Mathematical Control Theory

Lars Grüne
Mathematisches Institut
Fakultät für Mathematik und Physik
Universität Bayreuth
95440 Bayreuth
lars.gruene@uni-bayreuth.de
http://num.math.uni-bayreuth.de

Lecture Notes

Summer Semester 2023

Preface

These lecture notes were written for a course with the same name held in the summer semester 2023 at the University of Bayreuth, Germany. Chapters 1–7 deal with topics from linear control theory, while Chapters 8–14 provide an introduction into Model Predictice Control for nonlinear systems. This is the second edition of these lecture notes written entirely in English. Compared to the first edition several small corrections and additions were made.

Parts of the first part of these notes were written on the basis of the lecture notes [2], the textbooks [19] and [14], as well as the monograph [11], which were extensively used also when they are not explicitly cited. The chapters on Model Predictive Control are revised excerpts from the monograph [8]. I would like to thank Thomas Lorenz, Lisa Krügel, and Jan Zetzmann as well as, as usual, all students who reported errors and inaccuracies in these notes during the lecture.

The most recent version of these lecture notes is always available via my home page (Google: Lars Gruene).

Bayreuth, October 2023 Lars Grüne

Contents

Chapter 1

Basics

Control systems are dynamical systems in continuous or discrete time, which depend on a parameter $u \in \mathbb{R}^m$, which may change with time and/or depending on the state of the system. This parameter has different interpretations. It can be considered as a control input, i.e., as a value that can be actively controlled from the outside (e.g., acceleration of a vehicle, investment into a firm) or as a perturbation that acts on a system (e.g., uneven road surface, time-varying exchange rates). The mathematical area that studies these systems is called *control theory*. Here, "control" is not to be understood in the sense of supervision but rather in the sense of taking influence on a system from the outside. One also talks about *open-loop control* if u only depends on time, and about *closed-loop control* or *regulation* if u depends on the current state of the system. In addition to *Mathematical Control Theory* one also uses the term *Mathematical System Theory*.

1.1 Linear Control Systems

In this lecture we will consider control systems in continuous and discrete time. In continuous time, control systems are described by ordinary or partial differential equations. In this lecture we mostly limit ourselves to ordinary differential equations. In this case, the control system is given by the equation

$$\dot{x}(t) = f(t, x(t), u(t)). \tag{1.1}$$

The variable $t \in \mathbb{R}$ in this equation will always be interpreted as *time* and the notation $\dot{x}(t)$ is short for the derivative with respect to time $d/dt\, x(t)$. The quantity $x(t) \in \mathbb{R}^n$ is called the *state* of the equation at time t and $u(t) \in \mathbb{R}^m$ is called the *control value* or the *control input* at time t. The map $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is called *vector field*. Both f and the *control function* $u : \mathbb{R} \to \mathbb{R}^m$ have to fulfil certain regularity properties in order to guarantee that the solutions of (1.1) exist and are unique. We will not consider this problen at this stage in full generality, because in the first part of the lecture we will only look at a special case of control systems, which allow for a simpler treatment than the general case.

In discrete time the general model is given by the map

$$x(k+1) = f(k, x(k), u(k)). \tag{1.2}$$

In this equation $k \in \mathbb{N}$ is an abstract time index and $f : \mathbb{N} \times \mathbb{R}^n \times U \to \mathbb{R}^n$ is called the *transition map*. The abstract time index k usually stands for a real time $t_k \in \mathbb{R}$, often of the form $t_n = nT$ for a fixed $T > 0$. A discrete-time control system can ba obtained from the behaviour of a continuous-time model at the discrete time instants t_k — this procedure is called sampling and the resulting discrete-time system is called sampled-data system.[1]. In this case there are different ways do choose U. For instance, $u(k)$ could be a constant control value from $\mathbb{R}^m$, which is applied to the continuous-time system during the time interval $[t_k, t_{k+1})$. In this case $U = \mathbb{R}^m$ is a set of (vector valued) control values. The symbol $u(k)$ could, however, also stand for a time-varying control function, which is applied to the continuous-time system on the interval$[t_k, t_{k+1})$. In this case U is a set of functions.

Almost all results presented in this lecture hold for continuous-time and discrete-time control systems alike. However, we will usually only provide the proof for one of the two cases. In the first part of the lecture we will usually give the proofs for the continuous-time case, while in the second part we will usually provide proofs for discrete-time systems.

In the forst part of this lecture we will consider the following particular class of control systems.

Definition 1.1 A *linear time invariant* control system in continuous time is given by the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.3}$$

with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. In discrete time it is given by the equation

$$x(k+1) = Ax(k) + Bu(k) \tag{1.4}$$

with $A \in \mathbb{R}^{n\times n}$ and a linear map $B : U \to \mathbb{R}^n$. □

This class of control systems is particularly simple, since the right hand side is linear in x and u and, moreover, does not explicitly depend on t. Yet, it is already rich enough to describe a large number of real processes, e.g., in technical applications. Indeed, in engineering practice very often linear models are used, although not always in the form (1.3) (we will see an important extension later on in this lecture).

In order to illustrate why the class (1.3) can yield a sufficiently accurate modelling, we consider a model from mechanics, namely an inverted rigid pendulum fixed on a cart, cf. Figure 1.1.

The control u here is the acceleration of the cart. By means of physical laws an "exact"[2]

[1] A formal definition of the sampled-data system is contained in Section 8.2.

[2] The model (1.5) is not really exact, since it is already simplified: We have assumed that the pendulum is so light that it does not influence the motion of the cart. Moreover, a number of constants was chosen such that they cancel each other.

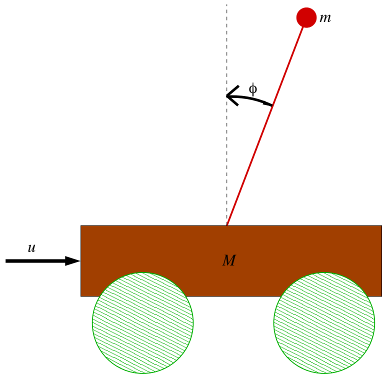


Figure 1.1: Schematic illustration of a pendulum on a cart

differential equation model can be derived .

$$\left.\begin{array}{rcl} \dot{x}_1(t) & = & x_2(t) \\ \dot{x}_2(t) & = & -kx_2(t) + g\sin x_1(t) + u(t)\cos x_1(t) \\ \dot{x}_3(t) & = & x_4(t) \\ \dot{x}_4(t) & = & u \end{array}\right\} =: f(x(t), u(t)) \qquad (1.5)$$

In this model the state vector $x \in \mathbb{R}^4$ consists of 4 components: x_1 represents the angle ϕ of the pendulum (cf. Fig. 1.1), which increases in counterclockwise direction, where $x_1 = 0$ corresponds to the upright pendulum. x_2 is the angular velocity, x_3 the position of the cart and x_4 its velocity. The constant k is a measure for the friction in the model (the larger k the more friction) and $g \approx 9.81m/s^2$ is the gravitational constant.

Obviously (1.5) is of the form (1.1). It is not of the form (1.3), though, since the nonlinear functions sin and cos cannot be written using the matrices A and B (note that A And B may only contain constant coefficients, i.e. the entries of these matrices may not depend on x).

Nevertheless, a linear model of the form (1.3) can be used in order to approximate (1.5) near certain points. This procedure, which is called *linearisation*, is possible near points $(x^*, u^*) \in \mathbb{R}^n \times \mathbb{R}^m$ in which $f(x^*, u^*) = 0$ holds. In these points we can obtain a system of the form (1.3) by defining A and B as

$$A := \frac{\partial f}{\partial x}(x^*, u^*) \quad \text{and} \quad B := \frac{\partial f}{\partial u}(x^*, u^*).$$

If f is continuously differentiable, we have

$$f(x + x^*, u + u^*) = Ax + Bu + o(\|x\| + \|u\|),$$

i.e., for $x \approx 0$ and $u \approx 0$ the values of $f(x + x^*, u + u^*)$ and $Ax + Bu$ are very similar. One can now prove that this similarity of values implies a similarity of the solutions of (1.1) and (1.3) (appropriately shifted), as long as they stay close to (x^*, u^*).[3]

[3] A mathematically exact formulation of the statement can be found as Satz 4.5 in [7].

For our example we apply linearisation to the equilibrium $(x^*, u^*) = (0, 0)$, which corresponds to the upright or inverted position of the pendulum. For this equilibrium, the above computation yields a system if the form (1.3) with

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ g & -k & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{1.6}$$

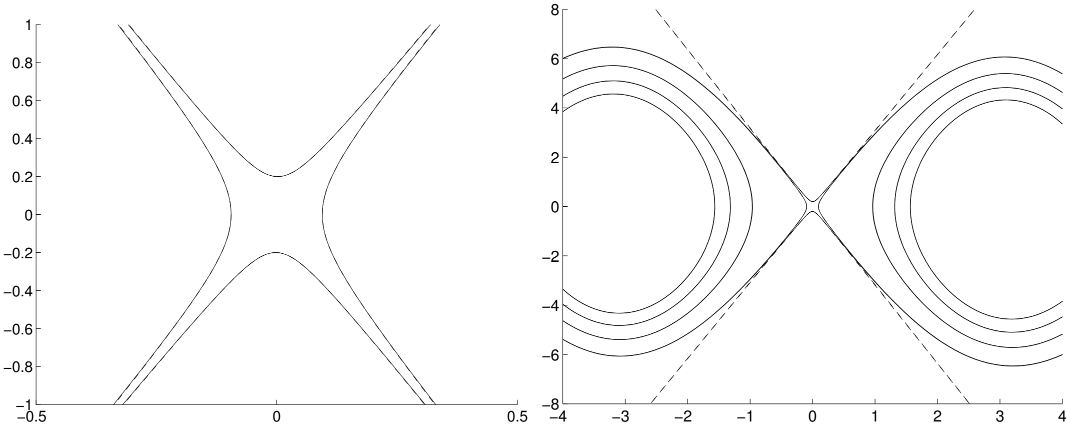


Figure 1.2: Comparison of the solutions of (1.5) (solid) with (1.3, 1.6) (dashed)

Figure 1.1 shows a comparison of the solutions of (1.5) (solid) with the solutions of (1.3, 1.6) (dashed), all for $u \equiv 0$ and with $k = 0.1$, $g = 9.81$, in two different neighbourhoods of 0. Depicted are four solution curves of the form

$$\left\{ \left(\begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right) \middle| \, t \in [-10, 10] \right\} \subset \mathbb{R}^2.$$

for each of the two equations. While in the small neighbourhood on the left hand side of the figure there is almost no visible difference between the curves, in the larger neighbourhood on the right hand side the curves differ significantly.

1.2 Existence and Uniqueness

Whenever differential equations are considered, existence and uniqueness must be clarified. We will first recall basic results for linear control systems (1.3) with $u \equiv 0$, i.e., for homogeneous linear differential equations.

To this end we introduce some notation.

For a matrix $A \in \mathbb{R}^{n\times n}$ with $[A]_{ij} \in \mathbb{R}$ we denote the entry in the i-th rowm and the j-th column. For $A \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}$ with At we denote the componentwise multiplikation, i.e., $[At]_{i,j} = [A]_{ij}t$. For $k \in \mathbb{N}_0$ the power A^k is inductively defined via $A^0 :=$ Id and $A^{k+1} := AA^k$.

Moreover, we need the following definition.

Definition 1.2 For a matrix $A \in \mathbb{R}^{n\times n}$ and a real number $t \in \mathbb{R}$ the matrix exponential is defined by

$$e^{At} := \sum_{k=0}^{\infty} A^k \frac{t^k}{k!}.$$

□

The convergence of this infinite series is to be understood componentwise, i.e. as

$$[e^{At}]_{ij} = \sum_{k=0}^{\infty} [A^k \frac{t^k}{k!}]_{ij}, \quad n \in \mathbb{N}_0.$$

Convergence of the components of this series — even absolutely, i.e., in modulus — follows from the comparison with the row sum norm

$$\alpha = \|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |[A]_{ij}|,$$

since $|[A^k]_{ij}| \leq \|A^k\|_\infty \leq \|A\|_\infty^k = \alpha^k$, also

$$\left| [A^k \frac{t^k}{k!}]_{ij} \right| = |[A^k]_{ij}| \left| \frac{t^k}{k!} \right| \leq \alpha^k \left| \frac{t^k}{k!} \right| = \frac{(\alpha|t|)^k}{k!}$$

and thus

$$|[e^{At}]_{ij}| \leq e^{\alpha|t|},$$

where the expression on the right hand side denotes the usual scalar exponential function.

Note that in general

$$[e^{At}]_{ij} \neq e^{[At]_{ij}},$$

where $e^{[At]_{ij}}$ is the scalar exponential.

From its definition, the matrix exponential satisfies

$$\text{(i)}\ \ e^{A0} = \mathrm{Id} \quad \text{and} \quad \text{(ii)}\ \ Ae^{At} = e^{At}A \tag{1.7}$$

The following lemma yields another important property.

Lemma 1.3 For arbitrary $A \in \mathbb{R}^{n\times n}$ the function $t \mapsto e^{At}$ is differentiable with

$$\frac{d}{dt}e^{At} = Ae^{At}$$

for any $t \in \mathbb{R}$.

Proof: Excercise

Theorem 1.4 Consider the linear differential equation

$$\dot{x}(t) = Ax(t) \tag{1.8}$$

with $x : \mathbb{R} \to \mathbb{R}^n$ and a given matrix $A \in \mathbb{R}^{n\times n}$.

Then for any *initial condition* of the form

$$x(t_0) = x_0 \tag{1.9}$$

with $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ there exists exactly one solution $x : \mathbb{R} \to \mathbb{R}^n$ of (1.8) that satisfies (1.9). We will denote this solution by $x(t; t_0, x_0)$. It is given by

$$x(t; t_0, x_0) = e^{A(t-t_0)}x_0. \tag{1.10}$$

Proof: We first show that the function $x(t) = e^{A(t-t_0)}x_0$ from (1.10) satisfies both the differential equation (1.8) and the initial condition (1.9). Lemma 1.3 yields

$$\frac{d}{dt}x(t) = \frac{d}{dt}e^{A(t-t_0)}x_0 = Ae^{A(t-t_0)}x_0 = Ax(t),$$

hence (1.8). Since (1.7)(i) we moreover obtain

$$x(t_0) = e^{A(t_0-t_0)}x_0 = e^{A0}x_0 = \mathrm{Id}x_0 = x_0,$$

i.e., (1.9).

Since this shows that (1.10) is a solution, this in particular proves existence of a solution.

It remains to show its uniqueness. To this end we first show that e^{At} is invertible with

$$(e^{At})^{-1} = e^{-At}. \tag{1.11}$$

For each $y_0 \in \mathbb{R}^n$ the function $y(t) = e^{-At}y_0$ solves the differential equation $\dot{y}(t) = -Ay(t)$. By the product rule we then obtain

$$\frac{d}{dt}(e^{-At}e^{At}x_0) = \frac{d}{dt}e^{-At}(e^{At}x_0) + e^{-At}\frac{d}{dt}e^{At}x_0 = -Ae^{-At}e^{At}x_0 + e^{-At}Ae^{At}x_0 = 0,$$

where in the last step we used (1.7)(ii). Thus, $e^{-At}e^{At}x_0$ is constant in t. This implies for all $t \in \mathbb{R}$ and all $x_0 \in \mathbb{R}^n$ the identity

$$e^{-At}e^{At}x_0 = e^{-A0}e^{A0}x_0 = \mathrm{Id}\,\mathrm{Id}\,x_0 = x_0,$$

and consequently

$$e^{-At}e^{At} = \mathrm{Id} \;\;\Rightarrow\;\; e^{-At} = (e^{At})^{-1}.$$

Using (1.11) we can now show uniqueness. Let $x(t)$ be an arbitrary solution of (1.8), (1.9). Then

$$\begin{aligned} \frac{d}{dt}(e^{-A(t-t_0)}x(t)) &= \frac{d}{dt}e^{-A(t-t_0)}(x(t)) + e^{-A(t-t_0)}\dot{x}(t) \\ &= -Ae^{-A(t-t_0)}x(t) + e^{-A(t-t_0)}Ax(t) \;=\; 0, \end{aligned}$$

where we again used (1.7)(ii). Hence, $e^{-A(t-t_0)}x(t)$ is constant in t, which for all $t \in \mathbb{R}$ implies

$$e^{-A(t-t_0)}x(t) = e^{-A(t_0-t_0)}x(t_0) = \mathrm{Id}x(t_0) = x_0.$$

Multiplying both sides of this identity with $e^{A(t-t_0)}$ and using (1.11) we get

$$x(t) = e^{A(t-t_0)}x_0.$$

Since $x(t)$ was an arbitrary solution, this shows uniqueness. □

A useful implication of this theorem is the following corollary.

Corollary 1.5 The matrix exponential e^{At} is the unique solution of the matrix differential equation

$$\dot{X}(t) = AX(t) \tag{1.12}$$

with $X : \mathbb{R} \to \mathbb{R}^{n\times n}$ and initial condition

$$X(0) = \mathrm{Id}. \tag{1.13}$$

□

Proof: With e_j we denote the j-th unit vector in $\mathbb{R}^n$. A simple computation reveals that a matrix valued function $X(t)$ is a solution of (1.12), (1.13) if and only iff $X(t)e_j$ is a solution of (1.8), (1.9) with $t_0 = 0$ and $x_0 = e_j$. With this observation the assertion follows immediately from Theorem 1.4. □

The following lemma summarises further properties of the matrix exponential.

Lemma 1.6 For $A, A_1, A_2 \in \mathbb{R}^{n\times n}$ and $s, t \in \mathbb{R}$ the following identities hold:

(i) $(e^{At})^{-1} = e^{-At}$

(ii) $e^{At}e^{As} = e^{A(t+s)}$

(iii) $e^{A_1t}e^{A_2t} = e^{(A_1+A_2)t}$ falls $A_1A_2 = A_2A_1$

(iv) For an invertible matrix $T \in \mathbb{R}^{n\times n}$ the equation

$$e^{T^{-1}ATt} = T^{-1}e^{At}T$$

holds.

Proof: (i) This was shown in the proof of Theorem 1.4.

(ii) With (i) it follows that both $e^{At}e^{As}e^{-As}$ and $e^{A(t+s)}e^{-As}$ solve the matrix valued initial value problem (1.12), (1.13). Sincce its solution is unique by Corollary 1.5 and e^{-As} is invertible, the claimed identity follows.

(iii) Using the assumption $A_1A_2 = A_2A_1$ one checks that both expressions solve the matrix initial value problem (1.12), (1.13) with $A = A_1 + A_2$. Hence the two expressions must coincide because of the uniqueness of the solution provided by Corollary 1.5.

(iv) One computes that both expressions solve the matrix initial value problem (1.12), (1.13) with $T^{-1}AT$ in place of A. Then again the assertion follows from the uniqueness of this solution established in Corollary 1.5. □

After these preparations we return to the linear control system (1.3). For the formulation of an existence und uniueness theorem we need to define a suitable function space $\mathcal{U}$ for the control function $u(\cdot)$. Certainly continuous functions would lead to an existence and uniqueness result, but this choice would be to restrictive, because throughout this lecture we will frequently need concatenations of control functions according to the following definition.

Definition 1.7 For two functions u_1, $u_2 : \mathbb{R} \to \mathbb{R}^m$ and $s \in \mathbb{R}$ we define the *concatenation at time s* as

$$u_1 \&_s u_2(t) := \begin{cases} u_1(t), & t < s \\ u_2(t), & t \geq s \end{cases}$$

□

Even if u_1 and u_2 are continuous, $u_1 \&_s u_2$ will in general not be continuous. We thus need a function space that is closed with respect to concatenation. There are several options for this. The most simple one is the following.

Definition 1.8 A function $u : \mathbb{R} \to \mathbb{R}^m$ is called *piecewise continuous*, if for any compact interval $[t_1, t_2]$ there exists a finite sequence of times $t_1 = \tau_1 < \tau_2 < \ldots < \tau_k = t_2$, such that $u|_{(\tau_i, \tau_{i+1})}$ is bounded and continuous for every $i = 1, \ldots, k-1$. We define $\mathcal{U}$ as the space of piecewise continuous functions from $\mathbb{R}$ to $\mathbb{R}^m$. □

Obviously, $\mathcal{U}$ is closed with respect to concatenation, but also with respect to addition and multiplikation (defining $(u_1 + u_2)(t) := u_1(t) + u_2(t)$ and $(u_1 \cdot u_2)(t) := u_1(t) \cdot u_2(t)$). In addition — and this is important for our purpose — the Riemann-Integral

$$\int_{t_1}^{t_2} u(t)dt$$

exists for functions $u \in \mathcal{U}$, since in every compact interval there are at most finitely many points of discontinuity.[4]

With this function space we can now formulate an existence and uniqueness result.

[4] An alternative to the space of piecewise constant functions is the space of Lebesgue measurable functions, where the integral is then chosen as the Labesgue integral. This space will be used for nonlinear control systems, cf. Chapter 8. For linear control systems the use of Lebesgue measurable control functions does not carry any advantage.

Theorem 1.9 Consider the linear control system (1.3)

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

with $x : \mathbb{R} \to \mathbb{R}^n$ and given matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$.

Then for any *initial condition* of the form (1.9)

$$x(t_0) = x_0$$

with $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$ and any piecewise continuous control function $u \in \mathcal{U}$ there exists a unique continuous function $x : \mathbb{R} \to \mathbb{R}^n$ that satisfies (1.9) and whose derivative exists and satisfies (1.3) for each t, in which u is continuous. This unique function is called the solution of (1.3), (1.9) and denoted by $x(t; t_0, x_0, u)$. It is given by the formula

$$x(t; t_0, x_0, u) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds. \tag{1.14}$$

Proof: We first check that (1.14) is indeed a solution in the sense described in the theorem. The map $t \mapsto \int_{t_0}^{t} g(s)ds$ is continuous for any Riemann-integrable function, and thus $x(t; t_0, x_0, u)$ is continuous in t. In those t where u is continuous we get

$$\begin{aligned}
&\frac{d}{dt}\left[e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds\right] \\
&= \frac{d}{dt}e^{A(t-t_0)}x_0 + \frac{d}{dt}\int_{t_0}^{t} e^{A(t-s)}Bu(s)ds \\
&= Ae^{A(t-t_0)}x_0 + \underbrace{e^{A(t-s)}Bu(s)|_{s=t}}_{=Bu(t)} + \int_{t_0}^{t} Ae^{A(t-s)}Bu(s)ds \\
&= A\left(e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds\right) + Bu(t),
\end{aligned}$$

i.e., (1.3). In addition we obtain

$$\underbrace{e^{A(t_0-t_0)}}_{=\mathrm{Id}}x_0 + \underbrace{\int_{t_0}^{t_0} e^{A(t_0-s)}Bu(s)ds}_{=0} = x_0,$$

i.e., (1.9).

It remains to show the uniqueness. To this end we consider two arbitrary solutions $x(t)$, $y(t)$ of (1.3), (1.9) in the sense of the theorem. Then

$$\dot{z}(t) = \dot{x}(t) - \dot{y}(t) = Ax(t) + Bu(t) - Ay(t) - Bu(t) = A(x(t) - y(t)) = Az(t)$$

for all t in which u is continuous. Since z is continuous, $\dot{z}$ can be extended continuously in the points of non continuity τ_i of u by $\dot{z}(\tau_i) = \lim_{t\to\tau_i} Az(t)$. We thus obtain a function that solves the differential equation $\dot{z}(t) = Az(t)$ for all $t \in \mathbb{R}$t. Since moreover

$$z(t_0) = x(t_0) - y(t_0) = x_0 - x_0 = 0$$

holds, z satisfies an initial value problem of the form (1.8), (1.9), whose unique solution according to Theorem 1.4 is given by $z(t) = e^{At}0 = 0$. Thus, $x(t) = y(t)$ for all $t \in \mathbb{R}$, proving uniqueness. □

A consequence from this theorem is the following corollary.

Corollary 1.10 The solutions of (1.3), (1.9) satisfy for all $t, s \in \mathbb{R}$ the equations

$$x(t; t_0, x_0, u) = x(t; s, x(s; t_0, x_0, u), u)$$

and

$$x(t; t_0, x_0, u) = x(t - s; t_0 - s, x_0, u(s + \cdot)).$$

Herein, $u(s + \cdot) \in \mathcal{U}$ is defined as $u(s + \cdot)(t) = u(s + t)$. From the combination of the two formulas for $t_0 = 0$ we also get

$$x(t; x_0, u) = x(t - s; x(s; x_0, u), u(s + \cdot)).$$

□

Proof: Follows immediately from (1.14). □

Remark 1.11 Another immediate consequence from the solution formula (1.14) is the identity

$$x(t; t_0, x_0, u) = x(t; t_0, x_0, 0) + x(t; t_0, 0, u). \tag{1.15}$$

This identity says that any solution is the superposition of an uncontrolled solution (i.e. with control 0) and a solution without unforced dynamics (i.e. with initial value 0). It is thus known as *superposition principle.* □

Remark 1.12 In the following chapters we often limit ourselves to the case $t_0 = 0$. In this case we use the shorter notation $x(t; x_0, u) = x(t; 0, x_0, u)$. □

Remark 1.13 One may consider the times $t_n = nT$ and a continuous-time control system with control functions that are constant with values $u_T(k)$ on the intervals $[t_k, t_{k+1})$. Then from the solution formula (1.14) explicit formulas for the matrices A_T and B_T for the corresonding sampled-data system

$$x_T(k + 1) = A_T x_T(k) + B_T u_T(k)$$

can be derived. Details will be worked out in an exercise. □

Chapter 2

Controllability

2.1 Definitions

An important aspect in the analysis of linear control systems of the form (1.3) is the question about its controllability. In its most general formulation, this concerns the question for which states x_0, $x_1 \in \mathbb{R}^n$ and times t_1 we can find a control function $u \in \mathcal{U}$ for which $x(t_1; x_0, u) = x_1$ holds. In other words: can we link the two states by a solution trajectory on a given time interval? Formally we define this property as follows.

Definition 2.1 Consider a linear control system (1.3).

A state $x_0 \in \mathbb{R}^n$ is called *controllable* to a State $x_1 \in \mathbb{R}^n$ at time $t_1 > 0$, if there exists $u \in \mathcal{U}$ with

$$x_1 = x(t_1; x_0, u).$$

In this case, the state x_1 is called *reachable* from x_0 at time t_1. □

The following lemma shows that it is sufficient to consider controllability for the case $x_0 = 0$.

Lemma 2.2 A state $x_0 \in \mathbb{R}^n$ is controllable to a state $x_1 \in \mathbb{R}^n$ at time $t_1 > 0$ if and only if the state $\tilde{x}_0 = 0$ is controllable to the state $\tilde{x}_1 = x_1 - x(t_1; x_0, 0)$ at time t_1.

Proof: Exercise.

This fact allows us to restrict the following definition of controllability and reachability to the case $x_0 = 0$.

Definition 2.3 Consider a linear control system (1.3).

(i) The *reachable set* (or *attainable set*) from a state $x_0 = 0$ at time $t \geq 0$ is given by

$$\mathcal{R}(t) = \{x(t; 0, u) \,|\, u \in \mathcal{U}\}.$$

(ii) The *controllable set* to a state $x_1 = 0$ at time $t \geq 0$ is given by

$$\mathcal{C}(t) = \{x_0 \in \mathbb{R}^n \,|\, \text{there exists } u \in \mathcal{U} \text{ with } x(t; x_0, u) = 0\}.$$

□

The relation between these sets is clarified by the following lemma.

Lemma 2.4 The reachable set $\mathcal{R}(t)$ fpr (1.3) equals the controllability set $\mathcal{C}(t)$ for the time-reversed system

$$\dot{z}(t) = -Az(t) - Bu(t). \tag{2.1}$$

Proof: By verifying that the followiong expressions satisfy the initial value problem, whic has a unique solution, one sees that the solutions of (1.3) and (2.1) satisfy

$$x(s, 0, u) = z(t - s, x(t, 0, u), u(t - \cdot))$$

for all $t, s \in \mathbb{R}$. Hence, if $x_1 \in \mathcal{R}(t)$ for (1.3) and $x(s, 0, u)$ is the corresponding solution, we obtain

$$z(0, x(t, 0, u), u(t - \cdot)) = x(t, 0, u) = x_1 \text{ and } z(t, x(t, 0, u), u(t - \cdot)) = x(0, 0, u) = 0,$$

implying $x_1 \in \mathcal{C}(t)$. The converse direction follows with analogous arguments. □

2.2 Analysis of controllability properties

We now want to clarify the structure of these sets. In this analysis we derive the technical auxiliary results for $\mathcal{R}(t)$ and only state the main results for both $\mathcal{R}(t)$ and $\mathcal{C}(t)$.

Lemma 2.5 (i) For all $t \geq 0$ the set $\mathcal{R}(t)$ is a subspace of $\mathbb{R}^n$.

(ii) $\mathcal{R}(t) = \mathcal{R}(s)$ for all $s, t > 0$.

Proof: (i) We have to show that for x_1, $x_2 \in \mathcal{R}(t)$ and $\alpha \in \mathbb{R}$ the relation $\alpha(x_1 + x_2) \in \mathcal{R}(t)$ holds. For x_1, x_2 in $\mathcal{R}(t)$ there exist control functions u_1, $u_2 \in \mathcal{U}$ with

$$x_i = x(t; 0, u_i) = \int_0^t e^{A(t-s)} B u_i(s) ds.$$

Hence, for $u = \alpha(u_1 + u_2)$ we obtain

$$\begin{aligned} x(t; 0, u) &= \int_0^t e^{A(t-s)} B u(s) ds \;=\; \int_0^t e^{A(t-s)} B \alpha(u_1(s) + u_2(s)) ds \\ &= \alpha \left(\int_0^t e^{A(t-s)} B u_1(s) ds + \int_0^t e^{A(t-s)} B u_2(s) ds \right) \;=\; \alpha(x_1 + x_2), \end{aligned}$$

implying $\alpha(x_1 + x_2) \in \mathcal{R}(t)$. This proves (i).

(ii) We give a direct proof here. Independently of this proof the statement also follows from Theorem 2.12.

We first show the auxiliary result

$$\mathcal{R}(t_1) \subseteq \mathcal{R}(t_2) \tag{2.2}$$

for $0 < t_1 < t_2$: If $y \in \mathcal{R}(t_1)$, then there exists $u \in \mathcal{U}$ with

$$x(t_1; 0, u) = y.$$

Then with the new control $\tilde{u} = 0\&_{t_2 - t_1} u(t_1 - t_2 + \cdot)$ Corollary 1.10 yields

$$x(t_2; 0, \tilde{u}) = x(t_2; t_2 - t_1, \underbrace{x(t_2 - t_1; 0, 0)}_{=0}, \tilde{u}) = x(t_2; t_2 - t_1, 0, \tilde{u}) = x(t_1; 0, u) = y,$$

which implies $y \in \mathcal{R}(t_2)$.

Next we show that for any $0 < t_1 < t_2$ the identity $\mathcal{R}(t_1) = \mathcal{R}(t_2)$ implies the identity $\mathcal{R}(t_1) = \mathcal{R}(t)$ for all $t \geq t_1$. In order to prove this, let $x \in \mathcal{R}(2t_2 - t_1)$, i.e. we assume that there is $u \in \mathcal{U}$ with $x = x(2t_2 - t_1, 0, u)$.

Since $x(t_2, 0, u) \in \mathcal{R}(t_2)$ and $\mathcal{R}(t_2) = \mathcal{R}(t_1)$, there exists a $v \in \mathcal{U}$ with $x(t_1, 0, v) = x(t_2, 0, u)$. Defining the control function $w = v\&_{t_1} u(t_2 - t_1 + \cdot)$, Corollary 1.10 yields

$$\begin{aligned} x(t_2, 0, w) &= x(t_2, t_1, \underbrace{x(t_1, 0, v)}_{=x(t_2,0,u)}, w) \\ &= x(t_2 + t_2 - t_1, t_1 + t_2 - t_1, x(t_2, 0, u), \underbrace{w(t_1 - t_2 + \cdot)}_{=u(\cdot)}) \\ &= x(2t_2 - t_1, 0, u) \;=\; x. \end{aligned}$$

Hence, we obtain $x \in \mathcal{R}(t_2)$ and consequently $\mathcal{R}(t_1) = \mathcal{R}(t_2) = \mathcal{R}(2t_2 - t_1) = \mathcal{R}(2(t_2 - t_1) + t_1)$. Repeating this construction inductively yields $\mathcal{R}(t_1) = \mathcal{R}(2^k(t_2 - t_1) + t_1)$ for all $k \in \mathbb{N}$ and thus because of (2.2) the claimed assertion $\mathcal{R}(t_1) = \mathcal{R}(t)$ for all $t \geq t_1$.

Now we show the assertion (ii): For this purpose, let $s > 0$ be arbitrary and consider an increasing sequence of times $0 < s_0 < \ldots < s_{n+1} = s$. Then according to (2.2) the sets $\mathcal{R}(s_0), \ldots, \mathcal{R}(s_{n+1})$ form an increasing sequence of $n + 2$ subspaces of $\mathbb{R}^n$. In particular, $\mathcal{R}(s_{k+1}) \neq \mathcal{R}(s_k)$ implies $\dim \mathcal{R}(s_{k+1}) \geq \dim \mathcal{R}(s_k) + 1$. Hence, if all the $\mathcal{R}(s_k)$ are pairwise different, we obtain $\dim \mathcal{R}(s_{n+1}) \geq n + 1$. This, however, contradicts $\mathcal{R}(s_{n+1}) \subseteq \mathbb{R}^n$, which means that at least two of the sets $\mathcal{R}(s_k)$ must coincide. Then the previous considerations imply $\mathcal{R}(t) = \mathcal{R}(s)$ for all $t \geq s$ and since $s > 0$ was arbitrary, this yields the assertion (ii). □

Remark 2.6 Since we just proved that the sets $\mathcal{R}(t)$ do not depend on t for $t > 0$, in the sequel we usually write $\mathcal{R}$ instead of $\mathcal{R}(t)$. □

Remark 2.7 The combination of Lemma 2.2 and Lemma 2.5 thus shows that the set of states that are reachable from an arbitrary initial state $x_0 \in \mathbb{R}^n$ at time $t > 0$ is the affine subspace

$$x(t; x_0, 0) + \mathcal{R},$$

whose dimension equals the dimension of $\mathcal{R}$. Note that this set is in general *not* independent of t. One exception is the case where $\mathcal{R} = \mathbb{R}^n$, since this implies $x(t; x_0, 0) + \mathcal{R} = \mathbb{R}^n$. In this case every state x_0 can be controlled into every other state x_1, which is why in this case we call the system *completely controllable* or, short, simply *controllable*. □

As we saw in the exercises, even for relatively simple control systems the direct computation of $\mathcal{R}$ and $\mathcal{C}$ by means of their definition can be challenging. As a remedy we will now derive a simple characterization of these sets. To this end, we need some tools from linear algebra.

Definition 2.8 (i) A subspace $U \subseteq \mathbb{R}^n$ is called A-invariant for a matrix $A \in \mathbb{R}^{n\times n}$, if $Av \in U$ for all $v \in U$ (or, briefly, $AU \subseteq U$).

(ii) For a subspace $V \subseteq \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, by

$$\langle A \,|\, V \rangle$$

we denote the smallest (with respect to its dimension) A-invariant subspace of $\mathbb{R}^n$ that contains V. □

Note that such a smallest subspace exists and is unique: on the one hand the space $\mathbb{R}^n$ itself is an A-invariant subspace that contains V. Since the dimension is finite, this implies that there also exists such a space with minimal dimension. On the other hand, assume there are several different such subspaces with minimal dimension. Then one easily checks that their intersection is again an A-invariant subspace that contains V. Since this intersection has a lower dimension than the spaces that were intersected, this contradict the minimality of their dimension.

Lemma 2.9 For a subspace $V \subseteq \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ the identity

$$\langle A \,|\, V \rangle = V + AV + \ldots + A^{n-1}V$$

holds.

Proof: "$\supseteq$": The A-invariance of $\langle A \,|\, V \rangle$ and the fact that $V \subseteq \langle A \,|\, V \rangle$ imply

$$A^k V \subseteq \langle A \,|\, V \rangle$$

for all $k \in \mathbb{N}_0$ and thus $\langle A \,|\, V \rangle \supseteq V + AV + \ldots + A^{n-1}V$.

"$\subseteq$": It suffices to show that $V + AV + \ldots + A^{n-1}V$ is A-invariant, since then $V \subseteq V + AV + \ldots + A^{n-1}V$ immediately implies $\langle A \,|\, V \rangle \subseteq V + AV + \ldots + A^{n-1}V$.

In order to prove A-invariance, consider the charakteristic polynomial of A

$$\chi_A(z) = \det(z\mathrm{Id} - A) = z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0.$$

By the theorem of Cayley-Hamilton, χ_A satisfies

$$\chi_A(A) = A^n + a_{n-1}A^{n-1} + \ldots + a_1A + a_0\mathrm{Id} = 0,$$

implying

$$A^n = -a_{n-1}A^{n-1} - \ldots - a_1A - a_0\mathrm{Id}.$$

Thus, any $v \in V + AV + \ldots + A^{n-1}V$ can be written as $v = v_0 + Av_1 + \ldots + A^{n-1}v_{n-1}$ for $v_0, \ldots, v_{n-1} \in V$. This implies

$$\begin{aligned} Av &= Av_0 + A^2v_1 + \ldots + A^nv_{n-1} \\ &= Av_0 + A^2v_1 - a_{n-1}A^{n-1}v_{n-1} - \ldots - a_1Av_{n-1} - a_0v_{n-1} \\ &= \tilde{v}_0 + A\tilde{v}_1 + \ldots + A^{n-1}\tilde{v}_{n-1} \end{aligned}$$

for suitable $\tilde{v}_0, \ldots, \tilde{v}_{n-1} \in V$. From this we obtain $Av \in V + AV + \ldots + A^{n-1}V$, i.e., A-invariance. □

We will now consider the special case in which $V = \operatorname{im} B$ is the image of the matrix B. In this case Lemma 2.9 says that

$$\langle A \,|\, \operatorname{im} B\rangle = \{Bx_0 + ABx_1 + \ldots + A^{n-1}Bx_{n-1} \,|\, x_0, \ldots, x_{n-1} \in \mathbb{R}^m\} = \operatorname{im}(B\; AB\; \ldots\; A^{n-1}B),$$

where $(B\; AB\; \ldots\; A^{n-1}B) \in \mathbb{R}^{n\times(m\cdot n)}$.

Definition 2.10 The matrix $(B\; AB\; \ldots\; A^{n-1}B) \in \mathbb{R}^{n\times(m\cdot n)}$ is called *controllability matrix* of the system (1.3). □

In the sequel for $t \in \mathbb{R}$ we use the notation

$$W_t := \int_0^t e^{A\tau}BB^T(e^{A\tau})^T d\tau.$$

Observe that $W_t \in \mathbb{R}^{n\times n}$ and W_t is thus a linear operator on $\mathbb{R}^n$. The matrix W_t is called *controllability Gramian* and is symmetrisch and positive semidefinite, because

$$x^TW_tx = \int_0^t \underbrace{x^Te^{A\tau}BB^T(e^{A\tau})^Tx}_{=\|B^T(e^{A\tau})^Tx\|^2\geq 0} d\tau \geq 0.$$

The image $\operatorname{im} W_t$ of this operator is described by the following lemma.

Lemma 2.11 For all $t > 0$ the identity $\langle A \,|\, \operatorname{im} B\rangle = \operatorname{im} W_t$ holds.

Proof: We show $\langle A \,|\, \operatorname{im} B\rangle^\perp = (\operatorname{im} W_t)^\perp$.

"$\subseteq$": Let $x \in \langle A \,|\, \operatorname{im} B\rangle^\perp$, i.e. $x^TA^kB = 0$ for all $k \in \mathbb{N}_0$. Then

$$x^Te^{At}B = \sum_{k=0}^{\infty} \frac{t^kx^TA^kB}{k!} = 0$$

and thus $x^T W_t = 0$, hence $x \in (\operatorname{im} W_t)^\perp$.

"$\supseteq$": Let $x \in (\operatorname{im} W_t)^\perp$ for some $t > 0$. Then

$$0 = x^T W_t x = \int_0^t \|B^T (e^{A\tau})^T x\|^2 d\tau,$$

which because of the continuity of the integrand implies $x^T e^{A\tau} B = (B^T (e^{A\tau})^T x)^T = 0$. Successively computing the derivatives of $x^T B e^{A\tau}$ with respect to τ yields

$$x^T A^k e^{A\tau} B = 0$$

for all $k \in \mathbb{N}_0$. For $\tau = 0$ this implies $x^T A^k B = 0$, i.e. $x \in (\operatorname{im} A^k B)^\perp$ for all $k \in \mathbb{N}_0$ and hence $x \in [\operatorname{im}(B\, AB\, \ldots\, A^{n-1}B)]^\perp = \langle A \,|\, \operatorname{im} B \rangle^\perp$. □

The following theorem is our main result on the structure of the reachable and controllable sets.

Theorem 2.12 For the system (1.3) and all $t > 0$ the identities

$$\mathcal{R}(t) = \mathcal{C}(t) = \langle A \,|\, \operatorname{im} B \rangle = \operatorname{im}(B\, AB\, \ldots\, A^{n-1}B)$$

hold.

Proof: The identity $\langle A \,|\, \operatorname{im} B \rangle = \operatorname{im}(B\, AB\, \ldots\, A^{n-1}B)$ was already shown in the computation before Definition 2.10. We show $\mathcal{R}(t) = \langle A \,|\, \operatorname{im} B \rangle$ for each $t > 0$ (which provides an alternative proof for the fact that $\mathcal{R}(t)$ is independent of t). The statement for $\mathcal{C}(t)$ then follows by time reversal with Lemma 2.4, since $\langle A \,|\, \operatorname{im} B \rangle = \langle -A \,|\, \operatorname{im} -B \rangle$.

"$\subseteq$": Let $x = x(t; 0, u) \in \mathcal{R}(t)$. Then the general solution formula states that

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

Now for all $\tau \in [0, t]$ the definition of $\langle A \,|\, \operatorname{im} B \rangle$ implies

$$e^{A(t-\tau)} B u(\tau) = \sum_{k=0}^{\infty} \frac{(t-\tau)^k}{k!} A^k B u(\tau) \in \langle A \,|\, \operatorname{im} B \rangle$$

and hence also $x \in \langle A \,|\, \operatorname{im} B \rangle$, since integration over elements from a subspace yields again an element from this subspace.

"$\supseteq$": Let $x \in \langle A \,|\, \operatorname{im} B \rangle$ and $t > 0$ be arbitrary. Then by Lemma 2.11 there exists $z \in \mathbb{R}^n$ with $x = W_t z$. If we define $u \in \mathcal{U}$ as $u(\tau) := B^T (e^{A(t-\tau)})^T z$ for $\tau \in [0, t]$, then we get

$$x(t; 0, u) = \int_0^t e^{A(t-\tau)} B B^T (e^{A(t-\tau)})^T z d\tau = W_t z = x,$$

and thus $x \in \mathcal{R}(t)$. □

Note that this proof is constructive: It provides an explicit formula for the control function u that steers 0 to x.

Corollary 2.13 (Kalman criterion) The system (1.3) is completely controllable if and only if

$$\mathrm{rg}(B\, AB\, \dots\, A^{n-1}B) = n.$$

In this case the matrix pair (A, B) is called *controllable.* □

If (A, B) is not controllable, then after suitable coordinate change of the state space $\mathbb{R}^n$ the pair (A, B) can be decomposed into a controllable and a non-controllable part, according to the following lemma.

Lemma 2.14 Let (A, B) be not controllable, i.e., $r := \dim\langle A \,|\, \mathrm{im}\, B\rangle < n$. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$, such that $\widetilde{A} = T^{-1}AT$ and $\widetilde{B} = T^{-1}B$ have the form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

with $A_1 \in \mathbb{R}^{r\times r}$, $A_2 \in \mathbb{R}^{r\times(n-r)}$, $A_3 \in \mathbb{R}^{(n-r)\times(n-r)}$, $B_1 \in \mathbb{R}^{r\times m}$ and the pair (A_1, B_1) is controllable. In particular, after the coordinate change with T the system has the form

$$\begin{aligned} \dot{z}_1(t) &= A_1z_1(t) + A_2z_2(t) + B_1u(t) \\ \dot{z}_2(t) &= A_3z_2(t) \end{aligned}$$

with $z_1(t) \in \mathbb{R}^r$, $z_2(t) \in \mathbb{R}^{n-r}$ and the z_1-subsystem is completely controllable.

Proof: Exercise.

Recall that the characteristic polynomial of a matrix does not change under coordinate transformations. This implies

$$\chi_A(z) = \det(z\mathrm{Id} - A) = \det(z\mathrm{Id} - \widetilde{A}) = \det(z\mathrm{Id} - A_1)\cdot\det(z\mathrm{Id} - A_3) = \chi_{A_1}(z)\cdot\chi_{A_3}(z).$$

This motivates the following definition.

Definition 2.15 We call χ_{A_1} the *controllable* and χ_{A_3} the *non-controllable part* of the charakteristic polynomial χ_A. □

The following theorem yields alternative characterizations of controllability, which do not require the explicit computation of the controllability matrix. Therein $(\lambda\mathrm{Id} - A \,|\, B) \in \mathbb{R}^{n\times(n+m)}$ denotes the matrix that results from writing the matrices $\lambda\mathrm{Id} - A$ and B beside each other.

Theorem 2.16 (Hautus criterion) The following conditions are equivalent:

(i) (A, B) is controllable

(ii) $\mathrm{rg}(\lambda\mathrm{Id} - A \,|\, B) = n$ for all $\lambda \in \mathbb{C}$

(iii) $\mathrm{rg}(\lambda\mathrm{Id} - A \,|\, B) = n$ for all eigenvalues $\lambda \in \mathbb{C}$ of A

Proof: We first prove "(ii) $\Leftrightarrow$ (iii)" and then "(i) $\Leftrightarrow$ (ii)".

"(ii) $\Rightarrow$ (iii)": immediately clear

"(ii) $\Leftarrow$ (iii)": Consider a $\lambda \in \mathbb{C}$ that is not an eigenvalue of A. Then $\det(\lambda \mathrm{Id} - A) \neq 0$, implying $\mathrm{rg}(\lambda \mathrm{Id} - A) = n$. This proves (ii), since $\mathrm{rg}(\lambda \mathrm{Id} - A \,|\, B) \geq \mathrm{rg}(\lambda \mathrm{Id} - A)$.

"(i) $\Leftrightarrow$ (ii)": We show this implication by contraposition, i.e. we prove "not (i) $\Leftrightarrow$ not (ii)".

"not (i) $\Leftarrow$ not (ii)": If (ii) does not hold, then there is a $\lambda \in \mathbb{C}$ with $\mathrm{rg}(\lambda \mathrm{Id} - A \,|\, B) < n$. Hence there is $p \in \mathbb{R}^n$, $p \neq 0$ with $p^T(\lambda \mathrm{Id} - A \,|\, B) = 0$, i.e.,

$$p^T A = \lambda p^T \text{ and } p^T B = 0.$$

The first identity implies $p^T A^k = \lambda^k p^T$ and thus

$$p^T A^k B = \lambda^k p^T B = 0$$

for $k = 0, \ldots, n-1$. Hence we obtain $p^T(B\, AB\, \ldots\, A^{n-1}B) = 0$, from which we can conclude $\mathrm{rg}(B\, AB\, \ldots\, A^{n-1}B) < n$. Thus, (A, B) is not controllable.

"not (i) $\Rightarrow$ not (ii)": If (A, B) is not controllabe, then by Lemma 2.14 there exists the transformation

$$\widetilde{A} = T^{-1}AT = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = T^{-1}B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

with coordinate transformation matrix T.

Now let $\lambda \in \mathbb{C}$ by an eigenvalue of A_3^T with eigenvector v. Then we get $v^T(\lambda \mathrm{Id} - A_3) = 0$, which for $w^T = (0, v^T)$ implies

$$w^T(\lambda \mathrm{Id} - \widetilde{A}) = \big(0^T(\lambda \mathrm{Id} - A_1) + v^T 0,\ 0^T(-A_2) + v^T(\lambda \mathrm{Id} - A_3)\big) = 0$$

and

$$w^T \widetilde{B} = \begin{pmatrix} 0^T B_1 \\ v^T 0 \end{pmatrix} = 0.$$

Using $p^T = w^T T^{-1} \neq 0$ we thus obtain

$$p^T(\lambda \mathrm{Id} - A \,|\, B) = w^T T^{-1}(\lambda \mathrm{Id} - A \,|\, B) = (w^T(\lambda \mathrm{Id} - \widetilde{A})T^{-1} \,|\, w^T \widetilde{B}) = 0,$$

which shows that (ii) does not hold. □

Remark 2.17 For discrete time systems (1.4) with $U = \mathbb{R}^m$ the conditions for complete controllability are completely identical. There is, however, one important difference: While controllability in continuous time implies controllability in arbitrary short time, in discrete time in the worst case one needs up to n time steps to reach a given state x_1. As an example consider the system

$$x(k+1) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(k)$$

with $x \in \mathbb{R}^2$ and $u \in \mathbb{R}$. Here we have $(B\, AB) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which has full rank, implying complete controllability. Yet, in order to control the system from $(0,0)^T$ to $(1,1)^T$ we need at least two time steps. In fact, in discrete time Lemma 2.5 only holds for times $s, t \geq n$. □

Chapter 3

Stability and stabilisation

In this chapter we will consider the problem of stabilising linear control systems. Before we address this problem, we have to clarify what we mean by stability.

3.1 Definitions

In this and in the following two sections we recall important results from the stability theory of linear time-invariant differential equations (1.8)

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

The exposition is relatively short. A more comprehensive treatment can be found in most textbooks on ordinary differential equations. We limit ourselves to the stability of equilibria.

Definition 3.1 A state $x^* \in \mathbb{R}^n$ is called *equilibrium* (also *steady state* or *fixed point*) of an ordinary differential equation, if the corresponding solution satisfies

$$x(t; x^*) = x^* \text{ for all } t \in \mathbb{R}.$$

□

We have already used equilibria without formal definition in the introductory chapter. It is easy to verify that a state x^* is an equilibrium of a time-invariant differential equation $\dot{x}(t) = f(x(t))$ if and only if $f(x^*) = 0$. Thus, for the linear differential equation (1.8) the point $x^* = 0$ is always an equilibrium. This equilibrium $x^* = 0$ will be studied in the following analysis.

Definition 3.2 Consider the equilibrium $x^* = 0$ of the linear differential equation (1.8).

(i) The equilibrium $x^* = 0$ is called *stable*, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that the inequality

$$\|x(t; x_0)\| \leq \varepsilon \text{ for all } t \geq 0$$

holds for all initial conditions $x_0 \in \mathbb{R}^n$ mit $\|x_0\| \leq \delta$.

(ii) The equilibrium $x^* = 0$ is called *locally asymptotically stable*, if it is stable and moreover

$$\lim_{t\to\infty} x(t; x_0) = 0$$

holds for all initial conditions x_0 from an open neighbourhood N of $x^* = 0$.

(iii) The equilibrium $x^* = 0$ is called *globally asymptotically stable*, if (ii) holds with $N = \mathbb{R}^n$.

(iv) The equilibrium $x^* = 0$ is called *locally* respectively *globally exponentially stable*, if there exist constants $c, \sigma > 0$ such that the inequality

$$\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \text{ for all } t \geq 0$$

holds for all x_0 from a neighbourhood N of $x^* = 0$, with $N = \mathbb{R}^n$ in the global case. □

Bemerkung 3.3 The stability property from (i) is also called “stability in the sense of Lyapunov”, since this concept was introduced at the end of the 19th century by the Russian mathematician Alexander M. Lyapunov. Note that the definitions immmediately lead to the implications

$$\text{(locally/globally) exponentially stable} \;\Rightarrow\; \text{(locally/globally) asymptotically stable} \;\Rightarrow\; \text{stable} .$$

The second implication follows directly from the definitions. The fact that exponential stability implies asymptotic stability can be seen as follows:
For a given $\varepsilon > 0$ property (i) follows with $\delta = \varepsilon/c$, because for $\|x_0\| \leq \delta$ this implies the inequality $\|x(t; x_0)\| \leq ce^{-\sigma t}\|x_0\| \leq c\|x_0\| \leq \varepsilon$. The convergence required in (ii) follows obviously from (iii). □

3.2 Eigenvalue criteria

The following theorem provides criteria for the matrix A which allow to check the stability properties of the equilibrium $x^* = 0$ (1.8) easily.

Theorem 3.4 Consider the linear time-invariant differential equation (1.8) for a matrix $A \in \mathbb{R}^{n\times n}$. Let $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$, $\lambda_l = a_l + ib_l$, be the eigenvalues of A, which are numbered such that each eigenvalue λ_l corresponds to a Jordan block J_l in the Jordan canonical form. Then:

(i) The equilibrium $x^* = 0$ is stable if and only if all eigevalues λ_l have non-positive real part $a_l \leq 0$ and for all eigenvalues with real part $a_l = 0$ the corresponding Jordan block J_l is one-dimensional.

(ii) The equilibrium $x^* = 0$ is locally asymptotically stabl if and only if all Eigenvalues λ_l have negative real part $a_l < 0$. In this case A is called a *Hurwitz matrix* or briefly *Hurwitz*.

Sketch of the proof: First one verifies that all stability properties are invariant under linear coordinate transformations $T \in \mathbb{R}^{n\times n}$, since the solutions $y(t; y_0)$ of the transformed system are given by

$$y(t; y_0) = T^{-1}x(t; Ty_0).$$

It is thus sufficient to check the stability properties for the Jordan canonical form of A given by

$$J = \begin{pmatrix} J_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_d \end{pmatrix}$$

with Jordan blocks of the form

$$J_l = \begin{pmatrix} \lambda_l & 1 & 0 & \cdots & 0 \\ 0 & \lambda_l & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \lambda_l & 1 \\ 0 & \cdots & \cdots & 0 & \lambda_l \end{pmatrix}, \tag{3.1}$$

with $j = 1, \dots, d$. We denote the solutions of $\dot{x}(t) = Jx(t)$ again with $x(t; x_0)$.

