Lecture Notes

“Mathematical Systems and Control Theory”

University of Hamburg
(winter term 2019/20)

Dr. Matthias Voigt
matthias.voigt@uni-hamburg.de

Preface

This document is based in large parts on the German lecture notes of Peter Benner who gave a similar course at the TU Chemnitz in winter term 2009/10. The usage of his LaTeX source code is highly appreciated. I believe that there are more errors and typos in this document, please send an email to

matthias.voigt@uni-hamburg.de

if you find any.

Many topics dicsussed in these notes can also be found in Chapters 3 and 4 of

> K. Zhou, J. C. Doyle, and K. Glover. Robust and Optimal Control, Prentice-Hall, Englewood Cliffs, NJ, 1996.

Further, more recent results discussed here will be cited throughout the lecture notes, so that you can read the original sources.

Contents

CHAPTER 1

Introduction

In this course we will consider dynamical systems that describe physical, technical, or economical processes. These should be manipulated with the help of input variables such that certain output variables show a certain desired behavior. Schematically, this is illustrated in Figure 1.1.

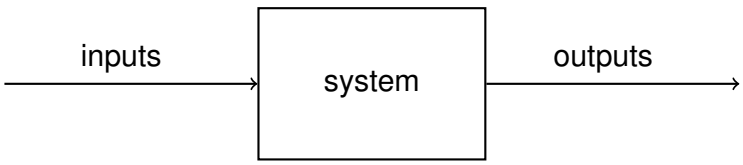


Figure 1.1: System description as black box.

Example: If the European central bank changes its base rate, then this influences developments at the German stock and financial markets. For example, the exchange rate between € and US-$ may rise or fall as well as share prices. Considering the stock market as a dynamical system, then the base rate can be viewed as an input to the system, whereas the German stock index DAX can be regarded as an output of the system.

Note that in this example, the system is not described by mathematical equations. Therefore, this system is a black box, since the internal variables (so-

called states) and how they are affected by the input is unknown. It is still possible to gain insight in the relationship between the inputs and outputs.

Here we will consider instationary (i. e. time-dependent) processes. Thus the inputs $u(\cdot)$ and outputs $y(\cdot)$ are functions of time. The internal variables of the system, called states, that are often not explicitly available, are denoted by $x(\cdot)$ and are a functions of time as well. We assume that the dynamic behavior of the system is described by (ordinary) differential equations of first order (Recall that higher-order systems can be reduced to first-order systems by a linearization.)

In general, the systems considered in this course can be described by the following definition.

Definition 1.1: A *(nonlinear) control system* (or *controlled system*) satisfies the following equations for (almost all) $t \in [t_0,\, t_\mathrm{f}]$, $t_0 < t_\mathrm{f} \leqslant \infty$:

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{(state equation)}, \tag{1.1}$$
$$x(t_0) = x^0 \in \mathcal{X} \quad \text{(initial condition)}, \tag{1.2}$$
$$y(t) = g(t, x(t), u(t)) \quad \text{(output equation)}. \tag{1.3}$$

Hereby,

$x : [t_0, t_f] \to \mathcal{X}$ is the *state (vector),*
$u : [t_0, t_f] \to \mathcal{U}$ is the *input* or *control (vector)*,
$y : [t_0, t_f] \to \mathcal{Y}$ is the *output (vector),*

and

$\mathcal{X} \subseteq \mathbb{R}^n$ is the *state space,*
$\mathcal{U} \subseteq \mathbb{R}^m$ is the *input space,*
$\mathcal{Y} \subseteq \mathbb{R}^p$ is the *output space.*

The number n is the *order* of the system (also *state-space dimension*, if $\mathcal{X} = \mathbb{R}^n$). The system is called *autonomous (time-invariant)*, if

$$f(t, x(t), u(t)) \equiv f(x(t), u(t)) \quad \text{and} \quad g(t, x(t), u(t)) \equiv g(x(t), u(t)),$$

i. e., for $u(t) \equiv 0$, $\dot{x}(t) = f(x(t))$ is an autonomous differential equation.

For the systems defined above, Figure 1.1 can be extended as in Figure 1.2. If one tries to model a physical, technical, or economical process by equations

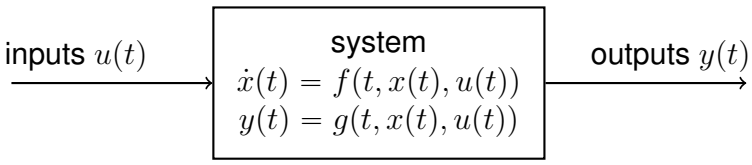


Figure 1.2: Nonlinear system as a black box.

of the form (1.1)–(1.3), the following aspects are of importance:

a) Which are the "free" input parameters (input/control variables)?

b) Which are the state variables?

c) Which variables can be measured or observed? (all/a few state variables or only derived quantities?)

d) What is the functional relationship?

e) Isa continuous-time modeling as in (1.1)–(1.3) appropriate or does one need a discrete-time model, i. e., a description of the dynamics by difference equations?

   Often mixed models are needed (so-called *hybrid systems*), since a few model variables may be described in continuous time, others only in discrete time.

f) Do the model variables behave in a deterministic or stochastic manner?

In the following we will mostly assume that we have continuous-time and deterministic model. Further, in this course we will only consider a simpler functional relationship between the variables in (1.1)–(1.3), namely we assume that f and g are affine linear functions.

Definition 1.2: A *linear control system* is given, if $\mathcal{X} = \mathbb{R}^n$, $\mathcal{U} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}^p$, and

$$f(t, x(t), u(t)) = A(t)x(t) + B(t)u(t),$$
$$g(t, x(t), u(t)) = C(t)x(t) + D(t)u(t),$$

where $A : [t_0, t_\mathrm{f}] \to \mathbb{R}^{n\times n}, B : [t_0, t_\mathrm{f}] \to \mathbb{R}^{n\times m}, C : [t_0, t_\mathrm{f}] \to \mathbb{R}^{p\times n}$, $D : [t_0, t_\mathrm{f}] \to \mathbb{R}^{p\times m}$ are sufficiently smooth matrix-valued functions.

For *autonomous systems* it holds that $A(t) \equiv A, B(t) \equiv B$, $C(t) \equiv$, and $D(t) \equiv D$. In this situation we talk about *linear time-invariant (LTI) systems*, if the systems fulfills the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(t_0) = x^0 \in \mathbb{R}^n, \tag{1.4}$$
$$y(t) = Cx(t) + Du(t). \tag{1.5}$$

A *linear time-varying (LTV) system* is given by

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x^0 \in \mathbb{R}^n, \tag{1.6}$$
$$y(t) = C(t)x(t) + D(t)u(t). \tag{1.7}$$

Analogously, for $k = 0, 1, 2, \ldots$ a *linear discrete-time system* is described by

$$x_{k+1} = A_k x_k + B_k u_k, \quad x_0 = x^0 \in \mathbb{R}^n, \tag{1.8}$$
$$y_k = C_k x_k + D_k u_k, \tag{1.9}$$

where $A_k \in \mathbb{R}^{n\times n}$, $B_k \in \mathbb{R}^{n\times m}$, $C_k \in \mathbb{R}^{p\times n}$, and $D_k \in \mathbb{R}^{p\times m}$. Again, in the time-invariant case it holds that $A_k \equiv A$, $B_k \equiv B$, $C_k \equiv C$, $D_k \equiv D$ etc.

Remark 1.3: For autonomous systems we can assume w. l. o. g. that $t_0 = 0$: If we move from x^0 to x^1 in the time interval $[t_0, t_1]$ with the control function $u(\cdot)$, then we could equivalently move from x^0 to x^1 in the time-interval $[0, t_1 - t_0]$, if we choose the control function $\widetilde{u}(t) := u(t + t_0)$ and consider the solution trajectory $\widetilde{x}(t) := x(t + t_0)$.

In the following we will assume that the control function $u(\cdot)$ lives in a function space U_{ad} of *admissible controls*. We assume further that $u(\cdot)$ is not subject to input constraints such as $a(\cdot) \leqslant u(\cdot) \leqslant b(\cdot)$ ("$\leqslant$" understood componentwise). This would lead to questions from linear or nonlinear optimization and is widely analysed in the optimal control of partial differential equations. This will not addressed in this course. Here the function space U_{ad} is the space of square-integrable functions on $[t_0, t_{\text{f}}]$ mapping to $\mathcal{U}$ denoted by $L_2([t_0, t_{\text{f}}]; \mathcal{U})$ or the space of piecewise continuous functions on $[t_0, t_{\text{f}}]$ mapping to $\mathcal{U}$ denoted by $PC([t_0, t_{\text{f}}]; \mathcal{U})$. Integration is understood in the Lebesgue sense.

To emphasize the dependence of the solution trajectories of the differential equations (1.1), (1.4), or (1.6) on the control function $u(\cdot)$, we write

$$x(t) = x(t; u),$$

where we assume that the solution of the corresponding initial value problem on the interval $[t_0,\, t_\mathrm{f}]$ exists for all $u \in U_\mathrm{ad}$ and is unique.

A central question in mathematical control theory is the following one:

> Given an initial condition x^0 and a target x^1, can we find a $\widehat{u} \in U_\mathrm{ad}$, such that for some $t_\mathrm{f} \geqslant t_1 \geqslant t_0$ it holds that $x(t_1; \widehat{u}) = x^1$?

A stronger question is the following one:

> Given an initial condition x^0, a target x^1 as well as $t_1 \leqslant t_f$, can we find a $\widehat{u} \in U_\mathrm{ad}$ such that $x(t_1; \widehat{u}) = x^1$?

Often the problem can be formulated in such a way that the target is $x^1 = 0$, so x can be interpreted as the deviation from some given reference trajectory. A weaker objective is then to find an *asymptotically stabilizing* control $\widehat{u} \in U_\mathrm{ad}$, that is, it holds that $\lim\limits_{t\to\infty} x(t; \widehat{u}) = 0$. Slightly modified, the question is whether in finite time one can enter an arbitrarily small neighborhood of zero.

Besides the existence of such control functions, the question of optimality plays an important role. For given $x_1 \in \mathcal{X}$ or given reference trajectory $x_\mathrm{ref}(\cdot)$ (e. g. $x_\mathrm{ref}(t) \equiv 0$, if the state describes the deviation from the reference trajectory), possible objective functionals are

$$\min_{u\in U_\mathrm{ad}} \left\{ t_1 \in [t_0,\, t_\mathrm{f}] \,\middle|\, x(t_1; u) = x^1 \right\}, \qquad \text{(\textit{time-optimal control})} \tag{1.10}$$

$$\min_{u\in U_\mathrm{ad}} \int_{t_0}^{t_\mathrm{f}} \|x(t) - x_\mathrm{ref}(t)\| \mathrm{d}t, \qquad \text{(\textit{minimum deviation control})} \tag{1.11}$$

$$\min_{u\in U_\mathrm{ad}, x(t;u)=x^1} \int_{t_0}^{t_\mathrm{f}} \|u(t)\|\, dt. \qquad \text{(\textit{energy minimizing control})} \tag{1.12}$$

Here $\|\cdot\|$ is an appropriate vector norm such as the Euclidean norm, but 1- and ∞-norms can be useful as well. Mixtures of these cost functionals appear quite often, in particular, we will have a closer look at combinations of (1.11) and (1.12), while (1.10) is subject of *optimal control theory*. In mathematical systems theory, the focus is often put on the input/output behavior, i. e., on $u(\cdot)$ and $y(\cdot)$. Therefore, the cost functionals above are often formulated in terms of $y(\cdot)$ instead of $x(\cdot)$, in particular, in the tracking problem, often $y_\mathrm{ref}(\cdot)$ is given instead of $x_\mathrm{ref}(\cdot)$.

Remark 1.4: Often in the literature the cost functional

$$\min_{u\in U_\mathrm{ad}} \int_{t_0}^{t_f} \|x(t) - x^1\| \mathrm{d}t \tag{1.13}$$

or respectively in combination with (1.12)

$$\min_{u\in U_{\text{ad}}} \int_{t_0}^{t_f} \|x(t) - x^1\| + \|u(t)\| \mathrm{d}t \tag{1.14}$$

are used. However, this often leads to trajectories of the state or control that are difficult to realize or that put high demands of the mechanics or electronics of the system since the control then tends to take only a high impact on the system at the end of the control interval. This can be improved by using energy-minimizing controls, but even better solutions are achieved by prescribing by using a reference trajectory $x_{\text{ref}}(\cdot)$ with $x_{\text{ref}}(t_f) = x^1$.

Here we will mostly deal with control functions $u(\cdot)$ that appear in the form of a feedback control. Thereby, the knowledge of the state or the output to steer the system to a desired state or to correct the deviation from the desired state. Hereby, we distinguish

- *state feedback:* $u(t) = u(t, x(t))$, in the linear case $u(t) = F(t)x(t)$ or $u(t) = Fx(t)$ in the time-invariant case with F, $F(t) \in \mathbb{R}^{m\times n}$;
- *output feedback:* $u(t) = u(t, y(t))$, in the linear case $y(t) = F(t)y(t)$ or $u(t) = Fy(t)$ in the time-invariant case with F, $F(t) \in \mathbb{R}^{m\times p}$.

The matrix F is galled *feedback matrix* or *gain* and which has to chosen appropriately in order to achieve the desired objectives. So one of the goals of this lecture is whether there exists such a feedback matrix and if yes, how it can be constructed.

In the linear case, plugging in the feedback into (1.6) (respectively, into (1.4)) leads to the following *closed-loop system*:

- for state feedback:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) = (A(t) + B(t)F(t))x(t).$$

- for output feedback:

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

The output feedback leads to a *closed-loop system* as in Figure 1.3.

The following example illustrates the questions and difficulties of mathematical control theory.

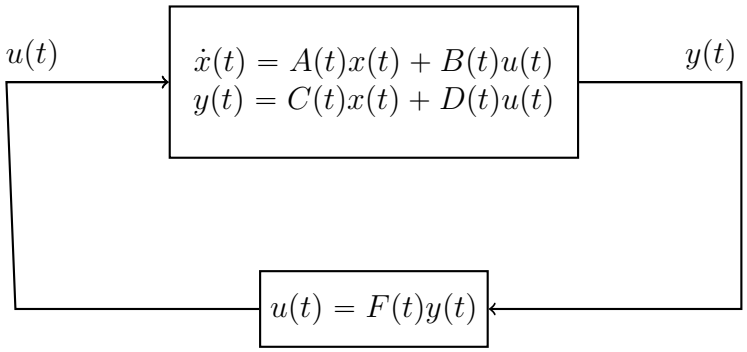


Figure 1.3: Closed-loop system.

Example: A major question in robotics is to control the position of a single–link rotational joint using a motor placed at the "pivot". As a mathematical we can use a pendulum to which we can apply a torque as external force $u(\cdot)$ to control the motion of the pendulum, see (1.4).

If we neglect friction and assume that the mass is concentrated at the tip of the pendulum, Newton's law for rotating objects yields

$$m\ddot{\theta}(t) + mg\sin\theta(t) = u(t)$$

describes the counterclockwise movement of the angle between the vertical axis and the pendulum subject to the control $u(\cdot)$. Scaling the variables to $m = 1$ and $g = 1$ (for simplicity), this is a first example of a nonlinear control system, if we set

$$x(t) := \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} := \begin{bmatrix} \theta(t) \\ \dot{\theta}(t) \end{bmatrix},$$
$$f(t, x(t), u(t)) := \begin{bmatrix} x_2(t) \\ -\sin(x_1(t)) + u(t) \end{bmatrix}, \quad g(t, x(t), u(t)) = x_1(t),$$

i. e., here we assume that only $\theta(t)$ can be measured but not the angular velocity $\dot{\theta}(t)$.

For $u(t) \equiv 0$, the stationary position $\theta = \pi$, $\dot{\theta} = 0$ is an unstable equilibrium, i. e., small perturbations will lead to an unstable motion. The objective now is to apply a torque (control u) to correct for deviations from this unstable equilibrium, so that the pendulum is kept in upright position.

Assuming small perturbations $\theta - \pi$ in the inverted pendulum problem, we have

$$\sin\theta = -(\theta - \pi) + o\big((\theta - \pi)^2\big).$$

(Here, $h(\theta) = o\big((\theta-\pi)^2\big)$ if $\lim_{\theta\to\pi} \frac{h(\theta)}{(\theta-\pi)^2} = 0$). This allows us to linearize the control system in order to obtain a linear control system for $\varphi(t) := \theta(t) - \pi$, namely

$$\ddot{\varphi}(t) - \varphi(t) = u(t).$$

This can be written as an LTI system, assuming only positions can be observed with

$$x(t) = \begin{bmatrix}\varphi(t)\\ \dot{\varphi}(t)\end{bmatrix}, \quad A = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}, \quad B = \begin{bmatrix}0\\ 1\end{bmatrix}, \quad C = \begin{bmatrix}1 & 0\end{bmatrix}, \quad D = 0.$$

Now the objective translates to: given initial values $x_1(0) = \varphi(0)$, $x_2(0) = \dot{\varphi}(0)$, find $u(\cdot)$ to bring $x(\cdot)$ to zero "as fast as possible". It is usually an additional goal to avoid overshoot and oscillating behavior as much as possible.

In the above example we have seen that that it is important to achieve that $\lim_{t\to\infty} x(t) = 0$ for all initial conditions $x(t_0) = x^0$ that are sufficiently close to the equilibrium $\bar{x} = 0$. Either this property is inherent in the system itself, then one does not have to do anything. Otherwise, as in the example above, we are interested in constructing a feedback such that the closed-loop system achieves this goal. Let us first define stability of an autonomous system.

Definition 1.5 (Stability of autonomous systems)**:** An equilibrium point $\bar{x}$ of the differential equation $\dot{x}(t) = f(x(t))$ (i. e., satisfying $f(\bar{x}) = 0$) is called

a) *stable*, if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|x(t_0) - \bar{x}\| < \delta \quad \Rightarrow \quad \|x(t) - \bar{x}\| < \varepsilon \quad \forall\, t \geqslant t_0.$$

b) *asymptotically stable*, if it is stable and if in addition, δ can be chosen such that

$$\|x(t_0) - \bar{x}\| < \delta \quad \Rightarrow \quad \lim_{t\to\infty} \|x(t) - \bar{x}\| = 0.$$

The importance of asymptotic stability is evident if $x(\cdot)$ is the deviation from a nominal path $r(\cdot)$, e. g. in Example 1 this deviation is

$$x(t) = \begin{bmatrix}\theta(t) - \pi\\ \dot{\theta}(t) - 0\end{bmatrix}.$$

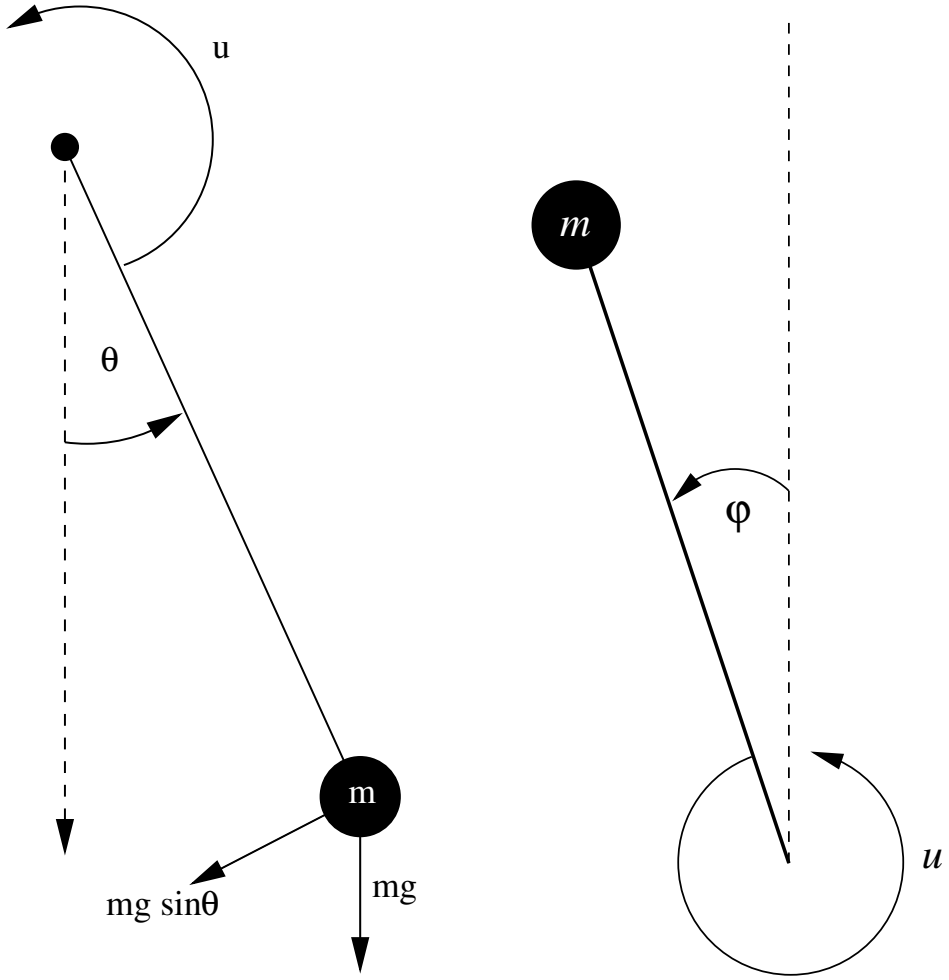
For nonlinear systems, (asymptotic) stability is not easy to check in general.

