar{x} \pm 3\sigma$ *confidence interval (magenta shaded region), and the estimation confidence region* $\hat{x} \pm 3\sigma_{kf}$ *(green shaded region). It can be seen that the state almost always is included in the Kalman-Filter based confidence region and the estimate by the Kalman Filter is superior to the one by the mean value and system covariance. This is particularly due to the low measurement noise with variance* $r = 0.1$.

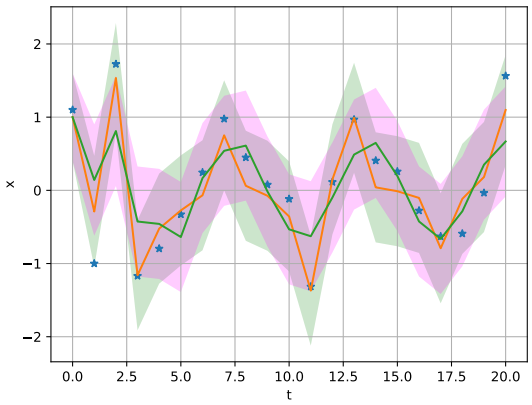


Figure 4.4: Illustration of the state evolution $x[k]$ (blue stars), Kalman-Filter-based estimate $\hat{\boldsymbol{x}}[k]$ (orange line), mean value $\bar{x}[k]$ (green line), $\hat{\boldsymbol{x}} \pm 3\sigma_{kf}$ interval (green shaded region), and $m \pm 3\sigma$ interval (magenta shaded region).

4.5 Joint State and Parameter estimation

In order to motivate the considerations in this section, consider the linear first order system with unknown parameter β

$$\dot{x}(t) = \lambda x(t) + \beta u(t), \quad x(0) = x_0 \tag{4.18}$$

$$y(t) = x(t) \tag{4.19}$$

with $\lambda, \beta \neq 0$. The parameter β can actually be obtained if y, u and $\dot{y}$ are known (if and only if $u(t) \neq 0$) from the equation

$$\beta = \frac{\dot{y}(t) - \lambda y(t)}{u(t)}.$$

Thus, β is *observable* or *identifiable* for all $u \neq 0$. Note that for $u = 0$ the *identifiability* of β is lost. For constant $u^* \neq 0$ the above calculation can be carried out in equilibrium conditions given that

On the other hand, knowing that β is constant, i.e. $\dot{\beta} = 0$ one can extend the system state and write an extended state-space model in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x \\ \beta \end{bmatrix} = \begin{bmatrix} \lambda x + \beta u \\ 0 \end{bmatrix} = \boldsymbol{f}(\boldsymbol{z}, u), \quad \boldsymbol{z} = \begin{bmatrix} x \\ \beta \end{bmatrix}$$

$$y = x = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ \beta \end{bmatrix} = \boldsymbol{c}^T \boldsymbol{z}$$

Note that this is a nonlinear system (due to the product $z_2 u$) Close to the equilibrium we can approximate the system dynamics using the linearization, i.e.

$$\tilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{z}^*, \quad \tilde{u} = u - u^*, \quad \tilde{y} = y - y^*, \quad \boldsymbol{z}^* = \begin{bmatrix} x^* \\ \beta \end{bmatrix}$$

$$\dot{\tilde{\boldsymbol{z}}} \approx \frac{\partial \boldsymbol{f}(\boldsymbol{z}^*)}{\partial \boldsymbol{z}} \tilde{\boldsymbol{z}} + \boldsymbol{b} u, \quad \tilde{\boldsymbol{z}}(0) = \tilde{\boldsymbol{z}}_0, \quad \boldsymbol{b} = \begin{bmatrix} \beta \\ 0 \end{bmatrix}$$

$$\tilde{y} = \boldsymbol{c}^T \tilde{\boldsymbol{z}}$$

Explicitely, for the case at hand this yields the system dynamics

$$\dot{\tilde{z}} = \begin{bmatrix} \lambda & u^* \\ 0 & 0 \end{bmatrix} z + \begin{bmatrix} \beta \\ 0 \end{bmatrix} \tilde{u}, \quad \tilde{z}(0) = \tilde{z}_0$$
$$\tilde{y} = \tilde{z}_1$$

with Kalman observability matrix $\mathcal{K}_O = \begin{bmatrix} 1 & 0 \\ \lambda & u^* \end{bmatrix}$ which has full rank as long as $u^* \neq 0$, i.e. the linearization of the extended system is completely observable as long as $u^* \neq 0$ (see discussion before!). Thus, for constant $u^* \neq 0$ the parameter β and the state x can be jointly reconstructed from the input-output data. This example shows the strict connection between identifiability and observability, while in this (nonlinear) context these are dependent on the input.