From the properties of the matrix exponential it follows that the solution

$$x(t; x_0) = e^{Jt}x_0$$

is of the form

$$x(t; x_0) = \begin{pmatrix} e^{J_1 t} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 0 & e^{J_d t} \end{pmatrix} x_0.$$

One further checks that

$$e^{J_l t} = e^{\lambda_l t} \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{m-1}}{(m-1)!} \\ 0 & 1 & t & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{t^2}{2!} \\ \vdots & \ddots & \ddots & 1 & t \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix},$$

where $e^{\lambda_l t}$ denotes the usual scalar exponential function, which satisfies

$$|e^{\lambda_l t}| = e^{a_l t} \begin{cases} \to 0, & a_l < 0 \\ \equiv 1, & a_l = 0 \\ \to \infty, & a_l > 0 \end{cases}$$

for $t \to \infty$.

The entries of $e^{J_l t}$ ar thus bounded if and only if the condition from (i) holds. Since, moreover, for any $k \in \mathbb{N}$ and any $\varepsilon > 0$ there is $c > 0$ with

$$e^{a_l t} t^k \leq c e^{(a_l+\varepsilon)t}, \tag{3.2}$$

the entries of $e^{J_l t}$ converge to 0 if and only if the condition from (ii) holds.

Via the matrix-vector multiplication $e^{Jt}x_0$ this property of the matrix entries carries over to the solution. Thus, the conditions in (i) and (ii) are equivalent to the respective stability conditions. □

In fact, the proof of (iii) shows global exponential stability, since the entries in (3.2) converge to 0 exponentially. The consequence of this fact is stated explicitly in the following theorem.

Satz 3.5 Consider the linear time-invariant differential equation (1.8) for a matrix $A \in \mathbb{R}^{n\times n}$ and let $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$, $\lambda_l = a_l + ib_l$, be the eigenvalues of A. Then the following four properties are equivalent.

(i) All eigenvalues λ_l have negative real part $a_l < 0$, i.e. the matrix is Hurwitz.

(ii) The equilibrium $x^* = 0$ is locally asymptotically stable.

(iii) The equilibrium $x^* = 0$ is globally exponentially stable. Here the constant $\sigma > 0$ from Definition 3.2(iv) can be chosen arbitrarily from the interval $(0, -\max_{l=1,\ldots,d} a_l)$.

(iv) The norm of the matrix exponential satisfies $\|e^{At}\| \leq ce^{-\sigma t}$ with σ as in (iii) and a constant $c > 0$ depending on the choice of σ.

Proof: (iii) $\Rightarrow$ (ii) follows from Remark 3.3, (ii) $\Rightarrow$ (i) follows from Theorem 3.4(iii) and (i) $\Rightarrow$ (iii) was shwn in the proof of Theorem 3.4(iii). Finally, (iii) $\Leftrightarrow$ (iv) follows immediately from the definition of the induced matrix norm (and holds for all norms in $\mathbb{R}^{n\times n}$ because they are all equivalent). □

Example 3.6 We consider the linear pendulum model from Chapter 1 for $u \equiv 0$ and neglecting the cart. The linearisation in the lower (= down hanging) equilibrium $x^* = \pi$ yields

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}$$

with eigenvalues

$$\lambda_{1/2} = -\frac{1}{2}k \pm \frac{1}{2}\sqrt{k^2 - 4g}.$$

Here $\sqrt{k^2 - 4g}$ is either complex or $< k$. In both cases we obtain $\mathrm{Re}\lambda_{1/2} < 0$ and thus exponential stability.

The linearisation in the upper (= upright or inverted) equilibrium $x^* = 0$ reads

$$A = \begin{pmatrix} 0 & 1 \\ g & -k \end{pmatrix}.$$

Here one computes the eigen values

$$\lambda_{1/2} = -\frac{1}{2}k \pm \frac{1}{2}\sqrt{k^2 + 4g},$$

of which the larger because of $\sqrt{k^2+4g} > k$ is always positive. Thus, we do not obtain stability. □

Remark 3.7 For discrete-time systems Theorem 3.5 remains essentially the same. However, in (i) the condition "real part $a_l < 0$" changes to "modulus $|\lambda_l| < 1$" and in (iv) the inequality $\|e^{At}\| \leq ce^{-\sigma t}$ becomes $\|A^k\| \leq ce^{-\sigma k}$. A matrix for which all eigenvalues satisfy the inequality $|\lambda_l| < 1$ is called *Schur-stable*. □

3.3 Lyapunov functions

In this section we will introduce an important tool for studying asymptotically stable differential equations, the so-called Lyapunov functions. Asymptotic and exponential stability only demand that the norm $\|x(t)\|$ of the solution tends to 0 for $t \to \infty$. For many analytical purposes it would, however, be much more convenient if the norm was strictly decreasing in t. This is, of course, not true in general. However, we can obtain strict monotonicity if we replace the norm $\|x(t)\|$ by a more general function. This is precisely the purpose of the Lyapunov function.

For linear systems we can restrict our consideration to so-called quadratic Lyapunov functions, as given by the following definition.

Definition 3.8 Let $A \in \mathbb{R}^{n\times n}$. A continuously diferentiable function $V : \mathbb{R}^n \to \mathbb{R}_0^+$ is called a *(quadratic) Lyapunov function* for A, if there are positive real constants $c_1, c_2, c_3 > 0$ such that the inequalities

$$c_1\|x\|^2 \leq V(x) \leq c_2\|x\|^2$$

and

$$DV(x)Ax \leq -c_3\|x\|^2$$

hold for all $x \in \mathbb{R}^n$. □

The following theorem shows that the existence of a quadratic Lyapunov function implies exponential stability of the corresponding differential equation.

Theorem 3.9 Let $A \in \mathbb{R}^{n\times n}$ be a matrix and $x(t;x_0)$ the solutions of the corresponding initial value problem (1.8), (1.9). Then, if there exists a quadratic Lyapunov funktion with constants $c_1, c_2, c_3 > 0$, then the solutions satisfy the estimate

$$\|x(t;x_0)\| \leq ce^{-\sigma t}\|x_0\|$$

for $\sigma = c_3/2c_2$ and $c = \sqrt{c_2/c_1}$, i.e., the equilibrium $x^* = 0$ is exponentially stable and the Matrix A ist Hurwitz.

Proof: From the condition on the derivative DV we conclude for $x = x(\tau, x_0)$ that

$$\frac{d}{dt}\bigg|_{t=\tau} V(x(t;x_0)) = DV(x(\tau;x_0))\dot{x}(\tau;x_0) = DV(x(\tau;x_0))Ax(\tau;x_0) \le -c_3\|x(\tau;x_0)\|^2$$

Since $-\|x\|^2 \le -V(x)/c_2$, for $\lambda = c_3/c_2$ this implies the inequality

$$\frac{d}{dt}V(x(t;x_0)) \le -\lambda V(x(t;x_0)).$$

This differential equation implies the inequality

$$V(x(t;x_0)) \le e^{-\lambda t}V(x_0),$$

(cf., e.g., the proof of [7, Satz 8.2]). Using the inequalities for $V(x)$ we thus obtain

$$\|x(t;x_0)\|^2 \le \frac{1}{c_1}e^{-\lambda t}V(x_0) \le \frac{c_2}{c_1}e^{-\lambda t}\|x_0\|^2$$

and hence by taking the square root on both sides

$$\|x(t;x_0)\| \le ce^{-\sigma t}\|x_0\|$$

for $c = \sqrt{c_2/c_1}$ and $\sigma = \lambda/2$. □

We will now look at the particular class of Lyapuniv functions, in which V is given by a bilinear form x^TPx with $P \in \mathbb{R}^{n\times n}$.

We recall that a matrix $P \in \mathbb{R}^{n\times n}$ is called *positive definite*, if $x^TPx > 0$ holds for all $x \in \mathbb{R}^n$ with $x \neq 0$. The following lemma summarises two properties of bilinear forms.

Lemma 3.10 For $P \in \mathbb{R}^{n\times n}$ it holds: (i) There exists a constant $c_2 > 0$ such that

$$-c_2\|x\|^2 \le x^TPx \le c_2\|x\|^2 \text{ for all } x \in \mathbb{R}^n.$$

(ii) P is positive definite if and only if there exists a constant $c_1 > 0$ with

$$c_1\|x\|^2 \le x^TPx \text{ for all } x \in \mathbb{R}^n.$$

Proof: The bilinearity implies for all $x \in \mathbb{R}^n$ with $x \neq 0$ and $y = x/\|x\|$ the identity

$$x^TPx = \|x\|^2y^TPy. \tag{3.3}$$

Since y^TPy is continuous in $y \in \mathbb{R}^n$, it assumes its minimum $c_{\min}$ and maximum $c_{\max}$ on the compact set $\{y \in \mathbb{R}^n \mid \|y\| = 1\}$.

(i) Inequality (i) now follows from (3.3) with $c_2 = \max\{c_{\max}, -c_{\min}\}$.

(ii) If P is positive definite, it follows that $c_{\min} > 0$ and (ii) follows with $c_1 = c_{\min}$. Conversely, positive definiteness of P follows immediately from (ii), thus we obtain the claimed equivalence. □

This leads us to the following conclusion.

Lemma 3.11 Let A, $P \in \mathbb{R}^{n\times n}$ and $c_3 > 0$ be such that the function $V(x) = x^TPx$ satisfies the inequality

$$DV(x)Ax \leq -c_3\|x\|^2$$

für alle $x \in \mathbb{R}^n$. Then P is positive definite if and only if A is Hurwitz. In this case V is a quadratic Lyapunov function.

Proof: If P is positive definite, Lemma 3.10(ii) immediately implies that V is a quadratic Lyapunov function, which by Theorem 3.9 yields that A is Hurwitz.

If P is not positive definite, then there exists $x_0 \in \mathbb{R}^n$ with $x_0 \neq 0$ and $V(x_0) \leq 0$. Since two different solutions of the differential equation cannot intersect, the solution $x(t;x_0)$ with $x_0 \neq 0$ can be 0. Thus the assumption on DV implies that $V(x(t;x_0))$ decreases monotonically for all $t \geq 0$. Particularly, there exists $c > 0$ such that $V(x(t;x_0)) \leq -c$ far all $t \geq 1$. Using the first estimate from Lemma 3.10(i) we then obtain that

$$\|x(t;x_0)\|^2 \geq c/c_2 > 0 \text{ for all } t \geq 1.$$

Hence $x(t;x_0)$ does not converge to 0, thus $x^* = 0$ is not exponentially stable and consequently A is not Hurwitz. □

We can rewrite the assumption on DV if we exploit the bilineare Form of the Lyapunov function.

Lemma 3.12 For a bilinear function $V(x) = x^TPx$ the following two statements are equivalent:

(i) $DV(x)Ax \leq -c_3\|x\|^2$ for all $x \in \mathbb{R}^n$ and a constant $c_3 > 0$

(ii) The matrix $C = -A^TP - PA$ is positive definite.

Proof: Since $x^TPy = y^TP^Tx$, we get $\frac{d}{dx}(x^TPy)Ax = \frac{d}{dx}(y^TP^Tx)Ax = y^TP^TAx = x^TA^TPy$. This implies by using the product rule

$$DV(x)Ax = x^TA^TPx + x^TPAx = x^T(A^TP + PA)x = -x^TCx.$$

Condition (i) is this equivalent to

$$x^TCx \geq c_3\|x\|^2 \text{ for all } x \in \mathbb{R}^n.$$

Because of Lemma 3.10 (ii) this condition is satisfied for some $c_3 > 0$ if and only if C is positive definite. □

The equation in Lemma 3.12 (iii) is known as *Lyapunov equation*. It is a natural idea to use this equation for the construction of Lyapunov functions. The question thus is whether for a given matrix A and a given positive definite matrix C we can find a positive definite matrix P that solves $A^TP + PA = -C$. The following lemma answers this question.

Lemma 3.13 For any matrix $A \in \mathbb{R}^{n\times n}$ and any positive definite matrix $C \in \mathbb{R}^{n\times n}$ the Lyapunov equation

$$A^TP + PA = -C \tag{3.4}$$

has a positive definite solution $P \in \mathbb{R}^{n\times n}$ if and only if A is Hurwitz. In this case, the solution P is unique.

Proof: If a positive definite solution P of (3.4) exists, then by Lemma 3.12 and 3.11 the function $V(x) = x^T P x$ is a quadratic Lyapunov Funktion. Thus, by Theorem 3.9 A is Hurwitz.

Assume conversely that A is Hurwitz and C is positive definite. We first show that (3.4) has a solution. Without loss of generality we can assume that A is in Jordan canonical form, since it is easily seen that P solves (3.4) if and only if $\widetilde{P} = (T^{-1})^T P T^{-1}$ solves

$$\widetilde{A}^T \widetilde{P} + \widetilde{P}\widetilde{A} = -(T^{-1})^T C T^{-1}$$

for $\widetilde{A} = TAT^{-1}$. We may thus assume that A is of the form

$$A = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \alpha_{n-1} & \beta_{n-1} \\ 0 & \cdots & \cdots & 0 & \alpha_n \end{pmatrix}, \tag{3.5}$$

where the α_i are the eigenvalues of A and the β_i are either 0 or 1. Writing the columns of P on top of each other into a large column vector $p \in \mathbb{R}^{n^2}$ and does the same for C and a vector c, equation (3.4) is equivalent to a linear system of equations Gleichungssystem

$$\overline{A}p = -c,$$

with $\overline{A} \in \mathbb{C}^{n^2 \times n^2}$. If A is of the form (3.5), by computing the coefficients one sees that $\overline{A}$ is a lower triangular matrix of the form

$$\overline{A} = \begin{pmatrix} \bar{\alpha}_1 & 0 & 0 & \cdots & 0 \\ * & \bar{\alpha}_2 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \bar{\alpha}_{n^2-1} & 0 \\ * & \cdots & \cdots & * & \bar{\alpha}_{n^2} \end{pmatrix}.$$

Here $*$ denotes arbitrary values while the values $\bar{\alpha}_i$ are of the form $\bar{\alpha}_i = \lambda_j + \lambda_k$, with λ_i being the eigenvalues of A. It is now known from linear algebra that

(i) the elements on the diagonal of a triangular matrix are its eigenvalues

(ii) a matrix is invertible if and only if none of its eigenvalues equals zero.

Since A is Hurwitz all λ_i have negative real part. Consequently, all the $\bar{a}_i$ have negative real part, too, and are thus non-zero. i.e. because of (i) and (ii) the matrix $\overline{A}$ is invertible. Hence, there is exactly one solution of the equation $\overline{A}p = c$ and thus exactly one solution P of the Lyapunov equation (3.4).

It remains to be shown that this solution P is positive definite ist. Because of Lemma 3.12 this matrix P satisfies all assumptions of Lemma 3.11. Since A is Hurwitz, by Lemma 3.11 P must be positive definite. □

The following theorem summarises the main result of this section.

Theorem 3.14 Consider $A \in \mathbb{R}^{n \times n}$. Then a quadratic Lyapunov function for the linear differential equation (1.8) exists if and only if $x^* = 0$ is exponentially stable, i.e. if A is Hurwitz.

Beweis: Assume a quadratic Lyapunov function V exists. Then by Theorem 3.9 the matrix A is Hurwitz.

Conversely, let A be Hurwitz. Then by Lemma 3.13 there exists a positive definite Matrix P that solves the Lyapunov equation (3.4) for a positive definite matrix C. Then, because of Lemma 3.12 and Lemma 3.11, $V(x) = x^T P x$ is a quadratic Lyapunov function. □

The existence of a quadratic Lyapunov function is thus a necessary and sufficient condition for the exponential stability of the equilibrium $x^* = 0$. It yields a characterization that is equivalent to the eigenvalue condition from Theorem 3.5.

Example 3.15 For the linearisation of the pendulum model in the lower equilibrium with

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}$$

is the bilinear Lyapunov function for $C = \mathrm{Id}$ given by the matrix

$$P = \begin{pmatrix} \frac{k^2+g^2+g}{2gk} & \frac{1}{2g} \\ \frac{1}{2g} & \frac{g+1}{2gk} \end{pmatrix}.$$

□

Remark 3.16 For discrete time systems the inequality in Definition 3.8 changes to

$$V(Ax) - V(x) \leq -c_3 \|x\|^2.$$

Due to this, the Lyapunov equation (3.4) becomes

$$A^T P A - P = -C. \tag{3.6}$$

With these changes, all results in this section remain valid. □

3.4 The stabilisation problem for linear control systems

We now have all the technical tools to tackle the stabilisation problem for linear control systems. In the exercises we have seen that a pre-computed control function $u(t)$ on a large time horizon does in general not work reliably for steering the system into a desired set point (for instance 0) and holding it there. Even the very small errors that occur in an accurate numerical simulation were enough to drive the system away from the desired point.

We therefore pursue a different approach. Instead of using an open-loop control that depends on t, we use a closed-loop control which computes the control value from the current state according to the formula $u(t) = F(x(t))$, for a function $F : \mathbb{R}^n \to \mathbb{R}^m$ that needs to be determined. Such a function, which assignes a control value to each state, is called a *feedback law* (also *state feedback law*, *(feedback) controller* or *regulator*). Since our system is linear. it appears natural to choose the feedback law also as a linear map, i.e., $u = Fx$ for a matrix $F \in \mathbb{R}^{m\times n}$. This has the advantage that the resulting *closed-loop system*

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

is a linear time-incvariant differential eqution, to which the theory from the previous sections applies.

To control the state of the system 0 and keep it there, we can solve the following stabilisation problem.

Definition 3.17 Consider a linear control system (1.3)

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

with matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. The *(feedback) stabilisation problem* for (1.3) consists in finding a linear map $F : \mathbb{R}^m \to \mathbb{R}^n$ (or, equivalently, the corresponding matrix $F \in \mathbb{R}^{m\times n}$), such that the equilibrium $x^* = 0$ is asymptotically stable for the resulting closed-loop system, i.e. for the linear ordinary differential equation

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

□

The following lemma is an easy consequence of our criteria for asymptotic stability.

Lemma 3.18 Consider two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then the matrix $F \in \mathbb{R}^{m\times n}$ solves the stabilisation problem, if and only if all eigenvalues of the matrix $A+BF \in \mathbb{R}^{n\times n}$ have negative real part.

Below we will investigate when — for given matrices A and B — such a matrix F exists and how it can be computed. Before this, we consider a simple example.

Example 3.19 As a simple and intuitively solvable example for a stabilisation problem we consider a (very simple) model for the regulation of heating. Assume we want to control the temperature x_1 at a fixed point in a room, where it can be measured. To simplify the problem we shift the temperature scale such that the desired temperature is $x_1^* = 0$[1]. The room contains a heating device with temperature x_2, which we can control. More precisely, we assume that the change of x_2 depends on the flow rate of the hot water through the heating device, which is controlled by u. This leads to the differential equation $\dot{x}_2(t) = u(t)$. In other words, the control value u causes the temperature to increase (for $u > 0$) or to decrease (if $(u < 0)$. For the temperature x_1 in the fixed point we assume that it satisfies

[1] One should thus think of x_1 as the deviation from the desired temperature rather than of an absolute value.

the differential equation $\dot{x}_1(t) = -x_1(t) + x_2(t)$. This means that for constant heating temperature x_2 we obtain

$$x_1(t) = e^{-t}x_1(0) + (1 - e^{-t})x_2(0),$$

i.e. the room temperature x_1 converges exponentially to the temperature x_2 of the heating device.

This modelling leads to the control system

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

A very intuitive regulation strategy is now the following: If $x_1 > x_1^* = 0$, we reduce the temperature in x_2, i.e. we choose $u < 0$. In the opposite case, i.e. if $x_1 < x_1^* = 0$, we increase the temperature x_2 by choosing $u > 0$. Since our feedback law should be linear, this can be achieved by setting $F(x) = -\lambda x_1$ for a $\lambda > 0$, or, in matrix notation $F = (-\lambda, 0)$ (observe that here we have $n = 2$ and $m = 1$, implying that F is a 1×2 matrix, i.e., a two-dimensional row vector). This gives us the closed-loop system

$$\dot{x}(t) = \begin{pmatrix} -1 & 1 \\ -\lambda & 0 \end{pmatrix} x(t).$$

Computing the eigenvalues of this matrix for $\lambda > 0$ reveals that all real parte are negative. Hence we have — inadvertently — solved the stabilization problem and consequently for arbitrary initial values the temperatures $x_1(t)$ and $x_2(t)$ converge to 0 exponentially fast. In particular, x_1 converges exponentially fast towards the desired temperature $x_1^* = 0$. This proves that our intuitively designed feedback controller achieves the desired result.

If we can measure the temperature x_2 at the heating device, then we could also choose $F(x) = -\lambda x_2$, or, in matrix notation, $F = (0, -\lambda)$ as feedback law. Again by computing the eigenvalues one sees that the closed-loop system is exponentially stable for all $\lambda > 0$. The bevaviour of the system for the two different feedback laws is wuite different, though. We will investigate this in the exercises. □

Remark 3.20 In practice, the state $x(t)$ of a system can often not be measured completely. Rather, one only has access to a vector $y = Cx$ of output values, for a matrix $C \in \mathbb{R}^{d \times n}$. In this case the feedback can only depend on the output vector y. The corresponding concept is called an *output feedback*.

This is actually similar to what we did in the example above, because we only used information from $x_1(t)$ or $x_2(t)$, in the feedback law, but not both. In the rest of this chapter we will assume that the entire vector $x(t)$ is accessible and can be used in the feedback law. The general case will then be addressed in Chapter 4. □

3.5 Solution of the stabilisation problem with one-dimensional control

In this section we investigate conditions under which we can find a solution for the stabilisation problem from Definition 3.17 with one-dimensional control. In particular we will give a necessary and sufficient condition on the matrices A und B in (1.3), under which the problem is solvable. The individual steps of this derivation provide a constructive method for computing the desired stabilising feedback law F.

In this derivation coordinate transformations will again play an important role. A coordinate change with transformation matrix $T \in \mathbb{R}^{n\times n}$ transforms the original control system

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.7}$$

into the form

$$\dot{x}(t) = \widetilde{A}x(t) + \widetilde{B}u(t) \tag{3.8}$$

with $\widetilde{A} = T^{-1}AT$ and $\widetilde{B} = T^{-1}B$. A feedback law F for (3.7) can be transformed into a feedback law for (3.8) via $\widetilde{F} = FT$; this follows immediately from the identity $T^{-1}(A + BF)T = \widetilde{A} + \widetilde{B}\widetilde{F}$ that the transformed system needs to satisfy.

We already saw in Lemma 2.14 that a pair (A, B) can be transformed into the form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

with controllable pair (A_1, B_1) and non-controllable rest.

Here we need yet another coordinate change, whch holds for controllable systems with one-dimensional control u. In this case we have $m = 1$, i.e., $B \in \mathbb{R}^{n\times 1}$. This means that the matrix B ist an n-dimensional column vector.

Lemma 3.21 Consider $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$. Then the pair (A, B) is controllable if and only if there is a coordinate transformation S with

$$\widetilde{A} = S^{-1}AS = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} \quad \text{und } \widetilde{B} = S^{-1}B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Here the values $\alpha_i \in \mathbb{R}$ are the coefficients of the characteristic polynomial of A written in the form $\chi_A(z) = z^n - \alpha_n z^{n-1} - \cdots - \alpha_2 z - \alpha_1$.

Proof: We start by proving that for matrices $\widetilde{A}$ in the form of the lemma the values α_i are indeed the coeffizients of the characteristic polynomial. We prove this claim by induction over n: For $n = 1$ the claim is immediately clear. For the induction step let $A_n \in \mathbb{R}^{n\times n}$ by of the form of the lemma and $A_{n+1} \in \mathbb{R}^{n\times n}$ be given by

$$A_{n+1} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & & & \\ \vdots & & A_n & \\ \alpha_0 & & & \end{pmatrix}.$$

If we compute $\det(z\mathrm{Id}_{\mathbb{R}^{n+1}} - A_{n+1})$ according to the first row, we obtain

$$\chi_{A_{n+1}} = z\chi_{A_n}(z) - \alpha_0 = z^{n+1} - \alpha_n z^n - \cdots - \alpha_1 z - \alpha_0,$$

which yields exactly the desired expression after renumbering the α_i.

Let us now assume that S exists. Then by a direct computation one sees that

$$\widetilde{R} = (\widetilde{B}\, \widetilde{A}\widetilde{B}\, \ldots\, \widetilde{A}^{n-1}\widetilde{B}) = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & \iddots & * \\ 0 & 1 & * & * \\ 1 & * & \cdots & * \end{pmatrix}, \tag{3.9}$$

where $*$ denotes arbitrary values. This matrix has full rank, since by reordering the rows (which does not change the rank of the matrix) we obtain an upper triangular matrix with only ones on the diagonal. This is obviously invertible and thus has full rank. This implies that $(\widetilde{A}, \widetilde{B})$ is controllable and since controllability persists under coordinate changes, the pair (A, B) is controllable, too.

Conversely, assume that (A, B) is controllable. Then the matrix $R = (B\, AB\, \ldots\, A^{n-1}B)$ is invertible and consequently R^{-1} exists. We now first show that the equation $R^{-1}AR = \widetilde{A}^T$ holds. To this end, we show the equivalent identity $AR = R\widetilde{A}^T$. Using the theorem of Cayley-Hamilton, this follows from the computation

$$\begin{aligned} AR &= A(B\, AB\, \ldots\, A^{n-1}B) = (AB\, A^2B\, \ldots\, A^{n-1}B\;\; A^nB) \\ &= (AB\, A^2B\, \ldots\, A^{n-1}B\;\; \alpha_n A^{n-1}B + \cdots + \alpha_1 B) \\ &= (B\, AB\, \ldots\, A^{n-1}B) \begin{pmatrix} 0 & \cdots & 0 & \alpha_1 \\ 1 & \cdots & 0 & \alpha_2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & \alpha_n \end{pmatrix} = R\widetilde{A}^T \end{aligned}$$

For $\widetilde{R}$ from (3.9) the analogous computation yields $\widetilde{R}^{-1}\widetilde{A}\widetilde{R} = \widetilde{A}^T$ and thus

$$\widetilde{A} = \widetilde{R}\widetilde{A}^T\widetilde{R}^{-1} = \widetilde{R}R^{-1}AR\widetilde{R}^{-1}.$$

The definitions of R and $\widetilde{R}$ moreover imply $R(1, 0, \ldots, 0)^T = B$ and $\widetilde{R}(1, 0, \ldots, 0)^T = \widetilde{B}$, hence $R\widetilde{R}^{-1}\widetilde{B} = B$. Thus, $S = R\widetilde{R}^{-1}$ is the desired transformation matrix. □

The form of the pair $(\widetilde{A}, \widetilde{B})$ achieved in 3.21 is called the *controllable canonical form*. Observe that the coordinate transformation S can be computed based on the knowledge of A, B and the coefficients of the characteristic polynomial of A.

Using the controllable canonical form we can now proceed to solving the stabilisation problem for $u \in \mathbb{R}$.

To this end, we first reformulate the stabilisation problem by means of characteristic polynomials. This can be done for arbitrary dimensions of the control u.

Definition 3.22 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. A polynomial χ is called *assignable* for the control system if there exists a linear feedback law $F \in \mathbb{R}^{m\times n}$ such that $\chi = \chi_{A+BF}$ holds for the characteristic polynomial χ_{A+BF} of the matrix $A + BF$. □

Sincce we know that the roots of the characteristic polynomial are exactly the eigenvalues of the corresponding matrix, Lemma 3.18 immediately yields the following characterisation.

Lemma 3.23 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then the stabilisation problem is solvable if and only if there exists an assignable polynomial, for which all roots in $\mathbb{C}$ have negative real part.

The following theorem shows the relation between controllability of (A, B) and assignability of polynomials.

Theorem 3.24 Consider a control system (1.3) wih matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$, i.e. with one-dimensional control. Then the following two properties are equivalent.

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

(ii) Every polynomial of the form $\chi(z) = z^n - \beta_n z^{n-1} - \cdots - \beta_2 z - \beta_1$ with $\beta_1, \ldots, \beta_n \in \mathbb{R}$ is assignable.

Proof: (i) ⇒ (ii): Let (A, B) be controllable and let S be the coordinate transformation from Lemma 3.21. We define

$$\widetilde{F} = (\beta_1 - \alpha_1 \;\; \beta_2 - \alpha_2 \;\; \ldots \;\; \beta_n - \alpha_n) \in \mathbb{R}^{1\times n}.$$

Then we obtain

$$\begin{aligned}
\widetilde{A} + \widetilde{B}\widetilde{F} &= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} (\beta_1 - \alpha_1 \;\; \beta_2 - \alpha_2 \;\; \ldots \;\; \beta_n - \alpha_n) \\
&= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix} + \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ \beta_1 - \alpha_1 & \beta_2 - \alpha_2 & \cdots & \beta_n - \alpha_n \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix}.
\end{aligned}$$

Now the second assertion of Lemma 3.21 yields that $\chi_{\widetilde{A}+\widetilde{B}\widetilde{F}} = \chi$. Hence, after transformation to original coordinates, $F = \widetilde{F}S^{-1}$ is the desired feedback matrix, since the characteristic polynomial of a matrix is invariant under coordinate transformations.

(ii) ⇒ (i): We show the implication "not (i) ⇒ not (ii)":

Let (A, B) be not controllable. Let T be the coordinate transformation from Lemma 2.14. Then for any arbitrary feedback law $\widetilde{F} = (F_1 \;\; F_2)$ we obtain

$$\widetilde{A} + \widetilde{B}\widetilde{F} = \begin{pmatrix} A_1 + B_1F_1 & A_2 + B_1F_2 \\ 0 & A_3 \end{pmatrix} =: \widetilde{D}.$$

The characteristic polynomial of this matrix satisfies

$$\chi_{\widetilde{D}} = \chi_{A_1+B_1F_1}\chi_{A_3}.$$

Hence (recalling that (A_1, B_1) is controllable) the assignable polynomials are of the form $\chi = \chi_k\chi_u$, where χ_k is an arbitrary normed polynomial of degree d and $\chi_u = \chi_{A_3}$. This is a strictly smaller set of polynomials than specified in (ii). Thus, (ii) cannot hold. □

Obviously, for the purpose of stabilization we do not need that *every* polynomial is assignable. We only need to find an assignable polynomial, whose roots all have negative real parts. The proof of Theorem 3.24 already suggests when this is possible.

Theorem 3.25 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$, i.e. with one-dimensional control. Let $A_1 \in \mathbb{R}^{d\times d}$, $A_2 \in \mathbb{R}^{d\times(n-d)}$, $A_3 \in \mathbb{R}^{(n-d)\times(n-d)}$, and $B_1 \in \mathbb{R}^{d\times 1}$ be the matrices from Lemma 2.14 with the convention that $A_1 = A$ and $B_1 = B$ in case (A, B) is controllable.

Then the assignable polynomials (1.3) are of the form $\chi = \chi_k\chi_{A_3}$, where χ_k is an arbitrary normed polynomial of degree d and χ_{A_3} is the characteristic polynomial of the matrix A_3, i.e. the uncontrollable part of the characteristic polynomial χ_A, cf. Definition 2.15. Here we use the convention $\chi_{A_3} = 1$ if $d = n$.

In particular, the stabilisation problem is solvable if and only if all eigenvalues of A_3 have negative real part (the eigenvalues of A_3 are also called the "uncontrollable eigenvalues"). In this case we call the pair (A, B) *stabilisable.*

Proof: The first statement follows immediately from the second part of the proof of Theorem 3.24. The statement about the stabilisation problem then follows from Lemma 3.23. □

Remark 3.26 All statements of this section also hold for discrete-time systems if the condition "the real part of the eigenvalue is less than 0" is replaced by "the modulus of the eigenvalue is less than 1". □

3.6 Solution of the stabilisation problem with multidimensional control

The results for multidimensional control $m > 1$ are completely analogous to those for one-dimensional control. A direct proof is, however, very technical, because we cannot use Lemma 3.21. We will thus reduce the multidimensinoal case to the one-dimensional case by using the following lemma, which is known as *Heymann's Lemma.*

Lemma 3.27 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Let the pair (A, B) be controllable and let $v \in \mathbb{R}^m$ be a vector with $\overline{B} = Bv \neq 0$. Then there exists a matrix $\overline{F} \in \mathbb{R}^{m\times n}$ such that the control system

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

with one-dimensional control $\bar{u}(t)$ is controllable.

Proof: By means of the recursive definition $x_{i+1} := Ax_i + Bu_i$ with appropriate u_i we first construct linearly independent vectors $x_1, \ldots, x_n \in \mathbb{R}^n$ which for all $l \in \{1, \ldots, n\}$ satisfy

$$Ax_i \in V_l \text{ für } i = 1, \ldots, l-1 \text{ with } V_l = \langle x_1, \ldots, x_l\rangle. \tag{3.10}$$

In order to construct these vectors, set $x_1 := \overline{B}$ (we can interpret the $n \times 1$ matrix $\overline{B}$ as column vector). Observe, that the property (3.10) is trivially satisfied for $l = 1$ and every $x_1 \neq 0$.

For $k \in 1, \ldots, n-1$ and given linearly independent vectors $x_1, \ldots, x_k$, which satisfy (3.10) for $l \in \{1, \ldots, k\}$, we now construct a vector x_{k+1}, such that $x_1, \ldots, x_k, x_{k+1}$ are linearly independent and (3.10) holds for $l \in \{1, \ldots, k+1\}$:

Case 1: $Ax_k \notin V_k$: set $u_k := 0 \in \mathbb{R}^m$ and $x_{k+1} := Ax_k$.

Case 2: $Ax_k \in V_k$: Since (3.10) holds, we obtain that V_k is A-invariant. From Chapter 2 we know that $\langle A \,|\, \mathrm{im}\, B\rangle = \mathrm{im}\, R$ with controllability matrix $R = (B\; AB\; \ldots\; A^{n-1}B)$ is the smallest A-invariant subspace that contains the image of B. Since (A, B) is controllable, we have $\langle A \,|\, \mathrm{im}\, B\rangle = \mathbb{R}^n$. Since V_k is now an A-invariant subspace with $\dim V_k = k < n$, it cannot contain the image of B. Hence, there is $u_k \in \mathbb{R}^m$ with $Ax_k + Bu_k \notin V_k$ and we set $x_{k+1} := Ax_k + Bu_k$.

We now construct the desired map $\overline{F}$ from the vectors $x_1, \ldots, x_n$. Since the x_i are linearly independent, the matrix $X = (x_1\; \ldots\; x_n)$ is invertible and we can define $\overline{F} := UX^{-1}$ for $U = (u_1, \ldots, u_n) \in \mathbb{R}^{m\times n}$. Here the $u_1, \ldots, u_{n-1}$ denote the control vectors used in the recursion defining the x_i, above, and $u_n := 0 \in \mathbb{R}^m$. This yields $\overline{F}x_i = u_i$ and thus $(A + B\overline{F})x_i = x_{i+1}$ for $i = 1, \ldots, n-1$. Since $\overline{B} = x_1$ we obtain

$$(\overline{B}\;\; (A + B\overline{F})\overline{B}\;\; \ldots\;\; (A + B\overline{F})^{n-1}\overline{B}) = X,$$

hence $(\overline{B}\;\; (A+B\overline{F})\overline{B}\;\; \ldots\;\; (A+B\overline{F})^{n-1}\overline{B})$ has rank n, implying that the pair $(A+B\overline{F}, \overline{B})$ is controllable. □

Wit this result, Theorems 3.24 and 3.25 are easily generalised to arbitrary control dimensions.

Theorem 3.28 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then the following two properties are equivalent.

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

(ii) Every polynomial of the form $\chi(z) = z^n - \beta_n z^{n-1} - \cdots - \beta_2 z - \beta_1$ mit $\beta_1, \ldots, \beta_n \in \mathbb{R}$ is assignable.

Proof: (i) ⇒ (ii): Let (A, B) be controllable and χ be given. Let $\overline{F} \in \mathbb{R}^{n\times m}$ and $\overline{B} \in \mathbb{R}^{n\times 1}$ be the matrices from Lemma 3.27 for some $v \in \mathbb{R}^m$ with $Bv \neq 0$ (note that such a $v \in \mathbb{R}^n$ exists since (A, B) is controllable, which implies $B \neq 0$). Then the pair $(A + B\overline{F}, \overline{B})$ is controllable and Theorem 3.24 immplies the existence of a feedback matrix $F_1 \in \mathbb{R}^{1\times n}$ with

$$\chi_{A+B\overline{F}+\overline{B}F_1} = \chi.$$

Since

$$A + B\overline{F} + \overline{B}F_1 = A + B\overline{F} + BvF_1 = A + B(\overline{F} + vF_1)$$

we can define the desired feedbacl law as $F = \overline{F} + vF_1$.

(ii) ⇒ (i): Completely identical to the proof of Theorem 3.24. □

Theorem 3.29 Consider a control system (1.3) with matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Let $A_1 \in \mathbb{R}^{d\times d}$, $A_2 \in \mathbb{R}^{d\times(n-d)}$, $A_3 \in \mathbb{R}^{(n-d)\times(n-d)}$ and $B_1 \in \mathbb{R}^{d\times m}$ be the matrices from Lemma 2.14 with the convention that $A_1 = A$ and $B_1 = B$ if (A, B) is controllable.

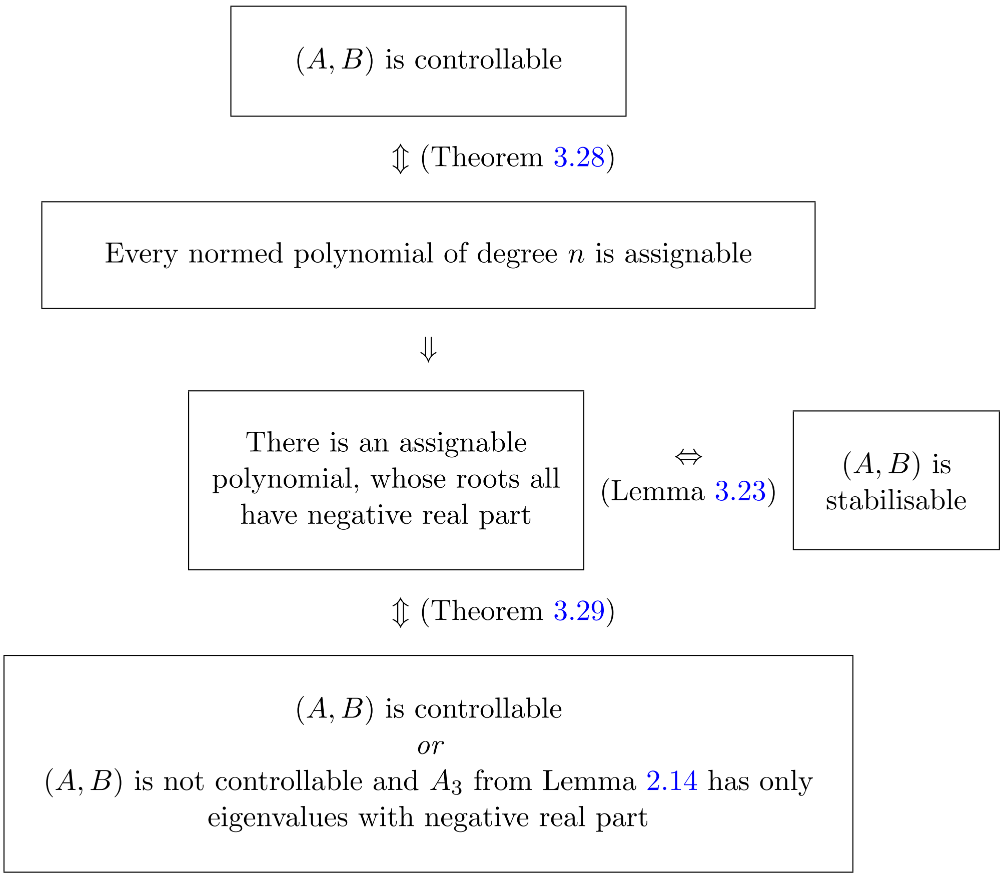
Then the assignable polynomials (1.3) are of the form $\chi = \chi_k \chi_u$, where χ_k is an arbitrary normed polynomial of degree d and χ_u is the characteristic polynomial of the matrix A_3. Here we use the convention $\chi_{A_3} = 1$ if $d = n$.

In particular, the stabilisation problem is solvable if and only if all eigenvalues of A_3 have negative real part. In this case we call the pair (A, B) *stabilisable.*

Proof: Completely analogous to the proof of Theorem 3.25. □

Bemerkung 3.30 Theorem 3.29 is called *pole shifting theorem*, because the roots of the characteristic polynomial are often called “poles” in control engineering (the reason will be explained later in Remark 5.15). This theorem describes how these roots can be shifted by means of a suitable choice of the feedback. □

We can now illustrate the main results for the stabilisation problem in the following schematic way:


(A, B) is controllable
$\Updownarrow$ (Theorem 3.28)
Every normed polynomial of degree n is assignable
$\Downarrow$
There is an assignable polynomial, whose roots all have negative real part
$\Leftrightarrow$ (Lemma 3.23)
(A, B) is stabilisable
$\Updownarrow$ (Theorem 3.29)
(A, B) is controllable
or
(A, B) is not controllable and A_3 from Lemma 2.14 has only eigenvalues with negative real part

If one replaces “negative real part” everywhere by “modulus less than 1”, then these statements remain valid for discrete-time systems.

3.7 Local stabilisation of nonlinear systems

In this section we show that a linear stabilizing feedback law can be used for the local stabilisation of a nonlinear control system. The basis for this fact is the following theorem from the theory of ordinary differential equations.

Theorem 3.31 Consider the nonlinear differential equation

$$\dot{x} = g(x) \tag{3.11}$$

with equilibrium $x^* \in \mathbb{R}^n$ and continuously differentiable vector field $g : \mathbb{R}^n \to \mathbb{R}^n$. Consider moreover the linearisation

$$\dot{y} = \widehat{A}y \qquad \text{with } \widehat{A} = \frac{d}{dx}g(x^*). \tag{3.12}$$

Then the equilibrium x^* is locally exponentially stable for equation (3.11) if and only if the equilibrium $y^* = 0$ is globally exponentially stable for equation (3.12).

A proof can be found, e.g., in [7, Satz 8.8].

Consider now the nonlinear control system

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

with equilibrium (x^*, u^*), i.e., $f(x^*, u^*) = 0$, and its linearisation

$$\dot{y} = Ay + Bv \qquad \text{with } A = \frac{\partial}{\partial x} f(x^*, u^*),\ B = \frac{\partial}{\partial u} f(x^*, u^*).$$

Recall from the introduction that f, A and B are related via $f(x,u) \approx A(x-x^*)+B(u-u^*)$, which implies that y and v are related to x and u via $y = x - x^*$ and $v = u - u^*$.

Let F be a stabilising feedback law for the linear control system. For the linear system this generates the control value $v = Fy$, which implies that using the above relation between x, y, u and v we obtain $u = u^* + F(x - x^*)$. Inserting this expression into f we obtain equation

$$\dot{x} = f(x, u^* + F(x - x^*)) =: g(x). \tag{3.13}$$

The linearisation of this equation is given by

$$\dot{y} = \widehat{A}y$$

with

$$\widehat{A} = \frac{d}{dx} g(x^*) = \frac{d}{dx}\bigg|_{x=x^*} f(x, u^* + F(x - x^*)) = \frac{\partial}{\partial x} f(x^*, u^*) + \frac{\partial}{\partial u} f(x^*, u^*)F = A + BF.$$

Since F stabilises the linear system exponentially, the equlibrium $y^* = 0$ is exponentially stable for (3.12) and Theorem 3.31 implies that x^* is a locally exponentially stable equilibrium for the nonlinear system with linear feedback law (3.13). The stabilising linear feedback law thus stabilises the nonlinear system locally in x^*.

Example 3.32 Consider the nonlinear inverted pendulum (1.5)

$$\left.\begin{array}{rcl} \dot{x}_1(t) & = & x_2(t) \\ \dot{x}_2(t) & = & -kx_2(t) + g\sin x_1(t) + u(t)\cos x_1(t) \\ \dot{x}_3(t) & = & x_4(t) \\ \dot{x}_4(t) & = & u(t) \end{array}\right\} =: f(x(t), u(t)).$$

The linearisation in $(x^*, u^*) = (0,0)$ reads

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ g & -k & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

cf. (1.6). In the exercises we computed a stabilising linear feedback law $F : \mathbb{R}^4 \to \mathbb{R}$ for this linear system. The corresponding matrix $F \in \mathbb{R}^{1\times 4}$ is

$$F = \Big(-\frac{g+k^2}{g^2} - \frac{4k}{g} - 6 - g,\ -\frac{k}{g^2} - \frac{4}{g} - 4 + k,\ \frac{1}{g},\ \frac{k}{g^2} + \frac{4}{g}\Big)$$

Figure (3.1) shows that this feedback law stabilises the nonlinear pendulum. The figure shows the components of the trajectory $x(t, x_0, F)$ for $x_0 = (1/2,\ 0,\ 0,\ 0)^T$.

□

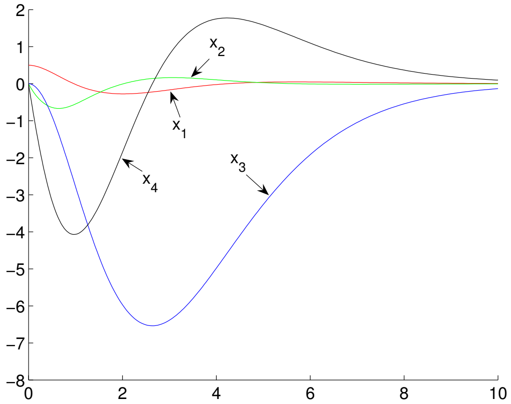


Figure 3.1: Solution of the nonlinear pendulum equation with stabilising linear feedback law

Chapter 4

Observability and observers

The solution for the stabilisation problem that we derived in the last chapter assumes that the whole state vector $x(t)$ is accessible for evaluation the control value $u(t) = Fx(t)$. In practical problems, this is hardly ever the case. Typically one can only access certain values $y(t) = C(x(t)) \in \mathbb{R}^k$, delivered by sensors, from which $u(t)$ must then be computed. Since in this part of this course we consider linear sytems, we assume that the measurement map $C : \mathbb{R}^n \to \mathbb{R}^k$ is also linear, i.e. given by a matrix $C \in \mathbb{R}^{k\times n}$.

Definition 4.1 A *linear control system with output* is given by[1] the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) \tag{4.1}$$

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

In this chapter we will derive conditions under which the stabilisation problem is solvable for (4.1) and show, how a feedback controller must be constructed in this case.

4.1 Observability and Duality

The most important question when analysing (4.1) is, how much "information" is contained in the output $y(t) = Cx(t)$. This is formalised by the following definitions.

Definition 4.2 (i) Two states $x_1, x_2 \in \mathbb{R}^n$ are called *distinguishable*, if there are $u \in \mathcal{U}$ and $t \geq 0$ with

$$Cx(t, x_1, u) \neq Cx(t, x_2, u).$$

(ii) The system (4.1) is called *observable*, if any two states $x_1, x_2 \in \mathbb{R}^n$ with $x_1 \neq x_2$ are distinguishable. □

The following lemma shows that by virtue of the linearity of the system, distinguishability can be expressed in a simpler way.

[1] Sometimes the extended variant $y(t) = Cx(t) + Du(t)$ with $D \in \mathbb{R}^{k\times m}$ is considered. The form we consider is obtained from this extended form by setting $D = 0$.

Lemma 4.3 Tow states $x_1, x_2 \in \mathbb{R}^n$ are distinguishable if and only if there exists $t \geq 0$ with

$$Cx(t, x_1 - x_2, 0) \neq 0.$$

Proof: The superposition principle (1.15) implies the identity

$$x(t, x_1, u) - x(t, x_2, u) = x(t, x_1 - x_2, 0),$$

which immediately implies the assertion since C is a linear map. □

This lemma implies that observability of (4.1) does not depend on u, and thus not on B. If the system (4.1) is observable, then we call the pair (A, C) *observable*.

Moreover, the lemma motivates the following definition.

Definition 4.4 (i) We call $x_0 \in \mathbb{R}^n$ *observable* if there is $t \geq 0$ with

$$Cx(t, x_0, 0) \neq 0$$

and *non-observable on* $[0, t]$ if

$$Cx(s, x_0, 0) = 0$$

for all $s \in [0, t]$.

(ii) We define the sets of *non-observable states on* $[0, t]$ for $t > 0$ as

$$\mathcal{N}(t) := \{x_0 \in \mathbb{R}^n \,|\, Cx(s, x_0, 0) = 0 \text{ for all } s \in [0, t]\}$$

and the set of *non-observable states* as

$$\mathcal{N} := \bigcap_{t>0} \mathcal{N}(t).$$

□

The following lemma clarifies the structure of these sets.

Lemma 4.5 For all $t > 0$ the identity

$$\mathcal{N} = \mathcal{N}(t) = \bigcap_{i=0}^{n-1} \ker(CA^i)$$

holds. In particular, $\mathcal{N}$ is a linear subspace, which moreover is A-invariant, i.e. it holds that $A\mathcal{N} \subseteq \mathcal{N}$.

Proof: A state $x_0 \in \mathbb{R}^n$ is contained in $\mathcal{N}(t)$ if and only if

$$0 = Cx(s, x_0, 0) = Ce^{As}x_0 \text{ for all } s \in [0, t]. \tag{4.2}$$

Now let $x_0 \in \bigcap_{i=0}^{n-1} \ker(CA^i)$. Then by the theorem of Cayley-Hamilton the identity $CA^i x_0 = 0$ holds for all $i \in \mathbb{N}_0$. The series representation of e^{As} then implies $Ce^{As}x_0 = 0$ for all $s \geq 0$ and thus (4.2), i.e., $x_0 \in \mathcal{N}(t)$.

Conversely, let $x_0 \in \mathcal{N}(t)$. Then by (4.2) we obtain $Ce^{As}x_0 = 0$. The i-th derivative of this expression in $s = 0$ then satisfies

$$CA^i x_0 = 0, \;\; i \in \mathbb{N}_0$$

and thus in particular $x_0 \in \ker CA^i$, $i = 0, \dots, n-1$. This implies $x_0 \in \mathcal{N}(t)$.

The A-invariance then follows from the expression for $\mathcal{N}$ using the theorem of Cayley-Hamilton. □

Obviously there is a certain similarity here with the controllability analysis, particularly with the sets $\mathcal{R}(t)$ und $\mathcal{R}$. We now show that this is more than just a superficial similarity. To this end, we need an appropriately defined dual system.

Definition 4.6 For a control system (4.1) defined by the matrices (A, B, C) the *dual system* is defined by the matrices (A^T, C^T, B^T). The dual system to

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \qquad\qquad x(t) \in \mathbb{R}^n,\, u(t) \in \mathbb{R}^m,\, y(t) \in \mathbb{R}^k$$

thus reads

$$\dot{x}(t) = A^T x(t) + C^T u(t), \quad y(t) = B^T x(t), \qquad x(t) \in \mathbb{R}^n,\, u(t) \in \mathbb{R}^k,\, y(t) \in \mathbb{R}^m.$$

□

In words, the dual system is obtained by transposing all matrices and swapping B and C, i.e., the input and the output matrix.

Theorem 4.7 For a control system (4.1) given by (A, B, C) and its dual system given by (A^T, C^T, B^T) we define

$$\begin{aligned} \mathcal{R} &= \langle A \,|\, \operatorname{im} B \rangle & \mathcal{N} &= \textstyle\bigcap_{i=0}^{n-1} \ker(CA^i) \\ \mathcal{R}^T &= \langle A^T \,|\, \operatorname{im} C^T \rangle & \mathcal{N}^T &= \textstyle\bigcap_{i=0}^{n-1} \ker(B^T (A^T)^i). \end{aligned}$$

Then the identities

$$\mathcal{R}^T = \mathcal{N}^\perp \quad \text{and} \quad \mathcal{N}^T = \mathcal{R}^\perp$$

hold. In particular, we obtain the equivalences

$$(A, B, C) \text{ controllable} \iff (A^T, C^T, B^T) \text{ observable}$$

$$(A, B, C) \text{ observable} \iff (A^T, C^T, B^T) \text{ controllable.}$$

Proof: Consider the matrix

$$M = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} \in \mathbb{R}^{(n \cdot k) \times n}.$$

For this matrix Lemma 4.5 implies

$$\mathcal{N} = \ker M.$$

In addition,

$$M^T = (C^T \; A^T C^T \; \ldots \; (A^T)^{n-1} C^T) \in \mathbb{R}^{n \times (n \cdot k)}$$

is the reachability matrix of the dual systems, cf. Definition 2.10, which yields $\mathcal{R}^T = \operatorname{im} M^T$. From linear algebra it is known that

$$\operatorname{im} M^T = (\ker M)^\perp.$$

This yields the first assertion since

$$\mathcal{R}^T = \operatorname{im} M^T = (\ker M)^\perp = \mathcal{N}^\perp.$$

By exchanging the two systems, the same derivation yields $\mathcal{R} = (\mathcal{N}^T)^\perp$, which implies the second assertion, since

$$\mathcal{R}^\perp = \left((\mathcal{N}^T)^\perp \right)^\perp = \mathcal{N}^T.$$

□

Thus, all statements for controllability can be carried over to observability. We do this explicitly for Corollary 2.13 and Lemma 2.14.

Definition 4.8 The matrix $(C^T, A^T C^T \; \ldots \; (A^T)^{n-1} C^T) \in \mathbb{R}^{n \times (k \cdot n)}$ is called *observability matrix* of the systems (1.3). □

Corollary 4.9 System (4.1) is observable if and only if

$$\operatorname{rg}(C^T, A^T C^T \; \ldots \; (A^T)^{n-1} C^T) = n.$$

□

Proof: This follows from Corollary 2.13 applied to the dual system. □

We now formulate the analogue to the decomposition

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

from Lemma 2.14.

Lemma 4.10 Let (A, C) be not observable, i.e., $\dim \mathcal{N} = l > 0$. Then there exists an invertible $T \in \mathbb{R}^{n\times n}$ such that $\widetilde{A} = T^{-1}AT$, $\widetilde{B} = T^{-1}B$ and $\widetilde{C} = CT$ are of the form

$$\widetilde{A} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad \widetilde{C} = (0 \;\; C_2)$$

with $A_1 \in \mathbb{R}^{l\times l}$, $A_2 \in \mathbb{R}^{l\times(n-l)}$, $A_3 \in \mathbb{R}^{(n-l)\times(n-l)}$, $B_1 \in \mathbb{R}^{l\times m}$, $B_2 \in \mathbb{R}^{(n-l)\times m}$ and $C_2 \in \mathbb{R}^{k\times(n-l)}$. Therein, the pair (A_3, C_2) is observable.