However, for LTI systems, one normally considers the zero equilibrium and its stability can be checked as follows. Note this in this situation one usually talks about stability of the system instead of stability of its zero equilibrium.

Proposition 1.6 (Stability of linear systems)**:** Let $\Lambda(A)$ denote the spectrum of $A \in \mathbb{R}^{n\times n}$. The linear time-invariant differential equation $\dot{x}(t) = Ax(t)$, $x(t_0) = x^0$ is

a) asymptotically stable $\Leftrightarrow$ $\Lambda(A) \subset \mathbb{C}^-$;

b) stable $\Leftrightarrow$ $\Lambda(A) \subset \overline{\mathbb{C}^-}$ and all imaginary eigenvalues of A are not defective.



(a) Pendulum as a mathematical model of a rotational joint

(b) Inverted pendulum: control to upright position

Figure 1.4: Example: Pendulum and inverted pendulum

CHAPTER 2

Analysis of Control Systems

2.1 Controllability

First, we ask the question whether, for a given $x^0 \in \mathbb{R}^n$, it is possible to reach a given target $x^1 \in \mathbb{R}^n$ with the help of a control function $u \in U_{\text{ad}}$. Since controllability is only affected by the state equation (1.1), we will ignore the output equation for now.

Definition 2.1 (Controllability)**:** Let $x^1 \in \mathcal{X} \subseteq \mathbb{R}^n$ be given.

a) The control system (1.1) with initial condition $x(t_0) = x^0 \in \mathcal{X} \subseteq \mathbb{R}^n$ is *controllable to* x^1 *in time* $t_1 > t_0$, if there exists a $u \in U_{\text{ad}}$ such that $x(t_1; u) = x^1$. Then the pair (t_1, x^1) is called *controllable from* (t_0, x^0).

b) The control system (1.1) with initial condition $x(t_0) = x^0 \in \mathcal{X} \subseteq \mathbb{R}^n$ is *controllable to* x^1 if there exists a $t_1 > t_0$ such that (t_1, x^1) is reachable from (t_0, x^0).

c) If for all $x^0 \in \mathcal{X}$, (t_0, x^0) is controllable to x^1 for all $x^1 \in \mathcal{X}$, then the control system (1.1) is called *(completely) controllable.*

d) The *controllability set with respect to* x^1 is defined by

$$\mathcal{C}(x^1, t_0) := \bigcup_{t_1 > t_0} \mathcal{C}(x^1, t_0, t_1),$$

where $\mathcal{C}(x^1, t_0, t_1) := \{x^0 \in \mathcal{X} \mid \exists u \in U_{\text{ad}} \text{ with } x(t_0; u) = x^0,\ x(t_1; u) = x^1\}$.

Analogously one defines the *reachability set with respect to* x^0, namely

$$\mathcal{R}(x^0, t_0, t_1) := \{x^1 \in \mathcal{X} \mid \exists u \in U_{\text{ad}} \text{ with } x(t_0; u) = x^0,\ x(t_1; u) = x^1\}$$

and

$$\mathcal{R}(x^0, t_0) := \bigcup_{t_1 > t_0} \mathcal{R}(x^0, t_0, t_1).$$

Thus, the controllability set contains all initial states that can be controlled to x^1, whereas the reachability set contains all states that can be controlled to from a given x^0.

In the following we will restrict ourselves to linear systems. We will see that for LTI systems, all controllability concepts will coincide and that $\mathcal{C} := \mathcal{C}(x^1, 0) = \mathbb{R}^n$ (for arbitrary x^1) is equivalent to controllability. Therefore we first need the solutions of the initial value problems (1.4) und (1.6). Here we are particularly interested in the *input-to-state mapping*

$$\mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times U_{\text{ad}} \to \mathbb{R}^n, \quad (t_0, x^0, t, u) \mapsto x(t),$$

which is given by following standard result form the theory of ordinary differential equations.

Theorem 2.2: a) Let Φ be the fundamental solution of $\dot{x}(t) = A(t)x(t)$, i. e. the solution of the homogeneous linear matrix differential equation

$$\frac{\partial}{\partial t}\Phi(t, s) = A(t)\Phi(t, s), \quad \Phi(s, s) = I_n. \tag{2.1}$$

Then for the unique solution of the differential equation (1.6) it holds that

$$x(t) = \Phi(t, t_0)x^0 + \int_{t_0}^{t} \Phi(t, s)B(s)u(s)\mathrm{d}s. \tag{2.2}$$

b) The unique solution of (1.4) satisfies $\Phi(t, s) = \mathrm{e}^{A(t-s)}$ and therefore,

$$x(t) = \mathrm{e}^{At}x^0 + \int_0^t \mathrm{e}^{A(t-s)}Bu(s)\mathrm{d}s = \mathrm{e}^{At}\left(x^0 + \int_0^t \mathrm{e}^{-As}Bu(s)\mathrm{d}s\right). \tag{2.3}$$

Proof. Exercise. □

As a consequence of Theorem 2.2 we obtain directly a representation of the *input-to-output mapping*

$$\mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times U_{\text{ad}} \to \mathbb{R}^p, \quad (t_0, x^0, t, u) \mapsto y(t).$$

Corollary 2.3: a) The unique solution of (1.7) satisfies

$$y(t) = C(t)\Phi(t, t_0)x^0 + C(t)\int_{t_0}^{t} \Phi(t, s)B(s)u(s)\mathrm{d}s + D(t)u(t). \tag{2.4}$$

b) The unique solution of (1.5) is given by

$$y(t) = C\mathrm{e}^{At}x^0 + \int_0^t C\mathrm{e}^{A(t-s)}Bu(s)\mathrm{d}s + Du(t) \tag{2.5}$$

$$= C\mathrm{e}^{At}\left(x^0 + \int_0^t \mathrm{e}^{-As}Bu(s)\mathrm{d}s\right) + Du(t). \tag{2.6}$$

In the following we will use that the fundamental solution Φ fulfills the *semi-group property*

$$\Phi(t, t) = I_n, \quad \Phi(t, s) = \Phi(t, \tau)\Phi(\tau, s) \tag{2.7}$$

for all $t,\ s,\ \tau \in \mathbb{R}$. Moreover, $\Phi(t, s)$ is invertible for all $t,\ s \in \mathbb{R}$ and it holds that

$$\Phi(t, s)^{-1} = \Phi(s, t). \tag{2.8}$$

To simplify the following considerations, from now on we assume that $U_{\text{ad}} = PC([t_0, \infty); \mathbb{R}^m)$. Analogously, the choice $U_{\text{ad}} = L_2([t_0, \infty); \mathbb{R}^m)$ would be feasible.

First we consider the target $x^1 = 0$ and the corresponding controllability sets with respect to $x^1 = 0$. This will not be a restriction in the case of linear systems.

Lemma 2.4: It holds that $x^0 \in \mathcal{C}(0, t_0, t_1)$ if and only if there exists a $u \in U_{\text{ad}}$ with

$$x^0 = -\int_{t_0}^{t_1} \Phi(t_0, s)B(s)u(s)\mathrm{d}s.$$

Proof. According to Theorem 2.2, $x^0 \in \mathcal{C}_0(t_0, t_1)$ is equivalent to

$$0 = x(t_1) = \Phi(t_1, t_0)x^0 + \int_{t_0}^{t_1} \Phi(t_1, s)B(s)u(s)\mathrm{d}s \tag{2.9}$$

$$= \Phi(t_1, t_0)\left(x^0 + \int_{t_0}^{t_1} \Phi(t_0, s)B(s)u(s)\mathrm{d}s\right), \tag{2.10}$$

where we have used the semi-group property of Φ. With the invertibility of Φ it follows that

$$0 = x^0 + \int_{t_0}^{t_1} \Phi(t_0, s)B(s)u(s)\mathrm{d}s$$

for a $u \in U_{\text{ad}}$ and hence the claim. □

The term of the Gramian (matrix) will play an important role in this course. First we give a definition.

Definition 2.5: For $G \in PC((-\infty, \infty); \mathbb{R}^{n\times m})$, the matrix

$$P(t_0, t_1) = \int_{t_0}^{t_1} G(t)G(t)^{\mathsf{T}}\mathrm{d}t$$

is called the (t_0, t_1)-*Gramian (matrix)* of G.

Obviously, the Gramian is positive semi-definite. Further properties are given in the following lemmas.

Lemma 2.6: It holds that

$$\ker P(t_0, t_1) = \left\{x \in \mathbb{R}^n \mid G(t)^{\mathsf{T}}x \equiv 0 \text{ on } [t_0, t_1]\right\}.$$

Proof. For an arbitrary $x \in \mathbb{R}^n$ it holds that

$$x^{\mathsf{T}}P(t_0, t_1)x = x^{\mathsf{T}}\int_{t_0}^{t_1} G(t)G(t)^{\mathsf{T}}\mathrm{d}t\, x = \int_{t_0}^{t_1} \underbrace{(G(t)^{\mathsf{T}}x)^{\mathsf{T}}(G(t)^{\mathsf{T}}x)}_{\geqslant 0\ \forall t}\mathrm{d}t \geqslant 0$$

Then $P(t_0, t_1)x = 0$, if and only if $G(t)^\mathsf{T} x \equiv 0$ on $[t_0, t_1]$. □

Lemma 2.7: Let G be as in Definition 2.5. Then the following statements are equivalent:

a) There exists a $u \in U_{\text{ad}}$ such that $x = \int_{t_0}^{t_1} G(t)u(t)\mathrm{d}t$.

b) It holds that $x \in \operatorname{im} P(t_0, t_1)$, i. e., there exists a $z \in \mathbb{R}^n$ with $x = P(t_0, t_1)z$.

Proof. First define

$$\mathcal{L} := \left\{ x \in \mathbb{R}^n \;\middle|\; \exists u \in U_{\text{ad}} \text{ with } x = \int_{t_0}^{t_1} G(t)u(t)\mathrm{d}t \right\}.$$

Because of the linearity of the integral and the vector space properties of U_{ad}, $\mathcal{L}$ is itself a subspace of $\mathbb{R}^n$, in particular, it is a vector space.

So we have to show that $\mathcal{L} = \operatorname{im} P(t_0, t_1)$. It is clear that $\operatorname{im} P(t_0, t_1) \subseteq \mathcal{L}$. (Simply set $u(t) = G(t)^\mathsf{T} z$ for $x = P(t_0, t_1)z$.)

Now let $x \in \mathcal{L} \cap \ker P(t_0, t_1)$. Then because of $x \in \mathcal{L}$ and Lemma 2.6

$$x^\mathsf{T} x = \int_{t_0}^{t} \underbrace{x^\mathsf{T} G(t)}_{\substack{\equiv 0, \text{ since} \\ x \in \ker P(t_0, t_1)}} u(t)\mathrm{d}t = 0$$

which results directly in $x = 0$. Therefore, one obtains $\dim \mathcal{L} \cap \ker P(t_0, t_1) = \{0\}$ and with the help of the dimension formula

$$\begin{aligned} n &\geqslant \dim(\mathcal{L} + \ker P(t_0, t_1)) = \dim(\mathcal{L}) + \dim(\ker P(t_0, t_1)) \\ &\geqslant \dim(\operatorname{im} P(t_0, t_1)) + \dim(\ker P(t_0, t_1)) = n. \end{aligned}$$

Overall, we get $\dim \mathcal{L} = \dim(\operatorname{im} P(t_0, t_1))$, hence $\mathcal{L} = \operatorname{im} P(t_0, t_1)$. □

If we set $G(t) = \Phi(t_0, t)B(t)$, then

$$P(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_0, t)B(t)B(t)^\mathsf{T}\Phi(t_0, t)^\mathsf{T}\mathrm{d}t \tag{2.11}$$

is called the (t_0, t_1)-*controllability Gramian* of the linear system (1.6)–(1.7). With this, one obtains the first characterization of the controllability set.

Theorem 2.8: Let $x^1 = 0$ and consider the LTV system (1.6)–(1.7) with $P(t_0, t_1)$ as in (2.11). Then the following assertions are satisfied:

a) $\mathcal{C}(0, t_0, t_1) = \operatorname{im} P(t_0, t_1)$;

b) $P(t_0, t_1)x = 0 \quad \Leftrightarrow \quad x^{\mathsf{T}}\Phi(t_0, t)B(t) \equiv 0$ on $[t_0, t_1]$.

Proof. a) Use Lemma 2.4 and Lemma 2.7.

b) Use Lemma 2.6.

□

A further very useful characterization of complete controllability of LTV systems is obtained by a more detailed analysis of the properties of the (t_0, t_1)-controllability Gramian $P(t_0, t_1)$. First recall the following property of the *adjoint equation* of $\dot{x}(t) = A(t)x(t)$ known from the theory of differential equations. This adjoint equation is given by

$$\dot{z}(t) = -A(t)^{\mathsf{T}} z(t). \tag{2.12}$$

If $\Phi(\cdot, \cdot)$ is the fundamental solution of $\dot{x}(t) = A(t)x(t)$, i. e., solution of the linear homogeneous matrix differential equation, (2.1), then

$$\Phi(t, s)^{-\mathsf{T}} = \Phi(s, t)^{\mathsf{T}}$$

is the fundamental solution of (2.12). [1] In particular, every solution of the initial value problem of (2.12) with $z(t_0) = z_0$ can be written as

$$z(t) = \Phi(t_0, t)^{\mathsf{T}} z_0. \tag{2.13}$$

Theorem 2.9: The following statements are equivalent:

a) The LTV system (1.6) is completely controllable.

b) Every solution of the adjoint equation (2.12) has the property

$$z(t)^{\mathsf{T}} B(t) \equiv 0 \text{ on } [t_0, \infty) \text{ for a } t_0 \in \mathbb{R} \quad \Rightarrow \quad z(t) \equiv 0. \tag{2.14}$$

c) For all $t_0 \in \mathbb{R}$, there exists a $t_1 \in \mathbb{R}$ such that $P(t_0, t_1)$ is positive definite.

[1] **Proof.** Let Ψ be the fundamental solution of (2.12), i. e. $\frac{\partial}{\partial t}\Psi(t, s) = -A(t)^{\mathsf{T}}\Psi(t, s)$, $\Psi(s, s) = I_n$. Then $\left(\frac{\partial}{\partial t}\Psi(t, s)^{\mathsf{T}}\right)\Phi(t, s) = -\Psi(t, s)^{\mathsf{T}}A(t)\Phi(t, s) = -\Psi(t, s)^{\mathsf{T}}\left(\frac{\partial}{\partial t}\Phi(t, s)\right)$, so $0 = \left(\frac{\partial}{\partial t}\Psi(t, s)^{\mathsf{T}}\right)\Phi(t, s) + \Psi(t, s)^{\mathsf{T}}\left(\frac{\partial}{\partial t}\Phi(t, s)\right) = \frac{\partial}{\partial t}\Psi(t, s)^{\mathsf{T}}\Phi(t, s)$. Hence, $\Psi(t, s)^{\mathsf{T}}\Phi(t, s)$ is constant and because of the initial condition it holds that $\Psi(t, s)^{\mathsf{T}}\Phi(t, s) = I_n$.

Proof. The proof follows by a ring closure argument.

a) ⇒ b): Assume that there exists a nontrivial solution of (2.12) with $z(t)^\mathsf{T} B(t) \equiv 0$ on $[t_0, \infty)$ for some $t_0 \in \mathbb{R}$, but $z(\widehat{t}) \neq 0$ for at least one $\widehat{t} \in \mathbb{R}$. Then $z(t_0) \neq 0$, since with (2.13) it holds that $z(\widehat{t}) = \Phi(t_0, \widehat{t})^\mathsf{T} z(t_0)$ and $\Phi(t_0, \widehat{t})$ is invertible.

Now we choose $x^0 \in \mathbb{R}^n$ such that $(x^0)^\mathsf{T} z(t_0) \neq 0$. Since (1.6) is completely controllable, by Definition 2.1 there exist a $t_1 > t_0$ and $u \in U_{\text{ad}}$ such that $x(t_1) \equiv x(t_1; u) = 0$ solves (1.6) with initial condition $x(t_0) = x^0$. With this it follows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big(x(t)^\mathsf{T} z(t)\big) &= \dot{x}(t)^\mathsf{T} z(t) + x(t)^\mathsf{T} \dot{z}(t) \\ &= x(t)^\mathsf{T} A(t)^\mathsf{T} z(t) + u(t)^\mathsf{T} \underbrace{B(t)^\mathsf{T} z(t)}_{\equiv 0 \text{ on } [t_0,\infty)} - x(t)^\mathsf{T} A(t)^\mathsf{T} z(t) \\ &= 0.\end{aligned}$$

Therefore, $x(t)^\mathsf{T} z(t)$ is constant and due to the initial conditions it holds that

$$x(t_1)^\mathsf{T} z(t_1) = x(t_0)^\mathsf{T} z(t_0) = (x^0)^\mathsf{T} z(t) \neq 0,$$

which is a contradiction to $x(t_1) = 0$.

b) ⇒ c): This step is proven in two parts. First we show the following statement:

> For all $t_0 \in \mathbb{R}$ there exists a $t_1 \in \mathbb{R}$ such that every nontrivial solution of the adjoint equation (2.12) has the property
>
> $$z(t)^\mathsf{T} B(t) \not\equiv 0 \text{ on } [t_0, t_1]. \tag{2.15}$$

Assume that this is not the case. This would mean that there exists a sequence $(t_k)_{k=1}^\infty$ with $t_k \to \infty$ for $k \to \infty$ and a sequence of solutions $(z_k(\cdot))_{k=1}^\infty$ of (2.12) with initial conditions $\|z_k(t_0)\| = 1$ for $k = 1, 2, \ldots$ such that

$$z_k(t)^\mathsf{T} B(t) \equiv 0 \text{ on } [t_0, t_k]. \tag{2.16}$$

We assume w. l. o. g. that $(z_k(t_0))_{k=1}^\infty$ is converging (otherwise, we could find a converging subsequence, since $\{z \in \mathbb{R}^n \mid \|z\| = 1\}$ is compact).

Let now $\widehat{z}_0 := \lim_{k\to\infty} z_k(t_0)$ and $\widehat{z}(\cdot)$ be solution of (2.12) with the initial condition $\widehat{z}(t_0) = \widehat{z}_0$. Then $\widehat{z}(t) \not\equiv 0$ (since $\|\widehat{z}(t_0)\| = 1$). Due to (2.14) it holds that $\widehat{z}(t)^\mathsf{T} B(t) \not\equiv 0$ on $[t_0, \infty)$. Thus there exists a $\widehat{t} > t_0$ with

$\widehat{z}(\widehat{t})^{\mathsf{T}} B(\widehat{t}) \neq 0$. Since the solution of (2.12) depends continuously on the initial condition, $(z_k(\cdot))_{k=1}^{\infty}$ converges uniformly to $\widehat{z}(\cdot)$. But then it holds that $z_k(\widehat{t})^{\mathsf{T}} B(\widehat{t}) \neq 0$ for sufficiently large k. But for $t_k \to \infty$ this leads to a contradiction to (2.16).

In the second step we show:

> If any nontrivial solution of (2.12) satisfies the property (2.15), then $P(t_0, t_1) > 0$.