Given the complete observability, x and λ can be jointly determined using e.g. a Luenberger or reduced order observer, or a Kalman-Bucy or Kalman Filter.

Luenberger observer for joint state and parameter estimation

$$\dot{\hat{z}} = \frac{\partial \boldsymbol{f}(\boldsymbol{z}^*)}{\partial \boldsymbol{z}} \tilde{\boldsymbol{z}} + \boldsymbol{b}\tilde{u} - L(\boldsymbol{c}^T \hat{\boldsymbol{z}} - y), \quad \hat{\boldsymbol{z}}(0) = \hat{\boldsymbol{z}}_0.$$

In particular, for the parameter estimation part, this yields

$$\dot{\hat{\beta}} = -l_\beta(\hat{x} - y), \quad \hat{\beta}(0) = \hat{\beta}_0$$
$$\Rightarrow \quad \hat{\beta}(t) = \hat{\beta}_0 - l_\beta \int_0^t (\hat{x}(\tau) - y(\tau)) \mathrm{d}\tau,$$

i.e. the parameter estimate follows a simple integration scheme.

Considering a general system dynamics of the form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{p}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$
$$\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{p})$$

with $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{u} \in \mathbb{R}^q, \boldsymbol{p} \in \mathbb{R}^s, \boldsymbol{y} \in \mathbb{R}^m$ and equilibrium (or operation) point $(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{p}^*)$. Close to this equilibrium the dynamics are approximated by

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}^*, \quad \tilde{\boldsymbol{u}} = \boldsymbol{u} - \boldsymbol{u}^*, \quad \tilde{\boldsymbol{p}} = \boldsymbol{p} - \boldsymbol{p}^*$$
$$\dot{\tilde{\boldsymbol{x}}} = A\tilde{\boldsymbol{x}} + B\tilde{\boldsymbol{u}} + E\tilde{\boldsymbol{p}}, \quad \tilde{\boldsymbol{x}}(0) = \tilde{\boldsymbol{x}}_0$$
$$\tilde{\boldsymbol{y}} = C\tilde{\boldsymbol{x}} + D\tilde{\boldsymbol{p}}$$

with $\boldsymbol{p}^*$ being an initial estimate of the paramter vector $\boldsymbol{p}$ and

$$A = \frac{\partial \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{p}^*)}{\partial \boldsymbol{x}}, \quad B = \frac{\partial \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{p}^*)}{\partial \boldsymbol{u}}, \quad E = \frac{\partial \boldsymbol{f}(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{p}^*)}{\partial \boldsymbol{p}},$$
$$C = \frac{\partial \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{p}^*)}{\partial \boldsymbol{x}}, \quad D = \frac{\partial \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{p}^*)}{\partial \boldsymbol{p}}$$

Introducing the joint, extended state vector

$$\tilde{\boldsymbol{z}} = \begin{bmatrix} \tilde{\boldsymbol{x}}^\mathsf{T} & \tilde{\boldsymbol{p}}^\mathsf{T} \end{bmatrix}^\mathsf{T}$$

this is written as

$$\begin{aligned}\dot{\tilde{\boldsymbol{z}}} &= \begin{bmatrix} A & E \\ 0 & 0 \end{bmatrix} \tilde{\boldsymbol{z}} + \begin{bmatrix} B \\ 0 \end{bmatrix} \tilde{\boldsymbol{u}}, \quad \tilde{\boldsymbol{z}}(0) = \tilde{\boldsymbol{z}}_0 \\ \tilde{\boldsymbol{y}} &= \begin{bmatrix} C & D \end{bmatrix} \tilde{\boldsymbol{z}}\end{aligned}$$

The observability of the complete state $\tilde{\boldsymbol{z}}$ (and thus the identifiability of the parameters $\boldsymbol{p}$ in particular) is ensured if the Kalman observability $\mathscr{K}_O$ matrix has rank $n+s$. This matrix is given for this system by

$$\mathscr{K}_O = \begin{bmatrix} C & D \\ CA & CE \\ CA^2 & CAE \\ \vdots & \vdots \\ CA^{n-1} & CA^{n-1}E \end{bmatrix} \in \mathbb{R}^{nm \times (n+s)}.$$

According to the observer theory developed before, if $\operatorname{rank}(\mathscr{K}_O) < n+s$ then the states and parameter can be jointly determined from the input-output data if the joint, extended system is detectable.