Beweis: Lemma 2.14 applied to the dual system (A^T, C^T) yields $\widehat{T}$ with

$$\widehat{T}^{-1}A^T\widehat{T} = \begin{pmatrix} \widehat{A}_1 & \widehat{A}_2 \\ 0 & \widehat{A}_3 \end{pmatrix}, \quad \widehat{T}^{-1}C^T = \begin{pmatrix} \widehat{C}_1 \\ 0 \end{pmatrix}.$$

For $S = (\widehat{T}^T)^{-1}$ this implies

$$S^{-1}AS = \begin{pmatrix} \widehat{A}_1^T & 0 \\ \widehat{A}_2^T & \widehat{A}_3^T \end{pmatrix}, \quad CS = \begin{pmatrix} \widehat{C}_1^T & 0 \end{pmatrix}.$$

By means of the additional coordinate transformation

$$Q = \begin{pmatrix} 0 & \mathrm{Id}_{\mathbb{R}^{n-l}} \\ \mathrm{Id}_{\mathbb{R}^l} & 0 \end{pmatrix}$$

the claimed decomposition follows with $T = SQ$ and

$$A_1 = \widehat{A}_3^T,\; A_2 = \widehat{A}_2^T,\; A_3 = \widehat{A}_1^T,\; C_2 = \widehat{C}_1^T.$$

We additionally give an alternative proof, which is more direct and does not resort to Lemma 2.14:

Let $v_1, \ldots, v_l$ by a basis of $\mathcal{N}$, i.e., $\mathcal{N} = \langle v_1, \ldots, v_l \rangle$. We pick $w_1, \ldots, w_{n-l}$ such that the v_i and w_j together form a basis of $\mathbb{R}^n$ and define $T := (v_1, \ldots, v_l, w_1, \ldots, w_{n-l})$. With e_i we denote the i-th unit vector in $\mathbb{R}^n$. Then $Te_i = v_i$, $i = 1, \ldots, l$, $Te_i = w_{i-l}$, $i = l+1, \ldots, n$, $T^{-1}v_i = e_i$, $i = 1, \ldots, l$ and $T^{-1}w_i = e_{i+l}$, $i = 1, \ldots, n-l$.

We first show that $\widetilde{A}$ has the desired structure. Suppose an entry in the 0-Block of $\widetilde{A}$ is not equal to 0. Then on the one hand

$$\widetilde{A}e_i \notin \langle e_1, \ldots, e_l \rangle = T^{-1}\mathcal{N}$$

for some $i \in \{1, \ldots, l\}$. On the other hand, A-invariance of $\mathcal{N}$

$$\widetilde{A}e_i = T^{-1}ATe_i = T^{-1}Av_i \in T^{-1}\mathcal{N},$$

which leads to a contradiction.

The structure of $\widetilde{C}$ follows from

$$\mathcal{N} = \bigcap_{i=0}^{n-1} \ker(CA^i) \subseteq \ker C.$$

This implies $v_i \in \ker C$ and thus $\widetilde{C}e_i = CTe_i = Cv_i = 0$. Hence, the first l columns of $\widetilde{C}$ must equal 0.

It remains to show observability of (A_3, C_2). To this end, note that for every $\tilde{x} \in \mathbb{R}^{n-l}$, $\tilde{x} \neq 0$ it holds that

$$C_2 A_3^i \tilde{x} = \widetilde{C}\widetilde{A}^i \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} = CA^i T \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix},$$

where in the first equation we used the structure of $\widetilde{A}$ and $\widetilde{C}$. Since

$$w := T \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \notin \mathcal{N},$$

there exists $i \in \{0, \dots, n-1\}$ with $CA^i w \neq 0$ and thus $C_2 A_3^i \tilde{x} \neq 0$. Since $\tilde{x} \neq 0$ was arbitrary,

$$\bigcap_{i=0}^{n-1} \ker(C_2 A_3^i) = \{0\}$$

follows, implying the observability of (A_3, C_2). □

Bemerkung 4.11 All statements in this section remain valid for discrete-time systems. The only result that changes is Lemma 4.5, which — analogous to controllability, cf. Remark 2.17 — only holds for $t \geq n$ in discrete time. □

4.2 Detectability

We have seen that (complete) controllability is sufficient but not necessary for being able to solve the stabilisation problem. The necessary and sufficient condition is stabilisability of (A, B), which according to Theorem 3.29 is the case if and only if all eigenvalues of the uncontrollable part A_3 of the matrix A have negative real parts.

This is similar for observability. In order to be able to sove the stabilisation problem for system (4.1) we do not need observability. It is sufficient to assume a weaker condition, which is given in the following definition.

Definition 4.12 The system (4.1) is called *detectable* (or *asymptotically observable*), if

$$\lim_{t\to\infty} x(t, x_0, 0) = 0 \quad \text{for all } x_0 \in \mathcal{N}.$$

□

This means that solutions for non-observable initial conditions and $u \equiv 0$ converge to 0. Intuitively spoken, the information about these initial conditions is not needed in the stabilisation problem, because the corresponding solutions converge to 0 anyway, and are thus asymptotically (and exponentially) stable.

The following lemma characterises detectability for the decomposition from Lemma 4.10.

Lemma 4.13 System (4.1) is detectable if and only if the matrix A_1 from Lemma 4.10 is Hurwitz, i.e., if it has only eigenvalues with negative real part.

Proof: First observe that detectability is preserved under coordinate changes. We can thus perform all computations in the basis given by 4.10.

In this basis, $\mathcal{N}$ is given by

$$\mathcal{N} = \left\{ x_0 \in \mathbb{R}^n \,\middle|\, x_0 = \begin{pmatrix} x_0^1 \\ 0 \end{pmatrix},\ x_0^1 \in \mathbb{R}^l \right\}.$$

From the form of the matrix $\widetilde{A}$ it then follows that allsolutions corresponding to initial values $x_0 \in \mathcal{N}$ can be written as

$$x(t, x_0, 0) = e^{\widetilde{A}t} x_0 = \begin{pmatrix} e^{A_1 t} x_0^1 \\ 0 \end{pmatrix}.$$

From detectability it now follows that $x(t, x_0, 0) \to 0$ for all $x \in \mathcal{N}$, i.e. $e^{A_1 t} x_0^1 \to 0$ for all $x_0^1 \in \mathbb{R}^l$. This is only possible if A_1 is Hurwitz.

Conversely, A_1 being Hurwitz implies the convergence $e^{A_1 t} x_0^1 \to 0$ for all $x_0^1 \in \mathbb{R}^l$, i.e. $x(t, x_0, 0) \to 0$ for all $x \in \mathcal{N}$ and thus detectability. □

The following theorem shows that detectability is dual to stabilisability.

Theorem 4.14 (A, C) is detectable if and only if (A^T, C^T) is stabilisable.

Proof: We denote the components of the decomposition from Lemma 4.10 applied to (A, C) by A_1, A_2, A_3, C_2. Likewise, we denote the components of the decomposition from Lemma 2.14 applied to (A^T, C^T) by $\widehat{A}_1, \widehat{A}_2, \widehat{A}_3, \widehat{C}_1$. The proof of Lemma 4.10 then implies $A_1 = \widehat{A}_3^T$.

By Lemma 4.13 detectability of (A, C) is equivalent to A_1 being Hurwitz. Likewise, by Theorem 3.29 stabilisability of (A^T, C^T) is equivalent to $\widehat{A}_3$ being Hurwitz. Since the eigenvalues of $\widehat{A}_3$ and $\widehat{A}_3^T = A_1$ coincide, this yields the claimed equivalence. □

Remark 4.15 In order to adapt these statements to discrete time, it suffices to change the eigenvalue conditions from "negative" to "modulus lesss than 1". □

4.3 Dynamic observers

A natural approach to solving the stabilisation problem for (4.1) is the choice $u(t) = Fy(t)$. This may work (cf. Example 3.19, where we considered $C = (0\ \ 1)$ and $C = (1\ \ 0)$). However, this approach may also fail, as the controllable and observable system (4.1) with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad C = (1\ \ 0)$$

shows, cf. the exercises. In fact, this system is not even stabilisable if we allow $F(y(t))$ to be an arbitrary continuous function $F : \mathbb{R} \to \mathbb{R}$.

For this reason we will not develop a method for stabilisation that always works if (4.1) is stabilisable abd detectable. The method works as follows for a system (4.1) with matrices (A, B, C):

(1) Design a stabilising linear feedback law for (A, B)

(2) Design an algorithm that computes an estimation $z(t) \approx x(t)$ from the measured output values $y(s)$, $s \in [0, t]$

(3) Control the system (4.1) via $u(t) = Fz(t)$.

Step (1) can be solved using the methods from Chapter 3. In this section we will consider Step (2) and in the following section we will then prove that for the proposed algorithm the method in Steps (1)–(3) indeed works.

The reason why the example above cannot be stabilized lies in the fact that observability does not require $Cx_0 \neq 0$ for $x_0 \neq 0$. Rather, it is only required that $Cx(t; t_0, x_0, 0) \neq 0$ for $t > 0$. Thus, in order to recognize that the estimate should attain a value $z(t) \neq 0$ (to which the feedback law can react), the algorithm in Step (2) must use the output over a certain time span, not merely its current value. We will achieve this by defining the estimate $z(t)$ as the solution of a suitably foormulated control system, in which — in addition to the control function — the output $y(t)$ of (4.1) acts as a second input. The following definition formalises this idea.

Definition 4.16 A *dynamic observer* (also called *Luenberger-observer*) for (4.1) is a linear control system of the form

$$\dot{z}(t) = Jz(t) + Ly(t) + Ku(t) \tag{4.3}$$

with $J \in \mathbb{R}^{n\times n}$, $L \in \mathbb{R}^{n\times k}$, $K \in \mathbb{R}^{n\times m}$, such that for all initial values x_0, $z_0 \in \mathbb{R}^n$ and all control functions $u \in \mathcal{U}$ the solutions $x(t, x_0, u)$ and $z(t, z_0, u, y)$ of (4.1), (4.3) with $y(t) = Cx(t, x_0, u)$ satisfy the estimate

$$\|x(t, x_0, u) - z(t, z_0, u, y)\| \leq ce^{-\sigma t}\|x_0 - z_0\|$$

for suitable constants $c, \sigma > 0$. □

In practice, the system (4.3) can, e.g., be solved numerically in order to determine the vaulues $z(t)$.

The following theorem clarifies when a dynamic observer exists. Its proof provides an explicit construction of the observer.

Theorem 4.17 A dynamic observer for system (4.1) exists if and only if the system is detectable.

Proof: "$\Leftarrow$" Since (4.1) is detectable, (A^T, C^T) is stabilisable. Wie can thus find a linear feedback law $\widehat{F} \in \mathbb{R}^{k\times n}$ such that $A^T + C^T\widehat{F}$ is Hurwitz. For $G = \widehat{F}^T$ the matrix $A + GC = (A^T + C^T\widehat{F})^T$ is then Hurwitz, too.

We now specify the matrices in (4.3) as $J = A + GC$, $L = -G$ and $K = B$, i.e.,

$$\dot{z}(t) = (A + GC)z(t) - Gy(t) + Bu(t).$$

Abbreviating $x(t) = x(t, x_0, u)$, $z(t) = z(t, z_0, u, y)$ and $e(t) = z(t) - x(t)$, for $e(t)$ we obtain the differential equation

$$\begin{aligned}\dot{e}(t) &= \dot{z}(t) - \dot{x}(t)\\ &= (A + GC)z(t) - Gy(t) + Bu(t) - Ax(t) - Bu(t)\\ &= (A + GC)z(t) - GCx(t) - Ax(t)\\ &= (A + GC)(z(t) - x(t)) \;\; = \;\; (A + GC)e(t)\end{aligned}$$

Since $A + GC$ is Hurwitz, we thus obtain

$$\|e(t)\| \leq ce^{-\sigma t}\|e(0)\|$$

for suitable $c, \sigma > 0$ and since $e(t) = z(t) - x(t)$ and $e(0) = z_0 - x_0$ this implies the desired estimate.

"⇒" Let $x_0 \in \mathcal{N}$ and $y(t) = Cx(t, x_0, 0) = 0$ for all $t \geq 0$. Setting $z_0 = 0$ then yields $z(t, z_0, 0, y) = z(t, 0, 0, 0) = 0$. Then the estimate from the definition of the dynamic observer yields

$$\|x(t, x_0, 0)\| = \|x(t, x_0, 0) - z(t, z_0, 0, y)\| \leq ce^{-\sigma t}\|x_0 - z_0\| = ce^{-\sigma t}\|x_0\| \to 0$$

for $t \to \infty$. It follows that $x(t, x_0, 0) \to 0$, implying detectability. □

4.4 Solution of the stabilisation problem with output

We will now analyse the method for stabilisation from the last section and prove that it works successfully if we use the dynamic observer (4.3) in Step (2).

From the Steps (1)–(3) with (4.3) in Step (2) we obtain the feedback equation

$$u(t) = Fz(t), \quad \dot{z}(t) = Jz(t) + Ly(t) + KFz(t). \tag{4.4}$$

This form of a feedback controller is called a *dynamic output feedback law*[2]. This is because $u(t)$ is computed from the *output* $y(t) = Cx(t)$ and the feedback controller has an "internal" *dynamic*, given by the differential equations for z.

[2] In contrast to this the feedback law $u(t) = Fx(t)$ constructed in Chapter 3 is called *static state feedback law.*

Definition 4.18 A dynamic output feedback law (4.4) solves the *stabilisation problem with output*, if the overall system of differential equations that is obtained by inserting (4.4) into (4.1), i.e.

$$\begin{aligned}\dot{x}(t) &= Ax(t) + BFz(t)\\ \dot{z}(t) &= Jz(t) + LCx(t) + KFz(t)\end{aligned}$$

with solutions $\binom{x(t)}{z(t)} \in \mathbb{R}^{2n}$ is exponentially stable. □

Theorem 4.19 Consider a control system (4.1) with matrices (A, B, C). Then the stabilisation problem with output is solvable in the sense of Definition 4.18 if and only if (A, B) is stabilisable and (A, C) is detectable.

In this case (4.4) together with the dynamic observer constructed in the proof of Theorem 4.17 and a stabilising feedback law $F \in \mathbb{R}^{m\times n}$ for (A, B) yields a stabilising dynamic feedback law.

Proof: "⇐": Let (A, B) be stabilisable and (A, C) be detectable. Further, let $F \in \mathbb{R}^{m\times n}$ be a stabilising feedback law for (A, B) and (4.3) be the dynamic observer constructed in the proof of Theorem 4.17. Then the system controlled by (4.4) becomes

$$\begin{aligned}\begin{pmatrix}\dot{x}(t)\\ \dot{z}(t)\end{pmatrix} &= \begin{pmatrix}A & BF\\ LC & J+KF\end{pmatrix}\begin{pmatrix}x(t)\\ z(t)\end{pmatrix}\\ &= \begin{pmatrix}A & BF\\ -GC & A+GC+BF\end{pmatrix}\begin{pmatrix}x(t)\\ z(t)\end{pmatrix}\\ &= T^{-1}\begin{pmatrix}A+BF & BF\\ 0 & A+GC\end{pmatrix}T\begin{pmatrix}x(t)\\ z(t)\end{pmatrix}.\end{aligned}$$

with

$$T = \begin{pmatrix}\mathrm{Id}_{\mathbb{R}^n} & 0\\ -\mathrm{Id}_{\mathbb{R}^n} & \mathrm{Id}_{\mathbb{R}^n}\end{pmatrix}, \qquad T^{-1} = \begin{pmatrix}\mathrm{Id}_{\mathbb{R}^n} & 0\\ \mathrm{Id}_{\mathbb{R}^n} & \mathrm{Id}_{\mathbb{R}^n}\end{pmatrix}.$$

Since exponential stability persists under coordinate transformations, it suffices to check that the matrix in the last line of this computation is Hurwitz. This is a block-triangular matrix, whose eigenvalues ar thus given by the eigenvalues of the blocks on the diagonal, i.e., of $A + BF$ and $A + GC$. Since $A + BF$ is Hurwizt by choice of F and $A + GC$ is Hurwitz by choice of G (in the proof of Theorem 4.17), we obtain only eigenvalues with negative real part. This yields the assertion.

"⇒": Using the coordinate transformation T from Lemma 2.14 the system becomes

$$\begin{aligned}\dot{x}^1(t) &= A_1x^1(t) + A_2x^2(t) + B_1Fz(t)\\ \dot{x}^2(t) &= A_3x^2(t)\\ \dot{z}(t) &= Jz(t) + LCx(t) + KFz(t)\end{aligned}$$

with $x(t) = T\binom{x^1(t)}{x^2(t)}$. Assume now that (A, B) is not stabilisable. Then A_3 has eigenvalues with positive real parts, the origin is thus not asymptotically stable for the equation $\dot{x}^2(t) = A_3 x^2(t)$ and consequently there is an initial value x_0^2 with $x^2(t, x_0^2) \not\to 0$. Thus, if we choose

$$x_0 = T \begin{pmatrix} x_0^1 \\ x_0^2 \\ z_0 \end{pmatrix} \in \mathbb{R}^{2n}$$

with arbitrary x_0^1, z_0, then $x(t, x_0, Fz) \not\to 0$ for any choice of the dynamic feedback law. This contradics the solvability of the stabilisation problem. Consequently, (A, B) must be stabilisable.

Detectability of (A, C) follows as in the proof of “⇒” in Theorem 4.17. □

Remark 4.20 The constructions and statements in this and in the preceding section hold analogously with the obvious modifications for discrete-time systems. □

Chapter 5

Analysis in frequency domain

A considerable part of modern control and systems theory developed out of electrical engineering, where the behaviour of electrical circuits with input and output signals is studied. An example for this may be an amplifier, which receives an input signal (from a microphone, a mobile phone etc.) and converts it into an output signal that is sent to a loudspeaker. Another example is an (analog) radio, in which the input signal (electromagnetic waves) are converted into an acoustic output signal. If we represent these devices by control systems, we can denote the input signal by u and the output signal by y. The changes the interpretation of these functions compared to the previous chapters: $u(t)$ is now an external signal (instead of a control function that we can determine) and $y(t)$ is and output signal that is supposed to satisfy certain properties (instead of the result of a measurement). Yet, this new intepretation does not change the mathematical description of the relation between u and y via the system (4.1). In these kind of applications, the initial condition is usually chosen as $x_0 = 0$. The interpretation of this choice is that until time $t = 0$ the system is at rest, and only afterwards it is influenced by the input signal $u(t)$, $t \geq 0$.

The two application example already indicate that frequencies play an important role in this interpretation. For this reason, in these applications u and y are usually not considered as functions of time but as functions depending on the frequency. To this end, we start by introducing the Laplace-Transformation.

5.1 Laplace transformation

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and $\mathbb{R}_0^+ = [0, \infty)$. By $L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m)$ we denote the space of all functions $u : \mathbb{R}_0^+ \to \mathbb{K}^m$ that are Lebesgue integrable on any compact interval in $\mathbb{R}_0^+$ and with $L^1(\mathbb{R}_0^+, \mathbb{K}^m)$ we denote the space of functions $u : \mathbb{R}_0^+ \to \mathbb{K}^m$, that are Lebesgue integrable on the whole half line $\mathbb{R}_0^+$. For $u \in L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m)$ and $\alpha \in \mathbb{R}$ define $u_\alpha : \mathbb{R}_0^+ \to \mathbb{K}^m$ via $u_\alpha(t) := u(t)e^{-\alpha t}$. Then we define the space of *α-exponentially integrable functions* as

$$\mathcal{E}_\alpha(\mathbb{K}^m) := \{u \in L^1_{loc}(\mathbb{R}_0^+, \mathbb{K}^m) \,|\, u_\alpha \in L^1(\mathbb{R}_0^+, \mathbb{K}^m)\}.$$

Example 5.1 The function $u(t) = e^t$ is continuous and is thus contained in $L^1_{loc}(\mathbb{R}_0^+, \mathbb{R})$. However, since

$$\int_0^t e^\tau d\tau = e^t - 1 \to \infty$$

for $t \to \infty$, it is not contained in $L^1(\mathbb{R}_0^+, \mathbb{R})$. For $\alpha > 1$ we obtain

$$\int_0^t u_\alpha(\tau) d\tau = \int_0^t e^\tau e^{-\alpha\tau} d\tau = \frac{1}{1-\alpha}(e^{(1-\alpha)t} - 1) \to \frac{1}{\alpha - 1}$$

for $t \to \infty$. Thus, the infinite Riemann integral exists and since $u_\alpha(t) \geq 0$ this implies that the infinite Lebesgue integral also exists. Consequently, $u(t) = e^t$ lies in $\mathcal{E}_\alpha(\mathbb{R})$ for all $\alpha > 1$. □

Definition 5.2 The functions in $\mathcal{E}_\alpha(\mathbb{K}^m)$ are called *Laplace-transformable*. For all $s \in \mathbb{C}_\alpha := \{s \in \mathbb{C} \,|\, \mathrm{Re}(s) > \alpha\}$ the (one-sided) *Laplace transform* of $u \in \mathcal{E}_\alpha(\mathbb{K}^m)$ is defined as

$$\hat{u}(s) := (\mathcal{L}u)(s) := \int_0^\infty u(t) e^{-st} dt.$$

The Laplace transform $\hat{u} = \mathcal{L}u$ is thus a function from $\mathbb{C}_\alpha$ to $\mathbb{C}^m$. □

Example 5.3 Laplace transforms of some functions from $\mathbb{R}_0^+$ to $\mathbb{R}$ with $a \in \mathbb{C}$, $\omega \in \mathbb{R}$, $m \in \mathbb{N}_0$:

$$\begin{array}{llcll}
(a) & u(t) = 1 & \Rightarrow & \hat{u}(s) = \dfrac{1}{s} & \text{for } \mathrm{Re}(s) > 0 \\
(b) & u(t) = \sin(\omega t) & \Rightarrow & \hat{u}(s) = \dfrac{\omega}{\omega^2 + s^2} & \text{for } \mathrm{Re}(s) > 0 \\
(c) & u(t) = \cos(\omega t) & \Rightarrow & \hat{u}(s) = \dfrac{s}{\omega^2 + s^2} & \text{for } \mathrm{Re}(s) > 0 \\
(d) & u(t) = e^{at} & \Rightarrow & \hat{u}(s) = \dfrac{1}{s - a} & \text{for } \mathrm{Re}(s) > \mathrm{Re}(a) \\
(e) & u(t) = e^{at}\sin(\omega t) & \Rightarrow & \hat{u}(s) = \dfrac{\omega}{\omega^2 + (s+a)^2} & \text{for } \mathrm{Re}(s) > \mathrm{Re}(a) \\
(f) & u(t) = e^{at}\cos(\omega t) & \Rightarrow & \hat{u}(s) = \dfrac{s - a}{\omega^2 + (s+a)^2} & \text{for } \mathrm{Re}(s) > \mathrm{Re}(a) \\
(g) & u(t) = \dfrac{t^m}{m!} e^{at} & \Rightarrow & \hat{u}(s) = \dfrac{1}{(s-a)^{m+1}} & \text{for } \mathrm{Re}(s) > \mathrm{Re}(a)
\end{array}$$

□

Remark 5.4 Although the integral in the definition of the Laplace transform is only defined for the values of $\mathrm{Re}(s)$ indicated in this table, the resulting expession for the transform may be defined for a larger set of values. For instance, in (d) the expression for $\hat{u}(s)$ is defined for all $s \neq a$. In the following we will admit all arguments $s \in \mathbb{C}$ of $\hat{u}$ for which the computed expression is well defined. □

The inverse of the Laplace transf is given by

$$(\mathcal{L}^{-1}\hat{u})(t) := \frac{1}{2\pi i}\int_{\beta-i\infty}^{\beta+i\infty} e^{st}\hat{u}(s)ds = \frac{e^{\beta t}}{2\pi i}\int_{-\infty}^{\infty} e^{i\omega t}\hat{u}(\beta+i\omega)d\omega.$$

More precisely, for all $u \in \mathcal{E}_\alpha(\mathbb{K}^m)$ and every $\beta > \alpha$ the identity $\mathcal{L}^{-1}\mathcal{L}u(t) = u(t)$ holds for almost all $t \in \mathbb{R}_0^+$. If u is continuous then it even holds for all $t \in \mathbb{R}_0^+$, cf. [11, Theorem A.3.19].

Below we list some important rules for computing the Laplace transform, for $a, a_1, a_2 \in \mathbb{R}$ und $u, u_1, u_2 \in \mathcal{E}_\alpha(\mathbb{K}^m)$. Further assumptions are indicated below the table.

$$\begin{array}{rrcl}
(i) & \mathcal{L}(a_1u_1 + a_2u_2)(s) & = & a_1\hat{u}_1(s) + a_2\hat{u}_2(s) \\
(ii) & \mathcal{L}(u(a\,\cdot))(s) & = & \frac{1}{a}\hat{u}\left(\frac{s}{a}\right), \quad \text{for } a > 0 \\
(iii) & \mathcal{L}(u(\cdot - a))(s) & = & e^{-sa}\hat{u}(s), \quad \text{for } a > 0 \\
(iv) & \mathcal{L}(e^{a\,\cdot}u)(s) & = & \hat{u}(s-a) \\
(v) & \mathcal{L}(\dot{u})(s) & = & s\hat{u}(s) - u(0) \\
(vi) & \mathcal{L}\left(\int_0^{\cdot} u(\tau)d\tau\right)(s) & = & \frac{1}{s}\hat{u}(s) \\
(vii) & \mathcal{L}(\cdot^k u)(s) & = & (-1)^k\frac{d^k\hat{u}}{ds^k}(s) \\
(viii) & \mathcal{L}(u_1 \star u_2)(s) & = & \hat{u}_1(s)\hat{u}_2(s) \\
(ix) & \lim\limits_{t\to 0, t>0} u(t) & = & \lim\limits_{s\to\infty} s\hat{u}(s)
\end{array}$$

In (iii) we assume that u is defined on $[-a,\infty)$ with $u(t) = 0$ for all $t \in [-a, 0]$. In (v) we assume that u is defined on $(-\varepsilon,\infty)$ for some $\varepsilon > 0$ and that u is differentiable in s. If u in (v) is discontinuous in 0, then $u(0)$ must be replaced by $\lim_{t\to 0,t<0} u(t)$. In $(viii)$, $u_1 \star u_2(t) = \int_0^t u_1(t-\tau)u_2(\tau)d\tau$ denotes the *convolution*.

5.2 The transfer function

The transfer function allows to express the input-output behaviour of a control system by means of the Laplace transform. Here the input-output behaviour denoted the map $u \mapsto y$ with $y(t) = Cx(t, 0, u)$, i.e., the function that assigns to the input function u the output of the solution of the control system with initial value $x_0 = 0$.

We now investigate how this map looks for the Laplace transformed signals. To this end, we again consider the system (4.1), i.e.,

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

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

Theorem 5.5 Consider the control system (4.1) and let $u \in \mathcal{U}$, $u \in \mathcal{E}_\alpha(\mathbb{R}^m)$ and $y(t) = Cx(t, 0, u)$. Then y is Laplace-transformable with

$$\hat{y}(s) = G(s)\hat{u}(s),$$

where $G(s) = C(s\mathrm{Id} - A)^{-1}B$.

Proof: According to (1.14) it holds that

$$y(t) = C\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau.$$

Since $u \in \mathcal{E}_\alpha(\mathbb{R}^m)$, u is exponentially bounded. Moreover, $\|e^{At}\|$ is exponentially bounded by $e^{\|A\|t}$. Hence, the integrand is exponentially bounded, thus also the integral and since x and y are continuous as results of an integration, we obtain $x \in \mathcal{E}_\alpha(\mathbb{R}^n)$, $y \in \mathcal{E}_\alpha(\mathbb{R}^k)$ for suitable (sufficiently large) $\alpha > 0$.

Applying the Laplace transform to (4.1), using the formulas (i) and (v) as well as $x_0 = 0$ we obtain

$$s\hat{x}(s) = A\hat{x}(s) + B\hat{u}(s), \qquad \hat{y}(s) = C\hat{x}(s)$$

for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) > \alpha$. The first equation is equivalent to

$$s\hat{x}(s) - A\hat{x}(s) = B\hat{u}(s) \quad \Leftrightarrow \quad (s\mathrm{Id} - A)\hat{x}(s) = B\hat{u}(s).$$

For all $s \in \mathbb{C}$ that are not eigenvalues of A (i.e. in particular for all s with sufficiently large real part), the matrix on the left hand side of this equation is invertible and it follows that

$$\hat{x}(s) = (s\mathrm{Id} - A)^{-1}B\hat{u}(s) \quad \Rightarrow \quad \hat{y}(s) = C\hat{x}(s) = C(s\mathrm{Id} - A)^{-1}B\hat{u}(s) = G(s)\hat{u}(x).$$

□

Definition 5.6 The function $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ from Theorem 5.5 is called *transfer function.* □

Remark 5.7 (i) From the representation

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

with the adjugate matrix $\mathrm{adj}(s\mathrm{Id} - A)$ it follows that $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ is a matrix valued function with rational entries, i.e. with entries of the form

$$g_{ij}(s) = \frac{p_{ij}(s)}{q_{ij}(s)} \tag{5.1}$$

with polynomials p_{ij}, q_{ij} of degree[1] $\deg p_{ij} < \deg q_{ij} \leq n$.

[1] For outputs of the form $y(t) = Cx(t) + Du(t)$ it holds that $G(s) = D + C(s\mathrm{Id} - A)^{-1}B$ and $\deg p_{ij} \leq \deg q_{ij} \leq n$.

(ii) The so-called *realisation theorie* deals with the question whether for a given function $G : \mathbb{C} \to \mathbb{C}^{k\times m}$ there exists a control system (4.1) such that G is its transfer function. One can show that for each proper[2] rational matrix function this is indeed the case. However, in general A, B, C are not unique.

(iii) Defining $g(t) := Ce^{At}B$, the solution formula yields

$$y(t) = \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau = \int_0^t g(t-\tau)u(\tau)d\tau = g \star u(t).$$

With the computation rule $(viii)$ of the Laplace transform we thus obtain

$$\hat{y}(s) = \mathcal{L}(g \star u)(s) = \hat{g}(s)\hat{u}(s).$$

Hence, the transfer function satisfies $G = \hat{g}$ (if we generalise the definition of the Laplace transform in the obvious way to matrix valued functions). □

Example 5.8 Wir consider the down-hanging and the inverted linearised pendulum, both without the variables of the cart, i.e.

$$A = \begin{pmatrix} 0 & 1 \\ -g & -k \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and, respectively,

$$A = \begin{pmatrix} 0 & 1 \\ g & -k \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In both cases we set $C = (1\ 0)$. This means that the output measures the position of the Pendulum.

For the down-hanging pendulum we thus obtain

$$(s\mathrm{Id} - A)^{-1} = \begin{pmatrix} s & -1 \\ g & s+k \end{pmatrix}^{-1} = \begin{pmatrix} \frac{s+k}{ks+s^2+g} & \frac{1}{ks+s^2+g} \\ \frac{-g}{ks+s^2+g} & \frac{s}{ks+s^2+g} \end{pmatrix}$$

and hence

$$G(s) = C(s\mathrm{Id} - A)^{-1}B = \frac{1}{ks+s^2+g}.$$

Analogously, for the inverted pendulum we get

$$G(s) = C(s\mathrm{Id} - A)^{-1}B = \frac{1}{ks+s^2-g}.$$

□

[2] Proper means that $\deg p_{ij} \leq \deg q_{ij}$ for all i, j.

5.3 Input-output stability

We now introduce a stability notion that fits the input-output perspective of the transfer function G.

Definition 5.9 A control system is called *input-output-stable* (briefly i/o-stable) if there exists a constant $K > 0$ such that for any function $u \in \mathcal{U}$ that is bounded on $\mathbb{R}_0^+$ the corresponding output

$$y(t) = C \int_0^t e^{A(t-\tau)} Bu(\tau) d\tau$$

with initial value $x_0 = 0$ satisfies the inequality $\quad \|y\|_\infty \leq K \|u\|_\infty$. □

Remark 5.10 (i) One can show that i/o stability is equivalent to the implication "$\|u\|_\infty < \infty \Rightarrow \|y\|_\infty < \infty$", the so-called *bounded input-bounded output (BIBO) stability.* In this form i/o stability is defined in many textbooks. The proof of this equivalence, however, needs several technical estimates that we avoid here for brevity of exposition. For our purposes the formulation from Definition 5.9 is better suited.

(ii) In order to distinguish the stability notion used in the previous chapters (A or the closed-loop system is exponentially stable, i.e. all eigenvalues of A or $A + BF$, respectively, have negative real part) from the notion of i/o stability, we also denote the "old" stability notion as *state stability.* □

A first necessary and sufficient condition for io stability is given by the following lemma.

Lemma 5.11 A system (4.1) is i/o-stable if and only if $g(t) = Ce^{At}B$ satisfies the inequality

$$g_{\max} := \int_0^\infty \|g(t)\| dt < \infty. \tag{5.2}$$

Proof: "$\Rightarrow$": Let the system by i/o-stable. We prove that

$$\int_0^\infty |\gamma_{ij}(t)| dt \leq K \tag{5.3}$$

for all component functions γ_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, m$ von $g = (\gamma_{ij})_{i=1,\ldots,k, j=1,\ldots,m}$, which implies (5.2).

In order to prove (5.3), for given $t > 0$ let u be given by $u(\tau) := \operatorname{sgn}(\gamma_{ij}(t - \tau)) e_j$ for $\tau \in [0, t]$. Then we obtain $[g(t - \tau)u(\tau)]_i = |\gamma_{ij}(t - \tau)|$. Defining $u(\tau) = 0$ for $\tau > t$, we obtain $\|u\|_\infty = 1$ and thus for the corresponding output $\|y\|_\infty \leq K$, and consequently also $|y_i(t)| \leq K$ for all $t \geq 0$. This implies

$$K \geq |y_i(t)| = \left| \int_0^t [g(t - \tau)u(\tau)]_i d\tau \right| = \left| \int_0^t |\gamma_{ij}(t - \tau)| d\tau \right| = \int_0^t |\gamma_{ij}(t-\tau)| d\tau = \int_0^t |\gamma_{ij}(\tau)| d\tau.$$

Since this holds for all $t \geq 0$, (5.3) follows.

"⇐": Let $g_{\max} < \infty$ and let u be an input signal with $\|u\|_\infty < \infty$. Then for all $t \geq 0$ the inequality

$$\|y(t)\| = \left\|\int_0^t g(t-\tau)u(\tau)d\tau\right\| \leq \int_0^t \|g(t-\tau)\|\|u(\tau)\|d\tau \leq \int_0^t \|g(t-\tau)\|d\tau\|u\|_\infty = g_{\max}\|u\|_\infty$$

holds. Thus, the system is i/o-stable with $K = g_{\max}$. □

Corollary 5.12 If (4.1) is state stable, i.e., if A is Hurwitz, then (4.1) is also i/o-stable. □

Proof: If (4.1) is state stable, then A is Hurwitz. Hence, by Theorem 3.5 the inequality $\|e^{At}\| \leq ce^{-\sigma t}$ holds for constants $c, \sigma > 0$ and all $t \geq 0$. This yields $\|g(t)\| \leq \|C\|ce^{-\sigma t}\|B\|$ and thus

$$\int_0^\infty \|g(t)\|dt \leq \int_0^\infty \|C\|ce^{-\sigma t}\|B\|dt = \frac{c\|C\|\|B\|}{\sigma} < \infty.$$

□

The converse of this result is obviously false. A simple counterexample is obtained by setting $C = 0$ setzen. Then, because of $y(t) \equiv 0$ for all $u \in \mathcal{U}$, the system is trivially i/o-stable with $K = 0$, regardless of whether A is Hurwitz or not.

Verifying the criterion (5.2) is in general difficult, since an infinite integral must be estimated. If, however, the transfer function G is known, then the criterion can be easily checked. To this end we say that $s^* \in \mathbb{C}$ is a pole of a rational (matrix) function G if s^* is a pole for at least one of its component functions. This, in turn, means that there are $j, k \in \mathbb{N}_0$ with $j < k$ such that s^* is a k-fold zero of the denominator polynomial and a j-fold zero of the enumerator polynomial (here $j = 0$ means that s^* is not a zero). Note that s^* is a pole of G if and only if $\|G(s)\|$ is unbounded in each neighbourhood of s^*.

Theorem 5.13 Consider a control system (4.1) with transfer function G. Then the system is i/o-stable if and only if all poles s^* of G lie in the open left complex half plane $\mathbb{C}^- = \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\}$, i.e. if they satisfy $\mathrm{Re}(s^*) < 0$.

Proof: "⇒": If the system is i/o-stable, then from Lemma 5.11 we know that $g_{\max} = \int_0^\infty \|g(t)\|dt < \infty$. Then for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) \geq 0$ we obtain

$$\|G(s)\| = \left\|\int_0^\infty g(t)e^{-st}dt\right\| \leq \int_0^\infty \|g(t)\|\underbrace{|e^{-st}|}_{\leq 1}dt \leq \int_0^\infty \|g(t)\|dt = g_{\max},$$

which means that G cannot have poles outside $\mathbb{C}^-$.

"⇐": Let $\gamma_{ij}(t)$ denote the components of the function $g(t) = Ce^{At}B$. From Remark 5.7 it follows that the entries of G are given by $g_{ij} = \hat{\gamma}_{ij}$. Now the series representation of the matrix exponential implies that the $\gamma_{ij}(t)$ are of the form

$$\gamma_{ij}(t) = \sum_{p=1}^{q} \mu_p e^{\lambda_p t}\frac{t^{k_p}}{k_p!},$$

where the λ_j are eigenvalues of A. From Example 5.3(g) we can thus conclude

$$g_{ij}(s) = \hat{\gamma}_{ij}(s) = \sum_{p=1}^{q} \mu_p \frac{1}{(s-\lambda_p)^{k_p+1}}.$$

This yields that the poles of G are given by λ_p. The assumption on the poles then implies that all λ_p lie in $\mathbb{C}^-$. This, in turn, implies that the integral $\int_0^\infty \gamma_{ij}(t)dt$ is finite for all i, j, hence also $\int_0^\infty \|g(t)\|dt < \infty$. Hence by Lemma 5.11 the system is i/o-stable. □

Example 5.14 For the pendulum this criterion allows to check easily that the down-hanging pendulum is i/o-stable, because the poles (i.e. the zeros of the denominator) are given by $-k/2 \pm \sqrt{k^2-4g}/2$. These numbers all have real part. Analogously for the inverted pendulum one sees that the poles are $-k/2 \pm \sqrt{k^2+4g}/2$. As one of these numbers has a positive real part, the inverted pendulum is not i/o-stable. □

Remark 5.15 (i) The proof shows that all poles of G are eigenvalues of A. This explains the name pole shifting theorem for Theorem 3.29.

(ii) In general not all eigenvalues of A are poles of G. On the one hand, each eigenvalue for which the corresponding eigenspace lies in $\mathcal{N}$ is missing, because the correponding solutions cannot be observed. On the other hand, the eigenvalues corresponding to eigenspaces that cannot be reached from $x_0 = 0$ are missing. These are the eigenspaces that are not contained in the reachable set $\mathcal{R}$.

If the system is controllable and observable, then all eigenvalues of A are poles of G. This can be confirmed comparing Example 5.14 with Example 3.6. If the system is stabilisable and detectable, then all unstable eigenvalues (i.e. those with positive real part) are poles of G. In this case, state stability is equivalent to i/o-stability. □

5.4 Feedback laws in frequency domain

In order to formulate a feedback law in frequency domain, we first need to extend this concept slightly. The this end we observe that both the static feedback law $u(t) = Fx(t)$ and the dynamic feedback law with $u(t) = Fz(t)$ and the differential equation $\dot{z}(t) = (J + KF)z(t) + Ly(t)$ are easily Laplace-transformered. We obtain the transfer functions

$$K(s) = F \quad \text{bzw.} \quad K(s) = F(s\mathrm{Id} - M)^{-1}L,$$

where in the first case we assume $C = \mathrm{Id}$ and in the second case we write $M = J + KF$. A closed loop can thus always be written as the coupling of two transfer functions G and K. For being consistent with the i/o concept, it would be desirable that this coupling is again a transfer function. This, however, requires an additional input for the closed-loop system, that we did not have so far, because the original input is "occupied" with $u = Fx$ or $u = Fz$, respetively. As a remedy we introduce a new, additional input $w(t)$, by replacing $Fx(t)$ or $Ly(t)$ by $F(x(t) + w(t))$ or $L(y(t) + w(t))$, respectively.

Theorem 5.16 Given two transfer functions G and K with appropriate dimensions, which are coupled via $\hat{y}(s) = G(s)\hat{u}(s)$ and $\hat{u}(s) = K(s)(\hat{y}(s) + \hat{w}(s))$. Then

$$\hat{y}(s) = (\mathrm{Id} - G(s)K(s))^{-1}G(s)K(s)\hat{w}(s)$$

holds for all $s \in \mathbb{C}$ for which $\mathrm{Id} - G(s)K(s)$ is invertible.

Proof: The two identities from the assumption imply

$$\hat{y}(s) = G(s)\hat{u}(s) = G(s)K(s)(\hat{y}(s) + \hat{w}(s)).$$

Rearranging the terms in this equation yields, that it is equivalent to

$$(\mathrm{Id} - G(s)K(s))\hat{y}(s) = G(s)K(s)\hat{w}(s)$$

This immediately yields the assertion. □

The feedback stabilisation problem in frequency domain can now be defined as the problem to find a transfer function K, such that $(\mathrm{Id} - G(s)K(s))^{-1}G(s)K(s)$ is i/o-stable, i.e., that it only exhibits poles in $\mathbb{C}^-$. In the special case that u and y are one-dimensional, a number of efficient computational techniques exists for this task. We will, however, not discuss them in detail here.

We will rather briefly discuss the role of the new additional input signal in the system. For this purpose we consider the simplest case of a stabilising static state feedback, i.e., $u = Fx$ and $C = \mathrm{Id}$. Then the solutions of closed-loop system with the additional input are given by

$$x(t) = e^{(A+BF)t}x_0 + \underbrace{\int_0^t e^{(A+BF)(t-\tau)}BFw(\tau)d\tau}_{=:v(t)}.$$

Exponential stability is now equivalent to the fact that $e^{(A+BF)t}$ converges to 0 as $t \to \infty$. This implies

$$\|x(t) - v(t)\| \leq ce^{-\sigma t}\|x_0\|,$$

i.e. the solution converges to $v(t)$. Stability thus ensures that the solution converges to a well defined limit function $v(t)$ that is independent of the initial value x_0 and only depends on the new input $w(t)$. This is a new interpretation of stability, which is equivalent to i/o stability and is thus also implied by the stability notions in the sense of Chapters 3 and 4. In the case $w \equiv 0$ the limit function satisfies $v \equiv 0$ and we are thus back in the situation of these chapters.

5.5 Graphical analysis

In this section we present two graphical representations of control systems that are very common in control engineering. These apply to systems with one-dimensionalem input and output, i.e., for $m = k = 1$. Obeerve that in this case the tranfer function G is a scalar function. Systems of this kind are called SISO systems (Single Input Single Output).

The Bode diagram

The Bode diagram[3] provides a graphical illustration of the relation between u and y. In particular, this representation explains why the consideration of the Laplace transform is called “analysis in frequency domain”. As a preparation we need the following theorem.

Theorem 5.17 Consider the transfer function $G : \mathbb{C} \to \mathbb{C}$ for an i/o-stable SISO system of the form (4.1). Then the output signal $y(t)$ corresponding to the input signal $u(t) = \sin(\omega t)$ converges for $t \to \infty$ to the function

$$y_\infty(t) = |G(i\omega)| \sin(\omega t + \varphi(\omega)),$$

where φ is an argument function[4] of $\omega \mapsto G(\omega i)$.

Proof: See [11, Proposition 2.3.22].

The values of the transfer function G on the imaginary axis $i\mathbb{R}$ — the so-called *frequency response* of G — thus has a very concrete meaning for the bevavious of the output $y(t)$ for sinusoidal inputs $u(t)$: The output signal is obtained by amplifying the input signal by $|G(i\omega)|$ and shifting its phase by $\varphi(\omega)$.

Figure 5.5 illustrates this for the model of the (down-hanging) pendulum with $k = 0.1$ and $g = 9.81$. Here we plot the numerically simulated output (red solid) for input $u(t) = \sin(\omega t)$ (black dashed) for $\omega = 4$. One sees that the output singal has an amplitude of about 0.16 and its phase is shifted by about π compared to the input signal; the pendulum thus oscillates oppositely compared to the cart and with lower amplitude. The corresponding transfer function satisfies $|G(i4)| = 0.1612$ and $\arg(G(i4)) = -3.077$, which confirmd this observation.

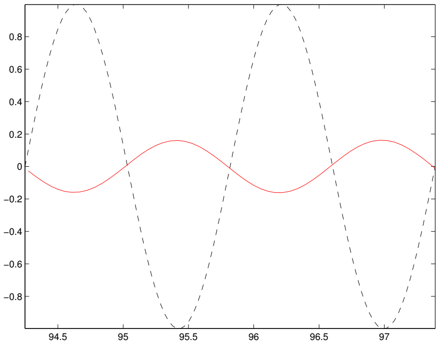


Figure 5.1: Input (black dashed) with frequency $\omega = 4$ and corresponding output (rot) for the down-hanging pendulum

[3] Hendrik Wade Bode (1905–1982), US-American electrical engineer

[4] Let I be an interval. A continuous function $\varphi : I \to \mathbb{R}$ is called *argument function* of a function $\gamma : I \to \mathbb{C} \setminus \{0\}$, if $\gamma(t) = |\gamma(t)| e^{i\varphi(t)}$ holds for all $t \in I$. We then write briefly $\varphi = \arg \gamma$.

This direct relation between the transfer function and the output signal means that, conversely, by measuring the amplitude and the phase shift of the output for sinusoidal input, we can reconstruct the values $G(i\omega) = |G(i\omega)|e^{\varphi(\omega)}$. The transfer function on the imaginary axis $i\mathbb{R}$ can thus be obtained experimentally from measurements.

This fact is particularly important because of a theorem from complex analysos: One can prove that the function G is uniquely determined by its values $G(i\omega)$ on $i\mathbb{R}$. More precisely, for i/o-stable systems (4.1) Cauchy's integral formula yields the expression

$$G(s) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{G(i\omega)}{i\omega - s} d\omega$$

for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 0$ (note that the absense of "$Du(t)$" in the expression for $y(t)$ in (4.1) is important here; otherwise the formula has to be modified). Since, moreover, $G(i\omega) \to 0$ holds for all $\omega \to \pm\infty$, the above integral can be approximated by an integral with compact integration interval. Consequently, fo an i/o-stabe system the tranfer function can be entirely reconstructed from measurement values for sinusoidal input signals, cf, [14, Abschnitt 6.5.3].

Graphically, these measurement values are depicted in the so-called Bode diagram, where logarithmic scales are used for the frequency and for the modulus $|G(i\omega)|$. Figure 5.2 shows this diagra for the down-hanging pendulum, again with $k = 0.1$ and $g = 9.81$.

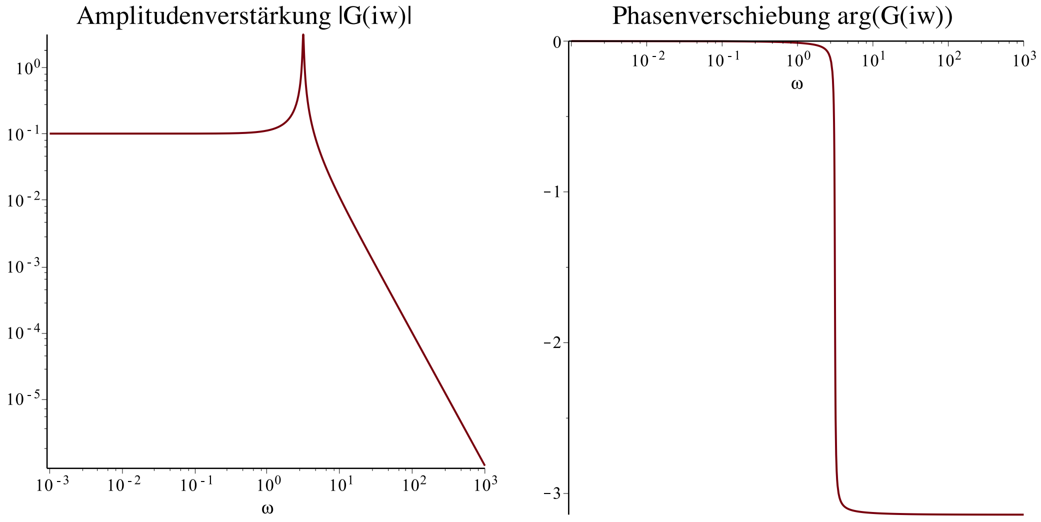


Figure 5.2: Bode diagram for the down-hanging pendulum

The left diagram says that the input signal is first weakly amplified and then, with increasing frequency up to about $\omega = 3$, the amplification increases and then decresase again for larger value of ω. Die phase remains almost unchanged for small values of ω, while after approximately $\omega = 3$ it is abruptly shifted by about $-\pi$. This behaviour is confirmed by the numerical simulations of the pendulum in Figure 5.3.

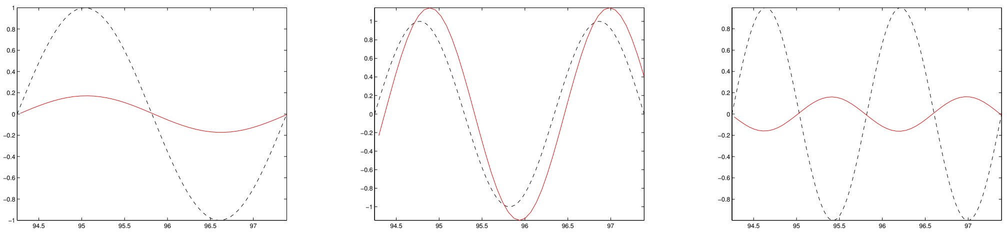


Figure 5.3: Input (black dashed) and output (red) for the down-hanging pendulum with $\omega = 2, 3, 4$ from left to right

The Nyquist diagram

The Nyquist diagram[5] serves for checking whether a closed-loop system is i/o-stable. Just like the Bode diagram the graph can be obtained entirely from measurement values and thus stability can be verified experimentally.

By Theorem 5.16, in the SISO case the transfer function is given by

$$G_{cl} := \frac{G(s)K(s)}{1 - G(s)K(s)}.$$

by Theorem 5.13 is is i/o-stable if and only if there are no poles in the closed right half plane of $\mathbb{C}$. A sufficient condition for this is that $F(s) := 1 - G(s)K(s)$ has no zeros in the closed right half plane, which is the case if and only if $G_0(s) := -G(s)K(s)$ does not attain the value -1 in the right half plane.

The Nyquist diagram[6] now depicts the values of $G_0(\omega i)$ graphically for $\omega \in (-\infty, \infty)$. In practice, this is realised approximately by plotting the values from $-R$ to R for a large $R \in \mathbb{R}$ in place of $\pm\infty$. Since $G(s)K(s)$ is the transfer function of the coupling of the feedback law and the system, the values of this product can be determined experimentally.

Figure 5.4 shows these figures for the inverted pendulum with $G(s) = 1/(ks + s^2 - g)$ for $k = 0.1$ and $g = 9.81$, and the static feedback law $K = -1$ (left) and $K = -10$ (right).

The consideration of the polynomials in the enumerator and the denominator of G_0 now yields the following stability criterion.

Nyquist criterion: Let $n^+ \in \mathbb{N}$ denote the number of poles of G_0 with positive real part. Moreover, we assume that G_0 does not have poles with real part equal to 0. Then the closed-loop system with transfer function G_{cl} is i/o-stable if and only if the curve in the Nyquist diagram (called the frequency response locus) $G(\omega i)$ for $\omega = -\infty \ldots, \infty$ winds around the point $-1 = -1 + 0i \in \mathbb{C}$ exactly n^+ times in counterclockwise direction. In case $n^+ = 0$ stability holds if and only if the the frequency response locus does not wind around the point -1 in clockwise direction.

In our example from Figure 5.4, since $K = \mathit{const}$ the transfer function G_0 has the same poles as G; hence there is a pole with positive real part and none with real part 0. Consequently,

[5] Harry Nyquist (1889–1976), US-American electrical engineer

[6] Here we only present the variant for $D = 0$. See, e.g., [14, Section 8.5] for the general case.

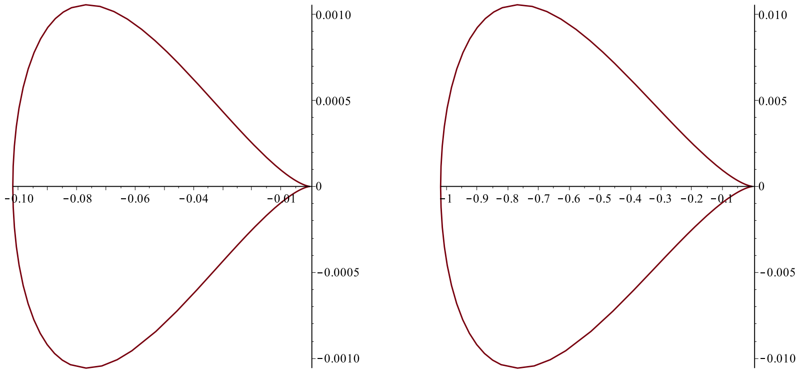


Figure 5.4: Nyquist diagram for the inverted pendulum with $K = -1$ (left) and $K = -10$ (right)

the frequency response locus must wind around -1 exactly once in counterclockwise direction. This is obviously not the case in the left figure for $K = -1$. Yet, ths condition is satisfied in the right figure for $K = -10$ (of course, the winding direction cannot be seen in this graph, but one can verify that it rund in counterclockwise direction, as required). The analysis in time domain yields that the corresponding closed-loop matrices for $K = -1$ and $K = -10$ are given by

$$A = \begin{pmatrix} 0 & 1 \\ g - K & -k \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 8.81 & -0.1 \end{pmatrix} \quad \text{bzw.} \quad A = \begin{pmatrix} 0 & 1 \\ -0.19 & -0.1 \end{pmatrix}.$$

The computation of the eigenvalues of this matrix confirms instability for $K = -1$ and stability $K = -10$. In fact, the treshold between instability and stability lies at $K = -9.81$.