Since we already have $P(t_0, t_1) \geqslant 0$, it remains to show that $\ker P(t_0, t_1) = \{0\}$. According to Theorem 2.8, $z_0 \in \ker P(t_0, t_1)$ is equivalent to

$$z_0^{\mathsf{T}} \Phi(t_0, t) B(t) \equiv 0 \text{ on } [t_0, t_1].$$

Because of (2.13), $z(t) = \Phi(t_0, t)^{\mathsf{T}} z_0$ is the solution of (2.12) with $z(t_0) = z_0$. Therefore, it holds that $z(t)^{\mathsf{T}} B(t) \equiv 0$ on $[t_0, t_1]$ and due to the first step, $z(t) \equiv 0$ on $[t_0, t_1]$. This implies $z_0 = 0$, hence $\ker P(t_0, t_1) = \{0\}$ is shown.

c) $\Rightarrow$ a): Let t_1 be chosen such that $P(t_0, t_1) > 0$. Then every pair (t_0, x^0) can be controlled to an arbitrary $x^1 \in \mathbb{R}^n$ in time t_1 with

$$u(t) := B(t)^{\mathsf{T}} \Phi(t_0, t)^{\mathsf{T}} v,$$

where $v \in \mathbb{R}^n$ can be determined as the solution of the equation

$$\begin{aligned} x^0 = x(t_0) &= \Phi(t_0, t_1) x^1 + \int_{t_1}^{t_0} \Phi(t_0, s) B(s) B(s)^{\mathsf{T}} \Phi(t_0, s)^{\mathsf{T}} v \mathrm{d}s \\ &= \Phi(t_0, t_1) x^1 - P(t_0, t_1) v. \end{aligned}$$

This equation follows from (2.2) by interchanging initial and final time and it has a unique solution due to the positive definiteness of $P(t_0, t_1)$.

□

For LTI systems where we can can w. l. o. g. assume $t_0 = 0$, we can say more. First, from the explicit formula of the fundamental solution of $\dot{x}(t) = Ax(t)$, it follows that the $(0, t_1)$-controllability Gramian can be written as

$$P(0, t_1) = \int_0^{t_1} \mathrm{e}^{-At} B B^{\mathsf{T}} \mathrm{e}^{-A^{\mathsf{T}} t} \mathrm{d}t.$$

Then it follows directly that $x \in \ker P(0, t_1)$ if and only if

$$B^{\mathsf{T}} \mathrm{e}^{-A^{\mathsf{T}} t} x \equiv 0 \text{ in } [0, t_1]. \tag{2.17}$$

In the following we will characterize controllability in terms of the properties of the following matrix.

Definition 2.10 (controllability matrix)**:** The *controllability matrix* of an LTI system is

$$\mathcal{K}(A,B) := \begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix} \in \mathbb{R}^{n\times nm}.$$

With this we obtain a characterization of the controllability set $\mathcal{C}(0,0,t)$ for LTI systems.

Theorem 2.11: For an LTI system (1.4) it holds that $\mathcal{C}(0,0,t) = \operatorname{im}\mathcal{K}(A,B)$ for all $t > 0$.

Proof. We show the statement indirectly by proving $\mathcal{C}(0,0,t)^\perp = (\operatorname{im}\mathcal{K}(A,B))^\perp$ for all $t > 0$.

From Theorem 2.8 a) it follows with $P(0,t) = P(0,t)^\mathsf{T} \geqslant 0$ that

$$\mathcal{C}(0,0,t)^\perp = (\operatorname{im} P(0,t))^\perp = \ker P(0,t).$$

Therefore, it remains to show that

$$\ker P(0,t) = (\operatorname{im}\mathcal{K}(A,B))^\perp = \ker \mathcal{K}(A,B)^\mathsf{T}, \tag{2.18}$$

or, in other words, $P(0,t)x = 0$ if and only if $x^\mathsf{T}\mathcal{K}(A,B) = 0$. From Theorem 2.8 b) resp. (2.17) we already know a property of the elements of the kernel of $P(0,t)$ which we want to use now. First we do some preliminary considerations. Let $\phi_A(x) = \sum_{j=0}^{n} \alpha_j x^j$ be the characteristic polynomial of A. Then the Theorem of Cayley-Hamilton states that $\phi_A(A) = 0$. Then because of $\alpha_n = 1$ with $\beta_j = -\alpha_j$ we get

$$A^n = \sum_{j=0}^{n-1} \beta_j A^j. \tag{2.19}$$

Thus it holds that $x^\mathsf{T}A^nB = \sum_{j=0}^{n-1} \beta_j x^\mathsf{T}A^jB$. By a repeated application of (2.19) as well as by summarizing all coefficients of $x^\mathsf{T}A^jB$ in $\beta_j^{(\nu)}$ we obtain the representation

$$x^\mathsf{T}A^{n+\nu}B = \sum_{j=0}^{n-1} \beta_j^{(\nu)} x^\mathsf{T}A^jB \quad \forall \nu \in \mathbb{N}_0. \tag{2.20}$$

From this we get the following chain of equivalences:

$$\begin{aligned} x^{\mathsf{T}}\mathcal{K}(A,B) = 0 \;&\Leftrightarrow\; x^{\mathsf{T}}A^jB = 0, \quad j = 0,\, 1,\, \ldots,\, n-1 \\ &\overset{(2.20)}{\Leftrightarrow}\; x^{\mathsf{T}}A^\nu B = 0 \quad \forall \nu \in \mathbb{N}_0 \\ &\Leftrightarrow\; 0 = \sum_{j=0}^{\infty} \frac{(-\tau)^j}{j!} x^{\mathsf{T}}A^jB = x^{\mathsf{T}}\mathrm{e}^{-A\tau}B \quad \forall \tau \in [0,t] \\ &\overset{(2.17)}{\Leftrightarrow}\; P(0,t)x = 0. \end{aligned}$$

So (2.18) follows and hence, the statement of the theorem. □

Theorem 2.11 shows that for a controllable LTI system it holds that $\mathcal{C}(0,0,t_1) = \mathcal{C}(0,0,t_2)$ for all $t_1,\, t_2 > 0$, in particular, that $\mathcal{C}(0,0) \equiv \mathcal{C}(0,0,t)$ for all $t > 0$ and therefore, all controllability concepts for LTI systems coincide. For simplicity, we will now write $\mathcal{C}$ instead of $\mathcal{C}(0,0)$. Further, from Theorem 2.8, it follows that $\operatorname{im}\mathcal{K}(A,B) \equiv \operatorname{im}P(0,t)$. So obviously, Theorem 2.9 implies that for a controllable LTI system it holds that $P(0,t) > 0$ for all $t > 0$. A very useful characterization of controllability for LTI systems is given by the so-called *Hautus-Popov test*.

Theorem 2.12 (Hautus-Popov lemma)**:** Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then the following statements are equivalent:

a) The pair (A,B) is controllable.

b) It holds that $\operatorname{rank}\mathcal{K}(A,B) = n$.

c) If $v \in \mathbb{C}^n\backslash\{0\}$ is a left eigenvector of A, then $v^{\mathsf{H}}B \neq 0$.

d) It holds that $\operatorname{rank}\begin{bmatrix} A-\lambda I & B \end{bmatrix} = n$ for all $\lambda \in \mathbb{C}$.

Proof. **a) ⇔ b):** This follows directly from Theorem 2.11 and the equivalence of controllability and $\mathcal{C} = \mathbb{R}^n$.

c) ⇔ d): The condition $v^{\mathsf{H}}\begin{bmatrix} A-\lambda I & B \end{bmatrix} = 0$ is true if and only if $v^{\mathsf{H}}A = \lambda v^{\mathsf{H}}$ and $v^{\mathsf{H}}B = 0$. So $\operatorname{rank}\begin{bmatrix} A-\lambda I & B \end{bmatrix} < n$, if and only if there exists a left eigenvector $v \in \mathbb{C}\backslash\{0\}$ of A that satisfies $v^{\mathsf{H}}B = 0$.

b) ⇒ d) Assume that $\operatorname{rank}\begin{bmatrix} A-\lambda I & B \end{bmatrix} < n$. Then there exists a $v \neq 0$ with $v^{\mathsf{H}}\begin{bmatrix} A-\lambda I & B \end{bmatrix}$, i. e. $v^{\mathsf{H}}A = \lambda v^{\mathsf{H}}$, $v^{\mathsf{H}}B = 0$. Then we get

$$v^{\mathsf{H}}A^jB = \lambda^j v^{\mathsf{H}}B = 0 \quad \text{for all } j \in \mathbb{N}_0.$$

This implies $v^{\mathsf{H}}\mathcal{K}(A,B) = 0$ which is a contradiction to b).

d) ⇒ b): Assume that it holds that $\operatorname{rank}\mathcal{K}(A,B) = r < n$. Then there exists an orthonormal basis $\{v_1, \ldots, v_r\}$ of $\mathcal{K} := \operatorname{im}\mathcal{K}(A,B)$. We extend this basis to an orthonormal basis of $\mathbb{R}^n$ by $\{v_{r+1}, \ldots, v_n\}$. This implies

$$(\operatorname{im}\mathcal{K}(A,B))^{\perp} = \operatorname{span}\{v_{r+1}, \ldots, v_n\}.$$

Define $V := \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix} \in \mathbb{R}^{n\times n}$. Since the columns of V are orthonormal, it holds that $VV^{\mathsf{T}} = I_n = V^{\mathsf{T}}V$. Moreover, $v_{r+j}^{\mathsf{T}}\mathcal{K}(A,B) = 0$ for $j = 1, \ldots, n-r$, in particular, it holds that $v_{r+j}^{\mathsf{T}}B = 0$ for $j = 1, \ldots, n-r$, i. e., $V^{\mathsf{T}}B = \left[\begin{smallmatrix} B_1 \\ 0 \end{smallmatrix}\right]$.

With the Theorem of Cayley-Hamilton (see the proof of Theorem 2.11 and (2.19)) it follows, that $A\mathcal{K} \subseteq \mathcal{K}$, i. e., $\mathcal{K}$ is an A-invariant subspace of $\mathbb{R}^n$. Since the columns of $V_1 := \begin{bmatrix} v_1 & \ldots & v_r \end{bmatrix}$ form a basis for this A-invariant subspace, there exists a $A_1 \in \mathbb{R}^{r\times r}$ with $\Lambda(A_{11}) \subseteq \Lambda(A)$ and $AV_1 = V_1A_{11}$. This implies

$$AV = V\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Now let $\widetilde{v} \neq 0$ be a left eigenvector A_{22}, i. e., $\widetilde{v}^{\mathsf{H}}A_{22} = \lambda\widetilde{v}^{\mathsf{H}}$ for some $\lambda \in \Lambda(A_{22}) \subseteq \Lambda(A)$. If one defines now $v := V \cdot \left[\begin{smallmatrix} 0 \\ \widetilde{v} \end{smallmatrix}\right]$, then $v \neq 0$ (since V is orthogonal and $\widetilde{v} \neq 0$) and it satisfies

$$\begin{aligned} v^{\mathsf{H}}B &= \begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix} V^{\mathsf{T}}B = \begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix}\begin{bmatrix} B_1 \\ 0 \end{bmatrix} = 0, \\ v^{\mathsf{H}}A &= \begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix} V^{\mathsf{T}}A = \begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix} V^{\mathsf{T}}AVV^{\mathsf{T}} = \begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} V^{\mathsf{T}} \\ &= \lambda\begin{bmatrix} 0 & \widetilde{v}^{\mathsf{H}} \end{bmatrix} V^{\mathsf{T}} = \lambda v^{\mathsf{H}}. \end{aligned}$$

Thus it holds that $v^{\mathsf{H}}\begin{bmatrix} A - \lambda I & B \end{bmatrix} = 0$ in contradiction to d).

□

Part c) of the theorem gives a practical test for controllability of an LTI system: compute all eigenvalues and left eigenvectors v_j of A and then check whether $v_j^{\mathsf{H}}B = 0$. This test can still be improved from the numerical point of view. Transforming the pair (A,B) to a staircase form instead will result in a numerically more stable scheme, since the accuracy of the eigenvectors may be very sensitive with respect to rounding errors, in particular, if A is almost defective. Moreover, the decision whether $v_j^{\mathsf{H}}B = 0$ is numerically difficult. The decomposition of (A,B) used in the proof of Theorem 2.12, namely

$$A = V\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} V^{\mathsf{T}}, \quad B = V\begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{2.21}$$

with orthogonal V and controllable (A_{11}, B_1) is called *(orthogonal) Kalman decomposition* of (A, B). By a change of basis $\widetilde{x} := V^{\mathsf{T}} x$ in state space, for an LTI system, one obtains the equivalent system

$$\begin{aligned} \dot{\widetilde{x}}_1(t) &= A_{11}\widetilde{x}_1(t) + A_{12}\widetilde{x}_2(t) + B_1 u(t), \\ \dot{\widetilde{x}}_2(t) &= \qquad\qquad A_{22}\widetilde{x}_2(t). \end{aligned}$$

Therefore, the components of $\widetilde{x}_2$ are already fixed by the initial condition $\widetilde{x}_2(0)$ and cannot be influences by the control; it holds that $\widetilde{x}_2(t) = \mathrm{e}^{A_{22}t}\widetilde{x}_2(0)$. Therefore, the components of $\widetilde{x}_2$ are called uncontrollable states, the right eigenvectors of A corresponding to eigenvalues in A_{22} are called *uncontrollable modes* of the LTI system.

Example (Example 1 revisited)**:** After a linearization and reduction to a system of first order, one obtains an LTI system with state space $\mathcal{X} = \mathbb{R}^2$ and

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

Then $\mathcal{K}(A, B) = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, thus $\operatorname{rank} \mathcal{K}(A, B) = 2$ and, according to Theorem 2.12 a), b) the system is controllable. Alternatively, one can use the Hautus-Popov test. We have $\Lambda(A) = \{-1, 1\}$, the left eigenvector associated with $\lambda_1 = 1$ is $v_1 = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ and $v_1^{\mathsf{H}} B = 1 \neq 0$; the left eigenvector associated with $\lambda_2 = -1$ is $v_2 = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ and $v_2^{\mathsf{H}} B = -1 \neq 0$. Hence, again controllability of the system is shown.

2.2 Stabilizability

Now we want to consider the weaker goal of reaching the given target only asymptotically. As a target we will take $x^1 = 0$. First we consider LTV systems.

Definition 2.13: The LTV system (1.6)–(1.7) is called *(asymptotically) stabilizable*, if for every initial state $x^0 \in \mathbb{R}^n$, there exists a $u \in U_{\mathrm{ad}}$ such that the solution of (1.6) satisfies

$$\lim_{t \to \infty} x(t; u) = 0.$$

A necessary condition for the stabilizability of LTV systems is provided by the following result.

Theorem 2.14: If the LTV system (1.6) is stabilizable and $z(\cdot)$ is a bounded, nontrivial solution of the adjoint equation (2.12) for $t \to \infty$, then

$$z(t)^\mathsf{T} B(t) \not\equiv 0 \quad \text{on } [t_0, \infty).$$

Proof. Assume that there exists a solution of the adjoint equation (2.12) such that

$$z(t)^\mathsf{T} B(t) = 0 \quad \forall t \in [t_0, \infty) \text{ and } \lim_{t\to\infty} \|z(t)\| < \infty.$$

Since by assumption z is non-trivial, it holds that $z(t_0) \neq 0$ such that we can find an initial state $x^0 \in \mathbb{R}^n$ with

$$(x^0)^\mathsf{T} z(t_0) \neq 0.$$

Analogously to the proof of Theorem 2.9 a) $\Rightarrow$ b), it follows that

$$x(t)^\mathsf{T} z(t) \equiv (x^0)^\mathsf{T} z(t_0) \neq 0 \quad \text{on } [t_0, \infty). \tag{2.22}$$

Let now $u \in U_{\text{ad}}$ be a stabilizing control for the solution of (1.6) with $x(t_0) = x^0$. Then with $\lim_{t\to\infty} x(t; u) = 0$ it also holds that

$$\lim_{t\to\infty} \|x(t; u)\| = 0.$$

Since $\|z(t)\|$ for $t \to \infty$ is bounded, there exists a sequence $(t_k)_{k=1}^\infty$ with $t_k \to \infty$ and

$$\lim_{k\to\infty} x(t_k)^\mathsf{T} z(t_k) = 0.$$

(Note: Because of the Cauchy-Schwarz inequality it holds that $\left|x(t_k)^\mathsf{T} z(t_k)\right| \leqslant \|x(t_k)\|\|z(t_k)\|$.) With this we have constructed a contradiction to (2.22). □

For checking stabilizability of an LTI system (resp. a matrix pair (A, B))) there exist similar characterizations as in Theorem 2.12.

Theorem 2.15 (Hautus test for stabilizability)**:** Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then the following statements are equivalent:

a) (A, B) is stabilizable.

b) There exists an $F \in \mathbb{R}^{m\times n}$ with $\Lambda(A + BF) \subset \mathbb{C}^-$.

c) In the Kalman decomposition (2.21) it holds that $\Lambda(A_{22}) \subset \mathbb{C}^-$.

d) If $v \neq 0$ is a left eigenvector of A associated with the eigenvalue λ with $\mathrm{Re}(\lambda) \geqslant 0$, then $v^{\mathsf{H}} B \neq 0$.

e) It holds $\mathrm{rank}\begin{bmatrix} A - \lambda I & B \end{bmatrix} = n$ for all $\lambda \in \mathbb{C}$ with $\mathrm{Re}(\lambda) \geqslant 0$.

Proof. Homework. Show the following ring closure a) $\Rightarrow$ d) $\Rightarrow$ c) $\Rightarrow$ e) $\Rightarrow$ b) $\Rightarrow$ a), similarly to the proof of Theorem 2.12. □

Example (Example 1 revisited)**:** For the system matrices

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

we obtain – as found above – $\Lambda(A) = \{-1, 1\}$, where $v_1 = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ is a left eigenvector associated with the only eigenvalue with nonnegative real part $\lambda_1 = 1$. We have $v_1^{\mathsf{T}} B = 1 \neq 0$, from which we infer stabilizability according to Theorem 2.15. Note that it is sufficient to evaluate $v_1^{\mathsf{H}} B$, since the eigenspace to λ_1 is onedimensional and therefore, every eigenvector to λ_1 is a nonzero scalar multiple of v_1.

Stabilizability could have been also checked with the following simple consequence of Theorems 2.12 and 2.15.

Corollary 2.16: A controllable LTI system is stabilizable.

2.3 Observability and Detectability

First we consider again an LTV system of the form (1.6)–(1.7) and ask the question, how much information of the state of the system can be obtained from the output equation (1.7). In practical applications, this is a very relevant question, since most often not the whole state is available for control design, but only observed or measured quantities. These could be a few of the state variables or derived quantities. For instance, in Example 1 we could only measure the position (first component of the state vector), but not the angular velocity (second component of the state vector).

Definition 2.17 (observability)**:** An LTV system is called *reconstructable (observable)*, if the following condition is satisfied:

If $x(\cdot)$ and $\widetilde{x}(\cdot)$ are solutions of (1.6) for the same control function $u \in U_{\text{ad}}$ and if

$$C(t)x(t) = C(t)\widetilde{x}(t) \quad \forall t \leqslant t_0 \quad (t \geqslant t_0),$$

then it holds that

$$x(t) = \widetilde{x}(t) \quad \forall t \leqslant t_0 \quad (t \geqslant t_0).$$

Thus, reconstructability means that systems with the same inputs and the same outputs in the past, also had the same states in the past. On the other hand, observability delivers the same statement for the future, where as reference time instance we take t_0. We will see later that for LTI systems, both concepts are equivalent. Statements about reconstructability and observability can be shown easily by making use of statements of a dual system. The following duality principle is also useful in many other considerations in systems and control.

Theorem 2.18 (duality)**:** An LTV system is reconstructable, if and only if

$$\dot{x}(t) = A(-t)^{\mathsf{T}}x(t) + C(-t)^{\mathsf{T}}u(t) \tag{2.23}$$

is controllable.