Note that the detectability is parameter dependent, and thus at this point some parameters must be known (at least in sign).

In the presence of process (i.e. model) and measurement uncertainties in form of stochastic fluctuations which can be modeled as white noise processes, the model can be extended to

$$\begin{aligned}\dot{\tilde{\boldsymbol{z}}} &= \begin{bmatrix} A & E \\ 0 & 0 \end{bmatrix} \tilde{\boldsymbol{z}} + \begin{bmatrix} B \\ 0 \end{bmatrix} \tilde{\boldsymbol{u}} + \boldsymbol{w}, \quad \tilde{\boldsymbol{z}}(0) = \tilde{\boldsymbol{z}}_0 \\ \tilde{\boldsymbol{y}} &= \begin{bmatrix} C & D \end{bmatrix} \tilde{\boldsymbol{z}} + \boldsymbol{v}\end{aligned}$$

with $\boldsymbol{w} \sim \mathscr{N}(0, Q)$ and $\boldsymbol{v} \sim \mathscr{N}(0, R)$ being Gaussian processes.

For this system a minimum covariance state and parameter estimator is given by the Kalman-Bucy (or the Kalman) Filter. This approach in particular makes sense if the parameter is actually not constant but fluctuates around a constant value and the fluctuations can be reasonably approximated by white noise variables.

Example 4.5. *Consider again the system* (4.18)*. The constant β can be estimated using a Luenberger observer for the extended model*

$$\begin{aligned}\dot{z}_1 &= \lambda z_1 + z_2 u, & z_1(0) &= x_0 \\ \dot{z}_2 &= 0, & z_2(0) &= \beta \\ y &= z_1.\end{aligned}$$

Considering the linearization about an arbitrary equilibrium state-input pair (x^*, u^*) of (4.18) *with $u(t) = u^* \neq 0$ being a constant input this observer reads*

$$\begin{aligned}\dot{\hat{z}}_1 &= \lambda \hat{z}_1 + \hat{z}_2 u^* - l_1(\hat{z}_1 - y), & \hat{z}_1(0) &= \hat{z}_{10} \\ \dot{\hat{z}}_2 &= -l_2(\hat{z}_1 - y), & \hat{z}_2(0) &= \hat{z}_{20},\end{aligned}$$

or in compact form

$$\dot{\hat{\boldsymbol{z}}} = A\hat{\boldsymbol{z}} - \boldsymbol{l}(\boldsymbol{c}^\intercal \hat{\boldsymbol{z}} - y), \qquad \hat{\boldsymbol{z}}(0) = \hat{\boldsymbol{z}}_0$$

with

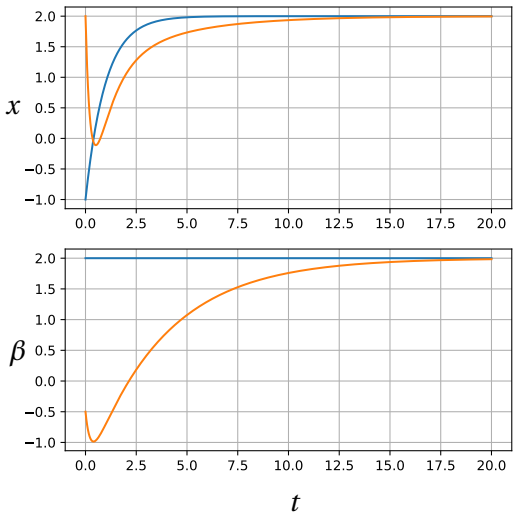
$$A = \begin{bmatrix} \lambda & u^* \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{c}^\intercal = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \boldsymbol{l} = \begin{bmatrix} l_1 \\ l_2 \end{bmatrix}.$$

The correction gain vector $\boldsymbol{l}$ can be assigned e.g. using the Ackerman formula

$$\boldsymbol{l} = T\tilde{\boldsymbol{l}}, \quad T = \begin{bmatrix} \hat{\boldsymbol{w}}, A\hat{\boldsymbol{w}} \end{bmatrix}, \quad \hat{\boldsymbol{w}} = \mathcal{K}_O^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \tilde{\boldsymbol{l}} = \begin{bmatrix} \bar{a}_0 - a_0 \\ \bar{a}_1 - a_1 \end{bmatrix}$$

with a_0, a_1 the coefficients of the characteristic polynomial of the matrix A and $\bar{a}_0, \bar{a}_1$ the respective desired coefficients of $A - \boldsymbol{l}\boldsymbol{c}^T$.