Remark 5.18 For discrete time-systems a consideration in frequency domain is also possible. Instead of the Laplace transform in discrete time one uses the so-called z-transform, which we will not discuss here in detail. □

Chapter 6

Optimal stabilisation

The method for the computation of stabilising feedback laws proposed in Chapter 3 has the disadvantage that — except for the eigenvalues — the dynamics of the closed-loop system cannot be influenced. For instance, it is often the case that large control values u require to spend a lot of energy (as in the pendulum model where u represents the acceleration of the cart), which one would like to avoid. The heating model, is another example. Here large overshoots, i.e., oscillations until the desired temperature is reached, should be avoided.

In this chapter we will therefore present an approach that allows to exert more influence on the behavior of the closed-loop system. This will be achieved by using optimisation techniques, in which the desired behaviour can be determined via the choice of the cost function. As in Chapter 3 we will assume that the whole state vector x is accessible for evaluating Fx. If this is not the case, a dynamic observer as described in Chapter 4 can be used. We restrict ourselves to optimal control problems that are linked to stabilisation problems. More general problems will be addressed in the context of model predictive control later in this course.

6.1 Foundations of optimal control

In this section we will derive basic results in optimal control that we will need for solving the optimal feedback stabilisation problem. Since the derivation of these results is the same for linear and nonlinear systems, we will present it in the more general nonlinear setting. This means, we consider nonlinear control systems of the form

$$\dot{x}(t) = f(x(t), u(t)). \tag{6.1}$$

We assume that $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous and that for each $R > 0$ there exists $L_R > 0$ such that the Lipschitz condition

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\| \tag{6.2}$$

holds for all x_1, $x_2 \in \mathbb{R}^n$ and all $u \in \mathbb{R}^m$ with $\|x_1\|$, $\|x_2\|$, $\|u\| \leq R$. Under this assumption the well-known existence and uniqueness theorem for ordinary differential equations can be modified in such a way that for each piecewise continuous control function $u \in \mathcal{U}$ and each initial value x_0 it yields the existence of a unique solution $x(t, x_0, u)$ with $x(0, x_0, u) = x_0$ (see also Theorem 8.1, below).

We now define the optimal control problem we want to consider in the following.

Definition 6.1 For a continuous and non-negative *cost function* $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_0^+$ we define the *cost functional*

$$J(x_0, u) := \int_0^\infty g(x(t, x_0, u), u(t))dt.$$

The optimal control problem is then given by the optimisation problem

$$\text{minimise } J(x_0, u) \text{ with respect to } u \in \mathcal{U} \text{ for each } x_0 \in \mathbb{R}^n.$$

The function

$$V(x_0) := \inf_{u \in \mathcal{U}} J(x_0, u)$$

is called the *optimal value function* of this optimal control problem. A pair $(x^*, u^*) \in \mathbb{R}^n \times \mathcal{U}$ with $J(x^*, u^*) = V(x^*)$ is called *optimal pair.* □

As function space $\mathcal{U}$ as before we use the space of piecewise continuous functions. In addition, we assume that u is bounded on each compact intervall and that the functions u are continuous on the right, i.e. that for all $t_0 \in \mathbb{R}$ the condition $\lim_{t \searrow t_0} u(t) = u(t_0)$ holds. Observe that the second assumption can be made without loss of generality, because the solution does not depend on the value of u in the points of discontinuity.

Bemerkung 6.2 In discrete time with dynamics

$$x(k+1) = f(x(k), u(k))$$

and initial condition $x(0) = x_0$ the cost functional reads

$$J(x_0, u) := \sum_{k=0}^{\infty} g(x(k, x_0, u), u(k)).$$

□

Note that the functional $J(x_0, u)$ need not be finite. Moreover, the infimum in the definition of V need not be a minimum.

The first theorem in this chapter now gives a characterisation of the optimal value function V.

Theorem 6.3 (Dynamic Programming Principle or **Bellman's principle of optimality)**
(i) For each $\tau > 0$ the optimal value function satisfies

$$V(x_0) = \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x(\tau,x_0,u)) \right\}.$$

(ii) For each $\tau > 0$ and an optimal pair (x^*, u^*) it holds that

$$V(x^*) = \int_0^\tau g(x(t,x^*,u),u^*(t))dt + V(x(\tau,x^*,u^*)).$$

Proof: (i) We first show

$$V(x_0) \le \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x(\tau,x_0,u))$$

for all $u \in \mathcal{U}$ and all $\tau > 0$. To this end, let $x_\tau = x(\tau,x_0,u)$, $\varepsilon > 0$ be arbitrary and $u_\tau \in \mathcal{U}$ be such that

$$J(x_\tau,u_\tau) \le V(x_\tau) + \varepsilon$$

holds. Let $\tilde{u} = u\&_\tau u_\tau(\cdot - \tau)$ (cf. Definition 1.7). Then

$$\begin{aligned} V(x_0) &\le \int_0^\infty g(x(t,x_0,\tilde{u}),\tilde{u}(t))dt \\ &= \int_0^\tau g(x(t,x_0,\tilde{u}),\tilde{u}(t))dt + \int_\tau^\infty g(x(t,x_0,\tilde{u}),\tilde{u}(t))dt \\ &= \int_0^\tau g(x(t,x_0,u),u(t))dt + \int_\tau^\infty g(\underbrace{x(t,x_0,\tilde{u})}_{=x(t-\tau,x_\tau,u_\tau)},u_\tau(t-\tau))dt \\ &= \int_0^\tau g(x(t,x_0,u),u(t))dt + \int_0^\infty g(x(t,x_\tau,u_\tau),u_\tau(t))dt \\ &= \int_0^\tau g(x(t,x_0,u),u(t))dt + J(x_\tau,u_\tau) \;\le\; \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x_\tau) + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the claimed inequality follows.

As the second step we show

$$V(x_0) \ge \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x(\tau,x_0,u)) \right\}.$$

To this end, consider again an arbitrary $\varepsilon > 0$. We choose u_0 such that $V(x_0) \ge J(x_0,u_0) - \varepsilon$

holds and abbreviate $x_\tau = x(\tau, x_0, u_0)$. Then

$$\begin{aligned}
V(x_0) &\geq \int_0^\infty g(x(t,x_0,u_0),u_0(t))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + \int_\tau^\infty g(x(t,x_0,u_0),u_0(t))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + \int_0^\infty g(x(t,x(\tau,x_0,u_0),u_0(\cdot+\tau)),u_0(t+\tau))dt - \varepsilon \\
&= \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + J(x(\tau,x_0,u_0),u_0(\cdot+\tau)) - \varepsilon \\
&\geq \int_0^\tau g(x(t,x_0,u_0),u_0(t))dt + V(x(\tau,x_0,u_0)) - \varepsilon \\
&\geq \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau g(x(t,x_0,u),u(t))dt + V(x(\tau,x_0,u)) \right\} - \varepsilon
\end{aligned}$$

which shows the claim, since $\varepsilon > 0$ was arbitrary.

(ii) From (i) we immediately get the inequality

$$V(x^*) \leq \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + V(x(\tau,x^*,u^*)).$$

The converse inequality follows from

$$\begin{aligned}
V(x^*) &= \int_0^\infty g(x(t,x^*,u^*),u^*(t))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + \int_\tau^\infty g(x(t,x^*,u^*),u^*(t))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + \int_0^\infty g(x(t,x(\tau,x^*,u^*),u^*(\cdot+\tau)),u^*(t+\tau))dt \\
&= \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + J(x(\tau,x^*,u^*),u^*(\cdot+\tau)) \\
&\geq \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + V(x(\tau,x^*,u^*)).
\end{aligned}$$

□

A consequence of this principle is the following corollary.

Corollary 6.4 Let (x^*, u^*) be an optimal pair. Then $(x(\tau, x^*, u^*), u^*(\cdot + \tau))$ is also an optimal pair for each $\tau > 0$. □

Proof: Exercise.

In words, Corollary 6.4 states that final pieces of optimal trajectories are optimal trajectories themselves.

All statements made so far also hold in discrete time (with analogous proofs). In discrete time the dynamic programming principle reads for all $K \in \mathbb{N}$

$$V(x_0) = \inf_{u\in\mathcal{U}} \left\{ \sum_{k=0}^{K-1} g(x(k,x_0,u),u(k)) + V(x(K,x_0,u)) \right\} \tag{6.3}$$

and the optimal pairs (x^*, u^*) satisfy

$$V(x^*) = \sum_{k=0}^{K-1} g(x(k, x^*, u), u^*(k)) + V(x(K, x^*, u^*)).$$

The next statement derives a partial differential equation for V from Theorem 6.3, by means of a clever limit process for $\tau \to 0$. This statement does not have a discrete-time counterpart.

Theorem 6.5 (Hamilton-Jacobi-Bellman differential equation)
Let g be continuous in x and u. Moreover, let $O \subseteq \mathbb{R}^n$ be open and such that $V|_O$ is finite.

(i) If V is continuously differentiable in $x_0 \in O$, then

$$DV(x_0) \cdot f(x_0, u_0) + g(x_0, u_0) \geq 0$$

holds for all $u_0 \in \mathbb{R}^m$.

(ii) If (x^*, u^*) is an optimal pair and V is continuously differentiable in $x_0 \in O$, then

$$\min_{u \in \mathbb{R}^m} \{DV(x^*) \cdot f(x^*, u) + g(x^*, u)\} = 0 \tag{6.4}$$

and the minimum is attained in $u^*(0)$. Equation (6.4) is called *Hamilton-Jacobi-Bellman equation.*

Proof: We first show the auxiliary identity

$$\lim_{\tau \searrow 0} \frac{1}{\tau} \int_0^\tau g(x(t, x_0, u), u(t)) dt = g(x_0, u(0))$$

for each $u \in \mathcal{U}$. Because of continuity of x and u (on the right) in t and since g is continuous, for any $\varepsilon > 0$ there is $t_1 > 0$ with

$$|g(x(t, x_0, u), u(t)) - g(x_0, u(0))| \leq \varepsilon$$

for all $t \in [0, t_1)$. For $\tau \in (0, t_1]$ this yields

$$\begin{aligned} \left| \frac{1}{\tau} \int_0^\tau g(x(t, x_0, u), u(t)) dt - g(x_0, u(0)) \right| &\leq \frac{1}{\tau} \int_0^\tau |g(x(t, x_0, u), u(t)) - g(x_0, u(0))| dt \\ &\leq \frac{1}{\tau} \int_0^\tau \varepsilon \; = \; \varepsilon \end{aligned}$$

and thus the statement for the limit, since $\varepsilon > 0$ was arbitrary.

Now both assertions follow:

(i) For $u(t) \equiv u_0 \in \mathbb{R}^m$ Theorem 6.3(i) implies

$$V(x_0) \leq \int_0^\tau g(x(t, x_0, u), u(t)) dt + V(x(\tau, x_0, u))$$

and thus

$$\begin{aligned} DV(x_0)f(x_0,u(0)) &= \lim_{\tau\searrow 0} \frac{V(x(\tau,x_0,u)) - V(x_0)}{\tau} \\ &\geq \lim_{\tau\searrow 0} -\frac{1}{\tau}\int_0^\tau g(x(t,x_0,u),u(t))dt \;=\; -g(x_0,u(0)), \end{aligned}$$

i.e., the first assertion.

(ii) From (i) we get

$$\inf_{u\in\mathbb{R}^m}\{DV(x^*)\cdot f(x^*,u) + g(x^*,u)\} \geq 0.$$

Theorem 6.3(ii) moreover implies

$$V(x^*) = \int_0^\tau g(x(t,x^*,u^*),u^*(t))dt + V(x(\tau,x^*,u^*)).$$

This yields

$$\begin{aligned} DV(x^*)f(x^*,u^*(0)) &= \lim_{\tau\searrow 0} \frac{V(x(\tau,x^*,u^*)) - V(x^*)}{\tau} \\ &= \lim_{\tau\searrow 0} -\frac{1}{\tau}\int_0^\tau g(x(t,x^*,u^*),u^*(t))dt \;=\; -g(x^*,u^*(0)), \end{aligned}$$

which implies the existence of the minimum in $u = u^*(0)$ and the claimed identity. □

Theorem 6.5 provides *necessary* optimality conditions, i.e., conditions that *must* be satisfied for the optimal value function or for an optimal pair, respectively — provided the optimal value function is continuously differentiable. In general, however, the necessary condition does not imply that a function is indeed an optimal value function or that a pair is an optimal pair. To this end, *sufficient* optimality conditions are needed. We will derive them in the following.

For this derivation we need additional assumptions, which can be formulated in different ways. Since we want to apply the theory to stabilisation problems, the following assumption is suitable for our purposes.

Definition 6.6 Assume that f satisfies $f(0,0) = 0$, i.e. the origin is an equilibrium of the control system for $u = 0$. Then we call the optimal control problem *null controlling*, if the implication

$$J(x_0,u) < \infty \quad\Rightarrow\quad x(t,x_0,u) \to 0 \text{ for } t \to \infty$$

holds. □

Now we can formulate the sufficient condition.

Theorem 6.7 (sufficient optimality condition)
Consider a null controlling optimal control problem. Let $W : \mathbb{R}^n \to \mathbb{R}_0^+$ be a continuously differentiable function with $W(0) = 0$, which satisfies the Hamilton-Jacobi-Bellman equation

$$\min_{u\in\mathbb{R}^m}\{DW(x)f(x,u) + g(x,u)\} = 0.$$

For given $x^* \in \mathbb{R}^n$ let $u^* \in \mathcal{U}$ a control function, such that for the corresponding solution $x(t, x^*, u^*)$ and all $t \geq 0$ the minimum in this equation for $x = x(t, x^*, u^*)$ is attained in $u = u^*(t)$.

Then (x^*, u^*) is an optimal pair and

$$V(x(t, x^*, u^*)) = W(x(t, x^*, u^*))$$

holds for all $t \geq 0$.

Proof: We prove the assertion for $t = 0$. Für $t > 0$ it follows by applying the proof to $(x(t, x^*, u^*), u^*(t + \cdot))$. Consider $u \in \mathcal{U}$ and let $x(t) = x(t, x^*, u)$ be the corresponding solution. We start by showing the inequality

$$J(x^*, u) \geq W(x^*).$$

In case $J(x^*, u) = \infty$ there is nothing to show. It thus suffices to consider the case $J(x^*, u) < \infty$. From the Hamilton-Jacobi-Bellman equation we can conclude

$$\frac{d}{dt} W(x(t)) = DW(x(t)) f(x(t), u(t)) \geq -g(x(t), u(t)),$$

and hence the fundamental theorem of calculus yields

$$W(x(T)) - W(x^*) = \int_0^T \frac{d}{dt} W(x(t)) dt \geq -\int_0^T g(x(t), u(t)) dt.$$

This implies

$$J(x^*, u) = \lim_{T \to \infty} \int_0^T g(x(t), u(t)) dt \geq \lim_{T \to \infty} \big(W(x^*) - W(x(T))\big) = W(x^*).$$

for all $T > 0$. Here the last identity follows since the problem is null controlling and $J(x^*, u) < \infty$. This implies $x(T) \to 0$ for $T \to \infty$ and thus by continuity of W and $W(0) = 0$ we obtain $W(x(T)) \to 0$.

Observe that this inequality in particular implies $V(x^*) = \inf_{u \in \mathbb{U}} J(x^*, u) \geq W(x^*)$. To conclude the proof it is thus sufficient to show

$$J(x^*, u^*) = W(x^*).$$

For the control u^* and the corresponding solution $x^* = x(t, x^*, u^*)$ the Hamilton-Jacobi-Bellman equation implies

$$\frac{d}{dt} W(x^*(t)) = DW(x^*(t)) f(x^*(t), u^*(t)) = -g(x^*(t), u^*(t)),$$

and analogously to above we get

$$J(x^*, u^*) = \lim_{T \to \infty} \int_0^T g(x^*(t), u^*(t)) dt = \lim_{T \to \infty} \big(W(x^*) - W(x(T))\big) = W(x^*).$$

□

Observe that both theorems in this section apply only if V or W, respectively, are differentiable. In the general nonlinear case, this assumption is relatively restrictive[1]. Moreover, it is in general quite difficult to compute V by solving this equation, even if V is differentiable.

In the linear case, however, the problem and the Hamilton-Jacobi-Bellman equation simplify considerably, such that an explicit solution as possible, as we will see in the following section.

6.2 The linear-quadratic problem

Now we return to the linear control system (1.3)

$$\dot{x}(t) = Ax(t) + Bu(t) =: f(x(t), u(t)).$$

In order to obtain an applicable solution theory, we also need to impose a suitable structure for the cost function $g(x, u)$.

Definition 6.8 A quadratic cost function $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_0^+$ is given by

$$g(x,u) = (x^T\, u^T) \begin{pmatrix} Q & N \\ N^T & R \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}$$

with $Q \in \mathbb{R}^{n\times n}$, $N \in \mathbb{R}^{n\times m}$ and $R \in \mathbb{R}^{m\times m}$, such that $G := \begin{pmatrix} Q & N \\ N^T & R \end{pmatrix}$ is symmetric and positive definite (briefly: spd). □

This is the reason for the name "linear-quadratic" optimal control problem: the dynamics is linear and the cost function is quadratic.

We first show that this problem is null-controlling.

Lemma 6.9 The linear-quadratic problem is null-controlling in the sense of Definition 6.6.

Proof: We first show the inequalities

$$g(x,u) \geq c_1\|x\|^2 \text{ and } g(x,u) \geq c_2\|f(x,u)\|^2 \tag{6.5}$$

for suitable constants $c_1, c_2 > 0$.

Since the matrix G is spd, Lemma 3.10 implies the inequality

$$g(x,u) \geq c_1 \left\| \begin{pmatrix} x \\ u \end{pmatrix} \right\|^2 \geq c_1\|x\|^2, \tag{6.6}$$

[1] The nonlinear theory of these equations uses the notion of "viscosity solutions", a generalised solution concept that is also meaningful if V is not differentiable.

i.e., the first estimate in (6.5). Since

$$\|f(x,u)\|^2 = (x^T, u^T) \left(\begin{array}{cc} A & A^T B \\ B^T A & B \end{array} \right) \left(\begin{array}{c} x \\ u \end{array} \right)$$

Lemma 3.10 moreover yields

$$\|f(x,u)\|^2 \leq c_3 \left\| \left(\begin{array}{c} x \\ u \end{array} \right) \right\|^2,$$

which together with (6.6) and $c_2 = c_1/c_3$ yields the second estimate in (6.5).

Consider now $x_0 \in \mathbb{R}^n$ and $u \in \mathcal{U}$ with

$$J(x_0, u) < \infty$$

and denote by $x(t) = x(t, x_0, u)$ the corresponding solution. We have to show that

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

To this end, assume that $x(t) \not\to 0$. This means that there exists $\varepsilon > 0$ and a sequence $t_k \to \infty$ such that $\|x(t_k)\| \geq \varepsilon$. Without loss of generality we can assume $t_{k+1} - t_k \geq \varepsilon/2$. Now we set $\delta = \varepsilon/4$ and distinguish two cases for $k \in \mathbb{N}$:

Case 1: $\|x(t)\| \geq \varepsilon/2$ for all $t \in [t_k, t_k + \delta]$. In this case for these t from (6.5) we get the inequality $g(x(t), u(t)) \geq c_1\varepsilon^2/4$ and consequently

$$\int_{t_k}^{t_k+\delta} g(x(t), u(t))dt \geq c_1\delta\varepsilon^2/4 = c_1\varepsilon^3/16.$$

Case 2: $\|x(t)\| < \varepsilon/2$ for a $t \in [t_k, t_k + \delta]$. In this case we get

$$\left\| \int_{t_k}^{t} f(x(\tau), u(\tau))d\tau \right\| = \|x(t_k) - x(t)\| \geq \|x(t_k)\| - \|x(t)\| \geq \varepsilon/2.$$

From the second estimate in (6.5) we obtain

$$g(x,u) \geq c_2\|f(x,u)\|^2 \geq \left\{ \begin{array}{ll} 0, & \|f(x,u)\| \leq 1 \\ c_2\|f(x,u)\|, & \|f(x,u)\| > 1 \end{array} \right\} \geq c_2(\|f(x,u)\| - 1)$$

and hence

$$\int_{t_k}^{t_k+\delta} g(x(\tau), u(\tau))d\tau \geq c_2 \int_{t_k}^{t_k+\delta} \|f(x(\tau), u(\tau))\| - 1 d\tau \geq c_2(\varepsilon/2 - \delta) \geq c_2\varepsilon/4.$$

Setting $\gamma = \min\{c_1\varepsilon^3/16,\, c_2\varepsilon/4\} > 0$ we obtain

$$J(x_0, u) = \int_0^\infty g(x(t), u(t))dt \geq \sum_{k=1}^\infty \int_{t_k}^{t_k+\delta} g(x(t), u(t))dt \geq \sum_{k=1}^\infty \gamma = \infty,$$

and thus a contradiction. □

We can thus use Theorem 6.7 for verifying the optimality of a solution of the linear-quadratic problem.

In order to find a candidate for the optimal value function, we make the ansatz

$$W(x) = x^T P x \tag{6.7}$$

for an spd matrix $P \in \mathbb{R}^{n\times n}$.

A priori we do not know whether this ansatz is justified — for now we simply assume is and investigate the consequences.

Lemma 6.10 If the linear-quadratic optimal control problem has an optimal value function of the form (6.7), then all optimal pairs (x^*, u^*) are of the form

$$u^*(t) = Fx(t, x^*, F)$$

where $F \in \mathbb{R}^{m\times n}$ is given by

$$F = -R^{-1}(B^T P + N^T),$$

and $x(t, x^*, F)$ is the solution of the closed-loop system with feedback law F, i.e.,

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

with initial condition $x(0, x^*, F) = x^*$.

Moreover, the origin is exponentially stable for the closed-loop system with feedback law F.

Proof: The optimal value function of the form (6.7) is continuously differentiable and satsfies $W(0) = 0$. This implies that both Theorem 6.5 and Theorem 6.7 are applicable.

If W is the optimal value function, then Theorem 6.5(ii) implies that the optimal control $u = u^*(t)$ for $x = x(t, x^*, u^*)$ minimises the expression

$$DW(x) \cdot f(x, u) + g(x, u). \tag{6.8}$$

Conversely, Theorem 6.7 yields that any control function, which minimises (6.8) along the corresponding solution trajectory generates an optimal pair. We thus have to show that the feedback law F generates such solutions and control functions and that the u^* specified in the theorem is the only control function that minimises (6.8).

For the linear-quadratic problem under consideration, the expression (6.8) to be minimised equals

$$\begin{aligned}
& DW(x) \cdot f(x, u) + g(x, u) \\
= \;& x^T P(Ax + Bu) + (Ax + Bu)^T Px + x^T Qx + x^T Nu + u^T N^T x + u^T Ru \\
= \;& 2x^T P(Ax + Bu) + x^T Qx + 2x^T Nu + u^T Ru \quad =: \quad h(u),
\end{aligned}$$

since P is symmetric. Since G is spd, R must be spd, too, and thus the second derivative of h with respect to u is spd. Thus, the function h is strictly convex in u. Thus, any zero of the derivative of h with respect to u is a global minimum. These derivatives are given by

$$\begin{array}{rrcl} & 0 & = & Dh(u) = 2x^TPB + 2x^TN + 2u^TR \\ \Leftrightarrow & -2u^TR & = & 2x^TPB + 2x^TN \\ \Leftrightarrow & -Ru & = & B^TPx + N^Tx \\ \Leftrightarrow & u & = & -R^{-1}(B^TPx + N^Tx) \ = \ Fx, \end{array}$$

which shows the claim.

Exponential stability of the closed-loop system follows from the Hamilton-Jacobi-Bellman equation. Because G is spd, by Lemma 3.10 we obtain

$$DW(x) \cdot f(x, Fx) = -g(x, Fx) \leq -c\|(x^T, (Fx)^T)^T\|^2 \leq -c\|x\|^2$$

for a suitable $c > 0$. Since, moreover, P is positive definite, $W(x)$ is a Lyapunov function and according to Lemma 3.11 exponential stabilily of the origin follows. □

If the optimal value function is of the form (6.7), then we obtain a particularly nice solution: We can not only compute the optimal control functions u^* explicitly, they are, moreover, given in feedback form and, as an (obviously intended) side effect the optimal feedback law stabilises the system.

We thus have to investigate when V can assume the form (6.7). The next lemma gives a sufficient condition for this fact, as well as a possibility for computing P.

Lemma 6.11 If the matrix $P \in \mathbb{R}^{n\times n}$ is an spd solution of the algebraic Riccati equation[2]

$$PA + A^TP + Q - (PB + N)R^{-1}(B^TP + N^T) = 0, \tag{6.9}$$

then the optimal value function of the problem is given by $V(x) = x^TPx$.

In particular, there exists at most one spd solution P of (6.9).

Proof: We start by showing that $W(x) = x^TPx$ solves the Hamilton-Jacobi-Bellman equation (6.4).

In the proof of Lemma 6.10 we already established the identity

$$\min_{u\in U}\{DW(x)\cdot f(x,u) + g(x,u)\} = DW(x)\cdot f(x,Fx) + g(x,Fx)$$

for the matrix $F = -R^{-1}(B^TP + N^T)$. Using

$$\begin{array}{l} F^TB^TP + F^TRF + F^TN^T \\ = \ -(N+PB)R^{-1}B^TP + (N+PB)R^{-1}RR^{-1}(B^TP+N^T) - (N+PB)R^{-1}N^T = 0 \end{array}$$

we obtain

$$\begin{array}{rl} & DW(x)\cdot f(x,Fx) + g(x,Fx) \\ = & x^T(P(A+BF) + (A+BF)^TP + Q + NF + F^TN^T + F^TRF)x \\ = & x^T(PA + A^TP + Q + (PB+N)F + \underbrace{F^TB^TP + F^TRF + F^TN^T}_{=0})x \\ = & x^T(PA + A^TP + Q + (PB+N)F)x \\ = & x^T(PA + A^TP + Q - (PB+N)R^{-1}(B^TP + N^T))x. \end{array}$$

[2] named after Jacopo Francesco Riccati, Italian mathematician, 1676–1754

If the algebraic Riccati equation (6.9) is satisfied, then this expression equals zero and the Hamilton-Jacobi-Bellman equation is satisfied.

In order to prove $V(x) = W(x)$, we nosw show that the assumptions of Theorem 6.7 are satisfied. Positive definiteness of P implies $W(x) \geq 0$ and $W(0) = 0$. As shown above, $W(x) = x^T Px$ solves the Hamilton-Jacobi-Bellman equation. Moreover, from the construction of u^* via the feedback law F in Lemma 6.10 it follows that it satisfies the assumptions on u^* in Theorem 6.7. Thus, $V(x) = W(x)$ follows from Theorem 6.7.

The uniqueness of the spd solution P follows from the fact that the proof of $V(x) = W(x)$ applies to any such solution. This implies $V(x) = x^T Px$ for all $x \in \mathbb{R}^n$, by which P is uniquely determined. □

Remark 6.12 Note that the uniqueness statement of this lemma only holds for spd, i.e., symmetric and positive definite solution matrices P. In general, the algebraic Riccati equation has more than one solution. However, at most one of these can be spd. □

Lemma 6.10 and 6.11 suggest the following strategy for solving the linear-quadratic problem:

> Find an spd solution P of the algebraic Riccati equation (6.9) and compute from this the optimal linear feedback law F according to Lemma 6.10.

This yields an optimal linear feedback law, which according Lemma 6.10 also solves the stabilisation problem.

The important question thus is: Under which assumptions can we prove the existence of a spd solution of the algebraic Riccati equation? The following theorem shows that this approach works under the weakest possible assumption on A and B.

Theorem 6.13 For the linear-quadratic optimal control problem the following statements are equivalent:

(i) The pair (A, B) is stabilisable.

(ii) The algebraic Riccati equation (6.9) has a unique spd solution P.

(iii) The optimal value function is of the form (6.7).

(iv) There exists an optimal linear feedback law, which stabilises the control system.

Proof: "(i) ⇒ (ii)": Consider the Riccati differential equation

$$\dot{P}(t) = P(t)A + A^T P(t) + Q - (P(t)B + N)R^{-1}(B^T P(t) + N^T)$$

with matrix-valued solution $P(t)$ that satisfies the initial condition $P(0) = 0$. From the theory of ordinary differential equations we know that this solution $P(t)$ exists for all t from a maximal existence interval $[0, t^*)$, i.e., t^* is chosen maximally. A direct computation shows that $P(t)^T$ is also a solution of this equation that satisfies $P(0)^T = 0$. Because of the uniqueness of the solution we obtain $P(t) = P(t)^T$, i.e. the solution is symmetric.

As the first step of the proof we show that this solution exists for all $t \geq 0$, i.e. that $t^* = \infty$. To this end we assume that $t^* < \infty$.

With analogous computations as in the proof of Lemma 6.10 one sees, that for all $t_1 - t \in [0, t^*)$ and all $u \in U$ the function $W(t, t_1, x) := x^T P(t_1 - t)x$ satisfies the inequality

$$\frac{\partial}{\partial t} W(t, t_1, x) + \frac{\partial}{\partial x} W(t, t_1, x) \cdot f(x, u) + g(x, u) \geq 0. \tag{6.10}$$

For any solution $x(t) = x(t, x_0, u)$ of the control system with arbitrary $u \in \mathcal{U}$ this implies

$$\frac{d}{dt} W(t, t_1, x(t)) = \frac{\partial}{\partial t} W(t, t_1, x(t)) + \frac{\partial}{\partial x} W(t, t_1, x(t)) \cdot f(x, u(t)) \geq -g(x(t), u(t)).$$

The fundamental theorem of calculus together with $W(t_1, t_1, x) = 0$ then yields

$$W(0, t_1, x_0) = -\int_0^{t_1} \frac{d}{dt} W(t, t_1, x) dt \leq \int_0^{t_1} g(x(t, x_0, u), u(t)) dt \tag{6.11}$$

for all $t_1 \in [0, t^*)$. Again analogously to the proof of Lemma 6.10 one checks that for the control value defined by $u = u^* = -R^{-1}(B^T P(t) + N^T)x$ one obtains equality in (6.10). With a similar derivation as above one sees that with the control function $u^*(t) = -R^{-1}(B^T P(t) + N^T)x(t, x_0, u^*)$ we get

$$W(0, t_1, x_0) = \int_0^{t_1} g(x(t, x_0, u^*), u^*(t)) dt. \tag{6.12}$$

Since G is spd and the solutions $x(t, x_0, u^*)$ are continuous, we get $W(0, t_1, x_0) > 0$ for $x_0 \neq 0$, implying that $P(t_1)$ is spd. With the particular choice $u \equiv 0$ inequality (6.11) implies that $W(0, t_1, x_0) = x^T P(t_1)x$ is uniformly bounded for all $t_1 \in [0, t^*)$. Now the symmetry of $P(t)$ implies that its entries satisfy

$$[P(t)]_{ij} = e_i^T P(t) e_j = \frac{1}{2}((e_i + e_j)^T P(t)(e_i + e_j) - e_i^T P(t) e_i - e_j^T P(t) e_j). \tag{6.13}$$

Hence, the entries of $P(t)$ are also uniformly bounded for $t \in [0, t^*)$. From the theory of ordinary differential equations it is known that if the right hand side of the equation is globally defined (which is the case for our equation, since the right hand side is defined for all $P \in \mathbb{R}^{n \times n}$) and $t^* < \infty$, then the norm of the solution must tend to infinity as $t \nearrow t^*$. This, however, is only possible if at least one entry of $P(t)$ grows unboundedly. Since here, however, all entries are bounded, $t^* < \infty$ is not possible.

The solution $P(t)$ is thus an spd matrix valued function for all $t \geq 0$. Moreover, (6.12) implies for all $s \geq t$ and all $x \in \mathbb{R}^n$ the inequality

$$x^T P(s) x \geq x^T P(t) x.$$

We now show that $P_\infty := \lim_{t\to\infty} P(t)$ exists. To this end we pick a stabilising feedback law F for the pair (A, B) and set $u_F(t) = Fx(t, x_0, F)$. Then from (6.11) and the estimate

$$g(x, Fx) \leq K\|x\|^2$$

we obtain the inequality

$$\begin{aligned} W(0, t_1, x_0) &\leq \int_0^{t_1} g(x(\tau, x_0, F), u_F(\tau))d\tau \\ &\leq \int_0^{t_1} K(Ce^{-\sigma t}\|x_0\|)^2 dt \\ &\leq \underbrace{\int_0^{\infty} KC^2 e^{-2\sigma t} dt}_{=\frac{KC^2}{2\sigma}=:D<\infty} \|x_0\|^2 \quad \leq \quad D\|x_0\|^2. \end{aligned}$$

This implies $x^T P(t) x \leq D\|x\|^2$ for all $t \geq 0$, thus for any fixed $x \in \mathbb{R}^n$ the expression $x^T P(t) x$ is bounded and monotone increasing. This implies that it converges for $t \to \infty$. Denoting the j-th basis vector as e_j and defining

$$l_{ij} = \lim_{t\to\infty} (e_i + e_j)^T P(t)(e_i + e_j) \quad \text{and} \quad l_j = \lim_{t\to\infty} e_j^T P(t) e_j,$$

from (6.13) we can conclude

$$\lim_{t\to\infty} [P(t)]_{ij} = \frac{1}{2}(l_{ij} - l_i - l_j).$$

This implies that the limit $P_\infty := \lim_{t\to\infty} P(t)$ exists. This matrix is symmetric and since

$$x^T P_\infty x \geq x^T P(t) x > 0 \quad \text{for all } x \neq 0 \text{ and all } t > 0$$

it is also positive definite.

We finally show that this P_∞ solves the algebraic Riccati equation. From the qualitative theory of ordinary differential equations it is known that $P(t) \to P_\infty$ implies that P_∞ is an equilibrium of the Riccati ODE.[3] This immediately implies that P_∞ solves the algebraic Riccati equation, from which the existence of an spd solution follows. Uniqueness then follows from Lemma 6.11.

"(ii) $\Rightarrow$ (iii)": Follows from Lemma 6.11

"(iii) $\Rightarrow$ (iv)": Follows from Lemma 6.10.

"(iv) $\Rightarrow$ (i)": Since a stabilising feedback law exists, the pair (A, B) is stabilisable. □

[3] see, e.g., Lemma 7.2 in [7]

Remark 6.14 The auxiliary function $W(t_0, t_1)$ used in the proof of "(i)⇒(ii)" is actually the optimal value function of the optimal control problem

$$\text{minimise } J(t_0, t_1, x_0, u) := \int_{t_0}^{t_1} g(x(t, t_0, x_0, u), u(t))dt$$

on the finite time horizon $[t_0, t_1]$. Here $x(t, t_0, x_0, u)$ denotes the solution of the control problem with initial time t_0 and initial value x_0, i.e., $x(t_0, t_0, x_0, u) = x_0$. □

This observation can even be further generalised, as we briefly sketch (without proofs):

For the linear-quadratic problem on finite time horizon with terminal cost $l(x) = x^T Lx$ for an spd matrix $L \in \mathbb{R}^n \times n$, i.e.

$$\text{minimise } J(t_0, t_1, x_0, u) := \int_{t_0}^{t_1} g(x(t, t_0, x_0, u), u(t))dt + l(x(t_1, t_0, x_0, u))$$

the optimal value function is given by

$$W(t_0, t_1) = x^T P(t_1 - t_0)x,$$

where $P(\cdot)$ solves the Riccati differential equation (as in the proof above), but now with initial condition $P(0) = L$.

Analogously to the infinite horizon problem, the optimal feedback law is given by

$$F(t) = -R^{-1}(B^T P(t_1 - t) + N^T),$$

but now it depends on the time t. The optimally controlled system on $[t_0, t_1]$ thus reads

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

Observe that for $t_1 \to \infty$ and t fixed, the time varying feedback law $F(t)$ converges to F from Lemma 6.10.

Remark 6.15 For discrete-time systems analogous results to the results in this chapter can be obtained. These result do not build upon the Hamilton-Jacobi-Bellman equation but rather on the optimality principle (6.3) for $K = 1$. This leads to the discrete-time algebraic Riccati equation

$$A^T PA - P - (A^T PB + N)(B^T PB + R)^{-1}(B^T PA + N^T) + Q = 0.$$

The formula for the optimal feedback law changes to $F = (B^T PB + R)^{-1}(B^T PA + N^T)$. □

6.3 Linear-quadratic output regulation

In the previous section we always assumed that the matrix G in the definition of the cost function $g(x,u)$ is positive definite. In the exercises we have seen that in general the LQ problem is not null-controlling and that the proposed solution may not work if this condition is violated.

Nevertheless, there are reasons to relax this condition. If we consider a control system with output (4.1) (cf. Chapter 4), i.e.

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

then it makes sense that the optimisation objective depends only on y and not on x. This means that we consider a cost function of the form $\tilde{g}(y,u)$. This can be achieved by choosing the submatrices Q and N of G of the form

$$Q = C^T\widetilde{Q}C,\ N = C^T\widetilde{N}$$

for matrices $\widetilde{Q}$ and $\widetilde{N}$ of appropriate dimension. Then we get

$$\begin{aligned} g(x,u) &= (x^T\,u^T)\underbrace{\begin{pmatrix} Q & N \\ N^T & R \end{pmatrix}}_{=:G}\begin{pmatrix} x \\ u \end{pmatrix} = (x^T\,u^T)\begin{pmatrix} C^T\widetilde{Q}C & C^T\widetilde{N} \\ \widetilde{N}^TC & R \end{pmatrix}\begin{pmatrix} x \\ u \end{pmatrix} \\ &= (y^T\,u^T)\underbrace{\begin{pmatrix} \widetilde{Q} & \widetilde{N} \\ \widetilde{N}^T & R \end{pmatrix}}_{=:\widetilde{G}}\begin{pmatrix} y \\ u \end{pmatrix} =: \tilde{g}(y,u). \end{aligned} \tag{6.14}$$

Here we choose $\widetilde{Q}$ and $\widetilde{N}$ such that $\widetilde{G}$ is spd. The matrix G is now no longer positive definite. Nevertheless the results from the previous sections can be carried over to this new G. To this end we must check where positive definiteness entered in the proofs:

(i) In Lemma 6.9 positive definiteness of G is used in order to show that the problem is null-controlling.

(ii) In Lemma 6.10 positive definiteness of the submatrix R is exploited implicitly, because the R^{-1} is used.

(iii) In the proof of the implication "(i)⇒(ii)" in Theorem 6.13 positive definiteness of G is exploited to prove that $P(t)$ is positiv definite.

Item (ii) is not an issue here, because R is still assumed to be positive definite. Items (i) and (iii) will be clarified in the sequel. The following lemma is essential for this.

Lemma 6.16 Let the pair (A,C) be observable. Then for any $t_1 > 0$ there is $c > 0$, such that for g from (6.14) the estimate

$$J(0,t_1,x_0,u) = \int_0^{t_1} g(x(t;x_0,u),u(t))dt \geq c\|x_0\|^2$$

holds for all $x_0 \in \mathbb{R}^n$ and all $u \in \mathcal{U}$.

Proof: From the general solution formula

$$x(t;x_0,u) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds = x(t;x_0,0) + x(t;0,u)$$

we can conslude that for all $\alpha > 0$

$$x(t;\alpha x_0,\alpha u) = \alpha x(t;x_0,u)$$

holds. This imlies for $x_0 \neq 0$ and $\alpha = \|x_0\|$ the identity

$$J(0,t_1,x_0,u) = \alpha^2 J(0,t_1,x_0/\alpha,u/\alpha) = \|x_0\|^2 J(0,t_1,x_0/\|x_0\|,u/\|x_0\|).$$

In order to show the assertion it thus suffices to prove the existence of $c > 0$ with

$$J(0,t_1,x_0,u) \geq c \quad \text{for all } x_0 \in \mathbb{R}^n \text{ with } \|x_0\| = 1 \text{ and all } u \in \mathcal{U}. \tag{6.15}$$

In order to prove (6.15) we first consider

$$J(0,t_1,x_0,0) = \int_0^{t_1} x(t;x_0,0)^T Q x(t;x_0,0)dt = \int_0^{t_1} y(t)^T \widetilde{Q} y(t)dt.$$

Since (A,C) is observable, Lemma 4.5 implies that for all $x_0 \neq 0$ there is $\tau \in [0,t_1]$ with $y(\tau) \neq 0$. Since $y(t)$ is continuous, we can conclude $y(t) \neq 0$ on an intervall around τ, which by means of the positive definiteness of $\widetilde{Q}$ implies the inequality $J(0,t_1,x_0,0) > 0$. Since $J(0,t_1,x_0,0)$ is continuous in x_0, it attains a minimum $c_0 > 0$ on the compact set $\{x_0 \in \mathbb{R}^n \,|\, \|x_0\| = 1\}$, implying

$$J(0,t_1,x_0,0) \geq c_0 \tag{6.16}$$

for all $x_0 \in \mathbb{R}^n$ with $\|x_0\| = 1$.

For estimating $J(0,t_1,x_0,u)$ we now choose an arbitrary $x_0 \in \mathbb{R}^n$ with $\|x_0\| = 1$ and an $\varepsilon > 0$. For control functions u with

$$\int_0^{t_1} u(t)^T R u(t)dt > \varepsilon \tag{6.17}$$

we then obtain $\int_0^{t_1} u(t)^T u(t)dt \geq k_1\varepsilon$, where $k_1 = 1/\|R\|$. Consequently, positive definiteness of $\widetilde{G}$ implies

$$J(0,t_1,x_0,u) = \int_0^{t_1} \underbrace{(y(t)^T u(t)^T)\widetilde{G}\begin{pmatrix} y(t) \\ u(t) \end{pmatrix}}_{\geq k_2 \left\| \binom{y(t)}{u(t)} \right\|^2 \geq k_2\|u(t)\|^2} dt \geq k_1 k_2 \varepsilon > 0 \tag{6.18}$$

with $k_2 = 1/\|\widetilde{G}^{-1}\|$. It thus remains to show the inequality for control functions $u \in \mathcal{U}$ with

$$\int_0^{t_1} u(t)^T R u(t)dt \leq \varepsilon. \tag{6.19}$$

Since R is positive definite, we get

$$\|u(t)\|^2 \leq c_1 u(t)^T R u(t)$$

for some $c_1 > 0$ and thus

$$\int_0^{t_1} \|u(t)\|^2 dt \le c_1 \varepsilon.$$

In addition we get

$$\|u(t)\| \le \begin{cases} \sqrt{\varepsilon}, & \|u(t)\|^2 \le \varepsilon \\ \|u(t)\|^2/\sqrt{\varepsilon}, & \|u(t)\|^2 > \varepsilon. \end{cases}$$

This implies

$$\int_0^{t_1} \|u(t)\| dt \le \int_0^{t_1} \max\{\sqrt{\varepsilon}, \|u(t)\|^2/\sqrt{\varepsilon}\} dt \le \int_0^{t_1} \sqrt{\varepsilon} + \|u(t)\|^2/\sqrt{\varepsilon} dt = (c_1 + t_1)\sqrt{\varepsilon}.$$

From the general solution formula we thus obtain the existence of a constant $c_2 > 0$ with

$$\|x(t; 0, u)\| \le c_2 \sqrt{\varepsilon} \tag{6.20}$$

for all $t \in [0, t_1]$. Similarly, the solution formula implies

$$\|x(t; x_0, 0)\| \le c_3 \|x_0\| = c_3 \tag{6.21}$$

for a suitable constant $c_3 > 0$ and all $t \in [0, t_1]$. In particular, this implies

$$\|x(t; x_0, u)\| \le c_4 \tag{6.22}$$

for $c_4 = c_2\sqrt{\varepsilon} + c_3$.

For the functional we thus obtain

$$J(0, t_1, x_0, u) \ge \int_0^{t_1} x(t; x_0, u)^T Q x(t; x_0, u) dt + 2\int_0^{t_1} x(t; x_0, u)^T N u(t) dt.$$

Because of (6.22), the second term satisfies the inequality

$$2\int_0^{t_1} x(t; x_0, u)^T N u(t) dt \ge -2c_4\|N\| \int_0^{t_1} \|u(t)\| dt \ge -2c_4 \|N\|(c_1 + t_1)\sqrt{\varepsilon} =: -c_5\sqrt{\varepsilon}.$$

From the estimate

$$(x_1 + x_2)^T Q (x_1 + x_2) = x_1^T Q x_1 + x_2^T Q x_2 + 2x_1^T Q x_2 \ge x_1^T Q x_1 + 2x_1^T Q x_2$$

for the first term, using $x_1(t) = x(t; x_0, 0)$, $x_2(t) = x(t; 0, u)$ and the Cauchy-Schwarz inequality we can conclude

$$\begin{aligned} \int_0^{t_1} x(t; x_0, u)^T Q x(t; x_0, u) dt &\ge \int_0^{t_1} x_1(t)^T Q x_1(t) + \int_0^{t_1} 2x_1(t)^T Q x_2(t) dt \\ &\ge c_0 - 2\|N\| \sqrt{\int_0^{t_1} \|x_1(t)\|^2 dt} \sqrt{\int_0^{t_1} \|x_2(t)\|^2 dt} \\ &\ge c_0 - 2\|N\| c_3 \sqrt{t_1 c_2^2 \varepsilon} \;\; =: \;\; c_0 - c_6\sqrt{\varepsilon}. \end{aligned}$$

Together this yields

$$J(0,t_1,x_0,u) \geq c_0 - c_7\sqrt{\varepsilon}$$

with $c_7 := c_5 + c_6$. Setting $\varepsilon = c_0^2/(2c_7)^2$ (implying $c_7\sqrt{\varepsilon} = c_0/2$), in the case (6.19) we finally arrive at

$$J(0,t_1,x_0,u) \geq c_0/2.$$

Together with the inequality (6.18) for the case (6.17) we thus obtain

$$J(0,t_1,x_0,u) \geq \max\{c_0/2,\ k_1k_2c_0^2/(4c_7)^2\} =: c$$

and consequently (6.15). □

Now we can clarify the items (i) and (iii) in der list above. We first look at item (i),i.e. we generalise Lemma 6.9 to the new cost function (6.14).

Lemma 6.17 Let the pair (A,C) be observable. Then the linear-quadratic problem with g from (6.14) is null-controlling.

Proof: We prove

$$x(t;x_0,u) \not\to 0 \quad\Rightarrow\quad J(x_0,u) = \infty.$$

Thus, consider a solution with $x(t;x_0,u) \not\to 0$. Then there exists a sequence of times $t_k \to \infty$ and an $\varepsilon > 0$ with $\|x(t_k;x_0,u)\| \geq \varepsilon$. Without loss of generality we may assume $t_{k+1} - t_k \geq 1$. Using Lemma 6.16, $x_k = x(t_k;x_0,u)$ and $u_k(\cdot) = u(t_k + \cdot)$ it then follows that

$$\int_{t_k}^{t_k+1} g(x(t;x_0,u),u(t))dt = \int_0^1 g(x(t;x_k,u_k),u_k(t))dt = J(0,1,x_k,u_k) \geq c\varepsilon^2.$$

This implies

$$\begin{aligned} J(x_0,u) &= \int_0^\infty g(x(t;x_0,u),u(t))dt \\ &\geq \sum_{k=1}^\infty \int_{t_k}^{t_k+1} g(x(t;x_0,u),u(t))dt \ \geq\ \sum_{k=1}^\infty \varepsilon^2 \ =\ \infty. \end{aligned}$$

□

It remains to address item (iii), i.e. to show that the proof of "(i)⇒(ii)" in Theorem 6.13 also works for g from (6.14). This is shown by the following theorem

Theorem 6.18 Let the pair (A,C) be observable. Then Theorem 6.13 also holds for the linear-quadratic problem with g from (6.14).

Proof: Using Lemma 6.17 in place of Lemma 6.9 we obtain all parts of the proof analogously to Theorem 6.13, except for "(i)⇒(ii)".

For proving "(i)⇒(ii)" the positive definiteness of G is used only in one place, i.e. for proving that

$$W(0,t_1,x_0) = \int_0^{t_1} g(x(t,x_0,u^*),u^*(t))dt$$

in equation (6.12) is positive for all $x_0 \neq 0$. This, however, follows from Lemma 6.16 also for g from (6.14) provided (A, C) is observable. Thus, the proof remains valid and the assertion follows. ∎

Remark 6.19 The corresponding Riccati equation reads

$$PA + A^T P + C^T \widetilde{Q} C - (PB + C^T \widetilde{N}) R^{-1} (B^T P + \widetilde{N}^T C)$$

and the optimal feedback law is given by

$$F = -R^{-1}(B^T P + \widetilde{N}^T C).$$

Observe that both $V(x) = x^T P x$ and Fx are in general *not* of the form $y^T \widetilde{P} y$ or $\widetilde{F} y$. In order to implement F for a control system of the form (4.1) in dependence of y, we still need an observer. □

Chapter 7

The Kalman Filter

Already in Chapter 4 we saw a possibility to compute the state $x(t)$ of a control system from the measured output $y(t) = Cx(t)$ via the dynamic observer $z(t)$. The focus of this considerations in this chapter was mainly asymptotic stability of the closed loop system, rather than the quality of the approximation $z(t) \approx x(t)$.

By means of the linear-quadratic optimal control from the last section we now want to develop a method that yields an — in an appropriate sense — optimal state estimation $z(t) \approx x(t)$.

The solution of this linear-quadratic state estimation problem is given by the so-called Kalman Filter (also called LQ estimator). This filter is nowadays contained in numerous technical devices, from the radar device via satellites to the smartphone. Here we consider the deterministic, continuous-time version of the Kalman filter on infinite time horizons, which builds on the results from the last chapter.

7.1 State estimation on infinite time horizon

We first consider the following, slightly differently formulated problem: Assume we are given a control system (4.1) with the modified notation $B = D$ and $u = v$, i.e.,

$$\dot{x}(t) = Ax(t) + Dv(t), \qquad y(t) = Cx(t), \tag{7.1}$$

where (A, C) is observable.

Let, moreover, $y_m : \mathbb{R} \to \mathbb{R}^l$ be a given function. The goal now is to use the solutions of (7.1) in order to find a constructively computable function $x^*(t)$, such that $y(t) = Cx^*(t)$ approximates the function $y_m(t)$. The interpretation of this problem setting is that $y_m(t) = Cx_m(t)$ represents measured output values generated by the solution x_m of a differential equation $\dot{x}_m = Ax_m$ with identical system matrix A as in (7.1). From these values the solution the state $x_m(t)$ shall be estimated as good as possible. We will explain in the next section, how this setting can be extended if $\dot{x}_m$ is generated by a control system that also includes a control function u.

The Kalman Filter, which we derive in the following, solves this problem optimally in the sense of an "indirect" least-squares approximation that proceeds in two steps.

In the *first step* we choose spd matrices $\widetilde{Q}$ and R of suitable dimension and compute for every $\tau \geq 0$ and every initial value x_0 with initial time $t_0 = \tau$ a control function $v : (-\infty, \tau] \to \mathbb{R}^n$, such that the solution $x_\tau(t) = x(t; \tau, x_0, v)$ minimises the functional

$$J_\tau(x_0, v) := \int_{-\infty}^{\tau} (Cx_\tau(t) - y_m(t))^T \widetilde{Q}(Cx_\tau(t) - y_m(t)) + v(t)^T Rv(t)dt. \tag{7.2}$$

We assume that the corresponding optimal value function

$$W_\tau(x_0) := \inf_{v \in \mathcal{U}} J_\tau(x_0, v)$$

is finite. In the *second step* we then choose $x^*(\tau)$ such that $W_\tau(x^*(\tau))$ becomes minimal, i.e. such that

$$W_\tau(x^*(\tau)) = \min_{x_0 \in \mathbb{R}^n} W_\tau(x_0)$$

holds.

This approach may at the first glance appear a little cumbersome. It leads, however, to a solution that is very easy to implement and that we will derive now.

First of all we transform the time in such a way, that the integration (7.2) is carried out from 0 to ∞, as in the linear quadratich problem from the last section.

To this end we set $x^\tau(t; x_0, v) := x(\tau - t; x_0, v)$ and $y_m^\tau(t) = y_m(\tau - t)$. Then, using the abbreviation $x^\tau(t) = x^\tau(t; x_0, v)$, for

$$J_\tau^-(x_0, v) := \int_0^\infty (Cx^\tau(t) - y_m^\tau(t))^T \widetilde{Q}(Cx^\tau(t) - y_m^\tau(t)) + v(t)^T Rv(t)dt \tag{7.3}$$

we obtain the identity $J_\tau^-(x_0, v) = J_\tau(x_0, v(\tau - \cdot))$. This implies in particular

$$W_\tau^-(x_0) := \inf_{v \in \mathcal{U}} J_\tau^-(x_0, v) = W_\tau(x_0).$$

Observe that $x^\tau(t; x_0, v)$ solves the control system

$$\dot{x}^\tau(t) = -Ax^\tau(t) - Dv(\tau - t).$$

Using a second transformation we can then bring (7.3) (almost) into the form of the known linear-quadratic output-regulation problem from definition 6.1 with g from (6.14):

For this purpose we extend the state $x \in \mathbb{R}^n$ of the system by a component $x_{n+1}(t) \equiv const$, i.e., $\dot{x}_{n+1}(t) \equiv 0$. This is achieved by setting

$$\bar{x} := \begin{pmatrix} x \\ x_{n+1} \end{pmatrix}, \quad \overline{A} := \begin{pmatrix} -A & 0 \\ 0 & 0 \end{pmatrix} \text{ und } \overline{D} := \begin{pmatrix} -D \\ 0 \end{pmatrix}.$$

If we define

$$\overline{Q}_\tau(t) := \begin{pmatrix} C^T\widetilde{Q}C & -C^T\widetilde{Q}y_m^\tau(t) \\ -y_m^\tau(t)^T\widetilde{Q}C & y_m^\tau(t)^T\widetilde{Q}y_m^\tau(t) \end{pmatrix}$$

and $g(t,\bar{x},v) := \bar{x}^T\overline{Q}_\tau(t)\bar{x} + v^TRv$, then for $\bar{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$ it follows that

$$g(t,\bar{x},v) = (Cx - y_m^\tau(t))^T\widetilde{Q}(Cx - y_m^\tau(t)) + v(t)^TRv(t)dt.$$

Consequently for $\bar{x}_0 = \begin{pmatrix} x_0 \\ 1 \end{pmatrix}$ and $\bar{x}^\tau(t,\bar{x}_0,v) = \begin{pmatrix} x^\tau(t,x_0,v) \\ 1 \end{pmatrix}$ we obtain

$$J_\tau^-(x_0,v) = \int_0^\infty g(t,\bar{x}^\tau(t;\bar{x}_0,v),v(t))dt =: \overline{J}_\tau(\bar{x}_0,v).$$

As usual, with $\overline{W}_\tau$ we denote the optimal value function. The problem is now of the usual linear-quadratic form with the exception that g depends explicitly on time. Yet, the equations that were used in the proof of Theorem 6.13 are still valid, if the time dependence is $\overline{M}(t)$ is taken care of appropriately. More precisely (for sake of brevity we state this without a proof), the following holds.