Proof. If one defines $z(t) := \widetilde{x}(t) - x(t)$, then reconstructability is nothing but

$$C(t)z(t) = 0 \quad \forall t \leqslant t_0 \quad \Rightarrow \quad z(t) = 0 \quad \forall t \leqslant t_0.$$

This is equivalent to:

$$z(\cdot) \not\equiv 0 \text{ solves } \dot{z}(t) = A(t)z(t) \quad \Rightarrow \quad C(t)z(t) \not\equiv 0 \quad \text{on } (-\infty, t_0].$$

If we replace t by $-t$, then this statement becomes

$$z(\cdot) \not\equiv 0 \text{ solves } \dot{z}(t) = A(-t)z(t) \quad \Rightarrow \quad C(-t)z(t) \not\equiv 0 \quad \text{on } [-t_0, \infty).$$

But with Theorem 2.9 this is equivalent to the controllability of (2.23). □

With this we can easily characterize reconstructability of LTV systems.

Theorem 2.19: An LTV system is reconstructable, if and only if for all $t_1 \in \mathbb{R}$ there exists a $t_0 < t_1$ such that the (t_0, t_1)-*reconstructability Gramian*

$$Q(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t, t_1)^\mathsf{T} C(t)^\mathsf{T} C(t) \Phi(t, t_1) \mathrm{d}t \tag{2.24}$$

is positive definite.

Proof. This is a consequence of Theorem 2.9 applied to the dual LTV system (2.23) and the duality principle from Theorem 2.19. □

For LTI systems, as a consequence of the duality principle and the Hautus-Popov lemma (Theorem 2.12) one obtains the following characterizations of observability and reconstructability. Since both terms only involve the matrices A and C we also talk about observability and reconstructability of the matrix pair $(A, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$ instead of the LTI system (1.4)–(1.5).

Corollary 2.20 (Hautus-Popov test)**:** Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. Then the following statements are equivalent:

a) The pair (A, C) is reconstructable.

b) The pair (A, C) is observable.

c) For the *observability matrix*

$$\mathcal{O}(A, C) := \mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np\times n} \tag{2.25}$$

it holds that $\operatorname{rank} \mathcal{O}(A, C) = n$.

d) If $v \neq 0$ is a right eigenvector of A, then $Cv \neq 0$.

e) It holds that

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n$$

for all $\lambda \in \mathbb{C}$.

If one only knows the output $y(t)$ for the design of a control, the problem of constructing an *output feedback control*

$$u(t) = Fy(t), \quad F \in \mathbb{R}^{m\times p},$$

arises in order to achieve the given goal. Due to the equivalence of observability and reconstructability of LTI systems, one mostly uses only the observability concept.

Remark 2.21: If one applies the Kalman decomposition (2.21) to the LTI system $\dot{x}(t) = A^{\mathsf{T}}x(t)+C^{T}u(t)$, then one obtains an orthogonal matrix $W \in \mathbb{R}^{n\times n}$ with

$$W^{\mathsf{T}}A^{\mathsf{T}}W = \begin{bmatrix} A_{11}^{\mathsf{T}} & A_{21}^{\mathsf{T}} \\ 0 & A_{22}^{\mathsf{T}} \end{bmatrix}, \quad W^{\mathsf{T}}C^{\mathsf{T}} = \begin{bmatrix} C_1^{\mathsf{T}} \\ 0 \end{bmatrix},$$

where $A_{11} \in \mathbb{R}^{r\times r}$, $C_1 \in \mathbb{R}^{p\times r}$ and $r = \dim(\mathcal{O}(A, C))$. This results in the *(orthogonal) observability Kalman decomposition*

$$W^{\mathsf{T}}AW = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \quad CW = \begin{bmatrix} C_1 & 0 \end{bmatrix}. \tag{2.26}$$

With the change of basis $\widetilde{x}(t) := W^{\mathsf{T}}x(t)$ and a partitioning analogously to (2.26) yields the system

$$\begin{aligned} \dot{\widetilde{x}}_1(t) &= A_{11}\widetilde{x}_1(t), \\ \dot{\widetilde{x}}_2(t) &= A_{21}\widetilde{x}_1(t) + A_{22}\widetilde{x}_2(t), \\ y(t) &= C_1\widetilde{x}_1(t). \end{aligned}$$

In other words, the state variables in $\widetilde{x}_2$ have no influence on the output. Therefore, they are called unobservable states, right eigenvectors of A corresponding to the eigenvalues of A_{22} are called unobservable modes.

Analogously to the weakening of controllabilty to stabilizability, observability can be weakened to detectability.

Definition 2.22 (detectablity)**:** The LTI system (1.4)–(1.5) is called *detectable*, if for every solution $z(\cdot)$ of $\dot{z}(t) = Az(t)$ with $Cz(t) \equiv 0$ it holds that $\lim_{t\to\infty} z(t) = 0$.

If in the definition one sets $z(\cdot) := x(\cdot)-\widetilde{x}(\cdot)$ for two solutions $x(\cdot)$, $\widetilde{x}(\cdot)$ of (1.4) for the same input function $u(\cdot)$, the detectability can be interpreted as follows: From $Cx(t) \equiv C\widetilde{x}(t)$ one cannot infer $x(t) \equiv \widetilde{x}(t)$, but $\lim_{t\to\infty}(x(t)-\widetilde{x}(t)) = 0$.

In other words, the non-observable part of the state is not known, but we can conclude its asymptotic behavior. With Theorem 2.19 and the observability Kalman decomposition (2.26), one obtains the following variant of the duality principle:

Corollary 2.23: The LTI system (1.4)–(1.5) is detectable, if and only if

$$\dot{x}(t) = A^{\mathsf{T}} x(t) + C^{\mathsf{T}} u(t)$$

is stabilizable.

Analogously to the Hautus-Popov test for stabilizability one obtains the following result.

Corollary 2.24 (Hautus-Popov test for detectability)**:** Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. Then the following statements are equivalent:

a) The pair (A, C) is detectable.

b) There exists a $G \in \mathbb{R}^{n\times p}$ with $\Lambda(A + GC) \subset \mathbb{C}^-$.

c) In the observability Kalman decomposition (2.26) it holds that $\Lambda(A_{22}) \subset \mathbb{C}^-$.

d) If v is a right eigenvector of A corresponding to the eigenvalue λ with $\mathrm{Re}(\lambda) \geqslant 0$, then $Cv \neq 0$.

e) It holds that

$$\operatorname{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n$$

for all $\lambda \in \mathbb{C}$ with $\mathrm{Re}(\lambda) \geqslant 0$.

If one wants to stabilize an LTI system and only the output $y(t)$ is available for control, then one wants to find an output feedback $u(t) = Fy(t)$ with $F \in \mathbb{R}^{m\times p}$ such that $\lim_{t\to\infty} x(t; u) = 0$. Note that the existence of such a feedback is not guaranteed, even if the system is both stabilizable and detectable.

CHAPTER 3

Stabilization, Lyapunov Equations, and Pole Placement

In this chapter we try to answer the question, how to determine a feedback matrix for LTI systems. From Theorem 2.15 we know that the computation of a stabilizing control function $u(\cdot)$ is possible with the help of state feedbacks. For that we need an $F \in \mathbb{R}^{m\times n}$ such that $\Lambda(A+BF) \subset \mathbb{C}^-$. With $u(t) := Fx(t)$ it follows that the solution trajectory of

$$\dot{x}(t) = Ax(t) + Bu(t) = Ax(t) + BFx(t) = (A+BF)x(t)$$

is asymptotically stable, if and only if F is stabilizing, i. e., $\Lambda(A+BF) \subset \mathbb{C}^-$.

Under certain assumptions, stabilizing state feedbacks can be obtained by the solution of linear-quadratic optimal control problems, see Chapter 4. First we want to discuss two simpler methods:

- Lyapunov's direct method: With the help of Lyapunov's stability theory and the solution of a linear system of equations, a stabilizing feedback can be computed directly. This method is a special case of a general theory for nonlinear systems which is based on the computation of *Lyapunov functions* for nonlinear systems of the form $\dot{x}(t) = f(x(t))$ (with $f(0) = 0$, i. e., $x = 0$ is an equilibrium of the dynamical system). There, one seeks a differentiable function $V : \mathcal{X} \to \mathbb{R}$ such that

  - $V(x) \geqslant 0$ and $V(x) = 0$, if and only if $x = 0$;
  - $\frac{\mathrm{d}}{\mathrm{d}t}V(x(t)) = \nabla V(x(t)) \cdot f(x(t)) < 0$ for all $x(t) \neq 0$.

  If one can find such a function, then Lyapunov's theorem says that $x = 0$ is an asymptotically stable equilibrium. By using energy-based modelling techniques leading to so-called port-Hamiltonian systems, a Lyapunov function can be obtained for free, however, it is not always possible to get a Lyapunov function in an easy way. Here will use Lyapunov functions in an implicit way. Namely, for an LTI system with state matrix A, we will consider Lyapunov equations of the form $AP + PA^T + I_n = 0$. Then the uncontrolled system is asymptotically stable, if and only if the solution matrix P is positive definite. In this case, $V(x) = x^{\mathsf{T}}Px$ is a Lyapunov function.

- pole placement: In general, the pole placement problem consists of computing a feedback matrix $F \in \mathbb{R}^{m\times n}$ such that $\Lambda(A + BF) = \mathcal{L}$ for a set $\mathcal{L} := \{\mu_1, \dots \mu_n\}$. If one chooses $\mathcal{L} \subset \mathbb{C}^-$, then the system is stabilized. However, without further conditions, this approach cannot be generalized to nonlinear or LTV systems. For instance, for LTV systems, the condition $\Lambda(A(t) + B(t)F(t)) \in \mathbb{C}^-$ for all $t \geqslant t_0$ is neither necessary, nor sufficient for stability of the closed-loop system. Locally, nonlinear systems can be approximated by LTV systems. I. e., for the stabilization of nonlinear, one should at least be able to stabilize LTV systems but not even this is sufficient. To achieve stabilization of a nonlinear system with the help of local stabilizations, further assumptions are necessary. A technique that achieves this goal is *model predictive control (MPC)*.

3.1 Lyapunov's Stability Theory

In this section, linear matrix equations will play an essential role. Thus, we will first look at a few important properties of such equations. Consider the *Sylvester equation*

$$AX + XB = W \tag{3.1}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$, $W \in \mathbb{R}^{n\times m}$ and the unknown matrix $X \in \mathbb{R}^{n\times m}$. This is a linear equation in the $n \cdot n$ unknowns x_{ij}, $i = 1, \dots, n$, $j = 1, \dots, m$. Thus there exists a representation of (3.1) in the form $Mx = w$ of a linear systems of equations in $\mathbb{R}^{mn}$. With the help of this representation we can directly obtain conditions for (unique) solvability of Sylvester equations. For this, we need the following definition.

Definition 3.1: Let $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{m\times q}$. Then the *Kronecker product (or tensor product)* of A and B is defined by

$$A \otimes B := \begin{bmatrix} a_{11}B & \dots & a_{1,p}B \\ \vdots & & \vdots \\ a_{n,1}B & \dots & a_{n,p}B \end{bmatrix} \in \mathbb{R}^{nm\times pq}.$$

Moreover, the vec operator is defined by $\mathrm{vec} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\cdot p}$ with

$$\mathrm{vec}(A) = \begin{bmatrix} a_{11} & \dots & a_{n,1} & a_{12} & \dots & a_{n,2} & \dots & a_{1,p} & \dots & a_{n,p} \end{bmatrix}^{\mathsf{T}}.$$

The following properties of the Kronecker products are directly obtained from Definition 3.1:

a) $(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha(A \otimes B)$ for all $\alpha \in \mathbb{R}$;

b) $(A + B) \otimes C = (A \otimes C) + (B \otimes C)$;

c) $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$;

d) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$;

e) $(A \otimes B)^{\mathsf{T}} = A^{\mathsf{T}} \otimes B^{\mathsf{T}}$;

f) $(A \otimes B)(C \otimes D) = AC \otimes BD$;

g) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, if A and B are both invertible.

An important property connects the Kronecker product and the vec operator, with its help we can "vectorize" a Sylvester equation

Lemma 3.2: For $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$ and $X \in \mathbb{R}^{n\times m}$ it holds that

$$\mathrm{vec}\,(AXB) = \left(B^{\mathsf{T}} \otimes A\right) \mathrm{vec}\,(X)\,.$$

Proof. Homework. □

Now we directly obtain the vectorized representation of the Sylvester equation (3.1).

Corollary 3.3: The Sylvester equation (3.1) is equivalent to

$$\left((I_m \otimes A) + \left(B^{\mathsf{T}} \otimes I_n\right)\right) \operatorname{vec}(X) = \operatorname{vec}(W), \tag{3.2}$$

i. e. X solves (3.1), if and only if $\operatorname{vec}(X)$ solves (3.2).

If one defines $M := (I_m \otimes A) + \left(B^{\mathsf{T}} \otimes I_n\right)$, then it is immediately clear that the Sylvester equation has a unique solution, if and only if M is nonsingular. A necessary and sufficient condition is that the M has no zero eigenvalues. Since due to the Kronecker product structure, the eigenvalues of M can be explicitly stated in terms of the eigenvalues of A and B, one obtains an easy to check condition for unique solvability of (3.1). The relation between the eigenvalues of A, B and M is stated as follows.

Theorem 3.4 (Theorem of Stephanos)**:** Let $p(x, y)$ be a complex polynomial in two variables, i. e., $p(x, y) = \sum_{j,k=1}^{\ell} \alpha_{jk} x^j y^k$ with x, y, $\alpha_{jk} \in \mathbb{C}$. For $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$ define a matrix-valued polynomial by replacing scalar multiplication by the Kronecker product, i. e.,

$$p(A, B) := \sum_{j,k=1}^{\ell} \alpha_{jk} (A^j \otimes B^k).$$

Then

$$\Lambda(p(A, B)) = \{p(\lambda, \mu) \ : \ \lambda \in \Lambda(A),\ \mu \in \Lambda(B)\}.$$

Proof. Homework. □

With this we can make statements about the solution of (3.1)

Theorem 3.5: Consider the Sylvester equation (3.1). Then

a) $\Lambda(M) = \Lambda\left((I_m \otimes A) + (B^{\mathsf{T}} \otimes I_n)\right) = \{\lambda + \mu \ : \ \lambda \in \Lambda(A),\ \mu \in \Lambda(B)\}$.

b) The Sylvester equation (3.1) and hence the linear system of equations (3.2) have a unique solution, if and only if $\Lambda(A) \cap \Lambda(-B) = \varnothing$.

Proof. Homework. □

Consider now the special case of (3.1) with $m = n$, $B = A^{\mathsf{T}}$ and $W = W^{\mathsf{T}}$.

With this one obtains the *Lyapunov equation*

$$AX + XA^{\mathsf{T}} = W. \tag{3.3}$$

Since the Lyapunov equation is symmetric, it follows directly that also X^{T} is a solution of (3.3). If the solution is unique, i. e., if according to Theorem 3.5 b), $\Lambda(A) \cap \Lambda(-A) = \varnothing$, then this unique solution is symmetric. A sufficient condition for unique solvability is that $\Lambda(A) \subset \mathbb{C}^-$, i. e., that A is Hurwitz[1]. In this case one even obtains an explicit solution formula which can also be generalized to (3.1), if A and B are both asymptotically stable.

Theorem 3.6: Let $\Lambda(A), \Lambda(B) \subset \mathbb{C}^-$. Then (3.1) has a unique solution that is given by

$$X = -\int_0^\infty \mathrm{e}^{At} W \mathrm{e}^{Bt} \mathrm{d}t. \tag{3.4}$$

Proof. The uniqueness of the solution follows directly from Theorem 3.5 b). Define now $Z : [0, \infty) \to \mathbb{R}^{n\times m}$ as the solution of the linear matrix-valued differential equation

$$\dot{Z}(t) = AZ(t) + Z(t)B \tag{3.5}$$

for the initial condition $Z(0) = W$. From the theory of linear homogeneous differential equations it follows that this initial value problem for the *Sylvester differential equation* (3.5) has a unique solution on $[0, \infty)$. This solution is $Z(t) = \mathrm{e}^{At} W \mathrm{e}^{Bt}$ as one can check easily: $Z(0) = W$ and

$$\begin{aligned} \dot{Z}(t) &= A\mathrm{e}^{At} W \mathrm{e}^{Bt} + \mathrm{e}^{At} W B \mathrm{e}^{Bt} \\ &= A\mathrm{e}^{At} W \mathrm{e}^{Bt} + \mathrm{e}^{At} W \mathrm{e}^{Bt} B \\ &= AZ(t) + Z(t)B. \end{aligned}$$

Here we have used that B and e^{Bt} commute. Since both A and B are Hurwitz, $\lim_{t\to\infty} \mathrm{e}^{At} = 0$, $\lim_{t\to\infty} \mathrm{e}^{Bt} = 0$ and thus,

$$Z_\infty := \lim_{t\to\infty} Z(t) = \lim_{t\to\infty} \mathrm{e}^{At} W \mathrm{e}^{Bt} = 0.$$

Integration of (3.5) over $[0, \infty)$ then gives

$$Z_\infty - Z(0) = A\int_0^\infty Z(t)\mathrm{d}t + \int_0^\infty Z(t)\mathrm{d}t B,$$

[1] This means that A is *asymptotically* stable.

therefore,

$$A\int_0^\infty Z(t)\mathrm{d}t + \int_0^\infty Z(t)\mathrm{d}tB = -W.$$

Thus, $-\int_0^\infty Z(t)\mathrm{d}t$ is a solution of the Sylvester equation (3.1). From the uniqueness of the solution it follows that $X = -\int_0^\infty Z(t)\mathrm{d}t = -\int_0^\infty \mathrm{e}^{At}W\mathrm{e}^{Bt}\mathrm{d}t$. □

If one considers an asymptotically stable LTI system $\dot{x}(t) = Ax(t) + Bu(t)$, then one directly obtain from Theorem 3.6 that the *controllability Gramian*

$$P = \int_0^\infty \mathrm{e}^{At}BB^\mathsf{T}\mathrm{e}^{A^\mathsf{T}t}\mathrm{d}t$$

is the unique solution of the Lyapunov equation

$$AP + PA^\mathsf{T} + BB^\mathsf{T} = 0. \tag{3.6}$$

From this we obtain a further criterion for controllability of LTI systems

Corollary 3.7: Let $\Lambda(A) \subset \mathbb{C}^-$. Then the pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ is controllable if and only if the solution of the Lyapunov equation (3.6) is positive definite.

Remark 3.8: Note that $P \neq \lim_{\tau\to\infty} P(0,\tau)$, since this limit is not defined if A is asymptotically stable. Thus P is not a "$[0,\infty]$-controllability Gramian".

Analogously, one obtains a characterization of observability for asymptotically stable LTI systems via the positive definiteness of the *observability Gramian* which is the unique solution of the Lyapunov equation

$$A^\mathsf{T}Q + QA + C^\mathsf{T}C = 0.$$

The following is the central result in Lyapunov's stability theory.

Theorem 3.9 (Lyapunov's theorem (1897))**:** Let $A, W \in \mathbb{R}^{n\times n}$ with $W = W^\mathsf{T}$ negative definite. Then the following holds:

a) If $\Lambda(A) \subset \mathbb{C}^-$, then the Lyapunov equation (3.3) has a unique solution X which is symmetric and positive definite.

b) If (3.3) has a solution $X > 0$, then A is asymptotically stable.

From Theorem 3.9 b) one obtains a test for asymptotic stability which is called *Lyapunov's direct method*[2]. For this one solves the Lyapunov equation $AX + XA^\mathsf{T} = -\alpha I_n$ for some $\alpha < 0$. If $X > 0$ (which can be checked by a Cholesky decomposition of X), then all solutions of $\dot{x}(t) = Ax(t)$ are asymptotically stable.

The following weaker version of Theorem 3.9 which goes back to Chen (1973) and Wimmer (1974), the definiteness of the right-hand side can be weakened under the additional assumption of controllability.

Theorem 3.10: Let the pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ be controllable. Then it holds that:

a) If $\Lambda(A) \subset \mathbb{C}^-$, then the Lyapunov equation (3.6) has a unique solution P. Moreover, it holds that $P = P^\mathsf{T} > 0$.

b) If (3.6) has a solution $P > 0$, then A is stable.

3.2 Stabilization with Lyapunov Equations

The following theorem, which goes back to Kleinman (1970) and Armstrong (1975) and uses previous ideas from Bass, results in a first stabilization method. In the following, by M^+ we denote the *(Moore-Penrose) pseudoinverse* of M, i. e., the unique matrix that satisfies the *Moore-Penrose conditions*

- $MM^+M = M$,
- $M^+MM^+ = M^+$,
- $(MM^+)^\mathsf{T} = MM^+$,
- $(M^+M)^\mathsf{T} = M^+M$.