Choosing the eigenvalues of $A - \boldsymbol{l}\boldsymbol{c}^T$ as $\lambda_1 = -1, \lambda_2 = -2$, and the parameters $\lambda = -1, \beta = 2$ and using the unit step-input $u(t) = h(t)$, the following result is obtained with $x(0) = -1, \hat{x}(0) = 2, \hat{\beta}(0) = -0.5$



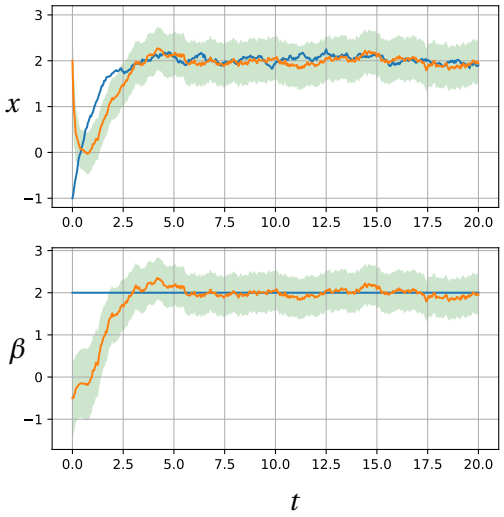
Example 4.6. *Considering again the motivating example* (4.18) *but with a fluctuating parameter $\beta = \bar{\beta} + w$ with $w \sim \mathcal{N}(0, q)$ and measurement $y = x + v$ with $v \sim \mathcal{N}(0, r)$ the best solution consists in designing a Kalman-Bucy Filter for the extended model*

$$\begin{aligned} \dot{\boldsymbol{z}} &= A\boldsymbol{z} + \boldsymbol{g}w, & \boldsymbol{z}(0) &\sim \mathcal{N}(\bar{\boldsymbol{z}}_0, \sigma^2), & w &\sim \mathcal{N}(0, q) \\ y &= \boldsymbol{c}^\intercal \boldsymbol{z} + v, & & & v &\sim \mathcal{N}(0, r), \end{aligned}$$

with $\boldsymbol{z} = [x, \bar{\beta}]^\intercal, \boldsymbol{g} = [u^*, 0]^\intercal$. According to (4.8) *the Kalman-Bucy Filter is then given*

$$\begin{aligned} \dot{\hat{\boldsymbol{z}}} &= A\hat{\boldsymbol{z}} - \boldsymbol{l}(\boldsymbol{c}^\intercal \hat{\boldsymbol{z}} - y), & \hat{\boldsymbol{z}}(0) &= \hat{\boldsymbol{z}}_0, \quad \boldsymbol{l} = P\boldsymbol{c}r^{-1} \\ \dot{P} &= AP + PA^\intercal + q^{-1}\boldsymbol{g}\boldsymbol{g}^\intercal - P\boldsymbol{c}r^{-1}\boldsymbol{c}^\intercal P, & P(0) &= P_0. \end{aligned}$$

Considering the values from the preceding example with $\lambda = -1, \beta = 2 + w, q = 0.5, r = 0.2, u(t) = h(t), x(0) = -1, \hat{\boldsymbol{z}} = [2, -0.5]^\intercal$ yields the following simulation result.


x
β
t
3
2
1
0
−1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

References

[AT67] M. Athans and E. Tse. „A direct derivation of the optimal linear filter using the maximum principle“. In: *IEEE Trans. Autom. Control* 12 (6) (1967), pp. 690–698 (cit. on pp. 60, 61).

[Gel78] A. Gelb. *Applied Optimal Estimation*. M.I.T. Press, Cambridge, 1978 (cit. on pp. 60, 64–66).

[Kal60] R. Kalman. „A New Approach to Linear Filtering and Prediction Problems“. In: *Transactions of the ASME–Journal of Basic Engineering* 82 (1960), pp. 35–45 (cit. on pp. 65, 66).

[KB61] R. E. Kalman and R. S. Bucy. „New results in linear filtering and prediction theory“. In: *Trans. ASME, J. Basic Engrg., ser. D* 83 (1961), pp. 95–108 (cit. on p. 60).

[Lof90] O. Loffeld. *Estimationstheorie II: Anwendungen – Kalman-Filter*. Oldenburg-Verlag, München, 1990 (cit. on p. 65).

[RW00] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes and Martingales*. Cambridge University Press, Cambridge, 2000 (cit. on p. 60).