Consider for $t \in [0,\sigma]$ the solution of the Riccati differential equation

$$\dot{\overline{P}}_{\tau,\sigma}(t) = \overline{P}_{\tau,\sigma}(t)\overline{A} + \overline{A}^T\overline{P}_\tau(t) + \overline{Q}_\tau(\sigma - t) - \overline{P}_{\tau,\sigma}(t)\overline{D}R^{-1}\overline{D}^T\overline{P}_{\tau,\sigma}(t) \qquad (7.4)$$

with initial condition $\overline{P}_{\tau,\sigma}(0) = 0$. Then the convergence

$$\overline{W}_\tau(\bar{x}) := \lim_{\sigma\to\infty} \bar{x}^T\overline{P}_{\tau,\sigma}(\sigma)\bar{x}$$

holds. Now we decompose $\overline{W}_{\tau,\sigma}(t)$ according to the definition of $\overline{A}$: Writing

$$\overline{P}_{\tau,\sigma}(t) = \begin{pmatrix} P_{\tau,\sigma}(t) & p_{\tau,\sigma}(t) \\ p_{\tau,\sigma}(t)^T & \alpha_{\tau,\sigma}(t) \end{pmatrix},$$

the shape of the matrices $\overline{A}$ and $\overline{D}$ implies that $P_{\tau,\sigma}(t)$ solves the equation

$$\dot{P}_{\tau,\sigma}(t) = -P_{\tau,\sigma}(t)A - A^TP_{\tau,\sigma}(t) + C^T\widetilde{Q}C - P_{\tau,\sigma}(t)DR^{-1}D^TP_{\tau,\sigma}(t).$$

This, however, is exactly the Riccati differential equation from the proof of Theorem 6.13. Moreover all the data and thus also $P_{\tau,\sigma}(t) = P(t)$ are independent of τ and σ. We can thus conclude

$$\lim_{\sigma\to\infty} P(\sigma) = P,$$

where P solves the algebraic Riccati equation

$$-PA - A^TP + C^T\widetilde{Q}C - PDR^{-1}D^TP = 0. \qquad (7.5)$$

Thus, with $\bar{x}_0^T = (x_0^T, 1)$ and $p_\tau = \lim_{\sigma\to\infty} p_{\tau,\sigma}(\sigma)$, $\alpha_\tau = \lim_{\sigma\to\infty}\alpha_{\tau,\sigma}(\sigma)$ we get

$$W_\tau(x_0) = \overline{W}_\tau(\bar{x}_0) = \lim_{\sigma\to\infty} \bar{x}_0^T\overline{P}_{\tau,\sigma}(\sigma)\bar{x}_0 = x_0^TPx_0 + 2x_0^Tp_\tau + \alpha_\tau.$$

The value $x^*(\tau)$ that we looked for in the second step of the method is thus (by derivating this expression and solving for x_0) given by

$$x^*(\tau) = -P^{-1}p_\tau = -Sp_\tau$$

for $S := P^{-1}$. By multiplication of (7.5) with S from the left and the right and with -1 it follows that S solves the so-called *dual Riccati equation*

$$AS + SA^T - SC^T\widetilde{Q}CS + DR^{-1}D^T = 0. \tag{7.6}$$

It remains to compute p_τ. From the Riccati differential equation (7.4) for $p_{\tau,\sigma}(t)$ we can deduce the differential equation

$$\dot{p}_{\tau,\sigma}(t) = -A^T p_{\tau,\sigma}(t) - P(t)DR^{-1}D^T p_{\tau,\sigma}(t) - C^T\widetilde{Q}y_m(\tau - \sigma + t)$$

with initial condition $q_{\tau,\sigma}(0) = 0$. This implies

$$\dot{p}_{\tau+s,\sigma+s}(t) = \dot{p}_{\tau,\sigma}(t)$$

and since these two solutions coincide in $t = 0$, they must coincide everywhere, i.e.,

$$p_{\tau+s,\sigma+s}(t) = p_{\tau,\sigma}(t).$$

hence we get

$$\begin{aligned} \left.\frac{d}{ds}\right|_{s=0} p_{\tau+s,\sigma+s}(\sigma+s) &= \dot{p}_{\tau,\sigma}(\sigma) \\ &= -A^T p_{\tau,\sigma}(\sigma) - P(\sigma)DR^{-1}D^T p_{\tau,\sigma}(\sigma) - C^T\widetilde{Q}y_m(\tau) \end{aligned}$$

and consequently with $\sigma \to \infty$

$$\frac{d}{d\tau}p_\tau = -A^T p_\tau - PDR^{-1}D^T p_\tau - C^T\widetilde{Q}y_m(\tau).$$

Finally, with (7.6) we obtain

$$\begin{aligned} \dot{x}^*(\tau) &= -S\frac{d}{d\tau}p_\tau \\ &= SA^T p_\tau + DR^{-1}D^T p_\tau + SC^T\widetilde{Q}y_m(\tau) \\ &= -SA^T S^{-1}x^*(\tau) - DR^{-1}D^T S^{-1}x^*(\tau) + SC^T\widetilde{Q}y_m(\tau) \\ &= (-SA^T - DR^{-1}D^T)S^{-1}x^*(\tau) + SC^T\widetilde{Q}y_m(\tau) \\ &= (AS - SC^T\widetilde{Q}CS)S^{-1}x^*(\tau) + SC^T\widetilde{Q}y_m(\tau) \\ &= Ax^*(\tau) - SC^T\widetilde{Q}(Cx^*(\tau) - y_m(\tau)) \\ &= Ax^*(\tau) + L(Cx^*(\tau) - y_m(\tau)) \end{aligned}$$

with $L = -SC^T\widetilde{Q}$.

This differential equation is the so-called *Kalman filter.* Its application works as follows: When $x^*(t)$ is known, then $x^*(s)$ for $s > t$ can be computed by solving the differential equation on the interval $[t, s]$ (analytically or numerically) from the data $y_m|_{[t,s]}$. The Kalman filter can thus be evaluated online in a recursive fashion.

Two properties of the Kalman filter are worth to be noted explicitly:

(i) The matrix L does not depend on y_m. For computing L one only needs to solve one of the two Riccati equations (7.5) or (7.6).

(ii) The matrix $A + LC$ is Hurwitz. This is because L^T is the LQ-optimal feedback law of the optimal control problem that corresponds to the dual Riccati equation (7.6). Thus, $A^T + C^T L^T$ is Hurwitz and consequently also $A + LC = (A^T + C^T L^T)^T$, as these matrices have identical eigenvalues.

7.2 The Kalman filter as observer

We now show how the Kalman filter can be used as an observer for the state of a general control system (4.1), i.e.,

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

with (A, C) being controllable. We assume that the initial value x_0 is unknown while the control function $u(t)$, $t \geq 0$, the output values $y(t) = Cx(t; x_0, u)$, $t \geq 0$, and an estimation z_0 of the initial value $x_0 are known.$ We now look for the curve $z(t)$, $t \geq 0$, with $z(0) = z_0$ in $\mathbb{R}^n$, such that the estimation error $Cz(t) - y(t)$ becomes as small as possible in an appropriate sense and such that $z(t)$ only depends on $y|_{[0,t]}$ (i.e. it can be computed from the data that is known at time t). The output $y(t)$ thus plays the role of the measurement $y_m(t)$ in the Kalman filter.

For solving the problem we make the ansatz

$$\dot{z}(t) = Az(t) + Bu(t) + v(t), \tag{7.7}$$

where now $v : \mathbb{R} \to \mathbb{R}^n$ shal be determined in such a way that $z(t)$ becomes a good estimate. In order to eliminate the term $Bu(t)$ from the equation, we define the *estimation error* $e(t) := z(t) - x(t)$. This satisfies the equation

$$\dot{e}(t) = Ae(t) + v(t), \tag{7.8}$$

i.e. it is determined by a control system (7.1) with $D = \mathrm{Id}$ and $x = e$. The error e hence plays the role of x in (7.1).

We now need to compute the counterpart of the measurement y_m for the e-system. We denote this quantity by e_m. In Section 7.1 we have minimized (in the notation of this section) the quantity $Ce(t) - e_m(t)$, here we want to minimize the quantity $Cz(t) - y(t)$. Hence, we require

$$Ce(t) - e_m(t) = Cz(t) - y(t) \tag{7.9}$$

which leads to

$$e_m(t) = y(t) + Ce(t) - Cz(t) = y(t) + Cz(t) - \underbrace{Cx(t)}_{=y(t)} - Cz(t) = 0.$$

The measurement values for the e-equation are thus constantly equal to zero. This is indeed reasonable, since the measured quantity $y_m(t) = y(t) = Cx(t)$ has already been incorporated into the equation for e via the definition of e.

If we now compute the feedback law L for the Kalman filter for (7.8) according to the previous section, because of $e_m \equiv 0$ the filter equation becomes

$$\dot{e}^*(t) = (A + LC)e^*(t).$$

This is equivalent to

$$\dot{z}(t) = Az(t) + Bu(t) + L(Cz(t) - y(t)) \tag{7.10}$$

and thus yields an online implementable observer equation (note the structural similarity with the dynamic observer from Chapter 4) for computing $z(t)$, which can be solved analytically or numerically. Observe that the optimal estimate e^* for e and the estimator z for x, respectively, are connected via $e^*(t) = z(t) - x(t)$. Thus, while we have derived the Kalman observer by applying the Kalman filter to the equation for e (7.8), for computing the estimate $z(t)$ we use the z-equation (7.10), because otherwise we would need the unknown state $x(t)$ in order to compute $z(t)$ from $e^*(t)$.

Since we do not have measurement values $y(t)$ for $t < 0$, we cannot compute the optimal initial value $e^*(0)$ as in the previous section. Moreover, even if we could compute it, it would not be of much use, since for (7.10) we would have to use the initial value $z(0) = e^*(0)+x_0$ — but x_0 is unknown. We thus use the estimate $z_0 \approx x_0$ as initial value in (7.10). Since $A-LC$ is Hurwitz, the estimation error $e^*(t)$ converges to 0 for $t \to \infty$, i.e. the approximation $z(t) \approx x(t)$ becomes better and better with increasing t. Since our approach is based on an LQ optimal control problem, we would, however, expect not only convergence but also that the estimate $z(t)$ starting in $z(0) = z_0$ is optimal in a suitable sense.

In order to see in which sense optimality holds, we extend $y(t)$ for $t < 0$ in such a way that $e^*(0) = z_0 - x_0$ and thus $z(0) = z_0$ becomes the solution of the Kalman filter. In other words, we generate "artificial" measurements for which the Kalman filter yields exactly the estimate z_0 at time $t = 0$. This is precisely the case if we synthese $y(t)$ via

$$y(t) = \begin{cases} Cx(t; z_0, 0), & t < 0 \\ Cx(t; x_0, u), & t \geq 0 \end{cases} \tag{7.11}$$

from the forward solution of (4.1) for x_0 and u and the backward solution for z_0 and $u \equiv 0$. For $v \equiv 0$ and $e(0) = 0$, from $e_m \equiv 0$ we then obtain

$$Ce(t) - e_m = 0$$

for all $t < 0$. This yields $J_0(0, 0) = 0$ for the objective (7.2), hence also $W_0(0) = 0$ and thus $e^*(0) = 0$. Consequently, it follows that $z(0) = z_0 - e^*(0) = z_0$.

The estimated value $z(t)$ computed using the initial estimation z_0 is thus exactly the terminal value of the solution of (7.7) that approximates the curve (7.11) in the best possible way in the sense of (7.2).

An important property of the Kalman filter is that is also yields good estimates in case of imprecise data $\tilde{y}(t) \approx y(t)$. This can be proved rigorously with stochastic methods.

The Kalman filter also exists in discrete time. In this case the differential equation (7.10) becomes a difference equation

$$z(k+1) = Az(k) + Bu(k) + L(Cz(k) - y(k)).$$

Since this is easier to implement than the differential equation (7.10) (which has to be solved numerically or analytically before the solution can be evaluated) and moreover requires only discrete-time measurements $y(k)$ (which are easier to acquire in practice than continuous-time measurements $y(t)$), in practical applications the discrete-time Kalman filter is often preferred.

Chapter 8

Nonlinear control systems

In this and in the subsequent chapters we will consider nonlinear control systems in continuous time

$$\dot{x}(t) = f(x(t), u(t)) \tag{8.1}$$

and, respectively, in discrete time

$$x(k+1) = f(x(k), u(k)), \tag{8.2}$$

written briefly as $x^+ = f(x, u)$. An example for a nonlinear control system in continuous time is the nonlinear pendulum that we already introduced in (1.5). While for continuous-time systems we chose the state and control space as $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, for discrete-time systems it does not significantly complicate the setting to allow for arbitrary metric spaces X and U for state and control.

In the following sections we briefly summarise some foundations about the solutions of such systems.

8.1 Continuous-time systems

In continuous time we consider control functions with values in $U \subset \mathbb{R}^m$. The function $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ then is a parameter dependent continuous vector field. The space of control functions is again denoted as $\mathcal{U}$, but we will allow for a larger space than in the previous sections. More precisely, we use control functions from $L^\infty(\mathbb{R}, U)$. Existence and uniqueness is then delivered by the following Theorem of Carathéodory.

Theorem 8.1 (Theorem of Carathéodory) Consider a control system with the following properties:

i) The space of control functions is given by

$$\mathcal{U} = L^\infty(\mathbb{R}, U) := \{u : \mathbb{R} \to U \,|\, u \text{ is measurable and essentially bounded}^{1}\}.$$

[1] i.e., bounded outside a set of Lebesgue measure 0

ii) The vector field $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ is continuous.

iii) For any $R > 0$ there exists a constant $L_R > 0$, such that the estimate

$$\|f(x_1, u) - f(x_2, u)\| \leq L_R \|x_1 - x_2\|$$

holds for all $x_1, x_2 \in \mathbb{R}^n$ and all $u \in U$ with $\|x_1\|, \|x_2\|, \|u\| \leq R$.

Then for any initial value $x_0 \in \mathbb{R}^n$, any initial time $t_0 \in \mathbb{R}$ and any control function $u \in \mathcal{U}$ there exists a (maximal) open interval I with $t_0 \in I$ and a unique absolutely continuous[2] function $x(t)$, which solves the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f(x(\tau), u(\tau))\, d\tau$$

for all $t \in I$.

Definition 8.2 We denote the unique function $x(t)$ from Theorem 8.1 with $x_u(t; t_0, x_0)$ and call it the *solution* of (8.1) with *initial value* $x_0 \in \mathbb{R}^n$ and *control function* $u \in \mathcal{U}$. In case $t_0 = 0$ we briefly write $x_u(t, x_0) = x_u(t; 0, x_0)$. □

The following observation justifies this definition: Since $x_u(t, x_0)$ is absolutely continuous, it is differentiable with respect to t for almost all $t \in I$. In particular, Theorem 8.1 and the fundamental theorem of calculus imply that $x_u(t, x_0)$ satisfies the differential equation (8.1) for almost all $t \in I$, i.e.

$$\dot{x}(t, x_0, u) = f(x(t, x_0, u), u(t))$$

holds for almost all $t \in I$.

Remark 8.3 In the following we always suppose that the assumptions (i)–(iii) of Theorem 8.1 are satisfied, but we will only mention this explicitly in important theorems. □

The proof of Theorem 8.1 (which we omit for sake of brevity) is similar to the proof of the respective theorem for continuous ordinary differential equations. It uses Banach's Fixed Point Theorem applied on a suitable function space. Together with an introduction into the necessary foundations of Lebesgue measure theory it can be found, e.g., in the book *Mathematical Control Theory* by E.D. Sontag [19, Appendix C].

Just as for continuous differential equations, uniqueness of solutions implies for all $t, s \in \mathbb{R}$ the relations

$$x_u(t; t_0, x_0) = x_u(t; s, x_u(s; t_0, x_0)) \tag{8.3}$$

(the so-called cocycle property) and

$$x_u(t; t_0, x_0) = x_{u(s+\cdot)}(t - s; t_0 - s, x_0),$$

which we already formulated for linear systems in Corollary 1.10. Setting $s = t_0$, the second equation in particular implies

$$x_u(t; t_0, x_0) = x_{u(t_0+\cdot)}(t - t_0, x_0). \tag{8.4}$$

[2] A function is called absolutely continuous if it can be written as an integral of an L^∞ function.

8.2 Sampled-data systems

As already mentioned in the first chapter, to every continuous time control system that satisfies the assumptions of Carathéodory's Theorem we can assign a corresponding discrete-time system, the so-called sampled-data system. This is obtained simply by looking at the state of the continuous-time system only at times kT, for $k \in \mathbb{N}$ and a fixed sampling time[3] $T > 0$. If we denote the continuous-time solution by $\hat{x}_{\hat{u}}(t, x_0)$, then the states $x(k)$ of the sampled-data system are given by

$$x(k) = \hat{x}_{\hat{u}}(kT, x_0).$$

Using (8.3) and (8.4) it follows that

$$x(k+1) = \hat{x}_{\hat{u}}((k+1)T; kT, \hat{x}_{\hat{u}}(kT, x_0)) = \hat{x}_{\hat{u}}((k+1)T; kT, x(k)) = \hat{x}_{\hat{u}(kT+\cdot)}(T, x(k)).$$

If for the control function $\hat{u}(\cdot)$ we define the functions $u(k) : [0, T] \to \mathbb{R}$ via

$$u(k)(t) := \hat{u}(kT + t), \ \ t \in [0, T]$$

then we obtain

$$x(k+1) = \hat{x}_{u(k)}(T, x(k)) =: f(x(k), u(k)), \tag{8.5}$$

which defines the discrete-time *sampled-data system*. In general, here the functions $u(k)$ satisfy $u(k) \in L^\infty([0, T], U)$. Yet, as already mentioned in Chapter 1, it is possible (and common engineering practice) to choose $u(k)$ from a smaller set of functions. A very common choice is to define $u(k)$ as a constant function. The corresponding continuous-time control function $\hat{u}$ is then piecewise constant. Sometimes the functions $u(k)$ are chosen as polynomials. In this case $\hat{u}$ is a piecewise polynomial function (but in general discontinuous at the boundary points kT of the sampling intervale).

In the remainder of this course we will work with discrete-time control systems, since for this class of systems model predictive control, which is the method on which we focus in the sequel, is easier to formulate and to analyse. Nevertheless, we will mention the particularities of sampled-data systems whenever appropriate.

[3] German: sampling = Abtastung, sampling time = Abtastzeit, sampled-data system = Abtastsystem

Chapter 9

Introduction to Model Predictive Control

In this introduction, we present the basics of Model Predictive Control (henceforth abbreviated as MPC) in an informal way. In particular, we introduce the central idea of iterative optimal control on a moving finite horizon.

MPC is a method for obtaining an approximately optimal feedback control for an optimal control problem on an infinite or indefinite time horizon. Feedback here means that the control at time k is of the form $u(k) = \mu(x(k))$ for a map $\mu : X \to U$. We have already seen how linear quadratic optimal control leads to an optimal feedback control. The decisive property that makes the approach via the Riccati equation computationally feasible is that the optimal value function V is of quadratic form $V(x) = x^T Px$. This means that we only have to determine the coefficients of the matrix P, whose number is of the order $O(n^2)$. However, as soon as the cost is nonquadratic, the dynamics is nonlinear or state and/or control constraints are introduced into the problem, the function V is no longer quadratic. This means that an exact representation by finitely many coefficients is in general no longer possible. The same holds for the optimal feedback law, which is in general a rather complicated function in x for which already the storage poses challenging problems, known as the "curse of dimensionality". This implies that the direct computation and storage of an approximately optimal feedback law is computationally intractable even for problems in moderate space dimensions, say 5–10.

In contrast to this, nowadays there exist powerful optimization algorithms which can compute single optimal trajectories in very short time, even for high dimensional systems like accurately discretized PDEs. The key idea of MPC is now to use this computational approach for obtaining a feedback law which is near optimal for infinite horizon problems.

In order to describe the idea of MPC, consider the discrete time model

$$x^+ = f(x, u) \tag{9.1}$$

where $f : X \times U \to X$ is a known and in general nonlinear map which assigns to a state x and a control value u the successor state x^+ at the next time instant and X and U are metric spaces. Starting from the current state $x(j)$, for any given control sequence

$u(0), \ldots, u(N-1)$ with *horizon length* $N \geq 2$, we can now iterate (9.1) in order to construct a prediction trajectory x_u defined by

$$x_u(0) = x(j), \quad x_u(k+1) = f(x_u(k), u(k)), \;\; k = 0, \ldots, N-1. \tag{9.2}$$

Proceeding this way, we obtain predictions $x_u(k)$ for the state of the system $x(j+k)$ for k time steps into the future, depending on the chosen control sequence $u(0), \ldots, u(N-1)$.

Now we use optimal control in order to determine $u(0), \ldots, u(N-1)$. To this end, we fix a cost function $\ell(x, u)$. This function may be very general. In the simplest case, X and U are vector spaces with norms and ℓ penalizes the distance of x to some "reference state" x_*; for simplicity we assume $x_* = 0$. Typically, one does not penalize the deviation of the state from the reference but also—if desired—the distance of the control values $u(k)$ to a reference control u_*, which here we also choose as $u_* = 0$. A common and popular choice for such a function is the quadratic function

$$\ell(x_u(k), u(k)) = \|x_u(k)\|^2 + \lambda \|u(k)\|^2,$$

where $\|\cdot\|$ denotes the norms[1] of the spaces X and U and $\lambda \geq 0$ is a weighting parameter for the control, which could also be chosen as 0 if no control penalization is desired. The purpose of MPC with a stage cost penalizing the distance to an equilibrium is that the optimal control should drive the system towards the reference state $x_* = 0$, in order to stabilize the system at this state, just as in the linear quadratic case. MPC with such stage costs is thus called *stabilizing MPC*. In contrast to this, MPC with more general cost function is often called *economic MPC*.

Regardless which cost function is used, the optimal control problem now reads

$$\text{minimize} \quad J_N(x(j), u(\cdot)) := \sum_{k=0}^{N-1} \ell(x_u(k), u(k))$$

with respect to all admissible[2] control sequences $u(0), \ldots, u(N-1)$ with x_u generated by (9.2).

Let us assume that this optimal control problem has a solution which is given by the minimizing control sequence $u^\star(0), \ldots, u^\star(N-1)$, i.e.,

$$\min_{u(0), \ldots, u(N-1)} J_N(x(j), u(\cdot)) = \sum_{k=0}^{N-1} \ell(x_{u^\star}(k), u^\star(k)).$$

In order to get the desired feedback value $\mu(x(j))$, we now set $\mu(x(j)) := u^\star(0)$, i.e., we apply the first element of the optimal control sequence. This procedure is sketched in Fig. 9.1.

We now apply this feedback law, i.e., the first element of $u^\star$, on the time interval from j to $j+1$. Thus we obtain

$$x(j+1) = f(x(j), \mu(x(j))) \tag{9.3}$$

[1] For simplicity of notation we use the same symbol for the in gereral different norms on X and U.

[2] The meaning of "admissible" will be defined in Sect. 11.2.

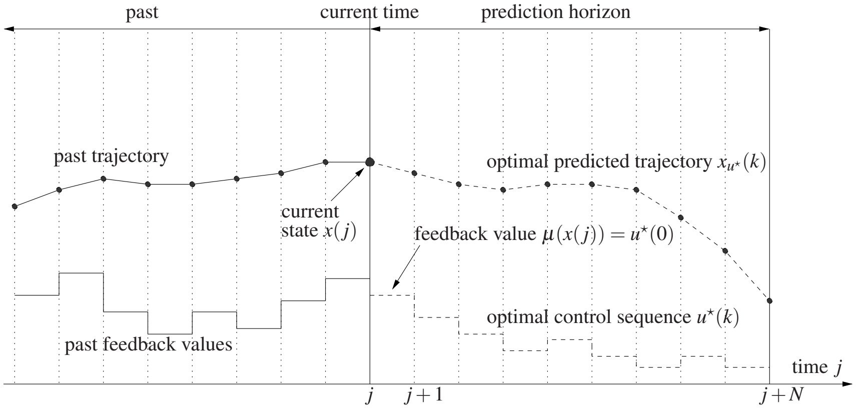

past
current time
prediction horizon
past trajectory
optimal predicted trajectory $x_{u^\star}(k)$
current state $x(j)$
feedback value $\mu(x(j)) = u^\star(0)$
optimal control sequence $u^\star(k)$
past feedback values
time j
j
$j+1$
$j+N$

Figure 9.1: Illustration of the MPC step at time j

System (9.3) is called the *MPC closed-loop system.*

At the following time instants $j+1, j+2, \ldots$ we repeat the procedure with the new measurements $x(j+1), x(j+2), \ldots$ in order to derive the feedback values $\mu(x(j+1)), \mu(x(j+2)), \ldots$. In other words, we obtain the feedback law μ by an *iterative online optimization* over the predictions generated by our model (9.1). This is the first key feature of model predictive control.

From the prediction horizon point of view, proceeding this iterative way the trajectories $x_u(k)$, $k = 0, \ldots, N$ provide a prediction on the discrete interval $j, \ldots, j+N$ at time j, on the interval $j+1, \ldots, j+N+1$ at time $j+1$, on the interval $j+2, \ldots, j+N+2$ at time $j+2$, and so on. Hence, the prediction horizon is moving and this *moving horizon* is the second key feature of model predictive control.

Regarding terminology, another term which is often used alternatively to *model predictive control* is *receding horizon control.* While the former expression stresses the use of model based predictions, the latter emphasizes the moving horizon idea. Despite these slightly different literal meanings, we prefer and follow the common practice to use these names synonymously. In addition, one often uses the term Nonlinear Model Predictive Control (NMPC) if one wants to indicate that our model (9.1) need not be a linear map.

9.1 Motivating examples

In this section we present three motivating examples (the corresponding numerical simulations and experiments will only be presented in the lectures), which show different phenomema which can be observed when using MPC.

The first example is the classical inverted pendulum, which is available as a real experiment at the Chair of Applied Mathematics. The cost function ℓ here penalizes the distance to the upright equilibrium. The ordinary differential equation system (which is similar to (1.5) but a little more complex in order to take into acount the motor dynamics) is sampled with sampling time $T = 50$ms. The video shows that this time is enough to solve the optimal control problem numerically in each sampling interval[3].

The second example is a very simple economic problem of optimal investment. Let $x \geq 0$ be the amount of capital invested in a company. The invested capital x yields a return of $Ax^\alpha - x$ in one time unit (e.g., a year), i.e., after one time step the amount of capital is Ax^α. The control u describes the amount of capital which is invested again in the next time step. Hence, the amount of money to be consumed is $Ax^\alpha - u$. The utility of consumption is measured by a classical logarithmic utility function $\ln(Ax^\alpha - u)$. We want to maximize this utility over several time steps, hence we want to minimize the cost function $\ell(x,u) = -\ln(Ax^\alpha - u)$. We note that this cost function is not of the form of a function which penalizes the distance from a reference point x_*. Numerical simultions for $A = 5$ and $\alpha = 0.34$ and state constraint set $\mathbb{X} = [0, 10]$ show that the finite horizon optimal solutions always end up at $x = 0$, i.e., at the end of the optimization horizon all money is spent (which is natural). However, for longer horizons the solutions spend quite some time in the vicinity of the point $x^e \approx 2.2344$ and the MPC closed-loop (9.3) converges to an equilibrium near this point. Further tests reveal that the limit point of the MPC closed-loop itself converges as $N \to \infty$.

There are many questions which arise from this behaviour: Why does the MPC closed-loop converge to a point far away from the endpoint of the finite horizon optimal trajectories? How do we characterize this point and its limit for $N \to \infty$? Is the MPC closed-loop trajectory approximately optimal in some sense? And how can we check whether an optimal control problem has such a behavior?

The third example is a simple partial differential equation control system governed by the 1d heat equation on $\Omega = (0, L)$. We consider the equation either with distributed control

$$\begin{array}{rcll} y_t(x,t) & = & y_{xx}(x,t) + \mu y(x,t) + \hat{u}(x,t) & \text{on } \Omega \times (0,\infty) \\ y(0,t) & = & y(L,t) \; = \; 0 & \text{on } (0,\infty) \\ y(x,0) & = & y_0(x) & \text{on } \Omega \end{array}$$

or with boundary control.

$$\begin{array}{rcll} y_t(x,t) & = & y_{xx}(x,t) + \mu y(x,t) & \text{on } \Omega \times (0,\infty) \\ y(0,t) & = & 0, \;\; y(L,t) \; = \; \hat{u}(t) & \text{on } (0,\infty) \\ y(x,0) & = & y_0(x) & \text{on } \Omega \end{array}$$

We set $\mu = 15$, which implies that $y \equiv 0$ is an unstable equilibrium for $u \equiv 0$. In order to stabilize this equilibrium, we consider the cost functions $\ell(y,u) = \|y\|_{L^2}^2 + \lambda\|u\|^2$ ("L^2-cost") and $\ell(y,u) = \|y_x\|_{L^2}^2 + \lambda\|u\|^2$ ("∇-cost"). As usual in MPC, it depends on the length of the horizon N whether the equilibrium $y \equiv 0$ is indeed stable. The simulations — all with sampling time $T = 0.01$ — show that depending on the parameters L and λ as well as on the type of the cost the minimal horizon length needed for stabilization differs significantly. This immediately leads to the question how we can estimate this minimal horizon length and whether we can tune, e.g., the stage cost ℓ such that this horizon becomes small.

[3] In practice, the state $x(j)$ must be computed from sensor data using a suitable observed, as, e.g., the Kalman filter or variants thereof. Also, in practice the MPC problem is initialized with the state $x(j-1)$ such that the time span until time j can be fully used in order to solve the optimal control problem. Both aspects will be neglected in the analysis of MPC schemes we will present in this lecture.

As we will see later, in all these examples we can prove that MPC yields approximately optimal infinite horizon trajectories. Hence, the problem on (rather short) finite horizons already contains enough information to compute near optimal solutions on an infinite horizon, a property that can be seen as a complexity reduction technique in time. In the subsequent analysis, we will in particular investigate the mechanisms behind this complexity reduction.

Chapter 10

Stability of discrete time nonlinear systems

10.1 Stability definitions

In the introduction, we already specified one of the goals of model predictive control, namely to control the state $x(n)$ of the system toward a reference point x_* and then keep it close to this point. In this section we formalize what we mean by "toward" and "close to" using concepts from stability theory of nonlinear systems. These concepts will also turn out to be useful for the analysis of MPC schemes in which ℓ does not penalize the distance to an equilibrium x_*.

We assume that the states $x(k)$ are generated by a difference equation of the form

$$x^+ = g(x) \tag{10.1}$$

for a not necessarily continuous map $g : X \to X$ via the usual iteration $x(k+1) = g(x(k))$. Similar to before, we write $x(k, x_0)$ for the trajectory satisfying the initial condition $x(0, x_0) = x_0 \in X$. Allowing g to be discontinuous is important for our MPC application, because g will later represent the MPC closed-loop system (9.3), i.e., $g(x) = f(x, \mu(x))$. Since μ is obtained as an outcome of an optimization algorithm, in general we cannot expect μ to be continuous and thus g will in general be discontinuous, too.

Nonlinear stability properties can be expressed conveniently via so-called comparison functions which were first introduced by Hahn in 1967 [10] and popularized in nonlinear control theory during the 1990s by Sontag, particularly in the context of input-to-state stability [17]. Although we mainly deal with discrete time systems, we stick to the usual continuous time definition of these functions using the notation $\mathbb{R}_0^+ = [0, \infty)$.

Definition 10.1 [Comparison functions] We define the following classes of comparison functions.

$$\mathcal{K} := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \mid \alpha \text{ is continuous \& strictly increasing with } \alpha(0) = 0\}$$

$$\mathcal{K}_\infty := \{\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \mid \alpha \in \mathcal{K},\ \alpha \text{ is unbounded}\}$$

$$\mathcal{L} := \{\delta : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \delta \text{ is continuous \& strictly decreasing with } \lim_{t\to\infty} \delta(t) = 0\}$$

$$\mathcal{KL} := \{\beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+ \,|\, \beta \text{ is continuous}, \beta(\cdot, t) \in \mathcal{K}\, \forall t \geq 0,\ \beta(r, \cdot) \in \mathcal{L}\, \forall r > 0\}.$$

□

Using this function, we can now introduce the concept of asymptotic stability. Here, for arbitrary $x_1, x_2 \in X$ we denote the distance from x_1 to x_2 by

$$|x_1|_{x_2} := d_X(x_1, x_2).$$

Furthermore, we use the ball

$$\mathcal{B}_\eta(x_*) := \{x \in X \,|\, |x|_{x_*} < \eta\}$$

and we say that a set $Y \subseteq X$ is *forward invariant* for (10.1) if $g(x) \in Y$ holds for all $x \in Y$.

Definition 10.2 [Asymptotic stability] Let $x_* \in X$ be an equilibrium for (10.1), i.e., $g(x_*) = x_*$. Then we say that x_* is *locally asymptotically stable* if there exist $\eta > 0$ and a function $\beta \in \mathcal{KL}$ such that the inequality

$$|x(n, x_0)|_{x_*} \leq \beta(|x_0|_{x_*}, n) \tag{10.2}$$

holds for all $x_0 \in \mathcal{B}_\eta(x_*)$ and all $n \in \mathbb{N}_0$.

We say that x_* is *asymptotically stable on a forward invariant set* Y with $x_* \in Y$ if there exists $\beta \in \mathcal{KL}$ such that (10.2) holds for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$ and we say that x_* is *globally asymptotically stable* if x_* is asymptotically stable on $Y = X$.

If one of these properties holds then β is called *attraction rate*. □

Note that asymptotic stability on a forward invariant set Y implies local asymptotic stability if Y contains a ball $\mathcal{B}_\eta(x_*)$. However, we do not necessarily require this property.

Asymptotic stability thus defined consists of two main ingredients:

(i) The smaller the initial distance from x_0 to x_* is, the smaller the distance from $x(n)$ to x_* becomes for all future n, or formally: for each $\varepsilon > 0$ there exists $\delta > 0$ such that $|x(n, x_0)|_{x_*} \leq \varepsilon$ holds for all $n \in \mathbb{N}_0$ and all $x_0 \in Y$ (or $x_0 \in \mathcal{B}_\eta(x_*)$) with $|x_0|_{x_*} \leq \delta$.

This fact is easily seen by choosing δ so small that $\beta(\delta, 0) \leq \varepsilon$ holds, which is possible since $\beta(\cdot, 0) \in \mathcal{K}$. Since β is decreasing in its second argument, for $|x_0|_{x_*} \leq \delta$ from (10.2) we obtain

$$|x(n, x_0)|_{x_*} \leq \beta(|x_0|_{x_*}, n) \leq \beta(|x_0|_{x_*}, 0) \leq \beta(\delta, 0) \leq \varepsilon.$$

(ii) As the system evolves, the distance from $x(n, x_0)$ to x_* becomes arbitrarily small, or formally: for each $\varepsilon > 0$ and each $R > 0$ there exists $N > 0$ such that $|x(n, x_0)|_{x_*} \leq \varepsilon$ holds for all $n \geq N$ and all $x_0 \in Y$ (or $x_0 \in \mathcal{B}_\eta(x_*)$) with $|x_0|_{x_*} \leq R$. This property easily follows from (10.2) by choosing $N > 0$ with $\beta(R, N) \leq \varepsilon$ and exploiting the monotonicity properties of β.

These two properties are known as (i) stability (in the sense of Lyapunov) and (ii) attraction. In the literature, asymptotic stability is often defined via these two properties. In fact, for continuous time (and continuous) systems (i) and (ii) are known to be equivalent to the continuous time counterpart of Definition 10.2, cf. [13, Sect. 3]. We conjecture that the arguments in this reference can be modified in order to prove that equivalence also holds for our discontinuous discrete time setting.

Asymptotic stability includes the desired properties of the MPC closed loop described earlier: whenever we are already close to the reference equilibrium we want to stay close; otherwise we want to move toward the equilibrium.

Asymptotic stability also includes that eventually the distance of the closed-loop solution to the equilibrium x_* becomes arbitrarily small. Occasionally, this may be too demanding. For instance, we will see that in general we cannot expect this behavior for stage costs ℓ which do not penalize the distance to x_*. In this case, one can relax the asymptotic stability definition to practical asymptotic stability as follows. Here we only consider the case of asymptotic stability on a forward invariant set Y.

Definition 10.3 [P-practically asymptotic stability] Let Y be a forward invariant set and let $P \subset Y$ be a subset of Y. Then we say that a point $x_* \in Y$ is *P-practically asymptotically stable on* Y if there exists $\beta \in \mathcal{KL}$ such that (10.2) holds for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$ with $x(n, x_0) \notin P$. □

Fig. 10.1 illustrates practical asymptotic stability (on the right) as opposed to "usual" asymptotic stability (on the left).

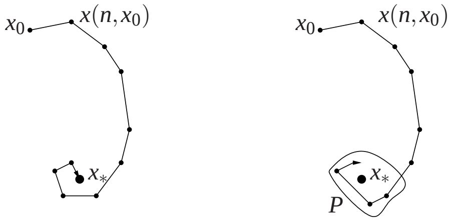


Figure 10.1: Sketch of asymptotic stability (left) as opposed to practical asymptotic stability (right)

This definition is typically used with P contained in a small ball around the equilibrium, i.e., $P \subseteq \mathcal{B}_\delta(x_*)$ for some small $\delta > 0$. In this case one obtains the estimate

$$|x(n, x_0)|_{x_*} \le \max\{\beta(|x_0|_{x_*}, n),\ \delta\} \tag{10.3}$$

for all $x_0 \in Y$ and all $n \in \mathbb{N}_0$, i.e., the system behaves like an asymptotically stable system until it reaches the ball $\mathcal{B}_\delta(x_*)$. Note that x_* does not need to be an equilibrium in Definition 10.3.

10.2 Lyapunov functions

In order to verify that our MPC controller achieves asymptotic stability we will utilize the concept of Lyapunov functions.

Definition 10.4 [Lyapunov function] Consider a system (10.1), a point $x_* \in X$ and let $S \subseteq X$ be a subset of the state space. A function $V : S \to \mathbb{R}_0^+$ is called a *Lyapunov function* on S if the following conditions are satisfied:

(i) There exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that

$$\alpha_1(|x|_{x_*}) \leq V(x) \leq \alpha_2(|x|_{x_*}) \tag{10.4}$$

holds for all $x \in S$.

(ii) There exists a function $\alpha_V \in \mathcal{K}$ such that

$$V(g(x)) \leq V(x) - \alpha_V(|x|_{x_*}) \tag{10.5}$$

holds for all $x \in S$ with $g(x) \in S$.

□

The following theorem shows that the existence of a Lyapunov function ensures asymptotic stability.

Theorem 10.5 [Asymptotic stability using Lyapunov functions] Let x_* be an equilibrium of (10.1) and assume there exists a Lyapunov function V on S. If S contains a ball $\mathcal{B}_\nu(x_*)$ with $g(x) \in S$ for all $x \in \mathcal{B}_\nu(x_*)$ then x_* is locally asymptotically stable with $\eta = \alpha_2^{-1} \circ \alpha_1(\nu)$. If $S = Y$ holds for some forward invariant set $Y \subseteq X$ containing x_* then x_* is asymptotically stable on Y. If $S = X$ holds then x_* is globally asymptotically stable.

Proof: The idea of the proof lies in showing that by (10.5) the function $V(x(n, x_0))$ is strictly decreasing in n and converges to 0. Then by (10.4) we can conclude that $x(n, x_0)$ converges to x_*. The function β from Definition 10.2 will be constructed from α_1, α_2 and α_V. In order to simplify the notation, throughout the proof we write $|x|$ instead of $|x|_{x_*}$.

First, if S is not forward invariant, define the value $\gamma := \alpha_1(\nu)$ and the set $\widetilde{S} := \{x \in S \,|\, V(x) < \gamma\}$. Then from (10.4) we get

$$x \in \widetilde{S} \Rightarrow \alpha_1(|x|) \leq V(x) < \gamma \Rightarrow |x| < \alpha_1^{-1}(\gamma) = \nu \Rightarrow x \in \mathcal{B}_\nu(x_*),$$

observing that each $\alpha \in \mathcal{K}_\infty$ is invertible with $\alpha^{-1} \in \mathcal{K}_\infty$.

Hence, for each $x \in \widetilde{S}$ inequality (10.5) applies and consequently $V(g(x)) \leq V(x) < \gamma$ implying $g(x) \in \widetilde{S}$. If $S = Y$ for some forward invariant set $Y \subseteq X$ we define $\widetilde{S} := S$. With these definitions, in both cases the set $\widetilde{S}$ becomes forward invariant.

Now we define $\alpha'_V := \alpha_V \circ \alpha_2^{-1}$. Note that concatenations of $\mathcal{K}$-functions are again in $\mathcal{K}$, hence $\alpha'_V \in \mathcal{K}$. Since $|x| \geq \alpha_2^{-1}(V(x))$, using monotonicity of α_V this definition implies

$$\alpha_V(|x|) \geq \alpha_V \circ \alpha_2^{-1}(V(x)) = \alpha'_V(V(x)).$$

Hence, along a trajectory $x(n, x_0)$ with $x_0 \in \widetilde{S}$, from (10.5) we get the inequality

$$V(x(n+1,x_0)) \leq V(x(n,x_0)) - \alpha_V(|x(n,x_0)|) \leq V(x(n,x_0)) - \alpha'_V(V(x(n,x_0))). \quad (10.6)$$

For the construction of β we need the last expression in (10.6) to be strictly increasing in $V(x(n,x_0))$. To this end we define

$$\tilde{\alpha}_V(r) := \min_{s\in[0,r]}\{\alpha'_V(s) + (r-s)/2\}.$$

Straightforward computations show that this function satisfies $r_2 - \tilde{\alpha}_V(r_2) > r_1 - \tilde{\alpha}_V(r_1) \geq 0$ for all $r_2 > r_1 \geq 0$ and $\min\{\alpha'_V(r/2), r/4\} \leq \tilde{\alpha}_V(r) \leq \alpha'_V(r)$ for all $r \geq 0$. In particular, (10.6) remains valid and we get the desired monotonicity when α'_V is replaced by $\tilde{\alpha}_V$.

We inductively define a function $\beta_1 : \mathbb{R}_0^+ \times \mathbb{N}_0 \to \mathbb{R}_0^+$ via

$$\beta_1(r,0) := r, \quad \beta_1(r,n+1) = \beta_1(r,n) - \tilde{\alpha}_V(\beta_1(r,n)). \quad (10.7)$$

By induction over n using the properties of $\tilde{\alpha}_V(r)$ and Inequality (10.6) one easily verifies the following inequalities:

$$\beta_1(r_2,n) > \beta_1(r_1,n) \geq 0 \text{ for all } r_2 > r_1 \geq 0 \text{ and all } n \in \mathbb{N}_0 \quad (10.8)$$
$$\beta_1(r,n_1) > \beta_1(r,n_2) > 0 \text{ for all } n_2 > n_1 \geq 0 \text{ and all } r > 0 \quad (10.9)$$
$$V(x(n,x_0)) \leq \beta_1(V(x_0),n) \text{ for all } n \in \mathbb{N}_0 \text{ and all } x_0 \in \widetilde{S} \quad (10.10)$$

From (10.9) it follows that $\beta_1(r,n)$ is monotone decreasing in n and by (10.8) it is bounded from below by 0. Hence, for each $r \geq 0$ the limit $\beta_1^\infty(r) = \lim_{n\to\infty}\beta_1(r,n)$ exists. We claim that $\beta_1^\infty(r) = 0$ holds for all r. Indeed, convergence implies $\beta_1(r,n) - \beta_1(r,n+1) \to 0$ as $n \to \infty$ which together with (10.7) yields $\tilde{\alpha}_V(\beta_1(r,n)) \to 0$. On the other hand, since $\tilde{\alpha}_V$ is continuous, we get $\tilde{\alpha}_V(\beta_1(r,n)) \to \tilde{\alpha}_V(\beta_1^\infty(r))$. This implies

$$\tilde{\alpha}_V(\beta_1^\infty(r)) = 0$$

which because of $\tilde{\alpha}_V(r) \geq \min\{\alpha_V(r/2), r/4\}$ and $\alpha_V \in \mathcal{K}$ is only possible if $\beta_1^\infty(r) = 0$.

Consequently, $\beta_1(r,n)$ has all properties of a $\mathcal{KL}$ function except that it is only defined for $n \in \mathbb{N}_0$. Defining the linear interpolation

$$\beta_2(r,t) := (n+1-t)\beta_1(r,n) + (t-n)\beta_1(r,n+1)$$

for $t \in [n, n+1)$ and $n \in \mathbb{N}_0$, we obtain a function $\beta_2 \in \mathcal{KL}$ which coincides with β_1 for $t = n \in \mathbb{N}_0$. Finally, setting

$$\beta(r,t) := \alpha_1^{-1} \circ \beta_2(\alpha_2(r), t)$$

we can use (10.10) in order to obtain

$$\begin{aligned}|x(n,x_0)| &\le \alpha_1^{-1}(V(x(n,x_0)) \le \alpha_1^{-1}\circ\beta_1(V(x_0),n)\\ &= \alpha_1^{-1}\circ\beta_2(V(x_0),n) \le \alpha_1^{-1}\circ\beta_2(\alpha_2(|x_0|,n) = \beta(|x_0|,n),\end{aligned}$$

for all $x_0 \in \widetilde{S}$ and all $n \in \mathbb{N}_0$. This is the desired inequality (10.2). If $\widetilde{S} = S = Y$ this shows the claimed asymptotic stability on Y and global asymptotic stability if $Y = X$. If $\widetilde{S} \neq S$, then in order to satisfy the local version of Definition 10.2 it remains to show that $x \in \mathcal{B}_\eta(x_*)$ implies $x \in \widetilde{S}$. Since by definition of η and γ we have $\eta = \alpha_2^{-1}(\gamma)$, we get

$$x \in \mathcal{B}_\eta(x_*) \Rightarrow |x| < \eta = \alpha_2^{-1}(\gamma) \Rightarrow V(x) \le \alpha_2(|x|) < \gamma \Rightarrow x \in \widetilde{S}.$$

This finishes the proof. □

Likewise, P-practical asymptotic stability can be ensured by a suitable Lyapunov function condition provided the set P is forward invariant.

Theorem 10.6 [P-practical asymptotic stability]
Consider forward invariant sets Y and $P \subset Y$ and a point $x_* \in P$. If there exists a Lyapunov function V on $S = Y \setminus P$ then x_* is P-practically asymptotically stable on Y.

Proof: The same construction of β as in the proof of Theorem 10.5 yields

$$|x(n,x_0)|_{x_*} \le \beta(|x|_{x_*},n) \qquad (10.2)$$

for all $n = 0,\ldots,n^*-1$, where $n^* \in \mathbb{N}_0$ is minimal with $x(n^*,x_0) \in P$. This follows with the same arguments as in the proof of Theorem 10.5 by restricting the times considered in (10.6) and (10.10) to $n = 0,\ldots,n^*-2$ and $n = 0,\ldots,n^*-1$, respectively.

Since forward invariance of P ensures $x(n,x_0) \in P$ for all $n \ge n^*$, the times n for which $x(n,x_0) \notin P$ holds are exactly $n = 0,\ldots,n^*-1$. Since these are exactly the times at which (10.2) is required, this yields the desired P-practical asymptotic stability. □

For continuous time systems $\dot{x} = g(x)$ all the concepts introduced in this section can be carried over directly. Particularly, the definitions of asymptotic and P-practical asymptotic stability are identical. In the definition of Lyapunov functions, condition (10.4) stays the same while condition (10.5) becomes

$$V(x(t,x_0)) \le V(x_0) - \int_0^t \alpha_V(|x(t,x_0)|_{x_*}).$$

This is equivalent to

$$\frac{V(x(t,x_0)) - V(x_0)}{t} \le -\frac{1}{t}\int_0^t \alpha_V(|x(t,x_0)|_{x_*})$$

and if V is continuously differentiable, then by letting $t \to 0$ one obtains the equivalent characterization

$$DV(x_0)g(x_0) \le -\alpha_V(|x_0|_{x_*}). \qquad (10.11)$$

Now it is obvious that this concept generalizes Definition 3.8, which we used in the linear case. With this definition of a Lyapunov function, all results in this section remain valid in the continuous time case.

Chapter 11

Model predictive control schemes

11.1 The MPC algorithm without terminal conditions

We start this chapter by formulating the basic MPC algorithm already sketched in Chapter 9 in a more rigorous way. Here, the stage cost $\ell : X \times U \to \mathbb{R}$ is a general function. In the case of sampled data systems we can take the continuous time nature of the underlying model into account by defining the stage cost ℓ as an integral over a continuous time running cost function $L : X \times U \to \mathbb{R}_0^+$ on a sampling interval. Using the continuous time solution $\hat{x}$ from (8.5), we can define

$$\ell(x, u) := \int_0^T L(\hat{x}(t, x, u), u(t))dt. \tag{11.1}$$

Defining ℓ this way, we can incorporate the intersampling behavior of the sampled data system, i.e., the behavior of the continuous time solution between two sampling times t_k and t_{k+1}, explicitly into our optimal control problem.

Given such a cost function ℓ and a prediction horizon length $N \geq 2$, we can now formulate the basic MPC scheme as an algorithm. In the optimal control problem ($\mathrm{OCP_N}$) within this algorithm we introduce a set of control sequences $\mathbb{U}^N(x_0) \subseteq U^N$ over which we optimize. This set may include constraints depending on the initial value x_0. Details about how this set should be chosen will be discussed in Sect. 11.2. For the moment we simply set $\mathbb{U}^N(x_0) := U^N$ for all $x_0 \in X$.

Algorithm 11.1 (Basic MPC algorithm)

At each time instant $j = 0, 1, 2 \ldots$:

(1) Measure the state $x(j) \in X$ of the system

(2) Set $x_0 := x(j)$, solve the optimal control problem

$$\boxed{\begin{aligned} &\text{minimize} \quad J_N(x_0,u(\cdot)) := \sum_{k=0}^{N-1} \ell(x_u(k,x_0),u(k)) \\ &\text{with respect to } \; u(\cdot) \in \mathbb{U}^N(x_0), \quad \text{subject to} \\ &x_u(0,x_0) = x_0, \quad x_u(k+1,x_0) = f(x_u(k,x_0),u(k)) \end{aligned}} \tag{OCP$_\mathrm{N}$}$$

and denote the obtained optimal control sequence by $u^\star(\cdot) \in \mathbb{U}^N(x_0)$.

(3) Define the MPC-feedback value $\mu_N(x(j)) := u^\star(0) \in U$ and use this control value in the next sampling period.

□

Observe that in this algorithm we have assumed that an optimal control sequence $u^\star(\cdot)$ exists. Sufficient conditions for this existence are briefly discussed after Definition 12.1, below.

The MPC closed loop system resulting from Algorithm 11.1 is given by (9.3) with state feedback law $\mu = \mu_N$, i.e.,

$$x^+ = f(x,\mu_N(x)). \tag{11.2}$$

The trajectories of this system will be denoted by $x_{\mu_N}(n)$ or, if we want to emphasize the initial value $x_0 = x_{\mu_N}(0)$, by $x_{\mu_N}(n,x_0)$.

During our theoretical investigations we will neglect the fact that computing the solution of (OCP$_\mathrm{N}$) in Step (2) of the algorithm usually needs some computation time τ_c which — in the case when τ_c is relatively large compared to the sampling period T — may not be negligible in a real time implementation.

In our abstract formulations of the MPC Algorithm 11.1 only the first element $u^\star(0)$ of the respective minimizing control sequence is used in each step, the remaining entries $u^\star(1),\ldots,u^\star(N-1)$ are discarded. In the practical implementation, however, these entries play an important role because numerical optimization algorithms for solving (OCP$_\mathrm{N}$) (or its variants) usually work iteratively: starting from an initial guess $u^0(\cdot)$ an optimization algorithm computes iterates $u^i(\cdot)$, $i = 1,2,\ldots$ converging to the minimizer $u^\star(\cdot)$ and a good choice of $u^0(\cdot)$ is crucial in order to obtain fast convergence of this iteration, or even to ensure convergence, at all. Here, the minimizing sequence from the previous time step can be efficiently used in order to construct such a good initial guess. Ways to implement this idea will be discussed in the excercises.

11.2 Constraints

One of the main reasons for the success of MPC (and MPC in general) is its ability to explicitly take constraints into account. Here, we consider constraints both on the control as well as on the state. To this end, we introduce a nonempty *state constraint set* $\mathbb{X} \subseteq X$ and for each $x \in \mathbb{X}$ we introduce a nonempty *control constraint set* $\mathbb{U}(x) \subseteq U$. Of course,

$\mathbb{U}$ may also be chosen independent of x. The idea behind introducing these sets is that we want the trajectories to lie in $\mathbb{X}$ and the corresponding control values to lie in $\mathbb{U}(x)$. This is made precise in the following definition.

Definition 11.2 [Admissibility] Consider a control system (8.2) and the state and control constraint sets $\mathbb{X} \subseteq X$ and $\mathbb{U}(x) \subseteq U$.

(i) The states $x \in \mathbb{X}$ are called *admissible states* and the control values $u \in \mathbb{U}(x)$ are called *admissible control values for* x. The elements of the set $\mathbb{Y} := \{(x,u) \in X \times U \,|\, x \in \mathbb{X}, u \in \mathbb{U}(x)\}$ are called *admissible pairs*.

(ii) For $N \in \mathbb{N}$ and an initial value $x_0 \in \mathbb{X}$ we call a control sequence $u \in U^N$ and the corresponding trajectory $x_u(k,x_0)$ *admissible for* x_0 *up to time* N, if

$$(x_u(k,x_0), u(k)) \in \mathbb{Y} \text{ for all } k = 0,\dots,N-1 \quad \text{and} \quad x_u(N,x_0) \in \mathbb{X}$$

holds. We denote the set of admissible control sequences for x_0 up to time N by $\mathbb{U}^N(x_0)$.