Theorem 3.11: Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ be stabilizable and $\beta \in \mathbb{R}$ with $\beta > \rho(A)$, where $\rho(A) := \max\{|\lambda| \ : \ \lambda \in \Lambda(A)\}$ is the spectral radius of A. If X is the unique solution of the Lyapunov equation

$$(A + \beta I_n)X + X(A + \beta I_n)^\mathsf{T} = 2BB^\mathsf{T}, \tag{3.7}$$

then $F := -B^\mathsf{T}X^+$ is a stabilizing feedback matrix for (A, B).

[2]The name "direct method" refers to the fact that no trajectories have to be computed to check stability

Proof. Let first (A, B) be controllable (and therefore, also stabilizable). Since $\beta > \rho(A)$ it holds that $\Lambda(A + \beta I_n) \subset \mathbb{C}^+$. To apply Theorem 3.10 a), $(-(A + \beta I_n), \sqrt{2}B)$ must be controllable. But this follows directly with the Hautus-Popov test, since

$$\begin{aligned} & n = \operatorname{rank}\begin{bmatrix} A - \lambda I_n & B \end{bmatrix} \quad \forall\, \lambda \in \mathbb{C} \\ \Leftrightarrow \quad & n = \operatorname{rank}\left[(A + \beta I_n) + \widetilde{\lambda} I_n \quad B \right] \begin{bmatrix} -I_n & 0 \\ 0 & \sqrt{2} I_m \end{bmatrix} \\ & \quad = \operatorname{rank}\begin{bmatrix} -(A + \beta I_n) - \widetilde{\lambda} I_n & \sqrt{2} B \end{bmatrix} \quad \forall\, \widetilde{\lambda} \in \mathbb{C}. \end{aligned}$$

With Theorem 3.10 b) it follows that (3.7) has a unique solution $X > 0$. Since X is invertible, the equivalence of (3.7) and

$$X^{-1}(A + \beta I_n) + (A + \beta I_n)^{\mathsf{T}} X^{-1} = 2X^{-1} B B^{\mathsf{T}} X^{-1}.$$

This results in

$$X^{-1}\left(A - B B^{\mathsf{T}} X^{-1}\right) + \left(A - B B^{\mathsf{T}} X^{-1}\right)^{\mathsf{T}} X^{-1} = -2\beta X^{-1}.$$

Since X and thus also X^{-1} are positive definite, the right-hand side of this Lyapunov equation is negative definite. Then with Theorem 3.9 b) it follows that $A - B B^{\mathsf{T}} X^{-1}$ is stable, i. e., $F = -B^{\mathsf{T}} X^{+}$ is a stabilizing feedback matrix, since for invertible matrices it holds that $X^{+} = X^{-1}$. Let now (A, B) be stabilizable. Due to Theorem 3.6 we know that (3.7) has a unique solution, which due to the representation (3.4) is positive semi-definite. Moreover, we know by Theorem 2.15 that (A, B) has a Kalman decomposition of the form

$$A = V \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} V^{\mathsf{T}}, \quad B = V \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where (A_{11}, B_1) is controllable, A_{22} is asymptotically stable and $V \in \mathbb{R}^{n \times n}$ is orthogonal. Left-multiplying (3.7) with V^{T} and right-multiplying it with V and partitioning

$$\widehat{X} = V^{\mathsf{T}} X V = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^{\mathsf{T}} & X_{22} \end{bmatrix}$$

as in the Kalman decomposition, then we obtain

$$\begin{aligned} & \begin{bmatrix} A_{11} + \beta I & A_{12} \\ 0 & A_{22} + \beta I \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^{\mathsf{T}} & X_{22} \end{bmatrix} \\ & \qquad + \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^{\mathsf{T}} & X_{22} \end{bmatrix} \begin{bmatrix} (A_{11} + \beta I)^{\mathsf{T}} & 0 \\ A_{12}^{\mathsf{T}} & (A_{22} + \beta I)^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} 2B_1 B_1^{\mathsf{T}} & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Then one obtain the following linear matrix equations

$$(A_{11}+\beta I)X_{11}+X_{11}(A_{11}+\beta I)^{\mathsf{T}}+A_{12}X_{12}^{\mathsf{T}}+X_{12}A_{12}^{\mathsf{T}}=2B_1B_1^{\mathsf{T}}, \tag{3.8}$$

$$(A_{22}+\beta I)X_{22}+X_{22}(A_{22}+\beta I)^{\mathsf{T}}=0. \tag{3.9}$$

The homogeneous equation (3.9) has the unique solution $X_{22}=0$ by Theorem 3.5 b). Since X and thus also $\widehat{X}$ are positive semi-definite, we get $X_{12}=0$. From the controllability of (A_{11},B_1) it follows that (3.8) has a unique solution $X_{11}>0$ and that $F_1=-B_1^{\mathsf{T}}X_{11}^{-1}$ is a stabilizing feedback matrix for (A_{11},B_1). If one sets $F:=\begin{bmatrix}F_1 & 0\end{bmatrix}V^{\mathsf{T}}$, then it holds that

$$V^{\mathsf{T}}(A+BF)V=\begin{bmatrix}A_{11}+B_1F_1 & A_{12}\\ 0 & A_{22}\end{bmatrix},$$

i. e., $\Lambda(A+BF)=\Lambda(A_{11}+B_1F_1)\cup\Lambda(A_{22})\subset\mathbb{C}^-$. Thus, F is a stabilizing feedback matrix for (A,B). Moreover, it holds that

$$F=\begin{bmatrix}-B_1^{\mathsf{T}}X_{11}^{-1} & 0\end{bmatrix}V^{\mathsf{T}}=-\begin{bmatrix}B_1^{\mathsf{T}} & 0\end{bmatrix}\begin{bmatrix}X_{11}^{-1} & 0\\ 0 & 0\end{bmatrix}V^{\mathsf{T}}=-B^{\mathsf{T}}V\begin{bmatrix}X_{11}^{-1} & 0\\ 0 & 0\end{bmatrix}V^{\mathsf{T}}.$$

Moreover, by simple calculations one can check that $V\left[\begin{smallmatrix}X_{11}^{-1} & 0\\ 0 & 0\end{smallmatrix}\right]V^{\mathsf{T}}$ fulfills the Moore-Penrose conditions with respect to $X=V\left[\begin{smallmatrix}X_{11} & 0\\ 0 & 0\end{smallmatrix}\right]V^{\mathsf{T}}$. Thus, $F=-B^{\mathsf{T}}X^{+}$. □

The proof above make use of the fact that the uncontrollable modes of the LTI system do not have to be stabilized such that the stabilization problem can be transferred to the controllable case. With this one obtains a complete proof of Theorem 2.15, since Theorem 3.11 delivers “c) ⇒ b)” under the condition that “a) ⇒ c)” has already been proven.

Algorithm 3.1 Bass algorithm

Input: Stabilizable pair $(A,B)\in\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m}$.
Output: Stabilizing feedback matrix $F\in\mathbb{R}^{m\times n}$, i. e., $\Lambda(A+BF)\subset\mathbb{C}^-$.
1: Set $\beta=2\|A\|_p$ for an easy to calculate norm, e. g., $p=1,\infty,\mathrm{F}$.
2: Solve (3.7).
3: Compute X^{+} and set $F=-B^{\mathsf{T}}X^{+}$.

The factor 2 in row 1 is a safety factor, that shall guarantee that the eigenvalues of $A+\beta I_n$ are sufficiently far away from the imaginary axis. The computation of the pseudoinverse X^{+} can be done by a spectral decomposition of the positive

semi-definite matrix X. An alternative to the stabilization by the Lyapunov (3.7) is the so-called *algebraic Bernoulli equation (ABE)*

$$A^{\mathsf{T}}X + XA - XBB^{\mathsf{T}}X = 0. \tag{3.10}$$

(which is a special algebraic Riccati equation). The following results have been shown in [BBQO07]:

Theorem 3.12: If $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ is stabilizable and $\Lambda(A) \cap \mathrm{i}\mathbb{R} = \varnothing$, then the following are equivalent:

a) The ABE (3.10) has a unique stabilizing positive semi-definite solution X_*, i. e., $X_* \geqslant 0$ and $\Lambda\big(A - BB^{\mathsf{T}}X_*\big) \subset \mathbb{C}^-$.

b) $\operatorname{rank} X_* = k$, where k is the number of unstable eigenvalues of A. With this it holds that $X_* = ZZ^{\mathsf{T}}$ with $Z \in \mathbb{R}^{n\times k}$.

c) It holds that $\Lambda\big(A - BB^{\mathsf{T}}X_*\big) = (\Lambda(A) \cap \mathbb{C}^-) \cup -(\Lambda(A) \cap \mathbb{C}^+)$.

So also the ABE can be used for the stabilization of linear time-invariant systems. The algorithm used in [BBQO07] for computing X, resp. Z is similarly expensive as the Bartels-Stewart algorithm for solving Lyapunov equations. Moreover, numerical experiments indicate, that stabilization properties of the ABE solution are often better than the ones of the Lyapunov equation.

3.3 Stabilization by Pole Placement

First we will show that a system is controllable if and only if for every set $\mathcal{L} := \{\mu_1, \dots \mu_n\} \subset \mathbb{C}$ which is closed with respect to complex conjugation, there exists a feedback matrix $F \in \mathbb{R}^{m\times n}$ with $\Lambda(A + BF) = \mathcal{L}$. Therefore, we introduce two normal forms for single input systems, that allow further characterizations of controllability. Note that these normal forms are only of theoretical interest, since they cannot be computed in a numerically stable way.

Let now

$$\phi_A(x) = x^n + \alpha_{n-1}x^{n-1} + \ldots + \alpha_1 x + \alpha_0 \tag{3.11}$$

be the characteristic polynomial of A. Further, we say that (A, B) and $\big(\widetilde{A}, \widetilde{B}\big)$ are *system equivalent*, if $\big(\widetilde{A}, \widetilde{B}\big)$ can be obtained from (A, B) by a change of basis in state space, i. e., if there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\big(\widetilde{A}, \widetilde{B}\big) = \big(T^{-1}AT, T^{-1}B\big).$$

Lemma 3.13: Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ and

$$A_{\mathrm{s}} = \begin{bmatrix} 0 & 0 & \dots & 0 & -\alpha_0 \\ 1 & 0 & \dots & 0 & -\alpha_1 \\ 0 & 1 & \dots & 0 & -\alpha_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -\alpha_{n-1} \end{bmatrix}, \quad B_{\mathrm{s}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.12}$$

Then it holds that

$$A\mathcal{K}(A, B) = \mathcal{K}(A, B)A_{\mathrm{s}}, \quad B = \mathcal{K}(A, B)B_{\mathrm{s}}, \tag{3.13}$$

where $\mathcal{K}(A, B)$ is the controllability matrix of (A, B). In particular, (A, B) is controllable, if and only if (A, B) and $(A_{\mathrm{s}}, B_{\mathrm{s}})$ are system equivalent.

Proof. Using the Theorem of Cayley-Hamilton it follows (cf. the proof of Theorem 2.11)

$$A^n = -\sum_{j=0}^{n-1} \alpha_j A^j.$$

Thus we obtain

$$\begin{aligned} A\mathcal{K}(A, B) &= \begin{bmatrix} AB & A^2B & \dots & A^{n-1}B & -\sum_{j=0}^{n-1} \alpha_j A^j B \end{bmatrix} \\ &= \mathcal{K}(A, B)A_{\mathrm{s}}. \end{aligned}$$

The second equation in (3.13) is immediate. Similarly as in Homework 3/1 b), one can show that $(A_{\mathrm{s}}, B_{\mathrm{s}})$ is controllable.

If $(A_{\mathrm{s}}, B_{\mathrm{s}})$ is system equivalent to (A, B), then (A, B) is controllable. On the other hand, if (A, B) is controllable, then $\mathcal{K}(A, B)$ is nonsingular by Theorem 2.12 and thus, by (3.13), (A, B) and $(A_{\mathrm{s}}, B_{\mathrm{s}})$ are system equivalent with $T := \mathcal{K}(A, B)$. □

Definition 3.14: The *controller normal form* of a pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ is given by

$$A_{\mathrm{CNF}} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \dots & -\alpha_{n-1} \end{bmatrix}, \quad B_{\mathrm{CNF}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{3.14}$$

It is important to note that *not* every pair (A, B) is system equivalent to its controller normal form as the following theorem shows.

Theorem 3.15: The pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ is controllable, if and only if there exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that $(A_{\mathrm{CNF}}, B_{\mathrm{CNF}}) = (S^{-1}AS, S^{-1}B)$.

Proof. Note first that ϕ_A is the characteristic polynomial of A_{CNF}. Thus, by Lemma 3.13, resp. (3.12), we have

$$A_{\mathrm{CNF}}\mathcal{K}(A_{\mathrm{CNF}}, B_{\mathrm{CNF}}) = \mathcal{K}(A_{\mathrm{CNF}}, B_{\mathrm{CNF}})A_{\mathrm{s}}, \quad B_{\mathrm{CNF}} = \mathcal{K}(A_{\mathrm{CNF}}, B_{\mathrm{CNF}})B_{\mathrm{s}}.$$

Now,

$$\mathcal{K}(A_{\mathrm{CNF}}, B_{\mathrm{CNF}}) = \begin{bmatrix} 0 & \dots & 0 & 1 \\ \vdots & \iddots & \iddots & * \\ 0 & \iddots & \iddots & \vdots \\ 1 & * & \dots & * \end{bmatrix}$$

which is nonsingular. Thus, $(A_{\mathrm{CNF}}, B_{\mathrm{CNF}})$ is system equivalent to $(A_{\mathrm{s}}, B_{\mathrm{s}})$. Thus Lemma 3.13 follows. □

Moreover, we need a property of the space $\operatorname{im}\mathcal{K}(A, B)$.

Lemma 3.16: Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ and $b_j = Be_j$, $j = 1, \dots, m$. Then it holds that

$$\operatorname{im}\mathcal{K}(A, B) = \operatorname{span}\{A^k b_j \ : \ k \in \mathbb{N}_0,\ j = 1, \dots, m\} =: \mathcal{K}.$$

In other words, $\operatorname{im}\mathcal{K}(A, B)$ is the smallest A-invariant subspace that contains $\operatorname{im} B$.

Proof. The statement follows from the Theorem of Cayley-Hamilton as in the proof of Theorem 2.11, since

$$A^{n+\nu} = \sum_{j=0}^{n-1} \beta_j^{(\nu)} A^j \quad \forall\, \nu \in \mathbb{N}_0.$$

Since every A-invariant subspace which contains $\operatorname{im} B$ also contains $\mathcal{K}$, the interpretation of $\operatorname{im}\mathcal{K}(A, B)$ as smallest A-invariant subspace that contains $\operatorname{im} B$. □

By further noting that for system equivalent matrix pairs (A, B) and $(\widetilde{A}, \widetilde{B}) = (T^{-1}AT, T^{-1}B)$ it holds that $\Lambda(A + BF) = \Lambda\big(\widetilde{A} + \widetilde{B}\widetilde{F}\big) = \mathcal{L}$ for $\widetilde{F} = FT$, we have all prerequisites to prove the theorem of pole placements. the proof follows the arguments of the proof of Theorem 13 in the book of Sontag, [Son98].

Theorem 3.17 (pole placement)**:** Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ and let the uncontrollable eigenvalues of (A, B) be $\{\lambda_{k+1}, \ldots, \lambda_n\}$. Then there exists a feedback matrix $F \in \mathbb{R}^{m\times n}$ such that $\Lambda(A + BF) = \mathcal{L}$, if and only if $\mathcal{L} = \{\mu_1, \ldots, \mu_k, \lambda_{k+1}, \ldots, \lambda_n\}$, where $\{\mu_1, \ldots, \mu_k\} \subset \mathbb{C}$ can be chosen arbitrarily as long a $\{\mu_1, \ldots, \mu_k\}$ is closed under complex conjugation. In case $m = 1$, F is unique.

In particular, (A, B) is controllable, if and only if for every set $\mathcal{L} = \{\mu_1, \ldots, \mu_n\}$ with $\mathcal{L} = \overline{\mathcal{L}}$, there exists an $F \in \mathbb{R}^{m\times n}$ with $\Lambda(A + BF) = \mathcal{L}$.

Proof. First, let (A, B) be not controllable. W. l. o. g., we can assume that (A, B) is in Kalman form, i. e.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad \text{with controllable pair } (A_1, B_1) \in \mathbb{R}^{k\times k} \times \mathbb{R}^{k\times m}.$$

For an arbitrary $F = \begin{bmatrix} F_1 & F_2 \end{bmatrix}$ is holds that

$$\Lambda(A + BF) = \Lambda(A_{11} + B_1F_1) \cup \Lambda(A_{22}) = \Lambda(A_{11} + B_1F_1) \cup \{\lambda_{k+1}, \ldots, \lambda_n\}.$$

Since $\Lambda(A_{22})$ are the uncontrollable eigenvalues, it is clear that $\mathcal{L}$ must attain the form within the theorem statement and that (A, B) will be controllable, if and only if $\mathcal{L}$ can be chosen arbitrarily. It remains to show that $\Lambda(A_{11} + B_1F_1)$ can be chosen arbitrarily by choosing an appropriate feedback matrix F. If for some set $\mathcal{L}_1 := \{\mu_1, \ldots, \mu_k\} = \overline{\mathcal{L}_1}$ one can find an $F_1 \in \mathbb{R}^{m\times k}$ with $\Lambda(A_{11} + B_1F_1) = \mathcal{L}_1$, then $F = \begin{bmatrix} F_1 & 0 \end{bmatrix}$ is the desired feedback matrix. According to the above considerations it remains to consider the case that (A, B) is controllable. In the case $m = 1$ we can further assume by Theorem 3.15 that (A, B) is in controller normal form. Then one immediately sees that with $F = \begin{bmatrix} f_1 & \ldots & f_n \end{bmatrix}$ with $f_j \in \mathbb{R}$ we get that

$$A + BF = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -\alpha_0 + f_1 & -\alpha_1 + f_2 & -\alpha_2 + f_3 & \ldots & -\alpha_{n-1} + f_n \end{bmatrix}.$$

Since

$$\phi_{A+BF}(x) = x^n + (\alpha_{n-1} - f_n)x^{n-1} + \ldots + (\alpha_1 - f_2)x + (\alpha_0 - f_1),$$

F is uniquely determined by $f_j = \alpha_{j-1} - \beta_{j-1}$, where

$$(x - \mu_1) \cdots (x - \mu_n) =: x^n + \beta_{n-1}x^{n-1} + \ldots + \beta_1 x + \beta_0.$$

Let now m be arbitrary. We will transfer this case to the case $m = 1$. For that, let $v \in \mathbb{R}^m$ with $b := Bv \neq 0$. We show now that there exists a $G \in \mathbb{R}^{m\times n}$ such that $(A + BG, b)$ is controllable. Then if $f \in \mathbb{R}^{1\times n}$ is the uniquely determined vector such that $\Lambda(A + BG + bf) = \mathcal{L}$, then $F := G + vf$ is the desired feedback matrix. It remains to show the existence of G. Define now a maximal set of linearly independent vectors $\mathcal{R} = \{x_1, \ldots, x_\ell\} \subset \mathbb{R}^n$ with

$$x_1 := b = Bv, \quad x_j - Ax_{j-1} \in \operatorname{im} B, \tag{3.15}$$

where $x_0 = 0$. Obviously, $\mathcal{R} \neq \varnothing$, since $b \in \mathcal{R}$ and $\dim \operatorname{span} \mathcal{R} = \ell \leqslant n$ and thus, "ℓ maximal" is well-defined. Note that (3.15) can also be formulated as follows:

$$x_j = Ax_{j-1} + Bu \text{ for some } u \in \mathbb{R}^m. \tag{3.16}$$

We show now that $\ell = n$. That ℓ has been chosen maximally, follows from

$$Ax_\ell + Bu \in \operatorname{span}\{x_1, \ldots, x_\ell\} \quad \forall\, u \in \mathbb{R}^m. \tag{3.17}$$

(Otherwise, with $x_{\ell+1} := Ax_\ell + Bu$ one would obtain a bigger set that satisfies the condition which would contradict the maximality of ℓ.) In particular, with $u = 0$ we get

$$Ax_\ell \in \operatorname{span}\{x_1, \ldots, x_\ell\}.$$

Then with (3.17) it follows that

$$\operatorname{im} B \subset \operatorname{span}\{x_1, \ldots, x_\ell\} - Ax_\ell = \operatorname{span}\{x_1, \ldots, x_\ell\}$$

and with (3.15)

$$Ax_k \in \operatorname{span}\{x_1, \ldots, x_\ell\}, \quad k = 1, \ldots, \ell.$$

With this, $\operatorname{span}\{x_1, \ldots, x_\ell\}$ is an A-invariant subspace which contains $\operatorname{im} B$. By Lemma 3.16, $\operatorname{im} \mathcal{K}(A, B)$ is the smallest subspace of $\mathbb{R}^n$ that fulfills this property. Thus, with the controllability of (A, B) we get

$$n = \dim \operatorname{im} \mathcal{K}(A, B) \leqslant \dim \operatorname{span}\{x_1, \ldots, x_\ell\}$$

and thus $\dim \operatorname{span}\{x_1, \ldots, x_\ell\} = \ell = n$. Let now $u_k \in \mathbb{R}^m$, $k = 1, \ldots, n-1$ be a sequence of vectors which generates $\mathcal{R}$ as in (3.16), i. e.,

$$x_k - Ax_{k-1} = Bu_{k-1}, \quad k = 2, \ldots, n.$$

If one chooses $u_n \in \mathbb{R}^m$ arbitrary and defines

$$X := \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad U := \begin{bmatrix} u_1 & \ldots & u_n \end{bmatrix} \in \mathbb{R}^{m\times n},$$

then $G := UX^{-1}$ is well-defined, since X is nonsingular due to the linear independence of $\{x_1, \ldots, x_n\}$. The matrix G satisfies

$$Gx_k := u_k, \quad k = 1, \ldots, n,$$

and with (3.15) one obtains

$$\mathcal{K}(A + BG, b) = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}.$$

Thus, $(A + BG, b)$ is controllable and the claim is shown. □

The proof of Theorem 3.15 motivates the term *"controller normal form"*, since the matrix F determining the controller can be just read off A_{CNF} in the case $m = 1$.

Remark 3.18: For computing the feedback matrix F in the case $m > 1$ one has to enforce uniqueness by imposing further constraints. This freedom should be exploited in order to achieve, e. g., maximum robustness of the closed-loop eigenvalues with respect to perturbations.

A suitable criterion for the numerical computation of F consists of making $A+BF$ diagonalizable and minimizing the condition number of its eigenvector matrix $X = X(F)$. We obtain the minimization problem

$$\min_{F\in\mathbb{R}^{m\times n}} \operatorname{cond}(X(F))$$

$$\text{subject to } (A + BF)X(F) = X(F) \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{bmatrix}.$$

The problem of robust pole placement was formulated and solved by Kautsky, Nichols, and Van Dooren in 1985 [KNVD85], where, besides the criterion above, further criteria have been analyzed to measure the sensitivity of the poles with respect to perturbations.

CHAPTER 4

Optimal Control

For a general nonlinear system as in Definition 1.1, i. e.,

$$\dot{x}(t) = f(t, x(t), u(t)), \quad x(t_0) = x^0, \tag{4.1}$$
$$y(t) = g(t, x(t), u(t)), \tag{4.2}$$

for $t \in [t_0, t_\mathrm{f}]$, we seek an *optimal control* $u \in U_\mathrm{ad}$ such that the cost functional $\mathcal{J} : U_\mathrm{ad} \to \mathbb{R}$ with

$$\mathcal{J}(u) = h_\mathrm{f}(x(t_\mathrm{f})) + \int_{t_0}^{t_\mathrm{f}} h(t, x(t), y(t), u(t))\mathrm{d}t \tag{4.3}$$

is minimized. Here, h is a suitably chosen function, which costs to the states, inputs, and outputs, while h_f measures the deviation from the desired terminal state.

To allow for reaching the desired state asymptotically, i. e., we demand a stabilization of the system, we also allow setting $t_\mathrm{f} = \infty$. However, first we consider the case $t_\mathrm{f} < \infty$, the case of an infinite time horizon is obtained by an asymptotic consideration.

Remark 4.1: By choosing $h \equiv 1$ and $h_\mathrm{f} \equiv 0$ and requesting $x(t_\mathrm{f}) = x^1$, we obtain the problem of minimizing t_f. This problem is called *time-optimal*

control, since we try to find the control function that steers the system into the desired state x^1 in the shortest possible time.

4.1 Necessary and Sufficient Optimality Conditions

The general approach to such optimal control problems is based on the Lagrange formalism: If one wants to solve a constrained optimization problem of the form

$$\min_{x\in\mathbb{R}^n} g(x) \quad \text{subject to} \quad f(x) = 0, \tag{4.4}$$

then one defines the *Lagrangian (function)*

$$\mathcal{L}(x,\lambda) = g(x) + \lambda^{\mathsf{T}} f(x),$$

with the *Lagrange multipliers* $\lambda \in \mathbb{R}^m$ and develops the necessary optimality conditions from

$$\mathcal{L}_x(x,\lambda) = 0, \quad \mathcal{L}_\lambda(x,\lambda) = 0.$$

Analogously, for dynamic constraints, one uses the *Hamilton principle* for which one defines a *Hamilton function*. For autonomous systems satisfying

$$h(x(t), y(t), u(t)) \equiv h(x(t), u(t)),$$

it is defined by

$$\mathcal{H}(x(t), u(t), \mu(t)) = h(x(t), u(t)) + \mu(t)^{\mathsf{T}} f(x(t), u(t)), \tag{4.5}$$

where $\mu : [t_0, t_\mathrm{f}] \to \mathbb{R}^n$ is the *costate function* corresponding to the Lagrange multipliers. Note that with this, the dynamic constraints can be expressed as

$$\mathcal{H}_\mu(x(t), u(t), \mu(t)) = \dot{x}(t).$$

The necessary optimality conditions then follow from the following theorem due to Pontryagin [PBGM62]. Here, this result is stated as a “minimum principle” as in [Pin93], where first $h_\mathrm{f} = 0$ is assumed.

Theorem 4.2 (Pontryagin’s maximum principle)**:** Let $u_* \in U_\mathrm{ad}$ and $x_*(t) := x(t; u_*)$ be the corresponding solution trajectory of (4.1). If u_* is optimal for (4.3), then u_* satisfies the necessary optimality conditions

a) $\mathcal{H}(x_*(t), u_*(t), \mu(t)) = \inf_{u \in U_{\mathrm{ad}}} \mathcal{H}(x_*(t), u(t), \mu(t))$ for all $t \in [t_0, t_\mathrm{f}]$.

b) The costate function satisfies the *adjoint equation*

$$\dot{\mu}(t) = -\mathcal{H}_x(x_*(t), u_*(t), \mu(t)),$$

(iii) $\mu(t_\mathrm{f}) = 0$ (*transversality conditions*).

Proof. See [MS82, Pin93, PBGM62]. □

The case $h_\mathrm{f} \not\equiv 0$ can be transferred such that the theorem above is applicable. Therefore, one uses that

$$\begin{aligned} h_\mathrm{f}(x(t_\mathrm{f})) - h_\mathrm{f}(x(t_0)) &= \int_{t_0}^{t_\mathrm{f}} \nabla h_\mathrm{f}(x(t)) \cdot \dot{x}(t) \mathrm{d}t \\ &= \int_{t_0}^{t_\mathrm{f}} \nabla h_\mathrm{f}(x(t)) \cdot f(x(t), u(t)) \mathrm{d}t, \end{aligned}$$

which leads to the modified running cost

$$\widetilde{h}(x(t), u(t)) = h(x(t), u(t)) + \nabla h_\mathrm{f}(x(t)) \cdot f(x(t), u(t)).$$

With this we get $\mathcal{J}(u) = h_\mathrm{f}(x(t_0)) + \int_{t_0}^{t_\mathrm{f}} \widetilde{h}(x(t), u(t)) \mathrm{d}t$. Since $h_\mathrm{f}(x(t_0))$ is constant, this term can be neglected in the optimization, i. e., one works with the modified cost functional

$$\widetilde{\mathcal{J}}(u) = \int_{t_0}^{t_\mathrm{f}} \widetilde{h}(x(t), u(t)) \mathrm{d}t.$$

If one replaces h by $\widetilde{h}$ in the Hamilton function, then one can apply Theorem 4.2 to $\widetilde{\mathcal{J}}$. However, note that the transversality condition changes to

(iii') $\mu(t_\mathrm{f}) = \nabla h_\mathrm{f}(x(t_\mathrm{f}))$.

Moreover, the necessary smoothness properties of h_f must be verified.

From now on, we consider again LTI systems as in (1.4)–(1.5), i. e.,

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(0) = x^0, \\ y(t) &= Cx(t), \end{aligned} \tag{4.6}$$

in the time interval $[0,\ t_\mathrm{f}]$. Hence, we have

$$\begin{aligned} f(t, x(t), u(t)) &= Ax(t) + Bu(t), \\ g(t, x(t), u(t)) &= Cx(t). \end{aligned}$$

Since $y(t) = Cx(t)$ and hence $h(t, x(t), y(t), u(t)) = h(t, x(t), Cx(t), u(t)) \equiv h(t, x(t), u(t))$, we will w. l. o. g. assume that $C = I_n$. By setting

$$h(t, x(t), u(t)) = \frac{1}{2} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q & S \\ S^{\mathsf{T}} & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix},$$
$$h_{\mathrm{f}}(x(t)) = \frac{1}{2} x(t)^{\mathsf{T}} M x(t),$$

then we obtain a quadratic cost functional $\mathcal{J}$ and thus the following problem setting:

Definition 4.3 (linear-quadratic optimal control problem)**:** The minimization problem

$$\min \mathcal{J}(u) = \frac{1}{2} \left(x(t_{\mathrm{f}})^{\mathsf{T}} M x(t_{\mathrm{f}}) + \int_0^{t_{\mathrm{f}}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} Q & S \\ S^{\mathsf{T}} & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \mathrm{d}t \right) \tag{4.7}$$

subject to

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x^0,$$

is called a *linear-quadratic optimal control problem*.

In control theory, such linear-quadratic optimal control problems are also called *linear-quadratic regulator problems*, or for short, *LQR problems*. In the sequel, we will make use of this abbreviation.

Remark 4.4: One can easily see that the assumption $C = I_n$ is indeed not a restriction. If the outputs should be weighted in the cost functional, then for a quadratic cost functional one can write

$$\begin{aligned} y(t)^{\mathsf{T}} Q_y y(t) + x(t)^{\mathsf{T}} Q_x x(t) &= x(t)^{\mathsf{T}} C^{\mathsf{T}} Q_y C x(t) + x(t)^{\mathsf{T}} Q_x x(t) \\ &= x(t)^{\mathsf{T}} \big(C^{\mathsf{T}} Q_y C + Q_x \big) x(t) \end{aligned}$$

Thus, with $Q := C^{\mathsf{T}} Q_y C + Q_x$ one would obtain a cost functional of the form (4.7).

The LQR cost functional weighs the following quantities:

- the deviation of the terminal state $x(t_{\mathrm{f}})$ from the target $\widehat{x}$ with the help of the term $x(t_{\mathrm{f}})^{\mathsf{T}} M x(t_{\mathrm{f}})$,
- the transient behavior of the state by $\int_0^{t_{\mathrm{f}}} x(t)^{\mathsf{T}} Q x(t) \mathrm{d}t$,

- the (energy) costs that have to be used for controlling, given by the term $\int_0^{t_\mathrm{f}} u(t)^\mathsf{T} R u(t) \mathrm{d}t$.

In Example 1, M would weigh the deviation of $\varphi(t_\mathrm{f}) = \pi$ and $\dot{\varphi}(t_\mathrm{f}) = 0$. The first term in the cost functional can be chosen to avoid an oscillatory transient behavior, while the last term in the cost functional assesses the input energy used to force the pendulum. Often, the mixed term $x(t)^\mathsf{T} S u(t)$ is not present. It occurs, e. g., if the original system has a feed-through term $Du(t)$ as in (1.5). Then, as in Remark 4.4, one has

$$\begin{aligned}
&y(t)^\mathsf{T} Q_y y(t) + x(t)^\mathsf{T} Q_x x(t) \\
&= x(t)^\mathsf{T} C^\mathsf{T} Q_y C x(t) + x(t)^\mathsf{T} Q_x x(t) + 2x(t)^\mathsf{T} C^\mathsf{T} Q_y D u(t) + u(t)^\mathsf{T} D^\mathsf{T} Q_y D u(t) \\
&= x(t)^\mathsf{T} \big(C^\mathsf{T} Q_y C + Q_x\big) x(t) + 2x(t)^\mathsf{T} C^\mathsf{T} Q_y D u(t) + u(t)^\mathsf{T} D^\mathsf{T} Q_y D u(t) \\
&=: x(t)^\mathsf{T} Q x(t) + 2x(t)^\mathsf{T} S u(t) + u(t)^\mathsf{T} R u(t).
\end{aligned}$$

By applying Pontryagin's maximum principle one obtains now the necessary optimality conditions. The theorem can be proven directly without using Theorem 4.2, while the proof structure follows a more general proof. In the following, let U_ad be set of functions that are piecewise continuous on $[0,\, t_\mathrm{f}]$, the proof for $U_\mathrm{ad} = L_2([0,\, t_\mathrm{f}], \mathbb{R}^m)$ is analogous.

Theorem 4.5: Consider the optimal control problem (4.7). Let $u_* \in U_\mathrm{ad}$ be an optimal control and let $x_*(t) = x(t; u_*)$ be the corresponding solution trajectory. Then there exists a costate function $\mu : [0,\, t_\mathrm{f}] \to \mathbb{R}^n$ such that x_*, μ, u_* solve the linear boundary value problem

$$\begin{bmatrix} I_n & 0 & 0 \\ 0 & -I_n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{\mu}(t) \\ \dot{u}(t) \end{bmatrix} = \begin{bmatrix} A & 0 & B \\ Q & A^\mathsf{T} & S \\ S^\mathsf{T} & B^\mathsf{T} & R \end{bmatrix} \begin{bmatrix} x(t) \\ \mu(t) \\ u(t) \end{bmatrix} \tag{4.8}$$

with boundary conditions

$$x(0) = x^0, \quad \mu(t_\mathrm{f}) = M x(t_\mathrm{f}) \tag{4.9}$$

Note that $\dot{u}$ enters the equation only formally and so u does not have to be differentiable.

Proof. The proof follows ideas from the calculus of variations, namely, the first variation has to vanish. Therefore, let u_* be the optimal control. Consider the first-order perturbation

$$u(t) := u_*(t) + \varepsilon v(t)$$

with $u \in U_{\text{ad}}$ and $\varepsilon \in \mathbb{R}$. Then the constraint in (4.7) becomes

$$\dot{x}(t) = Ax(t) + Bu_*(t) + \varepsilon Bv(t)$$

with the solution trajectory (see Chapter 2)

$$\begin{aligned} x(t) &= \mathrm{e}^{At}x^0 + \int_0^t \mathrm{e}^{A(t-s)}B\left(u_*(s) + \varepsilon v(s)\right)\mathrm{d}s \\ &= x_*(t) + \varepsilon \underbrace{\int_0^t \mathrm{e}^{A(t-s)}Bv(s)\mathrm{d}s}_{=:z(t)} \\ &= x_*(t) + \varepsilon z(t). \end{aligned}$$

Hereby, $z(\cdot)$ satisfies the linear, inhomogeneous differential equation

$$\dot{z}(t) = Az(t) + Bv(t), \quad z(0) = 0. \tag{4.10}$$

Now we introduce $\mu(t) \in \mathbb{R}^n$ and the Hamilton function $\mathcal{H}(x(t), u(t), \mu(t))$ by

$$\begin{aligned} \mathcal{H}(x(t), u(t), \mu(t)) = \frac{1}{2}\left(x(t)^{\mathsf{T}}Qx(t) + 2x(t)^{\mathsf{T}}Su(t) + u(t)^{\mathsf{T}}Ru(t)\right) \\ + \mu(t)^{\mathsf{T}}\left(Ax(t) + Bu(t)\right) \end{aligned}$$

Then we rewrite the cost functional as

$$\mathcal{J}(u) = \frac{1}{2}x(t_{\text{f}})^{\mathsf{T}}Mx(t_{\text{f}}) + \int_0^{t_{\text{f}}} \left(\mathcal{H}(x(t), u(t), \mu(t)) - \mu(t)^{\mathsf{T}}\dot{x}(t)\right)\mathrm{d}t.$$

Analogously, for u_* and x_* we obtain

$$\mathcal{J}(u_*) = \frac{1}{2}x_*(t_{\text{f}})^{\mathsf{T}}Mx_*(t_{\text{f}}) + \int_0^{t_{\text{f}}} \left(\mathcal{H}(x_*(t), u_*(t), \mu(t)) - \mu(t)^{\mathsf{T}}\dot{x}_*(t)\right)\mathrm{d}t.$$

Subtracting both equations gives

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(u_*) = \frac{1}{2}\left(x(t)^{\mathsf{T}}Mx(t) - x_*(t)^{\mathsf{T}}Mx_*(t)\right)\Big|_{t=t_{\text{f}}} \\ + \int_0^{t_{\text{f}}} \left(\mathcal{H}(x(t), u(t), \mu(t)) - \mathcal{H}(x_*(t), u_*(t), \mu(t))\right)\mathrm{d}t \\ + \int_0^{t_{\text{f}}} \mu(t)^{\mathsf{T}} \underbrace{\left(\dot{x}_*(t) - \dot{x}(t)\right)}_{=-\varepsilon\dot{z}(t)}\mathrm{d}t \end{aligned} \tag{4.11}$$

Now we consider the three terms on the right-hand side of (4.11) separately. By plugging in $x(t) = x_*(t) + \varepsilon z(t)$, for the first term we get

$$\begin{aligned} &x(t)^\mathsf{T} M x(t) - x_*(t)^\mathsf{T} M x_*(t) \\ &\quad = x_*(t)^\mathsf{T} M x_*(t) + 2\varepsilon x_*(t)^\mathsf{T} M z(t) + \varepsilon^2 z(t)^\mathsf{T} M z(t) - x_*(t)^\mathsf{T} M x_*(t) \\ &\qquad\qquad = 2\varepsilon x_*(t)^\mathsf{T} M z(t) + \mathcal{O}\big(\varepsilon^2\big). \end{aligned}$$

For the second term we get

$$\begin{aligned} &\mathcal{H}(x(t), u(t), \mu(t)) - \mathcal{H}(x_*(t), u_*(t), \mu(t)) \\ &\quad = \frac{1}{2}\left(x(t)^\mathsf{T} Q x(t) + 2x(t)^\mathsf{T} S u(t) + u(t)^\mathsf{T} R u(t)\right) + \mu(t)^\mathsf{T}(Ax(t) + Bu(t)) \\ &\qquad - \frac{1}{2}\left(x_*(t)^\mathsf{T} Q x_*(t) + 2x_*(t)^\mathsf{T} S u_*(t) + u(t)_*^\mathsf{T} R u_*(t)\right) \\ &\qquad\qquad - \mu(t)^\mathsf{T}(Ax_*(t) + Bu_*(t)). \end{aligned}$$

Now we set

$$u(t) = u_*(t) + \varepsilon v(t), \quad x(t) = x_*(t) + \varepsilon z(t)$$

and obtain (after a lengthy calculation)

$$\begin{aligned} &\mathcal{H}(x(t), u(t), \mu(t)) - \mathcal{H}(x_*(t), u_*(t), \mu(t)) \\ &\quad = \varepsilon\big(x_*(t)^\mathsf{T} Q z(t) + x_*(t)^\mathsf{T} S v(t) + u_*(t)^\mathsf{T} S^\mathsf{T} z(t) + u_*(t)^\mathsf{T} R v(t) \\ &\qquad + \mu(t)^\mathsf{T} A z(t) + \mu(t)^\mathsf{T} B v(t)\big) + \mathcal{O}\big(\varepsilon^2\big) \\ &\quad = \varepsilon\left(\big(x_*(t)^\mathsf{T} Q + u_*(t)^\mathsf{T} S^\mathsf{T} + \mu(t)^\mathsf{T} A\big) z(t)\right. \\ &\qquad \left. + \big(x_*(t)^\mathsf{T} S + u_*(t)^\mathsf{T} R + \mu(t)^\mathsf{T} B\big) v(t)\right) + \mathcal{O}\big(\varepsilon^2\big). \end{aligned}$$