(iii) A control sequence $u \in U^\infty$ and the corresponding trajectory $x_u(k,x_0)$ are called *admissible for* x_0 if they are admissible for x_0 up to every time $N \in \mathbb{N}$. We denote the set of admissible control sequences for x_0 by $\mathbb{U}^\infty(x_0)$.

(iv) A feedback law $\mu : X \to U$ is called *admissible* if $\mu(x) \in \mathbb{U}^1(x)$ holds for all $x \in \mathbb{X}$.

Whenever the reference to x or x_0 is clear from the context we will omit the additional "for x" or "for x_0". □

Since we can (and will) identify control sequences with only one element with the respective control value, we can consider $\mathbb{U}^1(x_0)$ as a subset of U, which we already implicitly did in the definition of admissibility for the feedback law μ, above. However, in general $\mathbb{U}^1(x_0)$ does not coincide with $\mathbb{U}(x_0) \subseteq U$ because using $x_u(1,x) = f(x,u)$ and the definition of $\mathbb{U}^N(x_0)$ we get $\mathbb{U}^1(x) := \{u \in \mathbb{U}(x) \mid f(x,u) \in \mathbb{X}\}$. With this subtle difference in mind, one sees that our admissibility condition (iv) on μ ensures both $\mu(n,x) \in \mathbb{U}(x)$ and $f(x,\mu(n,x)) \in \mathbb{X}$ whenever $x \in \mathbb{X}$.

Furthermore, our definition of $\mathbb{U}^N(x)$ implies that even if $\mathbb{U}(x) = \mathbb{U}$ is independent of x the set $\mathbb{U}^N(x)$ may depend on x for some or all $N \in \mathbb{N}_\infty$.

Often, in order to be suitable for optimization purposes these sets are assumed to be compact and convex. For our theoretical investigations, however, we do not need any regularity requirements of this type except that these sets are nonempty.

MPC is well suited to handle constraints because these can directly be inserted into Algorithm 11.1. In fact, since we already formulated the corresponding optimization problem ($\mathrm{OCP_N}$) with state dependent control value sets, the constraints are readily included if we use $\mathbb{U}^N(x_0)$ from Definition 11.2(ii) in ($\mathrm{OCP_N}$). However, when doing so we have to make sure that the constraints in ($\mathrm{OCP_N}$) can be satisfied for all j, i.e., that we do not optimize over an empty set because $\mathbb{U}^N(x_0) = \emptyset$. This is formalized in the following definition.

Definition 11.3 (i) An initial condition $x_0 \in \mathbb{X}$ is called *feasible* for ($\mathrm{OCP_N}$) if the constraints imposed in ($\mathrm{OCP_N}$) can be satisfied, i.e, if $\mathbb{U}^N(x_0) \neq \emptyset$.

(ii) A MPC algorithm 11.1 is called *recursively feasible* on a set $A \subseteq \mathbb{X}$ if each $x \in A$ is feasible for $(\mathrm{OCP}_\mathrm{N})$ and $x \in A$ implies $f(x, \mu_N(x)) \in A$ (implying that $f(x, \mu_N(x))$ is again feasible). □

One easily sees that recursive feasibility implies that $x_{\mu_N}(j)$ is feasible for all $j \in \mathbb{N}$ if $x_{\mu_N}(0) \in A$. In order to ensure recursive feasibility of $A = \mathbb{X}$ for Algorithm 11.1, we need the following assumption.

Assumption 11.4 [Viability] For each $x \in \mathbb{X}$ there exists $u \in \mathbb{U}(x)$ such that $f(x, u) \in \mathbb{X}$ holds. □

The property defined in this assumption is called *viability* or *weak (or controlled) forward invariance* of $\mathbb{X}$. It excludes the situation that there are states $x \in \mathbb{X}$ from which the trajectory leaves the set $\mathbb{X}$ for all admissible control values. Hence, it ensures $\mathbb{U}^N(x_0) \neq \emptyset$ for all $x_0 \in \mathbb{X}$ and all $N \in \mathbb{N}_\infty$. Thus, it ensures that any $x_0 \in \mathbb{X}$ is feasible for $(\mathrm{OCP}_\mathrm{N})$ and hence ensures that $\mu_N(x)$ is well defined for each $x \in \mathbb{X}$. We will see after the next example that viability of $\mathbb{X}$ also implies recursive feasibility and admissibility of the closed loop. Furthermore, a straightforward induction shows that under Assumption 11.4 any finite admissible control sequence $u(\cdot) \in \mathbb{U}^N(x_0)$ can be extended to an infinite admissible control sequence $\tilde{u}(\cdot) \in \mathbb{U}^\infty(x_0)$ with $u(k) = \tilde{u}(k)$ for all $k = 0, \dots, N-1$.

In order to see that the construction of a constraint set $\mathbb{X}$ meeting Assumption 11.4 is usually a nontrivial task, we consider the following Example.

Example 11.5 Consider

$$x^+ = f(x,u) = \begin{pmatrix} x_1 + x_2 + u/2 \\ x_2 + u \end{pmatrix},$$

which can be seen as a sampled-data model for a car on a one-dimensional road with position x_1, speed x_2 and piecewise constant acceleration u. Assume we want to constrain all variables, i.e., the position x_1, the velocity x_2 and the acceleration u to the interval $[-1, 1]$. For this purpose one could define $\mathbb{X} = [-1, 1]^2$ and $\mathbb{U}(x) = \mathbb{U} = [-1, 1]$. Then, however, for $x = (1, 1)^\top$, one immediately obtains

$$x_1^+ = x_1 + x_2 + u/2 = 2 + u/2 \geq 3/2$$

for all u, hence $x^+ \notin \mathbb{X}$ for all $u \in \mathbb{U}$. Thus, in order to find a viable set $\mathbb{X}$ we need to either tighten or relax some of the constraints. For instance, relaxing the constraint on u to $\mathbb{U} = [-2, 2]$ the viability of $\mathbb{X} = [-1, 1]^2$ is guaranteed, because then by elementary computations one sees that for each $x \in \mathbb{X}$ the control value

$$u = \begin{cases} 0, & x_1 + x_2 \in [-1, 1] \\ 2 - 2x_1 - 2x_2, & x_1 + x_2 > 1 \\ -2 - 2x_1 - 2x_2, & x_1 + x_2 < -1 \end{cases}$$

is in $\mathbb{U}$ and satisfies $f(x, u) \in \mathbb{X}$. A way to achieve viability without changing $\mathbb{U}$ is by tightening the constraint on x_2 by defining

$$\mathbb{X} = \{(x_1, x_2)^T \in \mathbb{R}^2 \mid x_1 \in [-1, 1], x_2 \in [-1, 1] \cap [-3/2 - x_1, 3/2 - x_1]\}, \qquad (11.3)$$

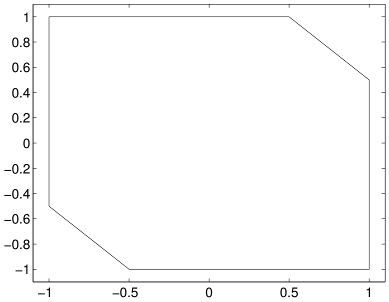


Figure 11.1: Illustration of the set $\mathbb{X}$ from (11.3)

see Fig. 11.5. Again, elementary computations show that for each $x \in \mathbb{X}$ and

$$u = \begin{cases} 1, & x_2 < -1/2 \\ -2x_2, & x_2 \in [-1/2, 1/2] \\ -1, & x_2 > 1/2 \end{cases}$$

the desired properties $u \in \mathbb{U}$ and $f(x,u) \in \mathbb{X}$ hold. □

This example shows that finding viable constraint sets $\mathbb{X}$ (and the corresponding $\mathbb{U}$ or $\mathbb{U}(x)$) is a tricky task already for very simple systems. Still, Assumption 11.4 significantly simplifies the subsequent analysis, cf. Theorem 11.6, below. For this reason we will impose this condition in our theoretical investigations for schemes without stabilizing terminal conditions. The assumption can be avoided if suitable terminal constraints are employed. We will discuss this extension of the scheme in Section 11.3.

The following theorem shows that the viability assumption ensures recursive feasibility of Algorithm 11.1 and that the resulting MPC closed loop satisfies the desired constraints.

Theorem 11.6 [Recursive Feasibility and Admissibility] Consider Algorithm 11.1 using $\mathbb{U}^N(x_0)$ from Def. 11.2(ii) in the optimal control problem ($\mathrm{OCP_N}$) for constraint sets $\mathbb{X} \subset X$, $\mathbb{U}(x) \subset U$, $x \in \mathbb{X}$, satisfying Assumption 11.4. Consider the MPC closed loop system (11.2). Then the MPC algorithm is recursively feasible on $A = \mathbb{X}$ and for any $x_{\mu_N}(0) \in \mathbb{X}$ the constraints are satisfied along the solution of (11.2), i.e.,

$$(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) \in \mathbb{Y} \tag{11.4}$$

for all $n \in \mathbb{N}$. Thus, the MPC-feedback μ_N is admissible in the sense of Definition 11.2(iv).

Proof: First, recall from the discussion after Assumption 11.4 that under this assumption the optimal control problem ($\mathrm{OCP_N}$) is feasible for each $x \in \mathbb{X}$, hence $\mu_N(x)$ is well defined for each $x \in \mathbb{X}$.

We now show that $x_{\mu_N}(n) \in \mathbb{X}$ implies $\mu_N(x_{\mu_N}(n)) \in \mathbb{U}(x_{\mu_N}(n))$ and $x_{\mu_N}(n+1) \in \mathbb{X}$. This implies recursive feasibility of $A = \mathbb{X}$, and admissibility follows by induction from $x_{\mu_n}(0) \in \mathbb{X}$.

The viability of $\mathbb{X}$ from Assumption 11.4 ensures that whenever $x_{\mu_N}(n) \in \mathbb{X}$ holds in Algorithm 11.1 then $x_0 \in \mathbb{X}$ is feasible for the respective optimal control problem $(\mathrm{OCP_N})$. Since the optimization is performed with respect to admissible control sequences only, also the optimal control sequence $u^\star(\cdot)$ is admissible for $x_0 = x_{\mu_N}(n)$. This implies $\mu_N(x_{\mu_N}(n)) = u^\star(0) \in \mathbb{U}^1(x_{\mu_N}(n)) \subseteq \mathbb{U}(x_{\mu_N}(n))$ and thus also

$$x_{\mu_N}(n+1) = f(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) = f(x(n), u^\star(0)) \in \mathbb{X},$$

i.e., $x_{\mu_N}(n+1) \in \mathbb{X}$. □

In the underlying optimization algorithms for solving $(\mathrm{OCP_N})$, usually the constraints cannot be specified via sets $\mathbb{X}$ and $\mathbb{U}(x)$. Rather, one uses so-called *equality* and *inequality constraints* in order to specify $\mathbb{X}$ and $\mathbb{U}(x)$ according to the following definition.

Definition 11.7 Given functions $G_i^S : X \times U \to \mathbb{R}$, $i \in \mathcal{E}^S = \{1, \ldots, p_g\}$ and $H_i^S : X \times U \to \mathbb{R}$, $i \in \mathcal{I}^S = \{p_g + 1, \ldots, p_g + p_h\}$ with $p_g, p_h \in \mathbb{N}_0$, we define the constraint sets $\mathbb{X}$ and $\mathbb{U}(x)$ via

$$\mathbb{X} := \left\{ x \in X \;\middle|\; \begin{array}{l} \text{there exists } u \in U \text{ with } G_i^S(x,u) = 0 \text{ for all } i \in \mathcal{E}^S \\ \text{and } H_i^S(x,u) \geq 0 \text{ for all } i \in \mathcal{I}^S \end{array} \right\}$$

and, for $x \in \mathbb{X}$

$$\mathbb{U}(x) := \left\{ u \in U \;\middle|\; \begin{array}{l} G_i^S(x,u) = 0 \text{ for all } i \in \mathcal{E}^S \text{ and} \\ H_i^S(x,u) \geq 0 \text{ for all } i \in \mathcal{I}^S \end{array} \right\}$$

Here, the functions G_i^S and H_i^S do not need to depend on both arguments. The functions G_i^S, H_i^S not depending on u are called *pure state constraints*, the functions G_i^S, H_i^S not depending on x are called *pure control constraints* and the functions G_i^S, H_i^S depending on both x and u are called *mixed constraints.* □

Observe that if we do not have mixed constraints then $\mathbb{U}(x)$ is independent of x.

The reason for defining $\mathbb{X}$ and $\mathbb{U}(x)$ via these (in)equality constraints is purely algorithmic: the plain information "$x_u(k, x_0) \notin \mathbb{X}$" does not yield any information for the optimization algorithm in order to figure out how to find an admissible $u(\cdot)$, i.e., a $u(\cdot)$ for which "$x_u(k, x_0) \in \mathbb{X}$" holds. In contrast to that, an information of the form "$H_i^S(x_u(k, x_0), u(k)) < 0$" together with additional knowledge about H_i^S (provided, e.g., by the derivative of H_i^S) enables the algorithm to compute a "direction" in which $u(\cdot)$ needs to be modified in order to reach an admissible $u(\cdot)$.

In our theoretical investigations we will use the notationally more convenient set characterization of the constraints via $\mathbb{X}$ and $\mathbb{U}(x)$ or $\mathbb{U}^N(x)$. In the practical implementation of our MPC method, however, we will use their characterization via the inequality constraints from Definition 11.7.

11.3 The MPC algorithm with terminal conditions

In this section we discuss an important variant of the basic MPC Algorithm 11.1. This algorithm adds a constraint on the terminal state $x_u(N, x_0)$ of the trajectory over which we optimize in ($\mathrm{OCP_N}$), as well as a weight on this term. This combination of constraint and weight on the terminal state is called *terminal conditions*. As we will see, under suitable assumptions on the terminal conditions, the behavior of the MPC closed-loop can significantly improve. The main disadvantage of terminal condition is that a rigorous derivation of a constraint and a weight meeting these assumptions can be very dificult for complex control systems.

The terminal constraint is of the form

$$x_u(N, x_0) \in \mathbb{X}_0 \text{ for a } \textit{terminal constraint set } \mathbb{X}_0 \subseteq \mathbb{X}. \tag{11.5}$$

Of course, in the practical implementation the constraint set $\mathbb{X}_0$ is again expressed via (in)equalities of the form given in Definition 11.7.

When using terminal constraints, the MPC-feedback law is only defined for those states x_0 for which the optimization problem within the MPC algorithm is feasible also for these additional constraints, i.e., for which there exists an admissible control sequence with corresponding trajectory starting in x_0 and ending in the terminal constraint set. Such initial values are again called *feasible* and the set of all feasible initial values form the feasible set. This set along with the corresponding admissible control sequences is formally defined as follows.

Definition 11.8 [Feasible set and admissible control sequences]
For $\mathbb{X}_0$ from (11.5) we define the *feasible set* for horizon $N \in \mathbb{N}$ by

$$\mathbb{X}_N := \{x_0 \in \mathbb{X} \mid \text{there exists } u(\cdot) \in \mathbb{U}^N(x_0) \text{ with } x_u(N, x_0) \in \mathbb{X}_0\}$$

and for each $x_0 \in \mathbb{X}_N$ we define the set of *admissible control sequences* by

$$\mathbb{U}^N_{\mathbb{X}_0}(x_0) := \{u(\cdot) \in \mathbb{U}^N(x_0) \mid x_u(N, x_0) \in \mathbb{X}_0\}.$$

□

Note that in $\mathbb{X}_N = \mathbb{X}$ and $\mathbb{U}^N_{\mathbb{X}_0}(x) = \mathbb{U}^N(x)$ holds if $\mathbb{X}_0 = \mathbb{X}$, i.e., if no additional terminal constraints are imposed.

The additional weight on the terminal state $x_u(N)$ is formalized by means of a terminal cost of the form $F(x_u(N, x_0))$ with $F : \mathbb{X}_0 \to \mathbb{R}$ in the optimization objective.

Together this leads to the following MPC algorithms extending the basic Algorithms 11.1. Note that compared to these basic algorithms only the optimal control problems are different, i.e., the part in the boxes in Step (2).

Algorithm 11.9 (MPC algorithm with terminal conditions)

At each time instant $j = 0, 1, 2 \ldots$:

(1) Measure the state $x(j) \in X$ of the system.

(2) Set $x_0 := x(j)$, solve the optimal control problem

$$\boxed{\begin{array}{l} \text{minimize} \quad J_N(x_0,u(\cdot)) := \displaystyle\sum_{k=0}^{N-1} \ell(x_u(k,x_0),u(k)) + F(x_u(N,x_0)) \\ \text{with respect to } \ u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0), \quad \text{subject to} \\ x_u(0,x_0) = x_0, \quad x_u(k+1,x_0) = f(x_u(k,x_0),u(k)) \end{array}} \tag{OCP$_{\mathrm{N,e}}$}$$

and denote the obtained optimal control sequence by $u^\star(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$.

(3) Define the MPC-feedback value $\mu_N(x(j)) := u^\star(0) \in U$ and use this control value in the next sampling period.

□

We end this section with three useful results on the sets of admissible control sequences from Definition 11.8.

Lemma 11.10 Let $x_0 \in \mathbb{X}_N$, $N \in \mathbb{N}$ and $K \in \{0,\dots,N\}$ be given.

(i) For each $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ we have $x_u(K,x_0) \in \mathbb{X}_{N-K}$.

(ii) For each $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ the control sequences $u_1 \in U^K$ and $u_2 \in U^{N-K}$ uniquely defined by the relation

$$u(k) = \begin{cases} u_1(k), & k = 0,\dots,K-1 \\ u_2(k-K), & k = K,\dots,N-1 \end{cases} \tag{11.6}$$

satisfy $u_1 \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)$ and $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(x_{u_1}(K,x_0))$.

(iii) For each $u_1(\cdot) \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)$ there exists $u_2(\cdot) \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(x_{u_1}(K,x_0))$ such that $u(\cdot)$ from (11.6) satisfies $u \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$.

Proof: (i) Using (8.4) we obtain the identity

$$x_{u(K+\cdot)}(N-K, x_u(K,x_0))) = x_u(N,x_0) \in \mathbb{X}_0,$$

which together with the definition of $\mathbb{X}_{N-K}$ implies the assertion.

(ii) The relation (11.6) together with (8.4) implies

$$x_u(k,x_0) = \begin{cases} x_{u_1}(k,x_0), & k = 0,\dots,K \\ x_{u_2}(k-K, x_{u_1}(K,x_0)), & k = K,\dots,N \end{cases} \tag{11.7}$$

For $k = 0,\dots,K-1$ this identity and (11.6) yield

$$u_1(k) = u(k) \in \mathbb{U}(x_u(k,x_0)) = \mathbb{U}(x_{u_1}(k,x_0))$$

and for $k = 0, \dots, N-K-1$ we obtain

$$u_2(k) = u(k+K) \in \mathbb{U}(x_u(k+K, x_0)) = \mathbb{U}(x_{u_2}(k, x_{u_1}(K, x_0))),$$

implying $u_1 \in \mathbb{U}^K(x_0)$ and $u_2 \in \mathbb{U}^{N-K}(x_{u_1}(K, x_0))$. Furthermore, (11.7) implies the equation $x_{u_2}(N-K, x_{u_1}(K, x_0)) = x_u(N, x_0) \in \mathbb{X}_0$ which proves $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x_{u_1}(K, x_0))$. This, in turn, implies that $\mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x_{u_1}(K, x_0))$ is nonempty, hence $x_{u_1}(K, x_0) \in \mathbb{X}_{N-K}$ and consequently $u_1 \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(n, x_0)$ follows.

(iii) By definition, for each $x \in \mathbb{X}_{N-K}(n+K)$ there exists $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(n+K, x)$. Choosing such a u_2 for $x = x_{u_1}(K, x_0) \in \mathbb{X}_{N-K}(n+K)$ and defining u via (11.6), similar arguments as in Part (ii), above, show the claim $u \in \mathbb{U}^N_{\mathbb{X}_0}(n, x_0)$. □

A straightforward corollary of this lemma is the following.

Corollary 11.11 For each $x \in \mathbb{X}_N$ the MPC-feedback law μ_N obtained from Algorithm 11.9 satisfies

$$f(x, \mu_N(x)) \in \mathbb{X}_{N-1}.$$

□

Proof: Since $\mu_N(x)$ is the first element $u^\star(0)$ of the optimal control sequence $u^\star \in \mathbb{U}^N_{\mathbb{X}_0}(x)$ we get $f(x, \mu_N(x)) = x_{u^\star}(1, x)$. Now Lemma 11.10(i) yields the assertion. □

The final result shows that with terminal conditions we can obtain Theorem 11.6 without having to assume viability of $\mathbb{X}$ — if in exchange we assume viability of the terminal constraint set $\mathbb{X}_0$.

Theorem 11.12 [Recursive Feasibility and Admissibility] Consider Algorithm 11.9 for constraint sets $\mathbb{X} \subset X$, $\mathbb{U}(x) \subset U$, $x \in \mathbb{X}$, and a terminal constraint set $\mathbb{X}_0$ which satisfies Assumption 11.4. Consider the MPC closed loop system (11.2). Then the MPC algorithm is recursively feasible on $A = \mathbb{X}_N$ and for $x_{\mu_N}(0) \in \mathbb{X}_N$ the constraints are satisfied along the solution of (11.2), i.e.,

$$(x_{\mu_N}(n), \mu_N(x_{\mu_N}(n))) \in \mathbb{Y} \tag{11.8}$$

for all $n \in \mathbb{N}$. Thus, the MPC-feedback μ_N is admissible in the sense of Definition 11.2(iv).

Proof: We show that under the viability assumption on $\mathbb{X}_0$ the inclusion $\mathbb{X}_{N-1} \subseteq \mathbb{X}_N$ holds. Then recursive feasibility follows from Corollary 11.11 and admissibility follows as in the proof of Theorem 11.6.

In order to show the inclusion $\mathbb{X}_{N-1} \subseteq \mathbb{X}_N$, consider $x \in \mathbb{X}_{N-1}$. Then there is an admissible control $u \in \mathbb{U}^{N-1}_{\mathbb{X}_0}(x)$, implying $x_u(N-1, x) \in \mathbb{X}_0$. Viability of $\mathbb{X}_0$ implies the existence of a control value $\tilde{u} \in \mathbb{U}(x_u(N-1, x))$ with $f(x_u(N-1, x), \tilde{u}) \in \mathbb{X}_0$. This implies that the control sequence

$$\hat{u} = (u(0), \dots, u(N-1), \tilde{u})$$

is admissible and satisfies $x_{\hat{u}}(N, x) = f(x_u(N-1, x), \tilde{u}) \in \mathbb{X}_N$. This implies $x \in \mathbb{X}_N$ and thus the desired inclusion. □

Chapter 12

Dynamic programming

This chapter repeats and extends some of the results from Section 6.1. As we will see, dynamic programming is not only important for deriving the Riccati equation but also as a basis for analyzing MPC schemes in the next chapters. We first consider finite horizon problems and then discuss infinite horizon problems.

12.1 Finite horizon problems

In this section we provide one of the classical tools in optimal control, the *dynamic programming principle*. We will formulate and prove the results in this section for $(\mathrm{OCP_{N,e}})$, since all other optimal control problems introduced above can be obtained as special cases of this problem. We will first formulate the principle for the open loop control sequences in $(\mathrm{OCP_{N,e}})$ and then derive consequences for the MPC-feedback law μ_N. The dynamic programming principle is often used as a basis for numerical algorithms. In contrast to this, here we will exclusively use the principle for analyzing the behavior of MPC closed loop systems. The reason for this is that the numerical effort of solving $(\mathrm{OCP_{N,e}})$ via dynamic programming usually grows exponentially with the dimension of the state of the system. In contrast to this, the computational effort of solving a single problem of type $(\mathrm{OCP_N})$ or $(\mathrm{OCP_{N,e}})$ scales much more moderately with the space dimension.

We start by defining some objects we need in the sequel.

Definition 12.1 Consider the optimal control problem $(\mathrm{OCP_{N,e}})$ with initial value $x_0 \in \mathbb{X}$ and optimization horizon $N \in \mathbb{N}_0$.

(i) The function

$$V_N(x_0) := \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} J_N(x_0, u(\cdot))$$

is called *optimal value function*.

(ii) A control sequence $u^\star(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ is called *optimal control sequence* for x_0, if

$$V_N(x_0) = J_N(x_0, u^\star(\cdot))$$

holds. The corresponding trajectory $x_{u^\star}(\cdot, x_0)$ is called *optimal trajectory*.

□

In our MPC Algorithms 11.1 and 11.9 we have assumed that an optimal control sequence $u^\star(\cdot)$ exists, cf. the comment after Algorithms 11.1. In general, this is not necessarily the case but under reasonable continuity and compactness conditions the existence of $u^\star(\cdot)$ can be rigorously shown. Examples of such theorems for a general infinite-dimensional state space can be found in Keerthi and Gilbert [12] or Doležal [3]. While for formulating and proving the dynamic programming principle we will not need the existence of $u^\star(\cdot)$, for all subsequent results we will assume that $u^\star(\cdot)$ exists, in particular when we derive properties of the MPC-feedback law μ_N. While we conjecture that most of the subsequent results in this lecture notes can be generalized to the case when μ_N is defined via an approximately minimizing control sequence, we decided to use the existence assumption because it considerably simplifies the presentation of the results in these lecture notes.

The following theorem introduces the *dynamic programming principle.* It gives an equation which relates the optimal value functions for different optimization horizons N and for different points in space.

Theorem 12.2 [Dynamic programming principle] Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$. Then for all $N \in \mathbb{N}$ and all $K = 1, \ldots, N$ the equation

$$V_N(x_0) = \inf_{u(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) \right. \\ \left. + \; V_{N-K}(x_u(K,x_0)) \right\} \qquad (12.1)$$

holds. If, in addition, an optimal control sequence $u^\star(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ exists for x_0, then we get the equation

$$V_N(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0), u^\star(k)) + V_{N-K}(x_{u^\star}(K,x_0)). \qquad (12.2)$$

In particular, in this case the "inf" in (12.1) is a "min".

Proof: First observe that from the definition of J_N for $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ we immediately obtain

$$J_N(x_0, u(\cdot)) = \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) \\ + \; J_{N-K}(x_u(K,x_0), u(\cdot + K)). \qquad (12.3)$$

Since $u(\cdot + K)$ equals $u_2(\cdot)$ from Lemma 11.10(ii) we obtain $u(\cdot + K) \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(x_u(K,x_0))$.

We now prove (12.1) by proving "≥" and "≤" separately. From (12.3) we obtain

$$\begin{aligned} J_N(x_0,u(\cdot)) &= \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \\ &\qquad + \; J_{N-K}(x_u(K,x_0),u(\cdot+K)) \\ &\geq \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_{N-K}(x_u(K,x_0)). \end{aligned}$$

Since this inequality holds for all $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$, it also holds when taking the infimum on both sides. Hence we get

$$\begin{aligned} V_N(x_0) &= \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} J_N(x_0,u(\cdot)) \\ &\geq \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right. \\ &\qquad\qquad \left. + \; V_{N-K}(x_u(K,x_0)) \right\} \\ &= \inf_{u_1(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u(k)) \right. \\ &\qquad\qquad \left. + \; V_{N-K}(x_{u_1}(K,x_0)) \right\}, \end{aligned}$$

i.e., (12.1) with "≥". Here in the last step we used the fact that by Lemma 11.10(ii) the control sequence u_1 consisting of the first K elements of $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$ lies in $\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)$ and, conversely, by Lemma 11.10(iii) each control sequence in $u_1(\cdot) \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)$ can be extended to a sequence in $u(\cdot) \in \mathbb{U}^N_{\mathbb{X}_0}(x_0)$. Thus, since the expression in braces does not depend on $u(K),\ldots,u(N-1)$, the infima coincide.

In order to prove "≤", fix $\varepsilon > 0$ and let $u^\varepsilon(\cdot)$ be an approximately optimal control sequence for the right hand side of (12.3), i.e.,

$$\begin{aligned} &\sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k,x_0),u^\varepsilon(k)) + J_{N-K}(x_{u^\varepsilon}(K,x_0),u^\varepsilon(\cdot+K)) \\ &\leq \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right. \\ &\qquad\qquad \left. + \; J_{N-K}(x_u(K,x_0),u(\cdot+K)) \right\} + \varepsilon. \end{aligned}$$

Now we use the decomposition (11.6) of $u(\cdot)$ into $u_1 \in \mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)$ and $u_2 \in \mathbb{U}^{N-K}_{\mathbb{X}_0}(x_{u_1}(K,x_0))$

from Lemma 11.10(ii). This way we can proceed

$$\begin{aligned}
&\inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right.\\
&\qquad\qquad\qquad\qquad \left. + \; J_{N-K}(x_u(K,x_0),u(\cdot+K)) \right\}\\
&= \inf_{\substack{u_1(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)\\ u_2(\cdot)\in\mathbb{U}^{N-K}_{\mathbb{X}_0}(x_{u_1}(K,x_0))}} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) \right.\\
&\qquad\qquad\qquad\qquad \left. + \; J_{N-K}(x_{u_1}(K,x_0),u_2(\cdot)) \right\}\\
&= \inf_{u_1(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_{u_1}(k,x_0),u_1(k)) \right.\\
&\qquad\qquad\qquad\qquad \left. + \; V_{N-K}(x_{u_1}(K,x_0)) \right\}
\end{aligned}$$

Now (12.3) yields

$$\begin{aligned}
V_N(x_0) &\le J(x_0,u^\varepsilon(\cdot))\\
&= \sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k,x_0),u^\varepsilon(k)) + J_{N-K}(x_{u^\varepsilon}(K,x_0),u^\varepsilon(\cdot+K))\\
&\le \inf_{u(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) \right.\\
&\qquad\qquad\qquad \left. + \; V_{N-K}(x_u(K,x_0)) \right\} + \varepsilon.
\end{aligned}$$

Since the first and the last term in this inequality chain are independent of ε and since $\varepsilon > 0$ was arbitrary, this shows (12.1) with "$\le$" and thus (12.1).

In order to prove (12.2) we use (12.3) with $u(\cdot) = u^\star(\cdot)$. This yields

$$\begin{aligned}
V_N(x_0) &= J(x_0,u^\star(\cdot))\\
&= \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + J_{N-K}(x_{u^\star}(K,x_0),u^\star(\cdot+K))\\
&\ge \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0),u^\star(k)) + V_{N-K}(x_{u^\star}(K,x_0))\\
&\ge \inf_{u(\cdot)\in\mathbb{U}^K_{\mathbb{X}_{N-K}}(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0),u(k)) + V_{N-K}(x_u(K,x_0)) \right\}\\
&= V_N(x_0),
\end{aligned}$$

where we used the (already proven) equality (12.1) in the last step. Hence, the two "$\geq$" in this chain are actually "$=$" which implies (12.2). □

The following corollary states an immediate consequence of the dynamic programming principle. It shows that tails of optimal control sequences are again optimal control sequences for suitably adjusted optimization horizon, time instant and initial value.

Corollary 12.3 If $u^\star(\cdot)$ is an optimal control sequence for initial value $x_0 \in \mathbb{X}_N$ and optimization horizon $N \geq 2$, then for each $K = 1, \dots, N-1$ the sequence $u^\star_K(\cdot) = u^\star(\cdot + K)$, i.e.,

$$u^\star_K(k) = u^\star(K+k), \quad k = 0, \dots, N-K-1$$

is an optimal control sequence for initial value $x_{u^\star}(K, x_0)$, time instant K and optimization horizon $N - K$. □

Proof: Inserting $V_N(x_0) = J_N(x_0, u^\star(\cdot))$ and the definition of $u^\star_k(\cdot)$ into (12.3) we obtain

$$V_N(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + J_{N-K}(x_{u^\star}(K, x_0), u^\star_K(\cdot))$$

Subtracting (12.2) from this equation yields

$$0 = J_{N-K}(x_{u^\star}(K, x_0), u^\star_K(\cdot)) - V_{N-K}(x_{u^\star}(K, x_0))$$

which shows the assertion. □

The next theorem relates the MPC-feedback law μ_N defined in the MPC Algorithms 11.1 and 11.9 to the dynamic programming principle. Here we use the argmin operator in the following sense: for a map $a : U \to \mathbb{R}$, a nonempty subset $\widetilde{U} \subseteq U$ and a value $u^\star \in \widetilde{U}$ we write

$$u^\star = \operatorname*{argmin}_{u \in \widetilde{U}} a(u) \tag{12.4}$$

if and only if $a(u^\star) = \inf_{u \in \widetilde{U}} a(u)$ holds. Whenever (12.4) holds the existence of the minimum $\min_{u \in \widetilde{U}} a(u)$ follows. However, we do not require uniqueness of the minimizer $u^\star$. In case of uniqueness equation (12.4) can be understood as an assignment, otherwise it is just a convenient way of writing "$u^\star$ minimizes $a(u)$".

Theorem 12.4 [Dynamic programming and MPC] Consider the optimal control problem ($\mathrm{OCP_{N,e}}$) with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$ and an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Then μ satisfies

$$\mu(x_0) = \operatorname*{argmin}_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_{N-1}(f(x_0, u))\} \tag{12.5}$$

if and only if μ satisfies

$$V_N(x_0) = \ell(x_0, \mu(x_0)) + V_{N-1}(f(x_0, \mu(x_0))), \tag{12.6}$$

where in (12.5) we interpret $\mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ as a subset of U, i.e., we identify the one element sequence $u = u(\cdot)$ with its only element $u = u(0)$. Moreover, if an optimal control sequence $u^\star$ exists then the MPC-feedback law $\mu(x_0) = \mu_N(x_0) = u^*(0)$ satisfies both (12.5) and (12.6).

Proof: Equation (12.6) follows from (12.5) by using (12.1) for $K = 1$ and the minimizing property of μ.

Conversely, assume (12.6). Inserting $x_u(1, x_0) = f(x_0, u)$ into the dynamic programming principle (12.1) for $K = 1$ we obtain

$$V_N(x_0) = \inf_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_{N-1}(1, f(x_0, u))\}\,. \tag{12.7}$$

This implies that the right hand sides of (12.6) and (12.7) coincide. Thus, the definition of argmin in (12.4) with $a(u) = \ell(x_0, u) + V_{N-1}(1, f(x_0, u))$ and $\widetilde{U} = \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ yields (12.5).

Finally, if $u^\star$ exists, then (12.6) (and thus also (12.5)) follows fur $\mu = \mu_n$ from the existence by inserting $u^\star(0) = \mu_N(x_0)$ and $x_{u^\star}(1, x_0) = f(x_0, \mu_N(x_0))$ into (12.2) for $K = 1$. □

Our final corollary in this section shows that we can reconstruct the whole optimal control sequence $u^\star(\cdot)$ using the feedback from (12.5).

Corollary 12.5 Consider the optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) with $x_0 \in \mathbb{X}$ and $N \in \mathbb{N}_0$ and consider admissible feedback laws $\mu_{N-k} : \mathbb{X} \to U$, $k = 0, \ldots, N-1$, in the sense of Definition 11.2(iv). Denote the solution of the closed loop system

$$x(0) = x_0, \quad x(k+1) = f(x(k), \mu_{N-k}(x(k))),\ k = 0, \ldots, N-1 \tag{12.8}$$

by $x_\mu(\cdot)$ and assume that the μ_{N-k} satisfy (12.5) with horizon $N-k$ instead of N and initial value $x_0 = x_\mu(k)$ for $k = 0, \ldots, N-1$. Then

$$u^\star(k) = \mu_{N-k}(x_\mu(k)), \quad k = 0, \ldots, N-1 \tag{12.9}$$

is an optimal control sequence for initial value x_0 and the solution of the closed loop system (12.8) is a corresponding optimal trajectory. □

Proof: Applying the control (12.9) to the dynamics (12.8) we immediately obtain

$$x_{u^\star}(k) = x_\mu(k), \quad k = 0, \ldots, N-1.$$

Hence, we need to show that

$$V_N(x_0) = J_N(x_0, u^\star) = \sum_{k=0}^{N-1} \ell(x_\mu(k), u^\star(k)) + F(x(N)).$$

Using (12.9) and (12.6) for $N-k$ instead of N and $x_0 = x_\mu(k)$ we get

$$V_{N-k}(x_\mu(k)) = \ell(x_\mu(k), u^\star(k)) + V_{N-k-1}(x_\mu(k+1))$$

for $k = 0, \ldots, N-1$. Summing these equalities for $k = 0, \ldots, N-1$ and eliminating the identical terms $V_{N-k}(x_\mu(k))$, $k = 1, \ldots, N-1$ on both sides we obtain

$$V_N(x_0) = \sum_{k=0}^{N-1} \ell(x_\mu(k), u^\star(k)) + V_0(x(N))$$

Since by definition of J_0 we have $V_0(x) = F(x)$, this shows the assertion. □

12.2 Infinite horizon problems

In this section we present the counterparts of the result from the previous section for infinite horizon problems. These are defined by as follows.

$$\boxed{\begin{array}{l} \text{minimize} \quad J_\infty(x_0, u(\cdot)) := \limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) \\ \text{with respect to } \; u(\cdot) \in \mathbb{U}^\infty(x_0), \quad \text{subject to} \\ x_u(0,x_0) = x_0, \quad x_u(k+1,x_0) = f(x_u(k,x_0), u(k)) \end{array}} \qquad (\mathrm{OCP}_\infty)$$

We assume that for all $x_0 \in \mathbb{X}$

$$-\infty < \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) = \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) < \infty,$$

which in particular implies that the optimal value function V_∞, as defined in the following definition, assumes finite values for all $x_0 \in \mathbb{X}$ and and that there is no admissible control sequence $\hat{u}(\cdot)$ for which

$$\liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x_{\hat{u}}(k,x_0), \hat{u}(k)) < V_\infty(x_0)$$

holds for some $x_0 \in \mathbb{X}$.

Definition 12.6 Consider the optimal control problem (OCP$_\infty$) with initial value $x_0 \in \mathbb{X}$.

(i) The function

$$V_\infty(x_0) := \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} J_\infty(x_0, u(\cdot))$$

is called *optimal value function.*

(ii) A control sequence $u^\star(\cdot) \in \mathbb{U}^\infty(x_0)$ is called *optimal control sequence* for x_0 if

$$V_\infty(x_0) = J_\infty(x_0, u^\star(\cdot))$$

holds. The corresponding trajectory $x_{u^\star}(\cdot, x_0)$ is called *optimal trajectory.*

□

The first result we state is the dynamic programming principle.

Theorem 12.7 [Dynamic programming principle] Consider the optimal control problem (OCP$_\infty$) with $x_0 \in \mathbb{X}$. Then for all $K \in \mathbb{N}$ the equation

$$V_\infty(x_0) = \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + V_\infty(x_u(K,x_0)) \right\} \tag{12.10}$$

holds. If, in addition, an optimal control sequence $u^\star(\cdot)$ exists for x_0, then we get the equation

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k,x_0), u^\star(k)) + V_\infty(x_{u^\star}(K,x_0)). \tag{12.11}$$

In particular, in this case the "inf" in (12.10) is a "min".

Proof: From the definition of J_∞ for $u(\cdot) \in \mathbb{U}^\infty(x_0)$ we immediately obtain

$$J_\infty(x_0, u(\cdot)) = \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + J_\infty(x_u(K,x_0), u(\cdot + K)), \tag{12.12}$$

where $u(\cdot + K)$ denotes the shifted control sequence defined by $u(\cdot + K)(k) = u(k+K)$, which is admissible for $x_u(K,x_0)$.

We now prove (12.10) by showing "$\geq$" and "$\leq$" separately: From (12.12) we obtain

$$\begin{aligned} J_\infty(x_0, u(\cdot)) &= \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + J_\infty(x_u(K,x_0), u(\cdot + K)) \\ &\geq \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + V_\infty(x_u(K,x_0)). \end{aligned}$$

Since this inequality holds for all $u(\cdot) \in \mathbb{U}^\infty$, it also holds when taking the infimum on both sides. Hence we get

$$\begin{aligned} V_\infty(x_0) &= \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} J_\infty(x_0, u(\cdot)) \\ &\geq \inf_{u(\cdot)\in\mathbb{U}^K(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + V_\infty(x_u(K,x_0)) \right\}, \end{aligned}$$

i.e., (12.10) with "$\geq$".

In order to prove "$\leq$", fix $\varepsilon > 0$ and let $u^\varepsilon(\cdot)$ be an approximately optimal control sequence for the right hand side of (12.12), i.e.,

$$\begin{aligned} &\sum_{k=0}^{K-1} \ell(x_{u^\varepsilon}(k,x_0), u^\varepsilon(k)) + J_\infty(x_{u^\varepsilon}(K,x_0), u^\varepsilon(\cdot + K)) \\ &\leq \inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)} \left\{ \sum_{k=0}^{K-1} \ell(x_u(k,x_0), u(k)) + J_\infty(x_u(K,x_0), u(\cdot + K)) \right\} + \varepsilon. \end{aligned}$$

Now we decompose $u(\cdot) \in \mathbb{U}^\infty(x_0)$ analogously to Lemma 11.10(ii) and (iii) into $u_1 \in \mathbb{U}^K(x_0)$ and $u_2 \in \mathbb{U}^\infty(x_{u_1}(K,x_0))$ via

$$u(k) = \begin{cases} u_1(k), & k = 0,\dots,K-1 \\ u_2(k-K), & k \geq K \end{cases}$$

This implies

$$\begin{aligned}
&\inf_{u(\cdot)\in\mathbb{U}^\infty(x_0)}\left\{\sum_{k=0}^{K-1}\ell(x_u(k,x_0),u(k))+J_\infty(x_u(K,x_0),u(\cdot+K))\right\}\\
&=\inf_{\substack{u_1(\cdot)\in\mathbb{U}^K(x_0)\\ u_2(\cdot)\in\mathbb{U}^\infty(x_{u_1}(K,x_0))}}\left\{\sum_{k=0}^{K-1}\ell(x_{u_1}(k,x_0),u_1(k))+J_\infty(x_{u_1}(K,x_0),u_2(\cdot))\right\}\\
&=\inf_{u_1(\cdot)\in\mathbb{U}^K(x_0)}\left\{\sum_{k=0}^{K-1}\ell(x_{u_1}(k,x_0),u_1(k))+V_\infty(x_{u_1}(K,x_0))\right\}
\end{aligned}$$

Now (12.12) yields

$$\begin{aligned}
V_\infty(x_0) &\le J_\infty(x_0,u^\varepsilon(\cdot))\\
&= \sum_{k=0}^{K-1}\ell(x_{u^\varepsilon}(k,x_0),u^\varepsilon(k))+J_\infty(x_{u^\varepsilon}(K,x_0),u^\varepsilon(\cdot+K))\\
&\le \inf_{u(\cdot)\in\mathbb{U}^K(x_0)}\left\{\sum_{k=0}^{K-1}\ell(x_u(k,x_0),u(k))+V_\infty(x_u(K,x_0))\right\}+\varepsilon,
\end{aligned}$$

i.e.,

$$\begin{aligned}
&V_\infty(x_0)\\
&\le \inf_{u(\cdot)\in\mathbb{U}^K(x_0)}\left\{\sum_{k=0}^{K-1}\ell(x_u(k,x_0),u(k))+V_\infty(x_u(K,x_0))\right\}+\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and the expressions in this inequality are independent of ε, this inequality also holds for $\varepsilon = 0$, which shows (12.10) with "$\le$" and thus (12.10).

In order to prove (12.11) we use (12.12) with $u(\cdot) = u^\star(\cdot)$. This yields

$$\begin{aligned}
V_\infty(x_0) &= J_\infty(x_0,u^\star(\cdot))\\
&= \sum_{k=0}^{K-1}\ell(x_{u^\star}(k,x_0),u^\star(k))+J_\infty(x_{u^\star}(x_0),u^\star(\cdot+K))\\
&\ge \sum_{k=0}^{K-1}\ell(x_{u^\star}(k,x_0),u^\star(k))+V_\infty(x_{u^\star}(K,x_0))\\
&\ge \inf_{u(\cdot)\in\mathbb{U}^K(x_0)}\left\{\sum_{k=0}^{K-1}\ell(x_u(k,x_0),u(k))+V_\infty(x_u(K,x_0))\right\}\\
&= V_\infty(x_0),
\end{aligned}$$

where we used the (already proved) equality (12.10) in the last step. Hence, the two "$\ge$" in this chain are actually "$=$" which implies (12.11). □

The following corollary states an immediate consequence from the dynamic programming principle. It shows that tails of optimal control sequences are again optimal control sequences for suitably adjusted initial value and time.

Corollary 12.8 If $u^\star(\cdot)$ is an optimal control sequence for (OCP_∞) with initial value x_0, then for each $K \in \mathbb{N}$ the sequence $u^\star_K(\cdot) = u^\star(\cdot + K)$, i.e.,

$$u^\star_K(k) = u^\star(K+k), \quad k = 0, 1, \ldots$$

is an optimal control sequence for initial value $x_{u^\star}(K, x_0)$ and initial time K. □

Proof: Inserting $V_\infty(x_0) = J_\infty(x_0, u^\star(\cdot))$ and the definition of $u^\star_K(\cdot)$ into (12.12) we obtain

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x_{u^\star}(k, x_0), u^\star(k)) + J_\infty(x_{u^\star}(x_0), u^\star_K(\cdot))$$

Subtracting (12.11) from this equation yields

$$0 = J_\infty(x_{u^\star}(x_0), u^\star_K(\cdot)) - V_\infty(x_{u^\star}(K, x_0))$$

which shows the assertion. □

The next two results are the analogues of Theorem 12.4 and Corollary 12.5 in the infinite horizon setting.

Theorem 12.9 Consider the optimal control problem $(\mathrm{OCP}_{\mathrm{N,e}})$ with $x_0 \in \mathbb{X}_N$ and $N \in \mathbb{N}_0$ and an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Then μ satisfies

$$\mu(x_0) = \operatorname*{argmin}_{u \in \mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)} \{\ell(x_0, u) + V_\infty(f(x_0, u))\} \tag{12.13}$$

if and only if μ satisfies

$$V_\infty(x_0) = \ell(x_0, \mu(x_0)) + V_\infty(f(x_0, \mu(x_0))), \tag{12.14}$$

where again we interpret $\mathbb{U}^1_{\mathbb{X}_{N-1}}(x_0)$ as a subset of U. Moreover, if an optimal control sequence $u^\star$ exists then the MPC-feedback law $\mu(x_0) = \mu_\infty(x_0) = u^*(0)$ satisfies both (12.13) and (12.14).

Proof: The proof is identical to the finite horizon counterpart Theorem 12.4. □

As in the finite horizon case, the following corollary shows that the feedback law (12.13) can be used in order to construct the optimal control sequence.

Corollary 12.10 Consider the optimal control problem (OCP_∞). Let $x_0 \in \mathbb{X}$ and consider an admissible feedback law $\mu : \mathbb{X} \to U$ in the sense of Definition 11.2(iv). Denote the solution of the closed loop system

$$x(0) = x_0, \quad x(k+1) = f(x(k), \mu_\infty(x(k))), \; k = 0, 1, \ldots \tag{12.15}$$

by x_μ, assume that μ_∞ satisfies (12.13) for initial values $x_0 = x_\mu(k)$ for all $k = 0, 1, \ldots$ and that

$$\lim_{k\to\infty} V_\infty(x_\mu(k)) \geq 0.$$

Then

$$u^\star(k) = \mu_\infty(x_{u^\star}(k, x_0)), \quad k = 0, 1, \ldots \tag{12.16}$$

is an optimal control sequence for initial time n and initial value x_0 and the solution of the closed loop system (12.15) is a corresponding optimal trajectory. □

Proof: From (12.16) for $x(n)$ from (12.15) we immediately obtain

$$x_{u^\star}(k) = x(k), \quad k = 0, 1, \dots.$$

Hence we need to show that

$$V_\infty(x_0) = J_\infty(x_0, u^\star).$$

Using (12.16) and (12.14) we get

$$V_\infty(x(k)) = \ell(x(k), u^\star(k)) + V_\infty(x(k+1))$$

for $k = 0, 1, \dots$. Summing these equalities for $k = 0, \dots, K-1$ for arbitrary $K \in \mathbb{N}$ and eliminating the identical terms $V_\infty(k, x_0)$, $k = 1, \dots, K-1$ on the left and on the right we obtain

$$V_\infty(x_0) = \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)) + V_\infty(x(K)).$$

Taking the upper limit for $K \to \infty$ and using $\lim_{k\to\infty} V(x_\mu(k)) \geq 0$ as well as the assumption on the lower limit after (OCP$_\infty$) implies that

$$\limsup_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)) \leq V_\infty(x_0) \leq \liminf_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k)),$$

which implies that the limit

$$\sum_{k=0}^{\infty} \ell(x(k), u^\star(k)) = \lim_{K\to\infty} \sum_{k=0}^{K-1} \ell(x(k), u^\star(k))$$

exists and equals $V_\infty(x_0)$. □

We note that the condition $\lim_{k\to\infty} V(x_\mu(k)) \geq 0$ is always satisfied when $\ell(x, u) \geq 0$ for all $x \in \mathbb{X}$, $u \in \mathbb{U}(x)$.

Corollary 12.10 implies that infinite horizon optimal control is nothing but MPC with $N = \infty$: Formula (12.16) for $k = 0$ yields that if we replace the optimization problem (OCP$_\mathrm{N}$) in Algorithm 11.1 by (OCP$_\infty$), then the feedback law resulting from this algorithm equals μ_∞. In fact, the infinite horizon problem can be seen as a discrete time nonlinear version of linear quadratic optimal control. Our last theorem (the only one that does not have a finite horizon counterpart in Section 12.1) shows that just like for linear quadratic optimal control, the optimal feedback law stabilizes an equilibrium, provided suitable inequalities are satisfied.

Theorem 12.11 [Asymptotic stability] Consider the optimal control problem (OCP$_\infty$) for the control system (8.2) and an equilibrium $x_* \in \mathbb{X}$. Assume that there exist $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{K}_\infty$ such that the inequalities

$$\alpha_1(|x|_{x_*}) \leq V_\infty(x) \leq \alpha_2(|x|_{x_*}) \quad \text{and} \quad \ell(x, u) \geq \alpha_3(|x|_{x_*}) \tag{12.17}$$

hold for all $x \in \mathbb{X}$ and $u \in U$. Assume furthermore that an optimal feedback μ_∞ exists, i.e., an admissible feedback law $\mu_\infty : \mathbb{X} \to U$ satisfying (12.13) for all $x \in \mathbb{X}$. Then this optimal feedback asymptotically stabilizes the closed loop system

$$x^+ = g(x) = f(x, \mu_\infty(x))$$

on $\mathbb{X}$ in the sense of Definition 10.2.

Proof: For the closed loop system, (12.14) and the last inequality in (12.17) yield

$$\begin{aligned} V_\infty(x) &= \ell(x, \mu_\infty(x)) + V_\infty(f(x, \mu_\infty(x))) \\ &\geq \alpha_3(|x|_{x_*}) + V_\infty(f(x, \mu_\infty(x))). \end{aligned}$$

Together with the first two inequalities in (12.17) this shows that V_∞ is a Lyapunov function on $\mathbb{X}$ in the sense of Definition 10.4 with $\alpha_V = \alpha_3$. Thus, Theorem 10.5 yields asymptotic stability on $\mathbb{X}$. □

Chapter 13

Analysis of general MPC schemes

13.1 Preliminaries

In this section we analyze the properties of the MPC closed-loop (11.2) for "general" stage costs ℓ. Of course, it is easy to see that ℓ cannot be completely general. Some properties must be met in order to obtain good closed-loop behavior and one of the main tasks in this chapter will be to figure out what these properties are. In the literature, this class of MPC schemes is often called "economic" MPC, because in practice the stage cost often models some economic goal, like maximal yield or minimum energy consumption. The next example is a very simply optimal control problem which falls into the last class.

Example 13.1 An example, which will serve as an illustration for all results in this section, is the 1d discrete-time system with dynamics and stage cost

$$x^+ = 2x + u \quad \text{and} \quad \ell(x,u) = u^2$$

and state and control constraint sets $\mathbb{X} = [-2,2]$ and $\mathbb{U}(x) = \mathbb{U} = [-3,3]$, i.e., $\mathbb{Y} = [-2,2] \times [-3,3]$.

The uncontrolled system is unstable, hence for initial values $x_0 \neq 0$ the solution will leave the admissible set $\mathbb{X}$ if no control is used. Hence, control action is needed in order to keep the system inside $\mathbb{X}$. Interpreting the stage cost $\ell(x,u) = u^2$ as the energy of the current control action, the control objective can be formulated as "keep the state inside $\mathbb{X}$ with minimal control effort". □

In what follows, two aspects will be investigated: the qualitative property of the MPC closed-loop trajectory (as, e.g., stability) and its quantitative properties measured in terms of the stage cost function. For the second purpose, three different quantities can be considered:

The first quantity is the infinite horizon closed-loop performance

$$J_\infty^{cl}(x_0,\mu) := \sum_{k=0}^{\infty} \ell(x_\mu(k), \mu(x_\mu(k))).$$

This would be the "natural" measure if we consider MPC as an approximation to an infinite horizon problem. However, as the infinite sum may not converge, we also look at other measures. We also consider the finite horizon closed-loop performance

$$J_K^{cl}(x_0,\mu) := \sum_{k=0}^{K-1} \ell(x_\mu(k),\mu(x_\mu(k))) \tag{13.1}$$

and the averaged infinite horizon performance

$$\overline{J}_\infty^{cl}(x_0,\mu) := \limsup_{K\to\infty} \frac{1}{K} J_K^{cl}(x_0,\mu).$$

Throughout this chapter by $(x^e,u^e)\in\mathbb{Y}$ we denote an equilibrium of the system, i.e., $f(x^e,u^e)=x^e$. Of particular interest are optimal equilibria according to the following definition.

Definition 13.2 An equilibrium $(x^e,u^e)\in\mathbb{Y}$ is called an *optimal equilibrium* if it yields the lowest value of the cost function among all admissible equilibria, i.e.,

$$\ell(x^e,u^e)\le\ell(x,u) \qquad \text{for all } (x,u)\in\mathbb{Y} \text{ with } f(x,u)=x.$$

□

Example 13.3 In Example 13.1, the equilibria are of the form $(x,-x)$ with cost $\ell(x,-x)=x^2$. Thus, the (unique) optimal equilibrium is given by $(x^e,u^e)=(0,0)$. □

The following lemma shows that an optimal equilibrium always exists when f and ℓ are continuous and $\mathbb{Y}$ is compact.

Lemma 13.4 If the constraint set $\mathbb{Y}\subset X\times U$ is compact, the maps $\ell:\mathbb{X}\times\mathbb{U}\to\mathbb{R}$ and $f:\mathbb{X}\times\mathbb{U}\to\mathbb{X}$ are continuous, and there exists an equilibrium in $\mathbb{Y}$, then there exists an optimal equilibrium, i.e., a pair $(x^e,u^e)\in\mathbb{Y}$ with $f(x^e,u^e)=x^e$ such that

$$\ell(x^e,u^e)=\inf\{\ell(x,u)\,|\,(x,u)\in\mathbb{Y}, f(x,u)=x\}.$$

Proof: Since pre-images of closed sets under continuous mappings are closed, the set $\{(x,u)\in\mathbb{Y}\,|\,f(x,u)=x\}$ is closed, hence compact, and nonempty. Thus, the continuous function ℓ attains a minimum. □

Hence, assuming the existence of an optimal equilibrium is not an overly restrictive assumption.

13.2 Averaged performance with terminal conditions

In this and in the following two sections we consider the MPC Algorithm 11.9 with optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$). We note that the terminal condition is only added to the open-loop functional $J_N(x_0, u)$ used in the MPC Algorithm 11.9 but not to the closed-loop performance index $J_K^{cl}(x, \mu)$ from (13.1), which is still defined without terminal cost or constraints according to (13.1). As before, the optimal value function is defined by

$$V_N(x) := \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x)} J_N(x, u(\cdot))$$

and we assume the existence of an optimal control sequence for each feasible initial condition x in order to synthesize the MPC feedback law μ_N according to Algorithm 11.9.

The following assumption links an equilibrium—which will later be chosen as an optimal equilibrium—to the terminal conditions. For its formulation, recall the definition of the feasible sets $\mathbb{X}_N$ from Definition 11.8(i).

Assumption 13.5 [Terminal conditions] (a) The set $\mathbb{X}_0$ is bounded and there is an equilibrium $(x^e, u^e) \in \mathbb{Y}$ with $x^e \in \mathbb{X}_0$ and $F(x^e) = 0$ such that for each $x \in \mathbb{X}_0$ there exists $u \in \mathbb{U}$ with $f(x, u) \in \mathbb{X}_0$ and

$$F(f(x, u)) \leq F(x) - \ell(x, u) + \ell(x^e, u^e)$$

(b) There exists $N_0 \in \mathbb{N}$ and $\eta > 0$ such that $\mathbb{X}_{N_0}$ contains the ball $\mathcal{B}_\eta(x^e)$. □

Condition (a) is a compatibility condition between the stage cost ℓ and the terminal cost F. The simplest way to satisfy condition (a) is by setting $\mathbb{X}_0 = \{x^e\}$ and $F \equiv 0$. However, using a terminal constraint set with only one point may cause convergence problems in the numerical optimization routine for solving ($\mathrm{OCP}_{\mathrm{N,e}}$). For ℓ with $\ell(x^e, u^e) = 0$ and $\ell(x, u) > 0$ otherwise, a systematic way to construct F with this property is via a linear quadratic approximation of the problem near x^e.

Observe that the requirement $F(x^e) = 0$ in Assumption 13.5(a) can be made without loss of generality because the inequality is invariant with respect to adding a constant to F. Assumption 13.5(b) is a nondegeneracy condition which prevents that the feasible sets $\mathbb{X}_N$ have empty interior for any $N \in \mathbb{N}$.

Under these assumptions we can formulate the first result.

Theorem 13.6 Consider the MPC Algorithm 11.9. Let Assumption 13.5(a) be satisfied, let $N \geq 2$ and assume V_N is bounded from below on $\mathbb{X}_N$. Then, for any $N \geq 2$ and any $x \in \mathbb{X}_N$ the averaged closed-loop performance satisfies the inequality

$$\overline{J}^{cl}_\infty(x, \mu_N) \leq \ell(x^e, u^e). \tag{13.2}$$

Proof: Let $\hat{x} \in \mathbb{X}_{N-1}$ and let $\hat{u}^\star$ be the optimal control sequence for this initial value with horizon $N-1$, i.e., $V_{N-1}(\hat{x}) = J_{N-1}(\hat{x}, \hat{u}^\star)$. Let $\tilde{u}$ be the control value from Assumption 13.5(a) for $\tilde{x} = x_{\hat{u}^\star}(N-1, \hat{x})$. Then, for the control sequence $u = (\hat{u}^\star(0), \ldots, \hat{u}^\star(N-1), \tilde{u})$ we obtain $x_u(N, \hat{x}) = f(\tilde{x}, \tilde{u})$ and thus Assumption 13.5(a) implies

$$\begin{aligned} V_N(\hat{x}) &\leq J_N(\hat{x}, u) \\ &= J_{N-1}(\hat{x}, \hat{u}^\star) - F(\tilde{x}) + \ell(\tilde{x}, \tilde{u}) + F(f(\tilde{x}, \tilde{u})) \\ &\leq V_{N-1}(\hat{x}) + \ell(x^e, u^e) \end{aligned}$$

Using the dynamic programming principle this inequality applied with $\hat{x} = f(x, \mu_N(x))$ implies

$$\ell(x, \mu_N(x)) = V_N(x) - V_{N-1}(f(x, \mu_N(x))) \leq V_N(x) - V_N(f(x, \mu_N(x))) + \ell(x^e, u^e)$$

and we can conclude

$$\begin{aligned} J_K^{cl}(x_0, \mu_N) &= \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k))) \\ &\leq \sum_{k=0}^{K-1} \Big[V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e, u^e) \Big] \\ &= V_N(x_0) - V_N(x_{\mu_N}(K)) + K\ell(x^e, u^e) \\ &\leq V_N(x_0) - M + K\ell(x^e, u^e), \end{aligned}$$

where $M \in \mathbb{R}$ is a lower bound on V_N. This yields

$$\overline{J}_\infty^{cl}(x_0, \mu_N) \leq \limsup_{K \to \infty} \left(\frac{V_N(x_0)}{K} - \frac{M}{K} + \ell(x^e, u^e) \right) = \ell(x^e, u^e).$$

□

We note that the boundedness assumption on V_N is satisfied if ℓ is continuous, $\mathbb{Y}$ is compact and F is bounded from below, because in this case both ℓ and F, and thus also V_N, are bounded from below.

Clearly, the estimate from Theorem 13.6 is particularly powerful if $\ell(x^e, u^e)$ is the best, i.e., the smallest possible value that $\overline{J}_\infty^{cl}(x_0, \mu_N)$ can attain. The next definition provides a property which is sufficient for this fact, as the subsequent Proposition 13.9 shows.

Definition 13.7 [Dissipativity and strict dissipativity] We say that an optimal control problem with stage cost ℓ is *strictly dissipative* at an equilibrium $(x^e, u^e) \in \mathbb{Y}$ if there exists a *storage function* $\lambda : \mathbb{X} \to \mathbb{R}$ bounded from below and satisfying $\lambda(x^e) = 0$, and a function $\rho \in \mathcal{K}_\infty$ such that for all $(x, u) \in \mathbb{Y}$ the inequality

$$\ell(x, u) - \ell(x^e, u^e) + \lambda(x) - \lambda(f(x, u)) \geq \rho(|x|_{x^e}) \tag{13.3}$$

holds.

We say that an optimal control problem with stage cost ℓ is *dissipative* at (x^e, u^e) if the same conditions hold with $\rho \equiv 0$. □

We note that the assumption $\lambda(x^e) = 0$ can be made without loss of generality because adding a constant to λ does not invalidate (13.3).

The classical physical interpretation of the storage function is that $\lambda(x)$ quantifies the amount of energy stored in the system at state x. The function $s(x,u) := \ell(x,u) - \ell(x^e,u^e)$ is called the *supply rate* and measures the (possibly negative) amount of energy supplied to the system via the input u at state x. With this interpretation, strict dissipativity then means that a certain amount of energy, quantified by $\rho(|x|_{x^e})$, is dissipated to the environment in each time step. Of course, in the context of general optimal control problems considered in this chapter the storage function and the supply rate need not have an energy interpretation.

Example 13.8 (i) Any optimal control problem with stage cost satisfying $\ell(x^e,u^e) = 0$ and $\ell(x,u) \geq \rho(|x - x^e|)$ is strictly dissipative with $\lambda \equiv 0$. Hence, MPC problems with stage cost penalizing the distance to a desired equilibrium $x_* = x^e$, as they typically appear in stabilization problems, are always strictly dissipative.

(ii) It is straightforward to check that Example 13.1 is dissipative with $\lambda \equiv 0$ and strictly dissipative with $\lambda(x) = -x^2/2$, both at $(x^e,u^e) = (0,0)$. Note that the storage function $\lambda = -x^2/2$ is bounded from below since $\mathbb{X}$ is bounded. Indeed, for an unbounded state constraint set $\mathbb{X}$ the system would not be strictly dissipative. In this example, the supply rate $s(x,u) = \ell(x,u) = u^2$ does have an energy interpretation and the storage function $\lambda(x)$ shows that the equilibrium (x^e,u^e) is the state in which the stored energy $\lambda(x)$ becomes maximal.

(iii) A somewhat more involved computation shows that the second example from Section 9.1 is strictly dissipative at $x^e = 1/\sqrt[\alpha-1]{\alpha A}$ with storage function $\lambda(x) = \alpha(x - x^e)/x^e$.

(iv) For linear quadratic problems with $Q = C^TC$ and $\mathbb{X} = \mathbb{R}$ it can be shown that strict dissipativity is equivalent to detectability of (A,C), i.e., there are no unobservable eigenvalues $\lambda \in \mathbb{C}$ with $|\lambda| \geq 0$. If $\mathbb{X} \subset \mathbb{R}$ is compact, then strict dissipativity is equivalent to the fact that no unobservable eigenvalues with $|\lambda| = 1$ exist. □

Proposition 13.9 For an optimal control problem (OCP_N) that is dissipative at (x^e,u^e), the point (x^e,u^e) is an optimal equilibrium and the inequality

$$\limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) \geq \ell(x^e,u^e) \tag{13.4}$$

holds for all $x \in \mathbb{X}$ and all admissible control sequences $u \in \mathbb{U}^\infty(x)$. □

Proof: Consider an arbitrary equilibrium $(x,u) \in \mathbb{Y}$. Then the identity $x = f(x,u)$ and (13.3) imply

$$\ell(x,u) - \ell(x^e,u^e) = \ell(x,u) - \ell(x^e,u^e) + \lambda(x) - \lambda(f(x,u)) \geq 0$$

which yields $\ell(x^e,u^e) \leq \ell(x,u)$ and thus (x^e,u^e) is an optimal equilibrium.

Moreover, using again (13.3) and denoting by M a lower bound on λ we have

$$\begin{aligned}\sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) &\geq \sum_{k=0}^{K-1} \ell(x^e,u^e) - \lambda(x_u(k,x)) + \lambda(x_u(k+1,x))\\ &= K\ell(x^e,u^e) - \lambda(x) + \lambda(x_u(K,x))\\ &\geq K\ell(x^e,u^e) - \lambda(x) + M\end{aligned}$$

for any $u \in \mathbb{U}^\infty(x)$. This yields

$$\limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) \geq \limsup_{K\to\infty}\left(\ell(x^e,u^e) - \frac{\lambda(x)-M}{K}\right) = \ell(x^e,u^e).$$

□

The property expressed by inequality (13.4) is known as *optimal operation at steady state*. It has been shown in [15] that under a controllability condition on the system the converse of Proposition 13.9 is also true, i.e., that optimal operation at a steady state implies dissipativity.

An immediate consequence of Proposition 13.9 is the following corollary.

Corollary 13.10 Consider the MPC Algorithm 11.9 with dissipative optimal control problem ($\mathrm{OCP_{N,e}}$). Then for all $x \in \mathbb{X}_N$

$$\overline{J}^{cl}_\infty(x,\mu_N) = \inf_{u\in\mathbb{U}^\infty(x)} \limsup_{K\to\infty}\frac{1}{K}\sum_{k=0}^{K-1}\ell(x_u(k,x),u(k)).$$

□

Hence, if dissipativity holds, then Theorem 13.6 ensures infinite horizon averaged optimality of the MPC closed loop.

Example 13.11 Since Example 13.1 is dissipative (see Example 13.8(ii)), the MPC closed loop must be infinite horizon averaged optimal. Indeed, as Fig. 13.1 shows, the closed-loop solution converges to the optimal equilibrium. Since the control (not shown in the figure) does the same, $\ell(x_{\mu_N}(k),\mu_N(x_{\mu_N}(k))) \to 0$ as $k\to\infty$ follows which implies $\overline{J}^{cl}_\infty(x,\mu_N) = 0$, which is clearly optimal since $\ell \geq 0$. □

13.3 Asymptotic stability with terminal conditions

One might conjecture that optimal operation at the steady state (x^e,u^e) implies that closed-loop solutions satisfying (13.2) must also converge to x^e. However, under the assumptions imposed in Theorem 13.6 and Proposition 13.9 this is not necessarily the case. To see this, it suffices to consider an optimal control problem with $\ell \equiv 0$. Such a problem clearly satisfies all assumptions (with terminal cost $F \equiv 0$ and storage function $\ell \equiv 0$), yet every trajectory

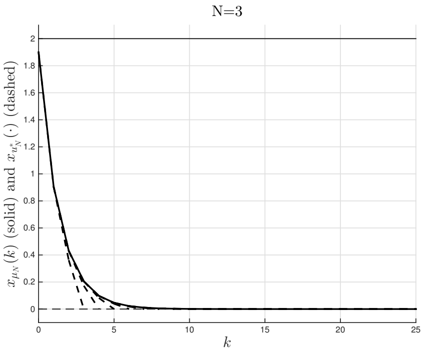


Figure 13.1: MPC closed-loop solution (solid) and open-loop predictions (dashed) for Example 13.1 with terminal constraint $\mathbb{X}_0 = \{0\}$ and horizon $N = 3$. The solid line at $x = 2$ indicates the upper bound of the admissible set $\mathbb{X}$

is an optimal trajectory and thus optimal trajectories obviously need not converge to x^e. In order to achieve this — and, in fact, even asymptotic stability of x^e — we need to assume strict dissipativity.

Under this assumption, we establish asymptotic stability by proving the existence of a Lyapunov function. This Lyapunov function will be built from the optimal value function of an auxiliary optimal control problem with stage cost

$$\tilde{\ell}(x,u) := \ell(x,u) - \ell(x^e,u^e) + \lambda(x) - \lambda(f(x,u)) \tag{13.5}$$

and terminal cost

$$\widetilde{F}(x) := F(x) + \lambda(x).$$

These costs are usually called *rotated* or *modified* costs. The name "rotated cost" stems from the fact that for linear f and strictly convex ℓ the graph of $\tilde{\ell}$ is obtained by rotating the graph of ℓ. The corresponding functional is given by

$$\widetilde{J}_N(x_0,u(\cdot)) = \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x_0),u(k)) + \widetilde{F}(x_u(N,x_0))$$

and the optimal value function by

$$\widetilde{V}_N(x_0) := \inf_{u(\cdot)\in\mathbb{U}^N_{\mathbb{X}_0}(x_0)} \widetilde{J}_N(x_0,u(\cdot)).$$

It is an easy exercise to check that the equalities $\tilde{\ell}(x^e,u^e) = 0$ and $\widetilde{F}(x^e) = 0$ and — under Assumption 13.5(a) — that the inequality

$$\widetilde{F}(f(x,u)) \leq \widetilde{F}(x) - \tilde{\ell}(x,u) \tag{13.6}$$

holds for each $x \in \mathbb{X}_0$ and the control u from Assumption 13.5(a). Moreover, for any $x \in \mathbb{X}_N$ and $u \in \mathbb{U}^N_{\mathbb{X}_0}(x)$ one easily verifies the identity

$$\widetilde{J}_N(x,u) = J_N(x,u) + \lambda(x) - N\ell(x^e,u^e). \tag{13.7}$$

Since the last two terms in (13.7) are independent of u, this implies that the optimal trajectories for J_N and $\widetilde{J}_N$ coincide and that the optimal value functions satisfy

$$\widetilde{V}_N(x) = V_N(x) + \lambda(x) - N\ell(x^e, u^e). \tag{13.8}$$

Since $\tilde{\ell}(x^e, u^e) = 0$ and $\widetilde{F}(x^e) = 0$, using the constant control $u \equiv u^e$ yields

$$\widetilde{V}_N(x^e) \leq \widetilde{J}_N(x^e, u) = 0 \quad \text{and thus} \quad V_N(x^e) \leq J_N(x^e, u) = N\ell(x^e, u^e) \tag{13.9}$$

using (13.8) and $\lambda(x^e) = 0$.

We now turn to show that $\widetilde{V}_N$ is a Lyapunov function for the MPC closed-loop system. For the rigorous proof of this property, we need the following continuity assumption on F, λ and V_N in x^e.

Assumption 13.12 [Continuity of F, λ and V_N at x^e] There exists γ_F, γ_λ and $\gamma_V \in \mathcal{K}_\infty$ such that the following properties hold.
(a) For all $x \in \mathbb{X}_0$ it holds that

$$|F(x) - F(x^e)| \leq \gamma_F(|x|_{x^e}).$$

(b) For all $x \in \mathbb{X}$ it holds that

$$|\lambda(x) - \lambda(x^e)| \leq \gamma_\lambda(|x|_{x^e}).$$

(c) For each $N \in \mathbb{N}$ and each $x \in \mathbb{X}_N$ it holds that

$$|V_N(x) - V_N(x^e)| \leq \gamma_V(|x|_{x^e}).$$

□

Note that γ_V in (c) is independent of N. We will comment at the end of this section on conditions under which (c) can be ensured.

Theorem 13.13 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem ($\mathrm{OCP}_{\mathrm{N,e}}$) and bounded $\mathbb{X}_0$. Let Assumptions 13.5(a) and 13.12 be satisfied. Then the optimal equilibrium x^e is asymptotically stable for the MPC closed loop on $\mathbb{X}_N$.

Proof: We show that the modified optimal value function $\widetilde{V}_N$ is a Lyapunov function for the closed-loop system in the sense of Definition 10.4 for $x_* = x^e$. Then the assertion follows from Theorem 10.5. To this end we first check an auxiliary inequality. As in the proof of Theorem 13.6, from Assumption 13.5(a) we obtain $\ell(x, \mu_N(x)) \leq V_N(x) - V_N(f(x, \mu_N(x))) + \ell(x^e, u^e)$ which we can rewrite as

$$V_N(x) \geq \ell(x, \mu_N(x)) + V_N(f(x, \mu_N(x))) - \ell(x^e, u^e). \tag{13.10}$$

Using (13.8) this implies

$$\begin{aligned}
\widetilde{V}_N(x) &= V_N(x) + \lambda(x) - N\ell(x^e, u^e)\\
&\geq \ell(x, \mu_N(x)) + V_N(f(x, \mu_N(x))) - \ell(x^e, u^e) + \lambda(x) - N\ell(x^e, u^e)\\
&= \ell(x, \mu_N(x)) + \widetilde{V}_N(f(x, \mu_N(x))) - \lambda(f(x, \mu_N(x))) - \ell(x^e, u^e) + \lambda(x)\\
&= \tilde{\ell}(x, \mu_N(x)) + \widetilde{V}_N(f(x, \mu_N(x))).
\end{aligned}$$

In order to check that $\widetilde{V}_N$ satisfies Definition 10.4, we now have to check the inequalities

$$\alpha_1(|x|_{x^e}) \leq \widetilde{V}_N(x) \leq \alpha_2(|x|_{x^e}) \quad \text{and} \quad \tilde{\ell}(x, u) \geq \alpha_3(|x|_{x^e}) \tag{13.11}$$

for α_1, α_2, $\alpha_3 \in \mathcal{K}_\infty$. The third inequality follows immediately from the definition of $\tilde{\ell}$ and strict dissipativity for $\alpha_3 = \rho$ from Definition 13.7. For the inequalities involving α_1 and α_2 we first need to establish a lower bound for $\widetilde{F}$.

To this end, for each $x \in \mathbb{X}_0$ we denote the control u from (13.6) by $\mu_0(x)$. Then (13.6) and strict dissipativity implies

$$\widetilde{F}(f(x, \mu_0(x))) \leq \widetilde{F}(x) - \tilde{\ell}(x, \mu_0(x)) \leq \widetilde{F}(x) - \rho(|x|_{x^e}).$$

By induction along the closed-loop solution for the feedback law μ_0 we then obtain

$$\widetilde{F}(x_{\mu_0}(K, x)) \leq \widetilde{F}(x) - \sum_{k=0}^{K-1} \rho(|x_{\mu_0}(k, x)|_{x^e}).$$

This implies that $x_{\mu_0}(K, x) \to x^e$ as $K \to \infty$, because otherwise the sum on the right hand side of this inequality grows unboundedly which implies $\widetilde{F}(x_{\mu_0}(K, x)) \to -\infty$ and contradicts Assumption 13.12(a) and (b) since $x_{\mu_0}(K, x)$ is contained in the bounded set $\mathbb{X}_0$. Again by Assumption 13.12(a) and (b) this implies $\widetilde{F}(x_{\mu_0}(K, x)) \to \widetilde{F}(x^e) = 0$ as $K \to \infty$ from which we can finally conclude

$$\widetilde{F}(x) \geq \lim_{K \to \infty} \sum_{k=0}^{K-1} \rho(|x_{\mu_0}(k, x)|_{x^e}) \geq \rho(|x|_{x^e}) \geq 0.$$

From this, the definitions of $\widetilde{J}_N$ and $\widetilde{V}_N$ immediately imply $\widetilde{V}_N(x) \geq \tilde{\ell}(x, \mu_N(x)) \geq \rho(|x|_{x^e})$ and thus the inequality for α_1 in (13.11) with $\alpha_1 = \rho$.

Together with (13.9) this implies $\widetilde{V}_N(x^e) = 0$ and the second inequality in (13.11) follows from (13.8) and Assumption 13.12(b) and (c) with $\alpha_2 = \gamma_\lambda + \gamma_V$. □

Observe that in the case of stabilizing stage costs according to Example 13.8(i), we obtain $\lambda \equiv 0$ and $\ell(x^e, u^e) = 0$, and thus $\widetilde{V}_N = V_N$. This implies that the optimal value function itself is a Lyapunov function.

We end this section by discussing sufficient conditions for the bound on V_N required in Assumption 13.12(c). In the case of equilibrium terminal conditions, i.e., $\mathbb{X}_0 = \{x^e\}$ and $F \equiv 0$, this property can be ensured by the condition that x^e is reachable from every $x \in \mathbb{X}_N$ with suitable bounded costs. In case ℓ and f are continuous, it is sufficient to assume that the control sequence steering x to x^e is sufficiently close to the constant control with value u^e. For details we refer to [1], particularly to part 2 of Assumption 2 in [1].

In case $\mathbb{X}_0$ contains a neighborhood of x^e, using Assumption 13.5(a) inductively yields the inequality

$$V_N(x) \le F(x) + N\ell(x^e, u^e)$$

while from (13.8) and $\widetilde{V}_N \ge 0$ we obtain

$$V_N(x) \ge -\lambda(x) + N\ell(x^e, u^e).$$

Since from (13.9) we moreover know $V_N(x^e) = N\ell(x^e, u^e)$, this implies Assumption 13.12(c) for $x \in \mathbb{X}_0$ provided Assumption 13.12(a) and (b) hold. For $x \in \mathbb{X}_N \setminus \mathbb{X}_0$ the inequality follows from boundedness of V_N which in turn follows from boundedness of ℓ along the optimal trajectories.

Example 13.14 According to Example 13.8, the optimal control problem from Example 13.1 is strictly dissipative. Moreover, one easily verifies that x^e is reachable in two steps from each $x \in \mathbb{X}$ with cost $4x^2$, which implies the upper bound on V_N for the terminal constraint set $\mathbb{X}_0 = \{0\}$. Hence, we expect the MPC closed loop to be asymptotically stable, which was already illustrated in Fig. 13.1. □

13.4 Non-averaged performance with terminal conditions

The averaged performance result from Theorem 13.6 provides a useful estimate for large times k. However, it also has two significant weaknesses. First, it does not provide an advantage over a stabilizing MPC algorithm. Indeed, for any combination of a continuous stage cost and a terminal condition for which the MPC closed-loop solution converges to x^e and the corresponding control sequence converges to u^e, the value $\ell(x_{\mu_N}(k), \mu_N(x_{\mu_N}(k)))$ converges to $\ell(x^e, u^e)$ from which $\overline{J}^{cl}_\infty(x, \mu_N) = \ell(x^e, u^e)$ follows. Hence, Theorem 13.6 only states that the economic MPC scheme does not perform worse than a stabilizing one. Second, the averaged estimate does not allow any statement about the finite time behavior of the closed-loop trajectory. Indeed, on any finite time interval of arbitrary length the closed-loop trajectory could behave arbitrarily bad as long as eventually it converges to the equilibrium. Clearly, this is not what we would expect an MPC closed-loop trajectory to do and it is also not consistent with what we see in numerical simulations, e.g., in Fig. 13.1. Hence, in this section we derive estimates for the non-averaged infinite and finite horizon performance $J^{cl}_\infty(x, \mu_N)$ and $J^{cl}_K(x, \mu_N)$, respectively. For the infinite horizon estimate the additional condition $\ell(x^e, u^e) = 0$ will be imposed, in order to make sure that the infinite sum in $J^{cl}_\infty(x, \mu_N)$ converges. For the finite horizon performance, such a condition is not needed. As we already know that — under the conditions of Theorem 13.13 — the equilibrium x^e is asymptotically stable, the finite horizon value $J^{cl}_K(x, \mu_N)$ measures the performance of the solution during the transient phase, i.e., until it reaches a small neighborhood of x^e. This is why we also call this value *transient performance.*

Since $J^{cl}_K(x, \mu_N)$ and $J^{cl}_\infty(x, \mu_N)$ do not involve any terminal constraints or costs, in our analysis we will also need to consider the optimal control problems $(\mathrm{OCP_N})$ and (OCP_∞) without terminal constraints and terminal costs. In order not to confuse these problems with those using terminal conditions, in this section we denote the functionals and the optimal value functions of the unconstrained problems $(\mathrm{OCP_N})$ and (OCP_∞) by J^{uc}_N, V^{uc}_N,

J_∞^{uc} and V_∞^{uc},respectively. We emphasize that we use the same stage cost ℓ in all problems. This implies that if one of the problems is strictly dissipative then all problems are. If this is the case, we also consider (OCP$_\mathrm{N}$) for the rotated cost $\tilde{\ell}$ and denote the corresponding functional by $\widetilde{J}_N^{uc}$. A straightforward computation reveals that J_N^{uc} and $\widetilde{J}_N^{uc}$ are related by the identity

$$\widetilde{J}_N^{uc}(x,u) = J_N^{uc}(x,u) + \lambda(x) - \lambda(x_u(N,x)) - N\ell(x^e,u^e). \tag{13.12}$$

Observe that compared to (13.7) the additional term $\lambda(x_u(N,x))$ appears here due to the absence of the terminal conditions.

In order to establish our theorems on transient performance, we will need a few preparatory results. The first statement shows that the finite horizon optimal trajectories most of the time stay close to the optimal equilibrium x^e.

Proposition 13.15 Assume that the optimal control problem (OCP$_\mathrm{N}$) is strictly dissipative with bounded storage function λ and $\rho \in \mathcal{K}_\infty$. Then for each $\delta > 0$ there exists $\sigma_\delta \in \mathcal{L}$ such that for all $N, P \in \mathbb{N}$, $x \in \mathbb{X}$ and $u \in \mathbb{U}^N(x)$ with $J_N^{uc}(x,u) \le N\ell(x^e,u^e) + \delta$, the set $\mathcal{Q}(x,u,P,N) := \{k \in \{0,\dots,N-1\} \,|\, |x_u(k,x)|_{x^e} \ge \sigma_\delta(P)\}$ has at most P elements. □

Proof: We fix $\delta > 0$ and claim that the assertion holds with $\sigma_\delta(P) := \rho^{-1}((2M+\delta)/P)$ where M is a bound on $|\lambda|$. To prove this claim, assume that there are N, P, x and u such that $J_N^{uc}(x,u) \le N\ell(x^e,u^e) + \delta$ but $\mathcal{Q}(x,u,P,N)$ contains at least $P+1$ elements. Then from (13.12) we can estimate

$$\widetilde{J}_N^{uc}(x,u) \le J_N^{uc}(x,u) + 2M - N\ell(x^e,u^e) \le 2M + \delta.$$

On the other hand, (13.3), (13.5) and the fact that $\mathcal{Q}(x,u,P,N)$ contains at least $P+1$ elements imply

$$\begin{aligned}\widetilde{J}_N^{uc}(x,u) \;&\ge\; \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x),u(k)) \;\ge\; \sum_{k=0}^{N-1} \rho(|x_u(k,x)|_{x^e}) \;\ge\; \sum_{\substack{k\in\{0,\dots,N-1\}\\ |x_u(k,x)|_{x^e} > \sigma_\delta(P)}} \rho(\sigma_\delta(P))\\ &\ge\; (P+1)\rho(\sigma_\delta(P)) \;\ge\; (P+1)\frac{2M+\delta}{P} \;>\; 2M+\delta\end{aligned}$$

which is a contradiction. □

We denote the property described by Proposition 13.15 as the *turnpike property*. For an illustration we refer to Fig. 13.2. In fact, there are various variants of the turnpike property known in optimal control, of which the one described by Proposition 13.15 is just a particular version.

We remark that the boundedness assumption on λ can be restrictive in case $\mathbb{X}$ is unbounded. However, for bounded subsets of the state constraint set $\mathbb{X}$ it is not a very strong assumption. Hence, it can be assumed to hold if either $\mathbb{X}$ itself is bounded or if near optimal trajectories are guaranteed to stay in a bounded subset of $\mathbb{X}$.

Example 13.16 Since Example 13.1 is strictly dissipative with bounded storage function (cf. Example 13.8), we expect the system to have the turnpike property. The numerical optimal trajectories depicted in Fig. 13.3 support this claim. □

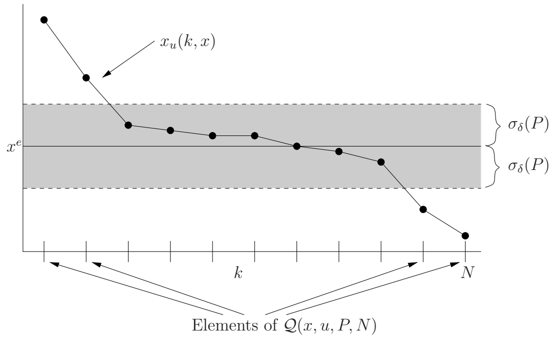


Figure 13.2: Illustration of the set $\mathcal{Q}(x,u,P,N)$ defined in Proposition 13.15

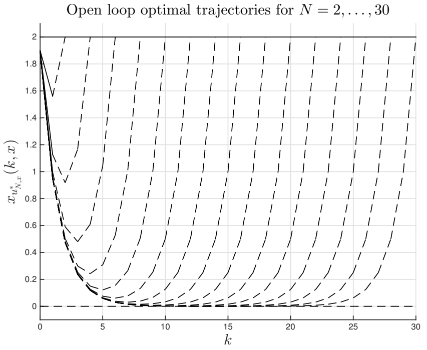


Figure 13.3: Open loop optimal trajectories (without terminal conditions) for Example 13.1 with different optimization horizons N. The turnpike property is clearly visible

Next we derive upper and lower bounds for V_∞^{uc}.

Lemma 13.17 Assume that the optimal control problem (OCP$_\mathrm{N}$) is strictly dissipative with bounded storage function λ, that $\ell(x^e,u^e) = 0$ and that Assumptions 13.5(a) and 13.12 hold. Then there is $C > 0$ such that the inequalities

$$-C \le V_\infty^{uc}(x) \le \gamma_V(|x|_{x^e}) \tag{13.13}$$

hold for all $x \in \bigcup_{N\in\mathbb{N}} \mathbb{X}_N$ with γ_V from Assumption 13.12(c).

Proof: For $x \in \mathbb{X}_N$, using the control sequence $u(k) = \mu_N(x_{\mu_N}(k,x))$ induced by the closed loop, from (13.10) with $\ell(x^e,u^e) = 0$ for any $K > 0$ we obtain

$$J_K^{uc}(x,u) \;=\; \sum_{k=0}^{K-1} \ell(x_u(k,x),u(k)) \le V_N(x) - V_N(x_u(K,x)).$$

By asymptotic stability of x^e for this solution we obtain $x_u(K,x) \to x^e$ and thus, since $V_N(x^e) = N\ell(x^e,u^e) = 0$ by (13.9), Assumption 13.12(c) yields $V_N(x_u(K,x)) \to 0$ as

$K \to \infty$. Using Assumption 13.12(c) and $V_N(x^e) = 0$, this implies

$$V_\infty^{uc}(x) \leq \limsup_{K\to\infty} J_K^{uc}(x,u) \leq V_N(x) \leq \gamma_V(|x|_{x^e}).$$

Moreover, the fact that $\widetilde{J}_N^{uc}(x,u) \geq 0$, (13.12) and the boundedness of λ imply $J_N^{uc}(x,u) \geq -C$ for some $C \geq 0$ and all x, u and N. This implies $V_\infty^{uc}(x) \geq -C$. □

Using the inequality ensured by this lemma we can prove an infinite horizon version of the turnpike property from Proposition 13.15.

Proposition 13.18 Assume that the optimal control problem $(\mathrm{OCP_N})$ is strictly dissipative, that $\mathbb{X}$ is bounded, that $\ell(x^e,u^e) = 0$ and that the inequalities (13.13) hold for all $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$. Then there exists $\sigma_\infty \in \mathcal{L}$ such that for all $P \in \mathbb{N}$, $x \in \mathbb{X}$ and $u \in \mathbb{U}^\infty(x)$ with $J_\infty^{uc}(x,u) \leq V_\infty^{uc}(x) + 1$, the set $\mathcal{Q}(x,u,P,\infty) := \{k \in \mathbb{N}_0 \,|\, |x_u(k,x)|_{x^e} \geq \sigma_\infty(P)\}$ has at most P elements. □

Proof: First note that by Lemma 13.17 and the assumption we get

$$J_\infty^{uc}(x,u) \leq \sup_{x\in\bigcup_{N\in\mathbb{N}}\mathbb{X}_N} V_\infty^{uc}(x) + 1 \leq \sup_{x\in\mathbb{X}} \gamma_V(|x|_{x^e}) + 1 =: \delta.$$

Now we can proceed as in the proof of Proposition 13.15: denoting by M a bound on $|\lambda|$, from (13.12) and $\ell(x^e,u^e) = 0$ we obtain

$$\widetilde{J}_\infty^{uc}(x,u) = \limsup_{K\to\infty} \widetilde{J}_K^{uc}(x,u) \leq \limsup_{K\to\infty} J_K^{uc}(x,u) + 2M \leq \delta + 2M.$$

Setting $\sigma_\infty(K) := \rho^{-1}((2M+\delta)/K)$, the assumption that $\mathcal{Q}(x,u,P,\infty)$ contains more than P elements then again yields a contradiction to this inequality. □

We note that this theorem implies $x_u(k,x) \to x^e$ as $k \to \infty$, because otherwise $\mathcal{Q}(x,u,P,\infty)$ would contain infinitely many elements for sufficiently large $P \in \mathbb{N}$. Using this fact we can improve the lower bound on V_∞^{uc} from Lemma 13.17.

Lemma 13.19 Under the assumptions of Proposition 13.18, the inequality $V_\infty^{uc}(x) \geq -\lambda(x)$ holds for all $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$.

Proof: Let $u \in \mathbb{U}^\infty(x)$ be such that $J_\infty^{uc}(x,u) \leq V_\infty^{uc}(x) + \varepsilon$ for an $\varepsilon \in (0,1)$. As explained above, Proposition 13.18 implies that $x_u(k,x) \to x^e$ as $k \to \infty$. The definition of V_∞^{uc} and (13.12) then imply that

$$\begin{aligned} V_\infty^{uc}(x) + \varepsilon &\geq \limsup_{K\to\infty} J_K^{uc}(x,u) \\ &= \limsup_{K\to\infty} \Big(-\lambda(x) + \underbrace{\widetilde{J}_K^{uc}(x,u)}_{\geq 0} + \underbrace{\lambda(x_u(K,x))}_{\to\lambda(x^e)=0} \Big) \geq -\lambda(x). \end{aligned}$$

This implies the assertion since $\varepsilon \in (0,1)$ was arbitrary. □

Our final preparatory result is needed for estimating the finite horizon transient performance. It thus concerns the optimal value of the problem with control functions u that steer a given initial value $x \in \mathbb{X}$ to the closed ball $\overline{\mathcal{B}}_\kappa(x^e)$ with radius $\kappa > 0$ around x^e. In order to simplify the notation, we briefly write

$$\mathbb{U}_\kappa^K(x) := \mathbb{U}^K_{\overline{\mathcal{B}}_\kappa(x^e)}(x) \tag{13.14}$$

using the notation from Definition 11.8 with $\overline{\mathcal{B}}_\kappa(x^e)$ in place of $\mathbb{X}_0$. We remark that Theorem 13.13 yields the existence of a $\beta \in \mathcal{KL}$ such that for all $x \in \mathbb{X}_N$ and all K with $\beta(|x|_{x^e}, K) \leq \kappa$ the control u obtained from the MPC feedback law via $u(k) = \mu_N(x_{\mu_N}(k,x))$ is contained in $\mathbb{U}_\kappa^K(x)$. This, in particular, shows that this set is nonempty for sufficiently large K.

The next lemma shows that the infimum of $J_K^{uc}(x,u)$ over $u \in \mathbb{U}_\kappa^K(x)$ and the corresponding approximately optimal trajectories behave similar to those of the infinite horizon problem. More precisely, part (a) of the following lemma is similar to Lemma 13.17, part (b) to Lemma 13.19 and part (c) to Proposition 13.18. Note that since we only consider finite horizon problems here, we do not need to assume $\ell(x^e,u^e) = 0$.

Lemma 13.20 Assume that the optimal control problem (OCP$_\mathrm{N}$) is strictly dissipative with bounded storage function λ and that Assumptions 13.5(a) and 13.12 hold. Fix $\kappa_0 > 0$ and let β be a $\mathcal{KL}$-function characterizing the asymptotic stability of the closed loop, whose existence is guaranteed by Theorem 13.13. Then for any $\kappa \in (0,\kappa_0]$, any $x \in \bigcup_{N\in\mathbb{N}_0} \mathbb{X}_N$ and $K_0 \in \mathbb{N}$ minimal with $\beta(|x|_{x^e}, K_0) \leq \kappa$, the following holds.

(a) For all $K \geq K_0$ the inequality

$$\inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) - K\ell(x^e,u^e) \leq \gamma_V(|x|_{x^e}) + \gamma_V(\kappa)$$

holds with $\gamma_V \in \mathcal{K}_\infty$ from Assumption 13.12(c).

(b) For all $K \in \mathbb{N}$ with $\mathbb{U}_\kappa^K(x) \neq \emptyset$ the inequality

$$-\gamma_\lambda(|x|_{x^e}) - \gamma_\lambda(\kappa) \leq \inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) - K\ell(x^e,u^e)$$

holds with γ_λ from Assumption 13.12(b).

(c) If in addition $\mathbb{X}$ is bounded then there exists $\sigma \in \mathcal{L}$ such that for all $K \geq K_0$, all $P \in \mathbb{N}$ and any $u \in \mathbb{U}_\kappa^K(x)$ with $J_K^{uc}(x,u) \leq \inf_{u\in\mathbb{U}_\kappa^K(x)} J_K^{uc}(x,u) + 1$ there is $k \leq \min\{P, K-1\}$ such that $|x_u(k,x)|_{x^e} < \sigma(\min\{P,K-1\})$.

Proof: (a) The proof of this inequality works similarly to the first part of the proof of Lemma 13.17. For $x \in \mathbb{X}_N$, we choose the control u obtained from the MPC feedback law via $u(k) = \mu_N(x_{\mu_N}(k,x))$. By Theorem 13.13 and the choice of K_0, this control lies in $\mathbb{U}_\kappa^K(x)$. As in the proof of Lemma 13.17, from (13.10) — now with $\ell(x^e,u^e) \neq 0$ — for this u we get

$$J_K^{uc}(x,u) \leq V_N(x) - V_N(x_u(K,x)) + K\ell(x^e,u^e)$$

and from Assumption 13.12(c) and $|x_u(K,x)|_{x^e} < \kappa$ we obtain the assertion.

(b) Let $\varepsilon > 0$ and take a control $u_\varepsilon \in \mathbb{U}^K_\kappa(x)$ with $\inf_{u\in\mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \varepsilon \geq J^{uc}_K(x,u_\varepsilon)$. Then by (13.12), Assumption 13.12(b) and $\lambda(x^e) = 0$, and recalling that strict dissipativity implies $\widetilde{J}^{uc}_K(x,u_\varepsilon) \geq 0$ we get

$$\begin{aligned}\inf_{u\in\mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \varepsilon &\geq J^{uc}_K(x,u_\varepsilon)\\ &= \underbrace{-\lambda(x)}_{\geq -\gamma_\lambda(|x|_{x^e})} + \underbrace{\widetilde{J}^{uc}_K(x,u_\varepsilon)}_{\geq 0} + \underbrace{\lambda(x_{u_\varepsilon}(K,x))}_{\geq -\gamma_\lambda(\kappa)} + K\ell(x^e,u^e)\\ &\geq -\gamma_\lambda(|x|_{x^e}) - \gamma_\lambda(\kappa) + K\ell(x^e,u^e).\end{aligned}$$

This implies (b) since $\varepsilon > 0$ was arbitrary.

(c) The assumptions and (a) imply that Proposition 13.15 can be applied with $\delta = \sup_{x\in\mathbb{X}} \gamma(|x|_{x^e}) + \gamma(\kappa_0) + 1$ for all $x \in \mathbb{X}$ and all $\kappa \in (0,\kappa_0]$. We set $\sigma = \sigma_\delta$ from this proposition. Since the set $\mathcal{Q}(x,u,\min\{P,K-1\},K)$ has at most $\min\{P,K-1\}$ elements, there exists at least one $k \in \{0,\dots,\min\{P,K-1\}\}$ with $k \notin \mathcal{Q}(x,u,\min\{P,K-1\},K)$, which thus satisfies $|x_u(k,x)|_{x^e} \leq \sigma(\min\{P,K-1\})$. □

We now have all the tools to prove the two main theorems of this section. The first theorem gives an upper bound for the non-averaged infinite horizon performance of the MPC closed-loop trajectory. We recall that when considering the infinite horizon problem we demand $\ell(x^e,u^e) = 0$. Taking into account the inequality $V^{uc}_\infty(x) \leq J^{cl}_\infty(x,\mu_N)$ which follows immediately from the definition of these functions, the theorem shows that economic MPC delivers an approximately (non-averaged) infinite horizon optimal closed-loop solution for which the approximation error tends to 0 as the horizon N tends to infinity.

Theorem 13.21 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem (OCP$_{\mathrm{N,e}}$). Assume that $\mathbb{X}$ is bounded, that $\ell(x^e,u^e) = 0$ and that Assumptions 13.5 and 13.12 hold. Then there exists $\delta_1 \in \mathcal{L}$ such that the inequalities

$$J^{cl}_\infty(x,\mu_N) \leq V_N(x) \leq V^{uc}_\infty(x) + \delta_1(N)$$

hold for all $x \in \mathbb{X}_N$.

Proof: In order to prove the first inequality, from (13.10) we obtain $\ell(x,\mu_N(x)) \leq V_N(x) - V_N(f(x,\mu_N(x)))$. This implies for any $K \in \mathbb{N}$

$$J^{cl}_K(x,\mu_N) = \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k,x),\mu_N(x_{\mu_N}(k,x))) \leq V_N(x) - V_N(x_{\mu_N}(K,x)). \quad (13.15)$$

Now from Theorem 13.13 we know that $|x_{\mu_N}(k,x)|_{x^e} \leq \beta(|x|_{x^e},k) \leq \beta(M,k) =: \nu(k)$, where $M := \max_{x,y\in\mathbb{X}} d(x,y)$. Note that $\nu \in \mathcal{L}$. Moreover, by (13.9) we have $V_N(x^e) = N\ell(x^e,u^e) = 0$ and from Assumption 13.12(c) we know the existence of $\gamma_V \in \mathcal{K}$ with $|V_N(x)| = |V_N(x) - V_N(x^e)| \leq \gamma_V(|x|_{x^e})$ for all $x \in \mathbb{X}$. Together this yields

$$|V_N(x_{\mu_N}(K,x))| \leq \gamma_V(\nu(K)).$$

Since $\gamma_V(\nu(K)) \to 0$ for $K \to \infty$, this inequality together with (13.15) yields the first inequality by letting $K \to \infty$.

For the second inequality, we note that it is sufficient to prove the inequality for all sufficiently large N, because by boundedness of V_N and V_∞^{uc}, for small N the inequality can always be satisfied by choosing $\delta_1(N)$ sufficiently large without violating the requirement $\delta_1 \in \mathcal{L}$. Consider σ_∞ from Proposition 13.18, pick N_0 and η from Assumption 13.5(b), choose N_1 such that $\sigma_\infty(N_1) < \eta$, fix $0 < \varepsilon < 1$ and choose an admissible control u_ε satisfying $J_\infty^{uc}(x, u_\varepsilon) \leq V_\infty^{uc}(x) + \varepsilon$. Then for $N \geq 2N_1$ we use Proposition 13.18 with $P = \lfloor N/2 \rfloor$. We thus obtain the existence of $k \in \{0, \dots, P-1\}$ such that $|x_{u_\varepsilon}(k,x)|_{x^e} < \sigma_\infty(P) \leq \sigma_\infty(N_1) < \eta$, implying $x_u(k,x) \in \mathbb{X}_{N_1} \subseteq \mathbb{X}_{N_2}$ and thus $u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_{N_2}}(x)$ for all $N_2 \geq N_1$. Particularly, this holds for $N_2 = N - k$, implying $u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)$. Now, from Assumption 13.12(c) applied to V_{N-k} we can conclude (again using $V_N(x^e) = 0$)

$$|V_{N-k}(x_{u_\varepsilon}(k,x))| \leq \gamma_V(\sigma_\infty(P)).$$

Moreover, Lemma 13.19 and the bound on λ yield

$$\begin{aligned} V_\infty^{uc}(x) + \varepsilon \geq J_\infty^{uc}(x, u_\varepsilon) \;&\geq\; J_k^{uc}(x, u_\varepsilon) + V_\infty(x_{u_\varepsilon}(k,x)) \\ &\geq\; J_k^{uc}(x, u_\varepsilon) - \lambda(x_{u_\varepsilon}(k,x)) \;\geq\; J_k^{uc}(x, u_\varepsilon) - \gamma_\lambda(\sigma_\infty(P)). \end{aligned}$$

Together with the dynamic programming principle (12.1) these inequalities imply

$$\begin{aligned} V_N(x) \;&=\; \inf_{u \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)} \{J_k^{uc}(x,u) + V_{N-k}(x_u(k,x))\} \;\leq\; J_k^{uc}(x, u_\varepsilon) + V_{N-k}(x_{u_\varepsilon}(k,x)) \\ &\leq\; V_\infty^{uc}(x) + \gamma_V(\sigma_\infty(P)) + \gamma_\lambda(\sigma_\infty(P)) + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this proves the assertion for $\delta_1(N) = \gamma_V(\sigma_\infty(\lfloor N/2 \rfloor)) + \gamma_\lambda(\sigma_\infty(\lfloor N/2 \rfloor))$. □

Since x^e is asymptotically stable for the MPC closed-loop trajectories, the closed-loop solutions converge towards x^e as $k \to \infty$. More precisely, given a time K, by Theorem 13.13 the solutions are guaranteed to satisfy $x_{\mu_N}(k,x) \in \overline{\mathcal{B}}_\kappa(x^e)$ for all $k \geq K$ and $\kappa = \beta(|x|_{x^e}, K)$ for β from Theorem 13.13. We denote the time span $\{0, \dots, K-1\}$ during which the system is (possibly) outside $\overline{\mathcal{B}}_\kappa(x^e)$ as *transient time* and the related finite horizon functional $J_K^{uc}(x,u)$ as *transient performance*. The next theorem then shows that among all possible trajectories from x to $\overline{\mathcal{B}}_\kappa(x^e)$, the MPC closed loop has the best transient performance up to error terms vanishing as $K \to \infty$ and $N \to \infty$. Again, in order to simplify the notation, we use $\mathbb{U}^K_\kappa(x)$ from (13.14). We remark that unlike the previous theorem here we do not need to assume $\ell(x^e, u^e) = 0$.

Theorem 13.22 Consider the MPC Algorithm 11.9 with strictly dissipative optimal control problem ($\mathrm{OCP_{N,e}}$). Assume that $\mathbb{X}$ is bounded and that Assumptions 13.5 and 13.12 hold. Then there exist $\delta_1, \delta_2 \in \mathcal{L}$ such that for all $x \in \mathbb{X}_N$ the inequality

$$J_K^{cl}(x, \mu_N) \leq \inf_{u \in \mathbb{U}^K_\kappa(x)} J_K^{uc}(x,u) + \delta_1(N) + \delta_2(K)$$

holds with $\kappa = \beta(|x|_{x^e}, K)$ and $\beta \in \mathcal{KL}$ characterizing the asymptotic stability of the closed loop guaranteed by Theorem 13.13.

Proof: We can without loss of generality assume $\ell(x^e, u^e) = 0$ because the claimed inequality is invariant under adding constants to ℓ. Moreover, similar to the proof of Theorem 13.21 it is sufficient to prove the inequality for all sufficiently large K and N, because by boundedness of all functions involved for small N and K the inequality can always be achieved by choosing $\delta_1(N)$ and $\delta_2(K)$ sufficiently large. As in the first step of the previous proof we obtain $|V_N(x_{\mu_N}(K,x))| \le \gamma_V(\nu(K))$. It is thus sufficient to show the existence of $\delta_1, \tilde{\delta}_2 \in \mathcal{L}$ with

$$V_N(x) \le \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x) + \delta_1(N) + \tilde{\delta}_2(K) \tag{13.16}$$

for all $x \in \mathbb{X}_N$ because then the assertion follows from (13.15) with $\delta_2 = \gamma_V \circ \nu + \tilde{\delta}_2$.

In order to prove (13.16), consider σ from Lemma 13.20(c), which we apply with $P = \lfloor N/2 \rfloor$ and pick $u_\varepsilon \in \mathbb{U}^K_\kappa(x)$ with $J^{uc}_K(x, u_\varepsilon) \le \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \varepsilon$ with an arbitrary but fixed $\varepsilon \in (0,1)$. This yields the existence of $k \in \{0, \dots, \lfloor N/2 \rfloor\}$, $k \le K-1$ with $|x_{u_\varepsilon}(k,x)|_{x^e} \le \sigma(\min\{P, K-1\})$. Since u_ε steers x to $\overline{\mathcal{B}}_\kappa(x^e)$, the shifted sequence $u_\varepsilon(k+\cdot)$ lies in $\mathbb{U}^{K-k}_\kappa(x_{u_\varepsilon}(k,x))$, implying that this set is nonempty. Hence, we can apply Lemma 13.20(b) in order to conclude $J^{uc}_{K-k}(x_{u_\varepsilon}(k,x), u_\varepsilon(k+\cdot)) \ge -\gamma_\lambda(\sigma(\min\{N, K-1\})) - \gamma_\lambda(\kappa)$. This implies

$$\begin{aligned} \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \varepsilon \ \ge \ J^{uc}_K(x, u_\varepsilon) \ &= \ J^{uc}_k(x, u_\varepsilon) + J^{uc}_{K-k}(x_{u_\varepsilon}(k,x), u_\varepsilon(k+\cdot)) \\ &\ge \ J^{uc}_k(x, u_\varepsilon) - \gamma_\lambda(\sigma(\min\{N, K-1\})) - \gamma_\lambda(\kappa) \end{aligned}$$

Moreover, by choosing N and K sufficiently large we can ensure $\sigma(\min\{P, K-1\} < \eta$ for η from Assumption 13.5(b), implying $u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_Q}(x)$ for all $Q \ge N_0$ and N_0 from Assumption 13.5(b). Particularly, choosing $N \ge 2N_0$ implies $N - k \ge N_0$ and thus $u_\varepsilon \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)$.

Using this relation, the inequality derived above, the dynamic programming principle (12.1) and Assumption 13.12(c) for V_{N-k} we obtain

$$\begin{aligned} V_N(x) \ &= \ \inf_{u \in \mathbb{U}^k_{\mathbb{X}_{N-k}}(x)} \{J^{uc}_k(x,u) + V_{N-k}(x_u(k,x))\} \ \le \ J^{uc}_k(x, u_\varepsilon) + V_{N-k}(x_{u_\varepsilon}(k,x)) \\ &\le \ \inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u) + \gamma_\lambda(\sigma(\min\{P, K-1\})) + \gamma_\lambda(\kappa) + \varepsilon \\ &\qquad + \gamma_V(\sigma(\min\{P, K-1\})). \end{aligned}$$

This shows the desired inequality (13.16) for

$$\delta_1(N) = \gamma_V(\sigma(\lfloor N/2 \rfloor)) + \gamma_\lambda(\sigma(\lfloor N/2 \rfloor))$$

and, using the choice of κ,

$$\tilde{\delta}_2(K) \ = \ \gamma_V(\sigma(K-1)) + \gamma_\lambda(\sigma(K-1)) \ + \ \gamma_\lambda(\beta(M,K))$$

with $M = \max_{x,y \in \mathbb{X}} d(x,y)$ and $\beta \in \mathcal{KL}$ characterizing the asymptotic stability of the closed loop. □

Example 13.23 Fig. 13.4 illustrates how $J^{cl}_K(x, \mu_N)$ depends on N for Example 13.1. The value $K = 30$ is so large that the effect of the term $\delta_2(K)$ is negligible and not visible in the figure, hence $J^{cl}_K(x, \mu_N)$ converges to $\inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x,u)$ for increasing N. □

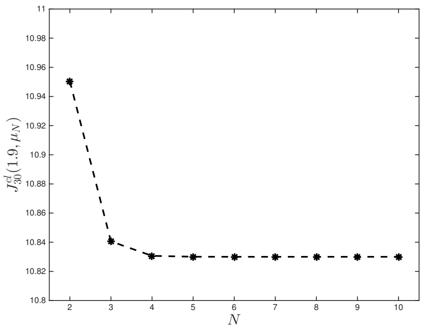


Figure 13.4: Value of $J_K^{cl}(x,\mu_N)$ for $K = 30$, $x = 1.9$ and varying N with terminal constraint $\mathbb{X} = \{0\}$

13.5 Averaged optimality without terminal conditions

In this and in the subsequent sections we discuss the case in which we do not impose terminal conditions on the problem, i.e., we consider the MPC Algorithm 11.1 with optimal control problem ($\mathrm{OCP_N}$). The corresponding functionals and optimal value functions will, as usual, be denoted by J_N and V_N and their infinite horizon counterparts by J_∞ and V_∞, i.e., we do not use the superscript notation J_N^{uc} etc. anymore in the sequel. The results are presented in parallel to Sects. 13.2–13.4.