Further, for the last term in (4.11), with partial integration we get

$$\begin{aligned} -\int_0^{t_\mathrm{f}} \varepsilon \mu(t)^\mathsf{T} \dot{z}(t)\mathrm{d}t &= -\varepsilon \mu(t)^\mathsf{T} z(t)\big|_0^{t_\mathrm{f}} + \varepsilon \int_0^{t_\mathrm{f}} \dot{\mu}(t)^\mathsf{T} z(t)\mathrm{d}t \\ &= -\varepsilon \mu(t_\mathrm{f})^\mathsf{T} z(t_\mathrm{f}) + \varepsilon \int_0^{t_\mathrm{f}} \dot{\mu}(t)^\mathsf{T} z(t)\mathrm{d}t, \end{aligned}$$

where we use that $z(0) = 0$ due to (4.10). Altogether we have

$$\begin{aligned} &\mathcal{J}(u) - \mathcal{J}(u_*) \\ &\qquad = \varepsilon\Bigg(\int_0^{t_\mathrm{f}} \big((x_*(t)^T Q + u_*(t)^\mathsf{T} S^T + \mu(t)^\mathsf{T} A + \dot{\mu}(t)^\mathsf{T}) z(t) \\ &\qquad\qquad + (x_*(t)^\mathsf{T} S + u_*(t)^\mathsf{T} R + \mu(t)^\mathsf{T} B) v(t)\big)\mathrm{d}t \\ &\qquad\qquad - \mu(t_\mathrm{f})^\mathsf{T} z(t_\mathrm{f}) + x_*(t_\mathrm{f})^\mathsf{T} M z(t_\mathrm{f})\Bigg) + \mathcal{O}\big(\varepsilon^2\big). \end{aligned} \tag{4.12}$$

A necessary condition for a minimum of $\mathcal{J}$ is that all directional derivatives of $\mathcal{J}$ from u_* vanish, i. e.,

$$0 = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\mathcal{J}(u_* + \varepsilon v) - \mathcal{J}(u_*) \right) \quad \text{for all } v \in U_{\text{ad}}.$$

If one chooses μ as solution of the linear inhomogeneous differential equation

$$\dot{\mu}(t) = -(A^{\mathsf{T}} \mu(t) + Q x_*(t) + S u_*(t)) \tag{4.13}$$

with the "terminal condition"

$$\mu(t_{\text{f}}) = M x_*(t_{\text{f}}) \tag{4.14}$$

then from (4.12) we necessarily get

$$x_*(t)^{\mathsf{T}} S + u_*(t)^{\mathsf{T}} R + \mu(t)^{\mathsf{T}} B \equiv 0 \quad \forall t \in [0, t_{\text{f}}] \tag{4.15}$$

If we then take (4.13), (4.14), (4.15) and the first equation in (4.6), then we get the two-point boundary value problem (4.8), (4.9). □

Remark 4.6: The boundary value problem (4.8) is also obtained, if Pontryagin's maximum principle is applied to the LQR problem. The first row of (4.8) corresponds to the constraint, i. e., $\mathcal{H}_\mu(x(t), u(t), \mu(t)) = \dot{x}(t)$. The second row follows from the adjoint equation (Theorem 4.2 (ii)), while the last row of (4.8) follows from the necessary condition for a minimum which is

$$\mathcal{H}_u(x(t), u(t), \mu(t)) = 0.$$

Note that u is considered unbounded here. The boundary conditions are exactly the initial value of the dynamic constraint in (4.7) and the transversality condition in the form (iii)' for non-vanishing h_{f}.

Further note that for the derivation of the necessary optimality conditions, no conditions on the matrices M, Q, R, S have been necessary.

To obtain sufficient optimality conditions, we basically use that $\mathcal{J}_{uu} \geqslant 0$ must hold for a minimum. To achieve this, we assume that M and $\left[\begin{smallmatrix} Q & S \\ S^{\mathsf{T}} & R \end{smallmatrix}\right]$ are at least symmetric and positive semi-definite (though one can also obtain sufficient conditions under much weaker conditions.) As sufficient optimality condition, we obtain the following result:

Theorem 4.7: Let x_*, μ, u_* be chosen such that $\left[x_*^\mathsf{T}, \mu^\mathsf{T}, u_*^\mathsf{T}\right]^T$ solves the linear boundary value problem (4.8), (4.9). Further let $\left[\begin{smallmatrix} Q & S \\ S^\mathsf{T} & R \end{smallmatrix}\right]$ and M be positive semi-definite. Then it holds that

$$\mathcal{J}(u) \geqslant \mathcal{J}(u_*) \tag{4.16}$$

for all $u \in U_{\text{ad}}$.

Proof. We proof the theorem as in convex optimization. Define

$$\Phi(s) := \mathcal{J}\left(su_* + (1-s)v\right).$$

Note that from the linearity of $\dot{x}(t) = Ax(t) + Bu(t)$, by setting $u := su_* + (1-s)v$ we obtain the solution trajectory $x = sx_* + (1-s)z$, where z is the solution of (4.6) corresponding to v. The claim of the theorem is equivalent to the statement that $\Phi(s)$ attains its minimum for $s = 1$ for all x, u that satisfy (1.4). Since $\Phi(s)$ is quadratic in s, $\Phi(s)$ has a minimum for $s = 1$, if and only if

$$\left.\frac{\mathrm{d}\Phi(s)}{\mathrm{d}s}\right|_{s=1} = 0, \qquad \left.\frac{\mathrm{d}^2\Phi(s)}{\mathrm{d}s^2}\right|_{s=1} \geqslant 0.$$

For each symmetric matrix K, we have the identity

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1}{2}(sq+(1-s)p)^\mathsf{T}K(sq+(1-s)p)\right)\Big|_{s=1} \\ = \left(sq^\mathsf{T}Kq - sp^\mathsf{T}Kq + (1-s)p^\mathsf{T}Kq - (1-s)p^\mathsf{T}Kp\right)\big|_{s=1} \\ = q^\mathsf{T}Kq - p^\mathsf{T}Kq = (q-p)^\mathsf{T}Kq.\end{aligned}$$

Thus, from the condition on the first derivative, we obtain the expression

$$\begin{aligned}\left.\frac{\mathrm{d}\Phi(s)}{\mathrm{d}s}\right|_{s=1} &= (x_*(t)-z(t))^\mathsf{T}Mx_*(t)\big|_{t=t_\mathrm{f}} + \int_0^{t_\mathrm{f}} (x_*(t)-z(t))^\mathsf{T}Qx_*(t) \\ &+ u_*(t)^\mathsf{T}S^\mathsf{T}(x_*(t)-z(t)) + (u_*(t)-v(t))^\mathsf{T}S^\mathsf{T}x_*(t) + (u_*(t)-v(t))^\mathsf{T}Ru_*(t)\mathrm{d}t\end{aligned} \tag{4.17}$$

Left-multiplying the second equation of (4.8) with $x_*(t)^\mathsf{T}$ and putting in the first and thereafter the third equation of (4.8), then one obtains

$$\begin{aligned}x_*(t)^\mathsf{T}Qx_*(t) &= -x_*(t)^\mathsf{T}A^\mathsf{T}\mu(t) - x_*(t)^\mathsf{T}Su_*(t) - x_*(t)^\mathsf{T}\dot{\mu}(t) \\ &= u_*(t)^\mathsf{T}B^\mathsf{T}\mu(t) - \dot{x}_*(t)^\mathsf{T}\mu(t) - x_*(t)^\mathsf{T}Su_*(t) - x_*(t)^\mathsf{T}\dot{\mu}(t) \\ &= -u_*(t)^\mathsf{T}S^\mathsf{T}x_*(t) - u_*(t)^\mathsf{T}Ru_*(t) - \dot{x}_*(t)^\mathsf{T}\mu(t) - x_*(t)^\mathsf{T}Su_*(t) \\ &\quad - x_*(t)^\mathsf{T}\dot{\mu}(t)\end{aligned} \tag{4.18}$$

Analogously, after left-multiplying with $z(t)^\mathsf{T}$ we obtain

$$\begin{aligned} z(t)^\mathsf{T} Q x_*(t) &= -z(t)^\mathsf{T} A^\mathsf{T} \mu - z(t)^\mathsf{T} S u_*(t) - z(t)^\mathsf{T} \dot{\mu}(t) \\ &= v(t)^\mathsf{T} B^\mathsf{T} \mu(t) - \dot{z}(t)^\mathsf{T} \mu(t) - z(t)^\mathsf{T} S u_*(t) - z(t)^\mathsf{T} \dot{\mu}(t) \\ &= -v(t)^\mathsf{T} S^\mathsf{T} x_*(t) - v(t)^\mathsf{T} R u_*(t) - \dot{z}^\mathsf{T} \mu(t) - z(t)^\mathsf{T} S u_*(t) \\ &\quad - z(t)^\mathsf{T} \dot{\mu}(t) \end{aligned} \tag{4.19}$$

Putting in (4.18), (4.19) into (4.17) yields

$$\begin{aligned} \left.\frac{\mathrm{d}\Phi(s)}{\mathrm{d}s}\right|_{s=1} &= (x_*(t) - z(t))^\mathsf{T} M x_*(t)\big|_{t=t_\mathrm{f}} \\ &\quad + \int_0^{t_\mathrm{f}} z(t)^\mathsf{T} \dot{\mu}(t) + \dot{z}(t)^\mathsf{T} \mu(t) - x_*(t)^\mathsf{T} \dot{\mu}(t) - \dot{x}_*(t)^\mathsf{T} \mu(t) \mathrm{d}t \\ &= (x_*(t) - z(t))^\mathsf{T} M x_*(t)\big|_{t=t_\mathrm{f}} + z(t)^\mathsf{T} \mu(t)\big|_{t=0}^{t=f_\mathrm{f}} - x_*(t)^\mathsf{T} \mu(t)\big|_{t=0}^{t=t_\mathrm{f}} \end{aligned}$$

Now by (4.9), $z(0) = x^0 = x_*(0)$ and $\mu(t_\mathrm{f}) = M x_*(t_\mathrm{f})$, so $\left.\frac{\mathrm{d}\Phi(s)}{\mathrm{d}s}\right|_{s=1} = 0$. Using the identity

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\left(\frac{1}{2}(sq + (1-s)p)^T K (sq + (1-s)p)\right) = (q-p)^\mathsf{T} K (q-p)$$

for a symmetric matrix K, we obtain for the second derivative of $\Phi(\cdot)$ that

$$\begin{aligned} \left.\frac{\mathrm{d}^2\Phi}{\mathrm{d}s^2}\right|_{s=1} &= (x_*(t) - z(t))^\mathsf{T} M (x_*(t) - z(t))\big|_{t=t_\mathrm{f}} \\ &\quad + \int_0^{t_\mathrm{f}} \begin{bmatrix} x_*(t) - z(t) \\ u_*(t) - v(t) \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_*(t) - z(t) \\ u_*(t) - v(t) \end{bmatrix} \mathrm{d}t \geqslant 0. \end{aligned}$$

Here, the nonnegativity follows from the positive semi-definiteness of M and $\left[\begin{smallmatrix} Q & S \\ S^\mathsf{T} & R \end{smallmatrix}\right]$. □

Now we have obtained a relation between the solution of the optimal control problem and the solution of the two-point boundary value problem. In principle, we could obtain the optimal control u_* of the LQR problem by solving (4.8) and (4.9). Significantly simpler and cheaper from the numerical point of view is the following approach.

4.2 Solution of the LQR Problem by Riccati Equations

4.2.1 The Finite Time Horizon Problem

The assumptions of Theorem 4.7 imply that $R \geqslant 0$ should be chosen in the cost functional. Moreover, often one even has $R > 0$. Otherwise, there would be costfree control parameters which is often not sensible. So in the following we restrict ourselves to a positive definite weight matrix for the control. In this case, R is invertible and the third equation in (4.8) can be resolved with respect to u. Ones obtains

$$u(t) = -R^{-1}\big(S^{\mathsf{T}}x(t) + B^{\mathsf{T}}\mu(t)\big). \tag{4.20}$$

and thus,

$$\begin{bmatrix}\dot{x}(t)\\ \dot{\mu}(t)\end{bmatrix} = H\begin{bmatrix}x(t)\\ \mu(t)\end{bmatrix}, \qquad x(0) = x^0, \quad \mu(t_{\mathrm{f}}) = Mx(t_{\mathrm{f}}), \tag{4.21}$$

where

$$\mathcal{H} = \begin{bmatrix} A - BR^{-1}S^{\mathsf{T}} & -BR^{-1}B^{\mathsf{T}} \\ -\big(Q - SR^{-1}S^{\mathsf{T}}\big) & -\big(A - BR^{-1}S^{\mathsf{T}}\big)^{\mathsf{T}}\end{bmatrix}. \tag{4.22}$$

In the sequel, we use the following abbreviations for better readability:

$$F := A - BR^{-1}S^{\mathsf{T}}, \quad G := BR^{-1}B^{\mathsf{T}}, \quad H := Q - SR^{-1}S^{\mathsf{T}}, \tag{4.23}$$

such that with (4.20), $Ax + Bu$ becomes $Ax - G\mu$ and we can write $\mathcal{H} = \left[\begin{smallmatrix} F & -G \\ -H & -F^{\mathsf{T}}\end{smallmatrix}\right]$.

Definition 4.8: A matrix $\mathcal{H} \in \mathbb{R}^{2n\times 2n}$ is called *Hamiltonian*, if

$$(\mathcal{H}J) = (\mathcal{H}J)^{\mathsf{T}}, \quad \text{where } J = \begin{bmatrix} 0 & I_n \\ -I_n & 0\end{bmatrix}. \tag{4.24}$$

From (4.24) it follows directly that Hamiltonian matrices have a certain explicit block structure.

Lemma 4.9: The matrix $H \in \mathbb{R}^{2n\times 2n}$ is Hamiltonian, if and only if

$$\mathcal{H} = \begin{bmatrix} F & -G \\ -H & -F^{\mathsf{T}}\end{bmatrix}, \qquad \text{where } G = G^{\mathsf{T}},\ H = H^{\mathsf{T}}.$$

This means that (4.21) is a boundary value problem for a linear differential equation with Hamiltonian coefficient matrix.

With the *ansatz* $\mu(t) = X(t)x(t)$, from (4.21) und with

$$\begin{aligned}\dot{\mu}(t) &= \dot{X}(t)x(t) + X(t)\dot{x}(t),\\ \mu(t_\mathrm{f}) &= X(t_\mathrm{f})x(t_\mathrm{f}),\end{aligned}$$

and the terminal condition $X(t_\mathrm{f}) = M$ we get

$$\begin{aligned}\dot{x}(t) &= Fx(t) - G\mu(t) = Fx(t) - GX(t)x(t) = (F - GX(t))x(t),\\ \dot{\mu}(t) &= -Hx(t) - F^\mathsf{T}\mu(t) = -Hx(t) - F^\mathsf{T}X(t)x(t)\\ &= \dot{X}(t)x(t) + X(t)\dot{x}(t)\\ &= \dot{X}(t)x(t) + X(t)Fx(t) - X(t)GX(t)x(t).\end{aligned}$$

From the latter equation we obtain

$$\dot{X}(t)x(t) = -\big(H + F^\mathsf{T}X(t) + X(t)F - X(t)GX(t)\big)x(t). \tag{4.25}$$

Thus, if $X(\cdot)$ satisfies the *Riccati differential equation*

$$\dot{X}(t) = -\mathcal{R}(X(t)) := -\big(H + F^\mathsf{T}X(t) + X(t)F - X(t)GX(t)\big),\ t \in [0,\, t_\mathrm{f}] \tag{4.26}$$

with the terminal condition

$$X(t_\mathrm{f}) = M, \tag{4.27}$$

then (4.25) is satisfied. One can show that (4.26), (4.27) has a unique solution on $[0,\, t_\mathrm{f}]$. Since with $X(\cdot)$ also $X(\cdot)^\mathsf{T}$ is a solution, together with uniqueness, it follows that $X(t) = X(t)^\mathsf{T}$ for all $t \in [0,\, t_\mathrm{f}]$. Further, one can show that $X(t)$ is positive definite on the whole time interval. The proofs can be found in [KK85]. Now one obtains the following result that can also be shown for LTV systems (see [KK85]).

Theorem 4.10: If $R > 0$, then under the assumptions of Theorem 4.7, one can show that the optimal control $u_*(\cdot)$ that solves the LQR problem is given by

$$u_*(t) = -R^{-1}\big(S^\mathsf{T} + B^\mathsf{T}X_*(t)\big)x_*(t) \quad \forall\, t \in [0,\, t_\mathrm{f}], \tag{4.28}$$

where $X_*(\cdot)$ is the unique solution of the Riccati differential equation (4.26) with terminal condition (4.27).

The “optimal costs” are

$$\mathcal{J}(u_*) = \frac{1}{2}\big(x^0\big)^\mathsf{T}X_*(0)x^0. \tag{4.29}$$

Proof. The form of the optimal control (4.28) follows directly from the boundary value problem (4.8)–(4.9) and by putting in $\mu(t) = X_*(t)x_*(t)$ in (4.20). The optimal costs are obtained by considering the value function

$$V(x(t)) := x(t)^\mathsf{T} X(t) x(t),$$

where $X(\cdot)$ is the solution of the Riccati differential equation (4.26) ist. By putting in $\dot{x}(t) = Ax(t) + Bu(t)$, (4.26) and (4.28) we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} V(x_*(t)) &= \dot{x}_*(t)^\mathsf{T} X_*(t) x_*(t) + x_*(t)^\mathsf{T} \dot{X}_*(t) x_*(t) + x_*(t)^\mathsf{T} X_*(t) \dot{x}_*(t) \\
&= 2\dot{x}_*(t)^\mathsf{T} X_*(t) x_*(t) + x_*(t)^\mathsf{T} \dot{X}_*(t) x_*(t) \\
&= 2x_*(t)^\mathsf{T} A^\mathsf{T} X_*(t) x_*(t) + 2u_*(t)^\mathsf{T} B^\mathsf{T} X_*(t) x_*(t) \\
&\quad - x_*(t)^\mathsf{T} (H + F^\mathsf{T} X_*(t) + X_*(t) F - X_*(t) G X_*(t)) x_*(t) \\
&= -2x_*(t)^\mathsf{T} X_*(t) B R^{-1} B^\mathsf{T} X_*(t) x_*(t) - x_*(t)^\mathsf{T} H x_*(t) \\
&\quad + x_*(t)^\mathsf{T} G x_*(t) \\
&= -x_*(t)^\mathsf{T} H x_*(t) - x_*(t)^\mathsf{T} G x_*(t).
\end{aligned}$$

With this, the cost functional for the optimal control can be written as

$$\begin{aligned}
\mathcal{J}(u_*) &= \frac{1}{2} x_*(t_\mathrm{f})^\mathsf{T} M x_*(t_\mathrm{f}) - \frac{1}{2} \int_0^{t_\mathrm{f}} \frac{\mathrm{d}}{\mathrm{d}t} V(x_*(t)) \mathrm{d}t \\
&= \frac{1}{2} \left(x_*(t_\mathrm{f})^\mathsf{T} M x_*(t_\mathrm{f}) - V(x_*(t))|_{t=0}^{t=t_\mathrm{f}} \right) \\
&= \frac{1}{2} \left(x_*(t_\mathrm{f})^\mathsf{T} M x_*(t_\mathrm{f}) - x_*(t_\mathrm{f})^\mathsf{T} X_*(t_\mathrm{f}) x_*(t_\mathrm{f}) + x_*(0)^\mathsf{T} X_*(0) x_*(0) \right) \\
&= \frac{1}{2} x_*(0)^\mathsf{T} X_*(0) x_*(0),
\end{aligned}$$

where we use that $X(t_\mathrm{f}) = M$. □

Remark 4.11: Note that the optimal control in $u_*(\cdot)$ in (4.28) is given as linear state feedback. So one obtains a closed-loop system, even though this is not directly clear from the boundary value problem (4.8)–(4.9). The ansatz $\mu(t) = X(t)x(t)$ is motivated by our goal to achieve a feedback control.

With this, one has an alternative for solving the LQR problem, namely be solving the "terminal value problem" for the Riccati differential equation. To do so, the equation can be vectorized using the vec operator and the Kronecker product and usind standard methods for initial value problems, where by a

transformation $t \to t_\mathrm{f} - t$, a terminal value problem can be turned into an initial value problem. On the other hand, it is much more advisable to use special methods for Riccati differential equations that exploit the given structure [CL90, Die92, KL85].

4.2.2 The Infinite Time Horizon Problem

As already discussed in the introduction, it is often sufficient to reach the target asymptotically. This leads to the question of an optimal stabilization with respect to the cost functional in (4.7) with $t_\mathrm{f} = \infty$, where we now set $M = 0$. To be able to achieve a stabilization, we must assume stabilizability. Since we aim again for a solution in terms of a feedback control, (4.20) motivates the ansatz $\mu(t) = Xx(t)$ for a constant matrix $X = X^\mathsf{T} \in \mathbb{R}^{n\times n}$. With this ansatz one obtains (analogously to the finite time-horizon case)

$$\begin{aligned} \dot{x}(t) &= (F - GX)x(t), \\ \dot{\mu}(t) &= -Hx(t) - F^\mathsf{T}\mu(t) = -Qx - F^\mathsf{T}Xx(t) \\ &= X\dot{x}(t) = X(F - GX)x(t). \end{aligned} \tag{4.30}$$

Now, the last equation is equivalent to

$$\left(H + F^\mathsf{T}X + XF - XGX\right) x(t) = 0 \quad \forall\, t \in [0,\, \infty). \tag{4.31}$$

Thus, if X satisfies the *algebraic Riccati equation (ARE)*

$$0 = \mathcal{R}(X) := H + F^\mathsf{T}X + XF - XGX, \tag{4.32}$$

then (4.31) is satisfied. However, note that in contrast to the Riccati differential equation, (4.32) has in general infinitely many solutions and even nonsymmetric solutions are possible. The structure of the solution set of (4.32) has been addressed in many research articles and is most completely described in [LR95]. We will see in the sequel, that we need a particular solution of the ARE. Since the solution trajectory of the state that is generated by our approach satisfies the linear homogeneous differential equation (4.30) such that the solution can be written as $x(t) = \mathrm{e}^{(F-GX)t}x^0$, we must necessarily have

$$\Lambda(F - GX) \subset \mathbb{C}^-,$$

otherwise, we would not achieve a stabilization of the system. This motivates the following definition.

Definition 4.12: A solution $X \in \mathbb{R}^{n\times n}$ of the ARE (4.32) is called *stabilizing*, if $\Lambda(F - GX) \subset \mathbb{C}^-$.

In other words, to compute a stabilizing feedback matrix $K \in \mathbb{R}^{m\times n}$ with the help of he LQR problem, we need a stabilizing solution of the ARE, since

$$\begin{aligned} F - GX &= F - BR^{-1}B^\mathsf{T}X \\ &= A - BR^{-1}\big(B^\mathsf{T}X + S^\mathsf{T}\big) \\ &= A + BK \quad \text{with } K := -R^{-1}\big(B^\mathsf{T}X + S^\mathsf{T}\big). \end{aligned}$$

It remains to discuss when and how a stabilizing solution can be computed. In the following, this solution will be denoted by X_*.

First, let X be an arbitrary solution of the ARE (4.32) and $T := \left[\begin{smallmatrix} I_n & 0 \\ X & I_n \end{smallmatrix}\right]$. Then for the Hamiltonian matrix $\mathcal{H}$ from (4.22) it holds that

$$\begin{aligned} T^{-1}\mathcal{H}T &= \begin{bmatrix} I_n & 0 \\ -X & I_n \end{bmatrix} \begin{bmatrix} F & -G \\ -H & -F^\mathsf{T} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ X & I_n \end{bmatrix} \\ &= \begin{bmatrix} F - GX & -G \\ -\mathcal{R}(X) & -(F - GX)^\mathsf{T} \end{bmatrix} = \begin{bmatrix} F - GX & -G \\ 0 & -(F - GX)^\mathsf{T} \end{bmatrix} \end{aligned}$$

This implies

$$\mathcal{H}\begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix}(F - GX), \tag{4.33}$$

i. e., $\Lambda(F - GX) \subset \Lambda(\mathcal{H})$ and the columns of $\left[\begin{smallmatrix} I_n \\ X \end{smallmatrix}\right]$ span an $\mathcal{H}$-invariant subspace. This is true for every solution of the ARE, for the stabilizing solution we need an n-dimensional $\mathcal{H}$-invariant subspace corresponding to the eigenvalues in the open left complex half-plane. First we discuss the question whether such a subspace actually exists. Therefore, we need a few properties of the spectrum of Hamiltonian matrices.

Lemma 4.13: If $\mathcal{H} \in \mathbb{R}^{2n\times 2n}$ is Hamiltonian and $\lambda \in \Lambda(\mathcal{H})$ with corresponding right eigenvector $x \in \mathbb{C}^{2n}$, then $-\overline{\lambda} \in \Lambda(\mathcal{H})$ with corresponding left eigenvector Jx, where $J = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$.

Proof. Homework. □

Since for real matrices, with λ also $\overline{\lambda}$ is an eigenvalue, eigenvalues of Hamiltonian matrices always occur in quadruples of λ, $\overline{\lambda}$, $-\lambda$, $\overline{\lambda}$, except if they are real

or imaginary, in which case they appear as pairs λ, $-\lambda$. Overall, the spectrum of a Hamiltonian matrix can be written as

$$\Lambda(\mathcal{H}) = \{\lambda_1, \ldots, \lambda_n\} \cup \{-\lambda_1, \ldots, -\lambda_n\} \tag{4.34}$$

with $\mathrm{Re}(\lambda_j) \leqslant 0$ for all $j \in \{1, \ldots, n\}$. To guarantee the existence of a stabilizing solution of the ARE, it follows from (4.33) that the corresponding Hamiltonian matrix may not have any eigenvalues on the imaginary axis in which case exactly n eigenvalues of $\mathcal{H}$ are in the left complex half-plane and an n-dimensional $\mathcal{H}$-invariant subspace associated with these eigenvalues exists. This can already be achieved with minimal requirements on the LQR problem as the following result shows.

Theorem 4.14: Let $\mathcal{H} = \begin{bmatrix} F & -G \\ -H & -F^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$ Hamiltonian, where (F, G) is stabilizable and (F, H) is detectable and $G,\ H \geqslant 0$. Then

$$\mathrm{Re}(\lambda) \neq 0 \quad \text{for all } \lambda \in \Lambda(H).$$

Proof. Assume that $\lambda = \imath\omega \in \Lambda(\mathcal{H})$. Because of Lemma 4.13 we can assume w. l. o. .g that $\omega \geqslant 0$. Let $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \neq 0$ with $x_1,\ x_2 \in \mathbb{C}^n$ an eigenvector associated with $\imath\omega$. Then we get

$$\mathcal{H} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \imath\omega \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{4.35}$$

Left-multiplying $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]^{\mathsf{H}}$ leads to

$$\begin{aligned} \begin{bmatrix} x_2^{\mathsf{H}} & x_1^{\mathsf{H}} \end{bmatrix} \mathcal{H} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \underbrace{x_2^{\mathsf{H}} F x_1}_{:=\alpha} - \underbrace{x_2^{\mathsf{H}} G x_2}_{:=\gamma} - \underbrace{x_1^{\mathsf{H}} H x_1}_{:=\beta} - \underbrace{x_2^{\mathsf{H}} F x_1}_{=\alpha} \\ &= \begin{bmatrix} x_2^{\mathsf{H}} & x_1^{\mathsf{H}} \end{bmatrix} \imath\omega \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \imath\omega \big(\underbrace{x_2^{\mathsf{H}} x_1}_{=:\zeta} + \underbrace{x_1^{\mathsf{H}} x_2}_{=\bar{\zeta}} \big) = 2\imath\omega\, \mathrm{Re}(\zeta). \end{aligned}$$

Since $\beta + \gamma \geqslant 0$ is the real part of the first expression, $\beta \geqslant 0$, $\gamma \geqslant 0$ and since the second expression is purely imaginary, we get $\beta = \gamma = 0$. Hence,

$$x_1^{\mathsf{H}} H = 0, \quad Hx_1 = 0 \quad \text{and} \quad x_2^{\mathsf{H}} G = 0, \quad Gx_2 = 0.$$

From the first equation in (4.35) one obtains

$$\imath\omega x_1 = Fx_1 - Gx_2 = Fx_1$$

and thus,

$$\begin{bmatrix} F - \imath\omega I_n \\ H \end{bmatrix} x_1 = 0. \tag{4.36}$$

Analogously, with the help of the second equation in (4.35) we obtain

$$x_2^{\mathsf{H}} \begin{bmatrix} F - \imath\omega I_n & G \end{bmatrix} = 0. \tag{4.37}$$

Since $x_1 \neq 0$ or $x_2 \neq 0$, (4.36) contradicts the assumed detectability or (4.37) contradicts the assumed stabilizability. □

The stabilizing solution of the ARE can now be obtained as follows:.

Lemma 4.15: Let $U = \begin{bmatrix} u_1, & \ldots, & u_n \end{bmatrix}$, $V = \begin{bmatrix} v_1, & \ldots, & v_n \end{bmatrix} \in \mathbb{R}^{n\times n}$ be such that $\operatorname{span}\{\left[\begin{smallmatrix} u_1 \\ v_1 \end{smallmatrix}\right], \ldots, \left[\begin{smallmatrix} u_n \\ v_n \end{smallmatrix}\right]\}$ is the n-dimensional $\mathcal{H}$-invariant subspace associated with $\{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{C}^-$ (with the same notation as in (4.34)). If U is invertible, then $X_* = VU^{-1}$ is the stabilizing solution of the ARE (4.32).

Proof. By the assumption we have

$$\begin{bmatrix} F & -G \\ -H & -F^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} Z, \quad \Lambda(Z) = \{\lambda_1, \ldots, \lambda_n\}.$$

Left-multiplying the first block row with U^{-1} gives

$$U^{-1}FU - U^{-1}GV = Z.$$

Then from the second block row we get

$$-HU - F^{\mathsf{T}}V = VZ = VU^{-1}FU - VU^{-1}GV.$$

Right-multiplying this equation with U^{-1}, then we get that $X_* = VU^{-1}$ solves the ARE. Moreover,

$$\operatorname{im} \begin{bmatrix} U \\ V \end{bmatrix} = \operatorname{im} \begin{bmatrix} U \\ V \end{bmatrix} U^{-1} = \operatorname{im} \begin{bmatrix} I_n \\ X_* \end{bmatrix}.$$

So X_* is the stabilizing solution of the ARE. □

The following results summarize some further properties of the stabilizing solution X_* of the ARE.

Lemma 4.16: Let $U = \begin{bmatrix} u_1, & \dots, & u_n \end{bmatrix}, V = \begin{bmatrix} v_1, & \dots, & v_n \end{bmatrix} \in \mathbb{R}^{n\times n}$ be such that $\operatorname{span}\{\left[\begin{smallmatrix} u_1 \\ v_1 \end{smallmatrix}\right], \dots, \left[\begin{smallmatrix} u_n \\ v_n \end{smallmatrix}\right]\}$ is the n-dimensional $\mathcal{H}$-invariant subspace associated with $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{C}^-$ (with the same notation as in (4.34)). Then $V^\mathsf{T}U$ is symmetric. If further, G and H are positive semi-definite, then $V^\mathsf{T}U \geqslant 0$.

Proof. By assumption we have

$$\mathcal{H}\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} Z \tag{4.38}$$

with $\Lambda(Z) = \{\lambda_1, \dots, \lambda_n\}$. Left-multiplying the first block row of (4.38) with V^T results in

$$V^\mathsf{T}FU - V^\mathsf{T}GV = V^\mathsf{T}UZ.$$

Transposing the second block row of (4.38) and right-multiplying with U gives

$$-U^\mathsf{T}HU - V^\mathsf{T}AU = Z^\mathsf{T}V^\mathsf{T}U. \tag{4.39}$$

Finally, adding (4.2.2) and (4.39) leads to

$$Z^\mathsf{T}V^\mathsf{T}U + V^\mathsf{T}UZ = -V^\mathsf{T}GV - U^\mathsf{T}HU. \tag{4.40}$$

This is a Lyapunov equation in the "unknown" $V^\mathsf{T}U$. Since by assumption Z is Hurwitz, with Theorem 3.6 we see that (4.40) has a unique solution and

$$V^\mathsf{T}U = \int_0^\infty \mathrm{e}^{Z^\mathsf{T}t}(V^\mathsf{T}GV + U^\mathsf{T}QU)\mathrm{e}^{Zt}\mathrm{d}t.$$

Since due to $\mathrm{e}^{Z^\mathsf{T}t} = (\mathrm{e}^{Zt})^\mathsf{T}$ and $G = G^\mathsf{T}$, $H = H^T$ the integrand is symmetric, $V^\mathsf{T}U$ is also symmetric. Moreover, if G and H are positive semi-definite, by Sylvester's law of intertia, this is also the case for the integrand and therefore, the entire right-hand side.

□

The next lemma shows that under our assumptions on the LQR problem. the invertibility of the matrix U in Lemma 4.15 is guaranteed.

Lemma 4.17: If U, V are as in Lemma 4.16, G, H positive semi-definite, and (F, G) is stabilizable, then U is invertible.

Proof. Suppose U is singular. Then there exists a vector $z \neq 0$ with $Uz = 0$. Left-multiplying the first block row of (4.38) with $(Vz)^\mathsf{T}$ and right-multiplying with z, then one gets

$$z^\mathsf{T} V^\mathsf{T} A \underbrace{Uz}_{=0} - z^\mathsf{T} V^\mathsf{T} G V z = z^\mathsf{T} V^\mathsf{T} U Z z. \tag{4.41}$$

Due to Lemma 4.16 it follows

$$z^\mathsf{T} V^\mathsf{T} G V z = -z^\mathsf{T} V^\mathsf{T} U Z z = -(Uz)^\mathsf{T} V Z z = 0.$$

Since G is symmetric positive semi-definite, we obtain $GVz = 0$. Thus the first block row of (4.38) right-multiplied with z yields $UZz = 0$. Since $z \in \ker U$ has be chosen arbitrarily, we get $Zz \in \ker U$ for all $z \in \ker U$ and hence, Z-invariance of $\ker U$. Thus, there exists an eigenvalue of Z, i. e., a λ_j ($1 \leqslant j \leqslant n$), and a vector $0 \neq z_j \in \ker U$ with $Zz_j = \lambda_j z_j$. Right-multiplication of the second block row of (4.38) with z_j gives

$$-Q \underbrace{Uz_j}_{=0} - F^\mathsf{T} V z_j = V Z z_j = \lambda_j V z_j,$$

i. e., $\left(\lambda_j I_n + F A^\mathsf{T}\right) V z_j = 0$. We have already shown $GVz = 0$ for arbitrary $z \in \ker U$. This holds particularly for z_j. With this it follows

$$(Vz_j)^\mathsf{T} \begin{bmatrix} \lambda_j I_n + F & G \end{bmatrix} = 0.$$

Due to the stabilizability of (F, G) with Theorem 2.15 we get $Vz_j = 0$. So we have

$$\begin{bmatrix} U \\ V \end{bmatrix} z_j = 0.$$

Since $z_j \neq 0$ it follows $\operatorname{rank} \left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right] < n$ which contradicts the assumption that the columns of $\left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right]$ span an n-dimensional $\mathcal{H}$-invariant subspace. □

With the two lemma above we can formulate the following result about the stabilizing solution of the ARE.

Theorem 4.18: Consider the ARE

$$0 = Q + F^\mathsf{T} X + X F - X G X \tag{4.42}$$

with $G = G^\mathsf{T} \geqslant 0$, $H = H^\mathsf{T} \geqslant 0$, and stabilizable (F, G). Assume further that the corresponding Hamiltonian matrix $\mathcal{H} = \left[\begin{smallmatrix} F & -G \\ -Q & -F^T \end{smallmatrix}\right]$ with spectrum as in

(4.34) has no imaginary eigenvalues and that the $\mathcal{H}$-invariant subspace with $\{\lambda_1, \ldots, \lambda_n\}$ be spanned by the columns of $\left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right]$ with $U, V \in \mathbb{R}^{n\times n}$. Then the ARE (4.42) has a unique stabilizing solution X_* which is symmetric and positive definite.

Proof. First, with Lemma 4.17 it holds that U is invertible. Thus there exists $X_* := VU^{-1}$ and the first block row of $\mathcal{H}\left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right] = \left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right] Z$, right-multiplied with U^{-1}, gives

$$F - GX_* = UZU^{-1}.$$

Thus, by Definition 4.12, X_* is stabilizing, since $\Lambda\big(UZU^{-1}\big) = \Lambda(Z) = \{\lambda_1, \ldots, \lambda_n\}$. The symmetry follows with Lemma 4.16, since with $V^\mathsf{T}U = U^\mathsf{T}V$ we obtain

$$X_* = VU^{-1} = U^{-\mathsf{T}}U^\mathsf{T}VU^{-1} = U^{-\mathsf{T}}V^\mathsf{T}UU^{-1} = \big(VU^{-1}\big)^\mathsf{T} = X_*^\mathsf{T}.$$

Under the given assumptions and with Lemma 4.16, $V^\mathsf{T}U \geqslant 0$. Due to the congruence

$$U^\mathsf{T}X_*U = V^\mathsf{T}U,$$

also $X_* \geqslant 0$ by Sylvester's inertia theorem. It remains to show uniqueness. Thus assume that X_* and $\widetilde{X}_*$ are two stabilizing solutions of the ARE (4.42), i. e.,

$$\begin{aligned} 0 &= H + F^\mathsf{T}X_* + X_*F - X_*GX_*, \\ 0 &= H + F^\mathsf{T}\widetilde{X}_* + \widetilde{X}_*F - \widetilde{X}_*G\widetilde{X}_*. \end{aligned}$$

Subtraction of both equations leads to

$$0 = \big(F - GX_*\big)^\mathsf{T}\big(X_* - \widetilde{X}_*\big) + \big(X_* - \widetilde{X}_*\big)\big(F - G\widetilde{X}_*\big).$$

This is a homogeneous Sylvester equation and since by assumption $\Lambda(F - GX_*)$, $\Lambda\big(F - G\widetilde{X}_*\big) \subset \mathbb{C}^-$, it follows from Theorem 3.6 that $X_* - \widetilde{X}_* = 0$, i. e., the uniqueness of the stabilizing solution. □

Bibliography

[BBQO07] S. Barrachina, P. Benner, and E. S. Quintana-Ortí. Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function. *Numer. Algorithms*, 46(4):351–368, 2007.

[CL90] C. Choi and A. J. Laub. Efficient matrix-valued algorithms for solving stiff Riccati differential equations. *IEEE Trans. Automat. Control*, 35(7):770–776, 1990.

[Die92] L. Dieci. Numerical integration of the differential Riccati equation and some related issues. *SIAM J. Numer. Anal.*, 29(3):781–815, 1992.

[KK85] H. W. Knobloch and H. Kwakernaak. *Lineare Kontrolltheorie*. Springer-Verlag, Berlin, 1985. In German.

[KL85] C. Kenney and R. B. Leipnik. Numerical integration of the differential matrix Riccati equation. *IEEE Trans. Automat. Control*, 30(10):962–970, 1985.

[KNVD85] J. Kautsky, N. K. Nichols, and P. Van Dooren. Robust pole assignment in linear state feedback. *Internat. J. Control*, 41(5):1129–1155, 1985.

[LR95] P. Lancaster and L. Rodman. *Algebraic Riccati Equations*. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1995.

[MS82] J. Macki and A. Strauss. *Introduction to Optimal Control Theory*. Springer-Verlag, 1982.

[PBGM62] L. S. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishenko. *The Mathematical Theory of Optimal Processes*. Interscience, New York, 1962.

[Pin93] E. R. Pinch. *Optimal Control and the Calculus of Variations*. Oxford University Press, Inc., Oxford, UK, 1993.

[Son98] E. D. Sontag. *Mathematical Control Theory*. Texts Appl. Math. Springer-Verlag, New York, NY, 2nd edition, 1998.