Since we do not impose any terminal conditions, we do not need Assumptions 13.5 and 13.12(a) anymore. However, we still need Part (b) and (a relaxed version of) Part (c) of Assumption 13.12, where the latter now refers to the optimal value function of the unconstrained problem ($\mathrm{OCP_N}$).

Assumption 13.24 [Continuity of λ and V_N at x^e] There exist γ_λ and $\gamma_V \in \mathcal{K}_\infty$ and $\omega \in \mathcal{L}$ such that the following properties hold.
(a) For all $x \in \mathbb{X}$ it holds that

$$|\lambda(x) - \lambda(x^e)| \leq \gamma_\lambda(|x|_{x^e}).$$

(b) For each $N \in \mathbb{N}$ and each $x \in \mathbb{X}$ it holds that

$$|V_N(x) - V_N(x^e)| \leq \gamma_V(|x|_{x^e}) + \omega(N).$$

□

Note that (b) implies viability of $\mathbb{X}$ which we assume for simplicity in this section. If desired, this condition could be relaxed (see [4] for details in a continuous time setting). One method of ensuring the continuity from (b) without requiring explicit knowledge of V_N is by assuming strict dissipativity and local controllability around x^e, see [16] or [6, Sect. 6].

We observe that Propositions 13.15 and 13.18 remain valid, as the assumptions, statements and proofs do not involve any terminal constraints or costs. Based on these two propositions, we can prove the following two auxiliary results, which lead to the main result of this section. In what follows, we denote by $u^\star_\infty$ and $u^\star_N$ the optimal control sequences for (OCP$_\infty$) and (OCP$_\mathrm{N}$), respectively, for initial value $x \in \mathbb{X}$.

Lemma 13.25 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x) = J_M(x, u^\star_N) + V_{N-M}(x^e) + R_1(x, M, N) \tag{13.17}$$

holds with $|R_1(x, M, N)| \le \gamma_V(\sigma_\delta(P)) + \omega(N - M)$ for all $x \in \mathbb{X}$, all $N \in \mathbb{N}$, all $P \in \mathbb{N}$ and all $M \notin \mathcal{Q}(x, u^\star_N, P, N)$, with σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$.

Proof: Observe that using the constant control $u \equiv u^e$ we can estimate $V_N(x^e) \le J_N(x^e, u) = N\ell(x^e, u^e)$. Thus, using Assumption 13.24 we get $J_N(x, u^\star_N) \le N\ell(x^e, u^e) + \gamma_V(|x|_{x^e}) + \omega(N)$, hence Proposition 13.15 applies to the optimal trajectory with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$. This in particular ensures $|x_{u^\star_N}(M, x)|_{x^e} \le \sigma_\delta(P)$ for all $M \notin \mathcal{Q}(x, u^\star_N, P, N)$. Now the dynamic programming principle (12.2) yields

$$V_N(x) = J_M(x, u^\star_N) + V_{N-M}(x_{u^\star_N}(M, x)).$$

Hence, (13.17) holds with $R_1(x, M, N) = V_{N-M}(x_{u^\star_N}(M, x)) - V_{N-M}(x^e)$. Then for any $P \in \mathbb{N}$ and any $M \notin \mathcal{Q}(x, u^\star_N, P, N)$ this implies $|R_1(x, M, N)| \le \gamma_V(|x_{u^\star_N}(M, x)|_{x^e}) + \omega(N - M) \le \gamma_V(\sigma_\delta(P)) + \omega(N - M)$ and thus the assertion. □

Lemma 13.26 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x) \le V_{N-1}(x) + \ell(x^e, u^e) + R_2(x, N)$$

holds with $|R_2(x, N)| \le \nu_2(|x|_{x^e}, N) = 2\gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + 2\omega(\lfloor N/2 \rfloor - 1)$ for all $x \in \mathbb{X}$, all $N \in \mathbb{N}$ and σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N - 1)$.

Proof: Given $x \in \mathbb{X}$, consider the optimal control $u^\star_{N-1}$ for horizon length $N - 1$ and σ_δ from Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e})$. Then Lemma 13.25 applied with $N - 1$ in place of N and $P = \lfloor N/2 \rfloor$ implies the existence of $M \in \{0, \dots, \lfloor N/2 \rfloor - 1\}$ with

$$V_{N-1}(x) = J_M(x, u^\star_{N-1}) + V_{N-M-1}(x^e) + R_1(x, M, N - 1)$$

with $|R_1(x, M, N - 1)| \le \gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + \omega(\lfloor N/2 \rfloor - 1)$. The construction in the proof of Lemma 13.25 moreover yields $|x_{u^\star_{N-1}}(M, x)|_{x^e} \le \sigma_\delta(\lfloor N/2 \rfloor)$. Using $u(k) = u^\star_{N-1}(k)$ for $k = 0, \dots, M - 1$ and $u(M + k) = u^\star_{N-M}(k)$ with the optimal control $u^\star_{N-M}$ for initial value $x_{u^\star_N}(M, x)$ and horizon $N - M$ for $k = M, \dots, N - 1$ yields

$$J_N(x, u) = J_M(x, u^\star_{N-1}) + V_{N-M}(x_{u^\star_N}(M, x)) = J_M(x, u^\star_{N-1}) + V_{N-M}(x^e) + \widehat{R}_1(x, M, N)$$

with $|\widehat{R}_1(x, M, N)| \le \gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + \omega(\lfloor N/2 \rfloor)$. Since for initial value x^e we can always stay at the equilibrium for one step and use the optimal control for initial value x^e for the remaining horizon, we obtain the inequality $V_{N-M}(x^e) \le \ell(x^e, u^e) + V_{N-M-1}(x^e)$. Together this yields

$$\begin{aligned}
V_N(x) \quad &\le \quad J_N(x, u) \quad = \quad J_M(x, u^\star_{N-1}) + V_{N-M}(x^e) + \widehat{R}_1(x, M, N) \\
&\le \quad J_M(x, u^\star_{N-1}) + \ell(x^e, u^e) + V_{N-M-1}(x^e) + \widehat{R}_1(x, M, N) \\
&= \quad V_{N-1}(x) + \ell(x^e, u^e) - R_1(x, M, N - 1) + \widehat{R}_1(x, M, N)
\end{aligned}$$

and thus the claim with $R_2(x,N) = \widehat{R}_1(x,M,N) - R_1(x,M,N-1)$. □

Now we can state the theorem on the infinite horizon average performance.

Theorem 13.27 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem (OCP$_\mathrm{N}$) with bounded storage function λ. Let Assumption 13.24 hold and assume V_N is bounded from below on $\mathbb{X}$. Then, for any $N \geq 2$ and any $x \in \mathbb{X}$ the averaged closed-loop performance satisfies the inequality

$$\overline{J}^{cl}_\infty(x,\mu_N) \leq \ell(x^e,u^e) + \delta_1(N) \tag{13.18}$$

with $\delta_1(N) \leq 2\gamma_V(\sigma_\delta(\lfloor N/2 \rfloor)) + 2\omega(\lfloor N/2 \rfloor - 1)$ for σ_δ from Proposition 13.15 with $\delta = \sup_{k\in\mathbb{N}} \gamma_V(|x_{\mu_N}(k)|_{x^e}) + \omega(N-1)$ and γ_V and ω from Assumption 13.24.

Proof: Abbreviate $x_{\mu_N}(k) = x_{\mu_N}(k,x)$. From the dynamic programming principle (12.2) and Lemma 13.26 applied with $x = x_{\mu_N}(k+1)$ we obtain

$$\begin{aligned}\ell(x_{\mu_N}(k),\mu_N(x_{\mu_N}(k))) &= V_N(x_{\mu_N}(k)) - V_{N-1}(x_{\mu_N}(k+1)) \\ &\leq V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e,u^e) + \underbrace{\nu_2(|x_{\mu_N}(k+1)|_{x^e},N)}_{=:\tilde\nu_2(k,N)}.\end{aligned}$$

Thus we obtain

$$\begin{aligned}\overline{J}^{cl}_\infty(x,\mu_N) &= \limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \ell(x_{\mu_N}(k),\mu_N(x_{\mu_N}(k))) \\ &= \limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} \Big(V_N(x_{\mu_N}(k)) - V_N(x_{\mu_N}(k+1)) + \ell(x^e,u^e) + \tilde\nu_2(k,N)\Big) \\ &= \ell(x^e,u^e) + \sup_{k\in\mathbb{N}} \tilde\nu_2(k,N) + \limsup_{K\to\infty} \frac{V_N(x_0) - V_N(x_{\mu_N}(K))}{K} \\ &\leq \ell(x^e,u^e) + \sup_{k\in\mathbb{N}} \tilde\nu_2(k,N) + \limsup_{K\to\infty} \frac{V_N(x_0) + M}{K} \;=\; \ell(x^e,u^e) + \sup_{k\in\mathbb{N}} \tilde\nu_2(k,N)\end{aligned}$$

where $-M$ is a lower bound on V_N on $\mathbb{X}$. This shows the claim with $\delta_1(N) = \sup_{k\in\mathbb{N}} \tilde\nu_2(k,N)$ which satisfies the stated bounds because σ_δ is increasing in δ. □

The difference between this and the corresponding result with terminal conditions is that we get the error term $\delta_1(N)$ on the right hand side of the estimate, which does, however, tend to 0 as $N \to \infty$ provided $\delta < \infty$. This is always the case for bounded state constraints $\mathbb{X}$. In case of unbounded $\mathbb{X}$, Theorem 13.34 from the next section can be used to obtain a bound for $|x_{\mu_N}(k)|_{x^e}$ which is independent of k.

Example 13.28 Fig. 13.5 shows $\overline{J}^{cl}_\infty(x,\mu_N)$ for Example 13.1 depending on N. The plot in the logarithmic scale shows that the value converges to the optimal value $\ell(0,0) = 0$ exponentially fast, hence the error $\delta_1(N)$ also vanishes exponentially fast. This is actually not a coincidence. However, an analysis of the rate of convergence is beyond the scope of this lecture. We refer to [9] for details. □

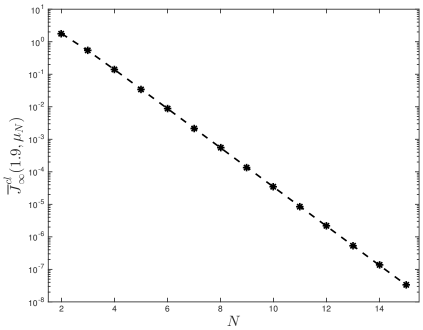


Figure 13.5: Value of $\overline{J}^{cl}_\infty(x, \mu_N)$ for $x = 1.9$ without terminal conditions depending on N

13.6 Asymptotic stability without terminal conditions

Now we turn to analyzing the stability properties of the MPC closed-loop solutions without terminal conditions. As in the case with terminal conditions, our goal is to assume strict dissipativity and to use the optimal value function for the modified stage cost $\tilde{\ell}$ from (13.5) as a Lyapunov function, but now without imposing terminal conditions. The crucial difference is that without terminal conditions the optimal trajectories of the original and the modified problem no longer coincide.

In order to see why, we refer to the optimal control problem (OCP$_\text{N}$) with stage cost $\tilde{\ell}$ as $(\widetilde{\text{OCP}}_\text{N})$ and, as before, denote the corresponding functional and the optimal value function by $\widetilde{J}_N$ and $\widetilde{V}_N$. Due to the fact that we no longer impose terminal conditions, the relations between V_N and $\widetilde{V}_N$ are not the same as in Sect. 13.3. For J_N and $\widetilde{J}_N$, instead of (13.7) we now have (13.12), which in the notation of this section reads

$$\widetilde{J}_N(x,u) = J_N(x,u) + \lambda(x) - \lambda(x_u(N,x)) - N\ell(x^e,u^e). \tag{13.19}$$

Unfortunately, in contrast to (13.7), this equation does not allow for an easy derivation of a relation between the optimal value functions of the form (13.8), because of the additional u-dependent term $\lambda(x_u(N,x))$ on the right hand side of (13.19). A first consequence of this fact is that the continuity Assumption 13.24(b) does not immediately carry over to $\widetilde{V}_N$. Hence, we need to introduce this as an independent assumption.

Assumption 13.29 [Continuity of $\widetilde{V}_N$ at x^e] There exists $\gamma_{\widetilde{V}} \in \mathcal{K}_\infty$ such that for each $N \in \mathbb{N}$ and each $x \in \mathbb{X}$ it holds that

$$|\widetilde{V}_N(x) - \widetilde{V}_N(x^e)| \le \gamma_{\widetilde{V}}(|x|_{x^e}).$$

□

In case strict dissipativity holds, $\tilde{\ell}$ is positive definite w.r.t. the equilibrium x^e, hence we obtain $\widetilde{V}_N(x^e) = 0$ and $\widetilde{V}_N(x) \ge 0$ for all $x \in \mathbb{X}$. Thus, the inequality in Assumption 13.29 is equivalent to $\widetilde{V}_N(x) \le \gamma_{\widetilde{V}}(|x|_{x^e})$ which can be guaranteed under conditions which guarantee that the system can be controlled to x^e with sufficiently low cost.

Unlike continuity, a straightforward check of Definition 13.7 (with storage function $\lambda \equiv 0$) shows that strict dissipativity carries over from $(\mathrm{OCP}_\mathrm{N})$ to $(\widetilde{\mathrm{OCP}}_\mathrm{N})$, even with the same ρ. Thus, in particular, all the previous lemmas that apply to $(\mathrm{OCP}_\mathrm{N})$ in case of strict dissipativity also apply to $(\widetilde{\mathrm{OCP}}_\mathrm{N})$. As a general rule, we denote all parameters, sets etc. referring to $(\widetilde{\mathrm{OCP}}_\mathrm{N})$ with a tilde, e.g., the set $\mathcal{Q}(x,u,N,P)$ from Proposition 13.15 will be denoted by $\widetilde{\mathcal{Q}}(x,u,N,P)$ when this proposition is applied to $(\widetilde{\mathrm{OCP}}_\mathrm{N})$.

As already mentioned above, from the definition we cannot directly deduce a simple relation like (13.8) between V_N and $\widetilde{V}_N$. The reason why we can still use $\widetilde{V}_N$ as an — at least practical — Lyapunov function lies in the fact that we can still establish an approximate version of (13.8). To this end, we first need the following preparatory lemma.

Lemma 13.30 If Assumption 13.24 and the assumptions of Proposition 13.15 hold, then the equation

$$V_N(x^e) = M\ell(x^e,u^e) + V_{N-M}(x^e) - R_3(x,P,N)$$

holds with $0 \le R_3(x,P,N) \le \gamma_V(\sigma_\delta(P)) + \omega(N-M) + \gamma_\lambda(\sigma_\delta(P))$ for all $N,P \in \mathbb{N}$ and for all $M \notin \mathcal{Q}(x,u_N^\star,N,P)$, where $u_N^\star \in \mathbb{U}^N(x^e)$ is the optimal control of $(\mathrm{OCP}_\mathrm{N})$ for initial value x^e and σ_δ is from Proposition 13.15 with $\delta = \omega(N-M)$.

Proof: The inequality $V_N(x^e) \le M\ell(x^e,u^e) + V_{N-M}(x^e)$ follows from the dynamic programming principle (12.1) using the control $u \equiv u^e$. For the opposite inequality consider the optimal control $u_N^\star \in \mathbb{U}^N(x^e)$ for initial value x^e. As in the proof of Lemma 13.25 we can apply Proposition 13.15 with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$ in order to conclude that for each $M \notin \mathcal{Q}(x,u_N^\star,N,P)$ we have

$$\begin{aligned}
V_N(x^e) &= \sum_{k=0}^{M-1} \ell(x_{u_N^\star}(k),u_N^\star(k)) + V_{N-M}(x_{u_N^\star}(M)) \\
&= -\lambda(x^e) + \lambda(x_{u_N^\star}(M)) + M\ell(x^e,u^e) + \sum_{k=0}^{M-1} \underbrace{\tilde\ell(x_{u_N^\star}(k),u_N^\star(k))}_{\ge 0} + V_{N-M}(x_{u_N^\star}(M)) \\
&\ge M\ell(x^e,u^e) + V_{N-M}(x^e) + \Big[V_{N-M}(x_{u_N^\star}(M)) - V_{N-M}(x^e)\Big] + \Big[\lambda(x_{u_N^\star}(M)) - \lambda(x^e)\Big] \\
&\ge M\ell(x^e,u^e) + V_{N-M}(x^e) - \gamma_V(\sigma_\delta(P)) - \omega(N-M) - \gamma_\lambda(\sigma_\delta(P))
\end{aligned}$$

which shows the claim. □

Now we can prove the approximate relation of the form (13.8) between $\widetilde{V}_N$ and V_N.

Lemma 13.31 If Assumptions 13.24 and 13.29 as well as the assumptions of Proposition 13.15 hold, then the equation

$$\widetilde{V}_N(x) = V_N(x) + \lambda(x) - V_N(x^e) + R_4(x,N)$$

holds with $|R_4(x,N)| \le \nu_4(|x|_{x^e},N)$ with

$$\begin{aligned}
\nu_4(|x|_{x^e},N) \;=\; \max\{&\gamma_V(\tilde\sigma_{\tilde\delta}(\lfloor N/3\rfloor)) + \gamma_V(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_{\widetilde{V}}(\tilde\sigma_{\tilde\delta}(\lfloor N/3\rfloor)) \\
&\quad + \gamma_\lambda(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_\lambda(\tilde\sigma_{\tilde\delta}(\lfloor N/3\rfloor)) + 3\omega(\lfloor N/3\rfloor), \\
&\gamma_{\widetilde{V}}(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_V(\sigma_\delta(\lfloor N/3\rfloor)) + \gamma_\lambda(\sigma_\delta(\lfloor N/3\rfloor)) \\
&\quad + 2\omega(\lfloor N/3\rfloor)\}
\end{aligned}$$

with σ_δ and $\tilde{\sigma}_{\tilde{\delta}}$ from Proposition 13.15 applied to $(\mathrm{OCP}_\mathrm{N})$ and $(\widetilde{\mathrm{OCP}}_\mathrm{N})$, respectively, with $\delta = \gamma_V(|x|_{x^e}) + \omega(N)$ and $\tilde{\delta} = \gamma_{\widetilde{V}}(|x|_{x^e})$.

Proof: Fix $x \in \mathbb{X}$ and let $u_N^\star$ and $\tilde{u}_N^\star \in \mathbb{U}^N(x)$ denote the optimal control minimizing $J_N(x,u)$ and $\widetilde{J}_N(x,u)$, respectively. We note that if $(\mathrm{OCP}_\mathrm{N})$ is strictly dissipative then $(\widetilde{\mathrm{OCP}}_\mathrm{N})$ is strictly dissipative, too, with bounded storage function $\lambda \equiv 0$ and same $\rho \in \mathcal{K}_\infty$. Moreover, $V_N(x) \leq N\ell(x^e,u^e) + \gamma_V(|x|_{x^e}) + \omega(N)$ and $\widetilde{V}_N(x) \leq N\tilde{\ell}(x^e,u^e) + \gamma_{\widetilde{V}}(|x|_{x^e})$, since $V_N(x^e) \leq N\ell(x^e,u^e)$ and $\widetilde{V}_N(x^e) = 0$. Hence, Proposition 13.15 applies to the optimal trajectories for both problems, yielding $\sigma_\delta \in \mathcal{L}$ and $\mathcal{Q}(x,u_N^\star,P,N)$ for $(\mathrm{OCP}_\mathrm{N})$ and $\tilde{\sigma}_{\tilde{\delta}}$ and $\widetilde{\mathcal{Q}}(x,\tilde{u}_N^\star,P,N)$ for $(\widetilde{\mathrm{OCP}}_\mathrm{N})$. For all $M \notin \widetilde{\mathcal{Q}}(x,\tilde{u}_N^\star,P,N)$ we can estimate

$$
\begin{aligned}
V_N(x) &\leq J_M(x,\tilde{u}_N^\star) + V_{N-M}(x_{\tilde{u}_N^\star}(M)) \\
&\leq J_M(x,\tilde{u}_N^\star) + V_{N-M}(x^e) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M) \\
&\leq \widetilde{J}_M(x,\tilde{u}_N^\star) - \lambda(x) + \lambda(x^e) + M\ell(x^e,u^e) + V_{N-M}(x^e) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) \\
&\qquad + \gamma_\lambda(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M) \\
&\leq \widetilde{V}_N(x) - \widetilde{R}_1(x,P,N) - \lambda(x) \\
&\qquad + V_N(x^e) + R_3(x,P,N) + \gamma_V(\tilde{\sigma}_{\tilde{\delta}}(P)) + \gamma_\lambda(\tilde{\sigma}_{\tilde{\delta}}(P)) + \omega(N-M),
\end{aligned}
$$

where we have applied the dynamic programming principle (12.1) in the first inequality, Proposition 13.15 for $(\widetilde{\mathrm{OCP}}_\mathrm{N})$ and Assumption 13.24(b) respectively Assumption 13.24(a) and (13.19) in the second and third inequality and Lemma 13.25 (applied to $(\widetilde{\mathrm{OCP}}_\mathrm{N})$, hence with remainder term denoted by $\widetilde{R}_1$) and Lemma 13.30 (applied to $(\mathrm{OCP}_\mathrm{N})$) in the last step. Moreover, $\lambda(x^e) = 0$ and $\widetilde{V}_N(x^e) = 0$ were used.

By exchanging the two optimal control problems and using the same inequalities as above, we get

$$
\begin{aligned}
\widetilde{V}_N(x) &\leq V_N(x) - R_1(x,P,N) + \lambda(x) - V_N(x^e) + \gamma_{\widetilde{V}}(\sigma_\delta(P)) + \gamma_\lambda(\sigma_\delta(P)) \\
&\qquad + \omega(N-M)
\end{aligned}
$$

for all $M \notin \mathcal{Q}(x,u_N^\star,P,N)$. Here we can omit the negative $-R_3$-term. Now, choosing $P = \lfloor N/3 \rfloor$, the union $\mathcal{Q}(x,\tilde{u}_N^\star,P,N) \cup \mathcal{Q}(x,u_N^\star,P,N)$ has at most $2N/3$ elements, hence there exists $M \leq 2N/3$ for which both inequalities hold. This yields $N - M \geq \lfloor N/3 \rfloor$ and thus

$$
\begin{aligned}
|R_1(x,P,N)| &\leq \gamma_V(\sigma_\delta(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor), \\
|\widetilde{R}_1(x,P,N)| &\leq \gamma_{\widetilde{V}}(\tilde{\sigma}_{\tilde{\delta}}(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor) \text{ and} \\
R_3(x,P,N) &\leq \gamma_V(\sigma_\delta(\lfloor N/3 \rfloor)) + \omega(\lfloor N/3 \rfloor) + \gamma_\lambda(\sigma_\delta(\lfloor N/3 \rfloor))
\end{aligned}
$$

which shows the claim. □

The following proposition shows in which sense $\widetilde{V}_N$ is a Lyapunov function for the system. This will be used in the subsequent theorem in order to prove semiglobal practical asymptotic stability of the closed loop.

Proposition 13.32 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem (OCP$_\mathrm{N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$ and let Assumptions 13.24 and 13.29 hold. Then for each $\Theta > 0$ there exists $\delta_1 \in \mathcal{L}$ such that for all $N \geq 2$ with $\delta_1(N) \leq \Theta$ the optimal value function $\widetilde{V}_N$ of $(\widetilde{\mathrm{OCP}}_\mathrm{N})$ is a Lyapunov function for the closed loop on $S = Y \setminus \mathbb{P}$ for the forward invariant sets $Y = \widetilde{V}_N^{-1}([0,\Theta])$ and $\mathbb{P} = \widetilde{V}_N^{-1}([0,\delta_1(N)])$. □

Proof: We have to check that Definition 10.4 is satisfied and that Y and P are forward invariant. The lower bound in (10.4) follows with $\alpha_1 = \rho$ because strict dissipativity implies $\tilde{\ell}(x,u) \geq \rho(|x|_{x^e})$ and thus

$$\widetilde{V}_N(x) = \inf_{u \in \mathbb{U}^N(x)} \sum_{k=0}^{N-1} \tilde{\ell}(x_u(k,x),u(k)) \geq \inf_{u \in \mathbb{U}^N(x)} \sum_{k=0}^{N-1} \rho(|x_u(k,x)|_{x^e}) \geq \rho(|x|_{x^e}).$$

The upper bound in (10.4) follows from Assumption 13.29 and $\widetilde{V}_N(x^e) = 0$ with $\alpha_2 = \gamma_{\widetilde{V}}$. In order to obtain inequality (10.5) we abbreviate $x^+ = f(x,\mu_N(x))$. Now, for all $x \in Y$ we obtain $\widetilde{V}_N(x) \leq \Theta$, which implies $|x|_{x^e} \leq \rho^{-1}(\Theta)$. In order to obtain a similar estimate for $|x^+|_{x^e}$, we observe that $\widetilde{V}_N(x) \leq \Theta$ implies $V_N(x) \leq \Theta - \lambda(x) + M + N\ell(x^e,u^e)$, where $M > 0$ denotes a bound on λ. Thus, Theorem 12.4 and strict dissipativity yield

$$\begin{aligned} V_{N-1}(x^+) &= V_N(x) - \ell(x,\mu_N(x)) \leq V_N(x) + \lambda(x) - \lambda(x^+) - \ell(x^e,u^e) \\ &\leq \Theta - \lambda(x^+) + M + (N-1)\ell(x^e,u^e). \end{aligned}$$

This implies

$$\widetilde{V}_{N-1}(x^+) \leq V_{N-1}(x^+) + \lambda(x^+) + M - (N-1)\ell(x^e,u^e) \leq \Theta + 2M$$

and we can conclude that $|x^+|_{x^e} \leq \rho^{-1}(\Theta + 2M)$. Hence, we can compute

$$\begin{aligned} \widetilde{V}_N(x^+) &= V_N(x^+) + \lambda(x^+) - V_N(x^e) + R_4(x^+,N) \\ &= V_{N-1}(x^+) + \ell(x^e,u^e) + \lambda(x^+) - V_N(x^e) + R_2(x^+,N) + R_4(x^+,N) \\ &= V_N(x) - \ell(x,\mu_N(x)) + \ell(x^e,u^e) + \lambda(x^+) - V_N(x^e) \\ &\qquad + R_2(x^+,N) + R_4(x^+,N) \\ &= \widetilde{V}_N(x) \underbrace{-\ell(x,\mu_N(x)) + \ell(x^e,u^e) + \lambda(x^+) - \lambda(x)}_{=-\tilde{\ell}(x,\mu_N(x))} \\ &\qquad + R_2(x^+,N) + R_4(x^+,N) - R_4(x,N). \end{aligned}$$

where we used Lemma 13.31 for $x = x^+$ for the first equality, Lemma 13.26 for the second, equation (12.6) for the third and Lemma 13.31 in the last step. Defining $\nu(N) = \nu_2(\rho^{-1}(\Theta + 2M),N) + 2\nu_4(\rho^{-1}(\Theta + 2M),N)$ with ν_2 and ν_4 from Lemma 13.26 and Lemma 13.31, respectively, we thus obtain

$$\widetilde{V}_N(x^+) \leq \widetilde{V}_N(x) - \rho(|x|_{x^e}) + \nu(N) \leq \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \qquad (13.20)$$

for $\chi := \rho \circ \alpha_2^{-1}(r)$. Now we set $\delta_1(N) = \max\{\chi^{-1}(2\nu(N)),\, \chi^{-1}(\nu(N)) + \nu(N)\}$. Then for all $x \in S = Y \setminus \mathbb{P}$ we obtain $\widetilde{V}_N(x) \geq \delta_1(N)$ and thus $\chi(\widetilde{V}_N(x)) \geq 2\nu(N)$ which implies

$$\widetilde{V}_N(x^+) \leq \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x))/2 \leq \widetilde{V}_N(x) - \chi(\alpha_1(|x|_{x^e}))/2$$

and thus (10.5) with $\alpha_V(r) = \chi(\alpha_1(r))/2$. This inequality also shows that all $x \in Y \setminus \mathbb{P}$ are mapped to Y, since $x \in Y \setminus \mathbb{P} = S$ implies $\widetilde{V}_N(x) \le \Theta$, hence $\widetilde{V}_N(x^+) < \widetilde{V}_N(x) \le \Theta$ and thus $x^+ \in Y$.

Finally, to prove forward invariance of $\mathbb{P}$ (which then also implies forward invariance of Y) we recall that $x \in \mathbb{P}$ if and only if $\widetilde{V}_N(x) \le \delta_1(N)$. Now we pick $x \in \mathbb{P}$ and distinguish two cases.

Case 1: $\chi(\widetilde{V}_N(x)) \ge \nu(N)$. In this case from (13.20) we obtain

$$\widetilde{V}_N(x^+) \le \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \le \widetilde{V}_N(x) \le \delta_1(N).$$

Case 2: $\chi(\widetilde{V}_N(x)) < \nu(N)$. In this case from (13.20) we obtain

$$\begin{array}{rcl} \widetilde{V}_N(x^+) & \le & \widetilde{V}_N(x) - \chi(\widetilde{V}_N(x)) + \nu(N) \;\le\; \widetilde{V}_N(x) + \nu(N) \\ & < & \chi^{-1}(\nu(N)) + \nu(N) \;\le\; \delta_1(N). \end{array}$$

Hence, in both cases we get $\widetilde{V}_N(x^+) \le \delta_1(N)$ and thus $x^+ \in \mathbb{P}$, which proves the forward invariance of $\mathbb{P}$. □

We note that for small values of N the inequality $\delta_1(N) \ge \Theta$ may hold, in which case the set S on which $\widetilde{V}_N$ is a Lyapunov function is empty.

The final theorem on practical asymptotic stability is now an easy consequence of Proposition 13.32. To this end, we use the following notion of semiglobal practical stability.

Definition 13.33 We call the MPC closed loop system (11.2) *semiglobally practically asymptotically stable with respect to the optimization horizon* N if there exists $\beta \in \mathcal{KL}$ such that the following property holds: for each $\delta > 0$ and $\Delta > \delta$ there exists $N_{\delta,\Delta} \in \mathbb{N}$ such that for all $N \ge N_{\delta,\Delta}$ and all $x \in \mathbb{X}$ with $|x|_{x_*} \le \Delta$ the inequality

$$|x_{\mu_N}(k,x)|_{x_*} \le \max\{\beta(|x|_{x_*}, k), \delta\}$$

holds for all $k \in \mathbb{N}_0$. □

Theorem 13.34 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem (OCP$_\mathrm{N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$ and let Assumptions 13.24 and 13.29 hold. Then the equilibrium x^e is semiglobally practically asymptotically stable on $\mathbb{X}$ with respect to the optimization horizon N.

Proof: Fixing $\Delta > \delta > 0$, the assertion follows immediately from Proposition 13.32 and Theorem 10.6 when choosing $\Theta = \alpha_2(\Delta)$ (implying $\overline{\mathcal{B}}_\Delta(x^e) \subset Y$) and $N_{\delta,\Delta} > 0$ so large that $\rho(\delta_1(N_{\delta,\Delta})) \le \delta$ holds for δ from Definition 13.33 and $\delta_1(N)$ from Proposition 13.32 (implying $\mathbb{P} \subset \overline{\mathcal{B}}_\delta(x^e)$). □

We will see in the next chapter that this result can be strengthened to "real" asymptotic stability for stabilizing stage cost under suitable additional assumptions.

Example 13.35 Fig. 13.6 shows the trajectories (open loop dashed, MPC closed loop solid) of Example 13.1 without terminal conditions for $N = 5$ and $N = 10$. One clearly sees the practical asymptotic stability of the closed loop and the turnpike phenomenon for the open-loop trajectories. □

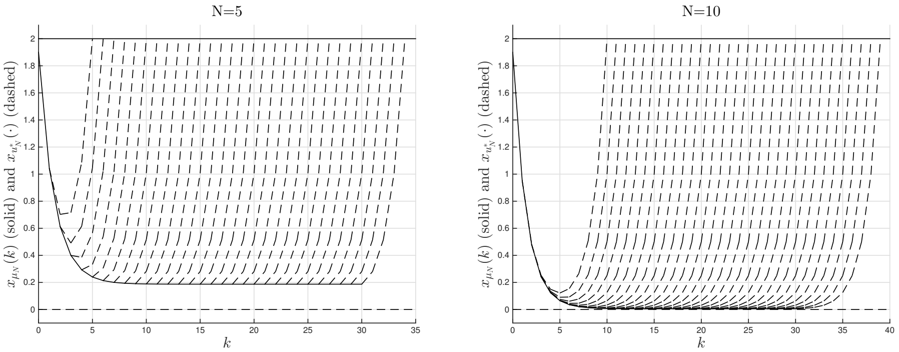


Figure 13.6: MPC closed-loop solution (solid) and open-loop predictions (dashed) for Example 13.1 without terminal conditions and horizon $N = 5$ (left) and $N = 10$ (right). The solid line at $x = 2$ indicates the upper bound of the admissible set $\mathbb{X}$

13.7 Non-averaged performance without terminal conditions

Our final results in this chapter concern the adaptation of the results from Sect. 13.4 to the case without terminal conditions. In order to generalize Theorem 13.21, we need the following continuity assumption on the infinite horizon optimal value function and the two subsequent auxiliary results.

Assumption 13.36 [Continuity of V_∞ at x^e] There exists $\gamma_{V_\infty} \in \mathcal{K}_\infty$ such that for each $x \in \mathbb{X}$ it holds that

$$|V_\infty(x) - V_\infty(x^e)| \leq \gamma_{V_\infty}(|x|_{x^e}).$$

□

Lemma 13.37 If Assumption 13.36 and the assumptions of Proposition 13.18 hold, then the equation

$$V_\infty(x) = J_M(x, u_\infty^\star) + V_\infty(x^e) + R_5(x, M) \tag{13.21}$$

holds with $|R_5(x, M)| \leq \gamma_{V_\infty}(\sigma_\infty(P))$ for all $x \in \mathbb{X}$, all $P \in \mathbb{N}$ and all $M \notin \mathcal{Q}(x, u_\infty^\star, P, \infty)$, where $u_\infty^\star \in \mathbb{U}^\infty(x)$ denotes the infinite horizon optimal control for initial value x and σ_∞ is from Proposition 13.18.

Proof: The dynamic programming principle (12.11) yields

$$V_\infty(x) = J_M(x, u_\infty^\star) + V_\infty(x_{u_\infty^\star}(M, x)).$$

Hence, (13.21) holds with $R_5(x, M) = V_\infty(x_{u_\infty^\star}(M, x)) - V_\infty(x^e)$. Then for any $P \in \mathbb{N}$ and $M \notin \mathcal{Q}(x, u_\infty^\star, P, \infty)$ we obtain $|R_5(x, M)| \leq \gamma_{V_\infty}(\|x_{u_\infty^\star}(M, x) - x^e\|) \leq \gamma_{V_\infty}(\sigma_\infty(P))$ and thus the assertion. □

Lemma 13.38 If Assumptions 13.24 and 13.36 and the assumptions of Propositions 13.15 and 13.18 hold, then the equation

$$J_M(x, u_\infty^\star) = J_M(x, u_N^\star) + R_6(x, M, N) \tag{13.22}$$

holds with $|R_6(x,M,N)| \leq \max\{\gamma_V(\sigma_\delta(P)) + \gamma_V(\sigma_\infty(P)) + 2\omega(N-M), \gamma_{V_\infty}(\sigma_\infty(P)) + \gamma_{V_\infty}(\sigma_\delta(P))\}$ for all $P \in \mathbb{N}$, all $x \in \mathbb{X}$ and all $M \in \{0,\dots,N\} \setminus (\mathcal{Q}(x,u_N^\star,P,N) \cup \mathcal{Q}(x,u_\infty^\star,P,\infty))$, with σ_∞ from Proposition 13.18 and σ_δ from Proposition 13.15 with $\delta = |x|_{x^e}$.

Proof: The finite horizon dynamic programming principle (12.1), (12.2) implies that $u = u_N^\star$ minimizes the expression $J_M(x,u) + V_{N-M}(x_u(M,x))$. Together with the error term R_1 from Lemma 13.25 and $\widehat{R}_1(x,M,N) = V_{N-M}(x_{u_\infty^\star}(M,x)) - V_{N-M}(x^e)$ this yields

$$\begin{aligned} J_M(x,u_N^\star) + V_{N-M}(x^e) &= J_M(x,u_N^\star) + V_{N-M}(x_{u_N^\star}(M,x)) - R_1(x,M,N) \\ &\leq\ J_M(x,u_\infty^\star) + V_{N-M}(x_{u_\infty^\star}(M,x)) - R_1(x,M,N) \\ &=\ J_M(x,u_\infty^\star) + V_{N-M}(x^e) - R_1(x,M,N) + \widehat{R}_1(x,M,N). \end{aligned}$$

Similar to the proof of Lemma 13.25 one sees that $|\widehat{R}_1(x,M,N)| \leq \gamma_V(\sigma_\infty(P)) + \omega(N-M)$ for all $M \notin \mathcal{Q}(x,u_\infty^\star,P,\infty)$.

Conversely, the infinite horizon dynamic programming principle (12.11) implies that $u_\infty^\star$ minimizes the expression $J_M(x,u_\infty^\star) + V_\infty(x_{u_\infty^\star}(M,x))$. Using the error terms R_5 from Lemma 13.37 and $\widehat{R}_5(x,M,N) = V_\infty(x_{u_N^\star}(M,x)) - V_\infty(x^e)$ we obtain

$$\begin{aligned} J_M(x,u_\infty^\star) + V_\infty(x^e) \quad &= \quad J_M(x,u_\infty^\star) + V_\infty(x_{u_\infty^\star}(M,x)) - R_5(x,M) \\ &\leq \quad J_M(x,u_N^\star) + V_\infty(x_{u_N^\star}(M,x)) - R_5(x,M) \\ &= \quad J_M(x,u_N^\star) + V_\infty(x^e) - R_5(x,M) + \widehat{R}_5(x,M,N). \end{aligned}$$

As in the proof of Lemma 13.25 one sees that Proposition 13.15 applies to $x_{u^\star}(\cdot,x)$ with $\delta = \gamma_V(|x|_{x^e})$. Hence, similar to the proof of Lemma 13.37 one obtains $|\widehat{R}_5(x,M,N)| \leq \gamma_{V_\infty}(\sigma_\delta(P))$ for all $M \notin \mathcal{Q}(x,u_N^\star,P,N)$. Together with the estimates for R_1 and R_5 from Lemma 13.25 and 13.37 this yields

$$\begin{aligned} |R_6(x,M,N)| \ &= \ |J_M(x,u_\infty^\star) - J_M(x,u_N^\star)| \\ &\leq\ \max\{|R_1(x,M,N)| + |\widehat{R}_1(x,M,N)|, |R_5(x,M)| + |\widehat{R}_5(x,M,N)|\} \\ &\leq\ \max\{\gamma_V(\sigma_\delta(P)) + \gamma_V(\sigma_\infty(P)) + 2\omega(N-M), \gamma_{V_\infty}(\sigma_\infty(P)) + \gamma_{V_\infty}(\sigma_\delta(P))\} \end{aligned}$$

and thus the claim. □

Now we can establish a version of Theorem 13.21 for economic MPC without terminal conditions. We will discuss after the proof how Theorem 13.39 relates to Theorem 13.21.

Theorem 13.39 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem (OCP$_\text{N}$) with bounded storage function λ, assume that $\ell(x^e,u^e) = 0$ and $\mathbb{X}$ is bounded and let Assumptions 13.24 and 13.36 hold. Then the inequality

$$J_K^{cl}(x,\mu_N) + V_\infty(x_{\mu_N}(K)) \leq V_\infty(x) + K\delta_1(N) \tag{13.23}$$

holds for all $K \in \mathbb{N}$ and all sufficiently large $N \in \mathbb{N}$ with

$$\begin{aligned} \delta_1(N) \quad := \quad & 2\gamma_V(\sigma_\delta(\lfloor (N-1)/8 \rfloor)) + 2\gamma_V(\sigma_\infty(\lfloor (N-1)/8 \rfloor)) \\ & + 2\gamma_{V_\infty}(\sigma_\delta(\lfloor (N-1)/8 \rfloor)) + 4\gamma_{V_\infty}(\sigma_\infty(\lfloor (N-1)/8 \rfloor)) + 4\omega(\lfloor N/2 \rfloor) \end{aligned}$$

with σ_∞ from Proposition 13.18 and σ_δ from Proposition 13.15 with $\delta = \sup_{x\in\mathbb{X}} |x|_{x^e}$.

Proof: We pick $x \in \mathbb{X}$ and abbreviate $x^+ := f(x, \mu_N(x))$. For the corresponding optimal control $u_N^\star$ Corollary 12.3 yields that $u_N^\star(\cdot + 1)$ is an optimal control for initial value x^+ and horizon $N-1$. Hence, for each $M \in \{1, \dots, N\}$ we obtain

$$\begin{aligned} \ell(x, \mu_N(x)) \;=\; V_N(x) - V_{N-1}(x^+) \;&=\; J_N(x, u_N^\star) - J_{N-1}(x^+, u_N^\star(\cdot + 1)) \\ &=\; J_M(x, u_N^\star) - J_{M-1}(x^+, u_N^\star(\cdot + 1)), \end{aligned}$$

where the last equality follows from the fact that the omitted terms in the sums defining $J_M(x, u_N^\star)$ and $J_{M-1}(x^+, u_N^\star(\cdot+1))$ coincide. Using Lemma 13.37 for N, x and M and for $N-1$, x^+ and $M-1$, respectively, yields

$$\begin{aligned} V_\infty(x) - V_\infty(x^+) \;&=\; J_M(x, u_\infty^\star) + V_\infty(x^e) + R_5(x, M) \\ &\qquad - J_{M-1}(x^+, u_\infty^\star) - V_\infty(x^e) - R_5(x^+, M-1) \\ &=\; J_M(x, u_\infty^\star) - J_{M-1}(x^+, u_\infty^\star) + R_5(x, M) - R_5(x^+, M-1). \end{aligned}$$

Putting the two equations together and using Lemma 13.38 yields

$$\ell(x, \mu_N(x)) = V_\infty(x) - V_\infty(x^+) + R_7(x, M, N). \tag{13.24}$$

with

$$R_7(x, M, N) = -R_6(x, M, N) + R_6(x^+, M-1, N-1) - R_5(x, M) + R_5(x^+, M-1).$$

From Lemma 13.37 and 13.38 we obtain the bound

$$\begin{aligned} |R_7(x, M, N)| \;&\leq\; 2\gamma_V(\sigma_\delta(P)) + 2\gamma_V(\sigma_\infty(P)) + 2\gamma_{V_\infty}(\sigma_\delta(P)) + 4\gamma_{V_\infty}(\sigma_\infty(P)) \\ &\qquad +4\omega(N-M) \end{aligned}$$

provided we choose $M \in \{1, \dots, N\}$ with $M \notin \mathcal{Q}(x, u_N^\star, P, N) \cup \mathcal{Q}(x, u_\infty^\star, P, \infty)$ and $M-1 \notin \mathcal{Q}(x^+, u_N^\star(\cdot+1), P, N-1) \cup \mathcal{Q}(x^+, u_\infty^\star(\cdot+1), P, \infty)$. Since each of the four $\mathcal{Q}$ sets contains at most P elements, their union contains at most $4P$ elements and hence if $N > 8P$ then there is at least one such M with $M \leq N/2$.

Thus, choosing $P = \lfloor (N-1)/8 \rfloor$ yields the existence of $M \leq N/2$ such that

$$|R_7(x, M, N)| \leq \delta_1(N). \tag{13.25}$$

Applying (13.24), (13.25) for $x = x_{\mu_N}(k, x)$, $k = 0, \dots, K-1$, we can conclude

$$\begin{aligned} J_K^{cl}(x, \mu_N) &= \sum_{k=0}^{K-1} \ell(x_{\mu_N}(k, x), \mu_N(x_{\mu_N}(k, x))) \\ &\leq \sum_{k=0}^{K-1} \Big(V_\infty(x_{\mu_N}(k, x)) - V_\infty(x_{\mu_N}(k+1, x)) + \delta_1(N) \Big) \\ &\leq V_\infty(x) - V_\infty(x_{\mu_N}(K, x)) + K\delta_1(N). \end{aligned}$$

This proves the claim. □

The interpretation of (13.23) is as follows. If we follow the MPC closed-loop trajectory up to some time K and then continue by using the infinite horizon optimal trajectory starting at $x_{\mu_N}(K,x)$, then the value of the overall trajectory exceeds the infinite horizon optimal value by at most $K\delta(N)$. Although seemingly different, it is indeed closely related to Theorem 13.21 because of the following fact: the inequality from Theorem 13.21 holds for all $x \in \mathbb{X}$ if and only if

$$J_K^{cl}(x,\mu_N) + V_\infty(x_{\mu_N}(K,x)) \leq V_\infty(x) + \delta_1(N) \tag{13.26}$$

holds for all $x \in \mathbb{X}$ and all $K \geq 1$. This is because $J_K^{cl}(x,\mu_N) + V_\infty(x_{\mu_N}(K,x)) \leq J_\infty^{cl}(x,\mu_N)$ for all $K \geq 1$, hence Theorem 13.21 implies (13.26). Conversely, since the assumptions of Theorem 13.21 imply $V_\infty(x_{\mu_N}(K,x)) \to V_\infty(x^e) = 0$ for $K \to \infty$, the validity of (13.26) for all $K \geq 1$ implies the inequality from Theorem 13.21 by letting $K \to \infty$. Comparing (13.23) with (13.26) one immediately sees the difference between the case with and without terminal conditions: without terminal conditions we get the additional factor K in front of the error term, which implies that for large K the error may increase and that for $K \to \infty$ and fixed N the solution may be far from optimal. A numerical illustration of this effect can be found in Example 13.41(iii), below. However, note that the estimate from Theorem 13.27 shows that the averaged value still behaves well for $K \to \infty$, hence the behavior of the trajectories cannot completely deteriorate.

Finally, we formulate and prove the counterpart of Theorem 13.22 for the case without terminal conditions. To this end, recall the definition of $\mathbb{U}_\kappa^K(x)$ from (13.14).

Theorem 13.40 Consider the MPC Algorithm 11.1 with strictly dissipative optimal control problem (OCP$_\mathrm{N}$) with bounded storage function λ and $\rho \in \mathcal{K}_\infty$, let $\mathbb{X}$ be bounded and let Assumptions 13.24 and 13.29 hold. Then there exist $\delta_1, \delta_2, \delta_3 \in \mathcal{L}$ such that for all $x \in \mathbb{X}$ the inequality

$$J_K^{cl}(x,\mu_N) \leq \inf_{u \in \mathbb{U}_\kappa^K(x)} J_K(x,u) + \delta_1(N) + K\delta_2(N) + \delta_3(K)$$

holds with $\kappa = \max\{\beta(|x|_{x^e}, K), \rho^{-1}(\delta_1(N))\}$, with δ_1 from Proposition 13.32 and β characterizing the semiglobal practical asymptotic stability in Theorem 13.34.

Proof: First observe that the assumptions of this theorem include those of Theorem 13.34. Hence, from the proof of Theorem 13.34 we obtain the identity

$$\tilde{\ell}(x,\mu_N(x)) = \widetilde{V}_N(x) - \widetilde{V}_N(f(x,\mu_N(x))) + R_2(x,N) + R_4(f(x,\mu_N(x)),N) + R_4(x,N)$$

with $|R_2(x,N) + R_4(f(x,\mu_N(x),N) + R_4(x,N)| \leq \nu_2(a,N) + 2\nu_4(a,N) =: \delta_2(N)$, with ν_2 and ν_4 from Lemma 13.26 and 13.31, respectively, and $a = \sup_{x\in\mathbb{X}} |x|_{x^e}$. Summing this cost along the closed-loop trajectory yields

$$\sum_{k=0}^{K-1} \tilde{\ell}(x_{\mu_N}(k,x), \mu_N(x_{\mu_N}(k,x))) \leq \widetilde{V}_N(x) - \widetilde{V}_N(x_{\mu_N}(K)) + K\delta_2(N). \tag{13.27}$$

Now the dynamic programming principle (12.1) and Assumption 13.29 yield for all $K \in \{1,\ldots,N\}$ and all $u \in \mathbb{U}_\kappa^K(x)$

$$\widetilde{J}_K(x,u) = \underbrace{\widetilde{J}_K(x,u) + \widetilde{V}_{N-K}(x_u(K,x))}_{\geq \widetilde{V}_N(x)} - \underbrace{\widetilde{V}_{N-K}(x_u(K,x))}_{\leq \gamma_{\widetilde{V}}(\kappa)} \geq \widetilde{V}_N(x) - \gamma_{\widetilde{V}}(\kappa). \tag{13.28}$$

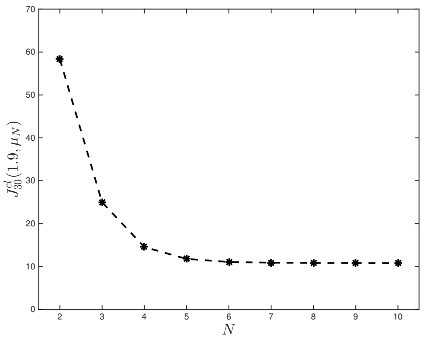


Figure 13.7: Value of $J^{cl}_K(x, \mu_N)$ for $K = 30$, $x = 1.9$ and varying N without terminal conditions

Due to the non-negativity of $\tilde{\ell}$, for $K \geq N$ we get $\widetilde{J}_K(x, u) \geq \widetilde{V}_N(x)$ for all $u \in \mathbb{U}^K(x)$. Hence (13.28) holds for all $K \in \mathbb{N}$. Moreover, we have $\widetilde{V}_N(x) \geq 0$. Using (13.27), (13.28) and (13.12) and the definition of δ_2, for all $u \in \mathbb{U}^K_\kappa(x)$ we obtain

$$\begin{aligned} J^{cl}_K(x, \mu_N(x)) \;&=\; \sum_{k=0}^{K-1} \tilde{\ell}(x_{\mu_N}(k, x), \mu_N(x_{\mu_N}(k, x))) - \lambda(x) + \lambda(x_{\mu_N}(K, x)) \\ &\leq\; \widetilde{V}_N(x) - \widetilde{V}_N(x_{\mu_N}(K, x)) + K\delta_2(N) - \lambda(x) + \lambda(x_{\mu_N}(K, x)) \\ &\leq\; \widetilde{J}_K(x, u) + \gamma_{\widetilde{V}}(\kappa) - \widetilde{V}_N(x_{\mu_N}(K, x)) + K\delta_2(N) - \lambda(x) + \lambda(x_{\mu_N}(K, x)) \\ &=\; J_K(x, u) + \gamma_{\widetilde{V}}(\kappa) - \widetilde{V}_N(x_{\mu_N}(K, x)) + K\delta_2(N) - \lambda(x_u(K, x)) + \lambda(x_{\mu_N}(K, x)) \\ &\leq\; J_K(x, u) + \gamma_{\widetilde{V}}(\kappa) + K\delta_2(N) + 2\gamma_\lambda(\kappa). \end{aligned}$$

Using the definition of κ we can estimate and define

$$\begin{aligned} \gamma_{\widetilde{V}}(\kappa) + 2\gamma_\lambda(\kappa) \;&\leq\; \underbrace{\sup_{x \in \mathbb{X}} \gamma_{\widetilde{V}}(\beta(|x|_{x^e}, K)) + 2\gamma_\lambda(\beta(|x|_{x^e}, K))}_{=:\delta_3(K)} \\ &\quad + \underbrace{\gamma_{\widetilde{V}}(\rho^{-1}(\delta(N))) + 2\gamma_\lambda(\rho^{-1}(\delta(N)))}_{=:\delta_1(N)} \end{aligned}$$

which finishes the proof. □

Example 13.41 (i) Fig. 13.7 illustrates how $J^{cl}_K(x, \mu_N)$ depends on N in Example 13.1. As in Fig. 13.4, the value $K = 30$ is so large that the effect of the term $\delta_2(K)$ is negligible and not visible in the figure, hence $J^{cl}_K(x, \mu_N)$ converges to $\inf_{u \in \mathbb{U}^K_\kappa(x)} J^{uc}_K(x, u)$ for increasing N.

(ii) We note that the error estimate depends on the bound on the storage function λ which enters in several of the previous estimates. This dependence is actually visible when computing $J^{cl}_K(x, \mu_N)$ via numerical simulations. In Example 13.1 the bound on λ increases with increasing $\mathbb{X}$ (cf. Example 13.8). Fig. 13.8 shows that increasing the state constraint set from $\mathbb{X} = [-2, 2]$ to $\mathbb{X} = [-3, 3]$ indeed considerably increases the error, although the

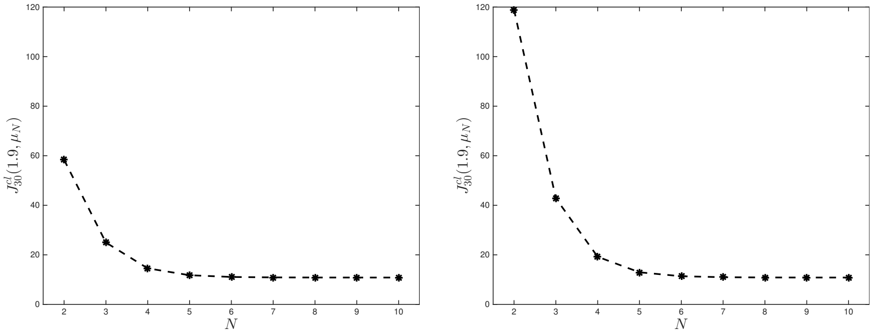


Figure 13.8: Value of $J^{cl}_K(x, \mu_N)$ for $K = 30$, $x = 1.9$ and varying N without terminal conditions for $\mathbb{X} = [-2, 2]$ on the left and $\mathbb{X} = [-3, 3]$ on the right

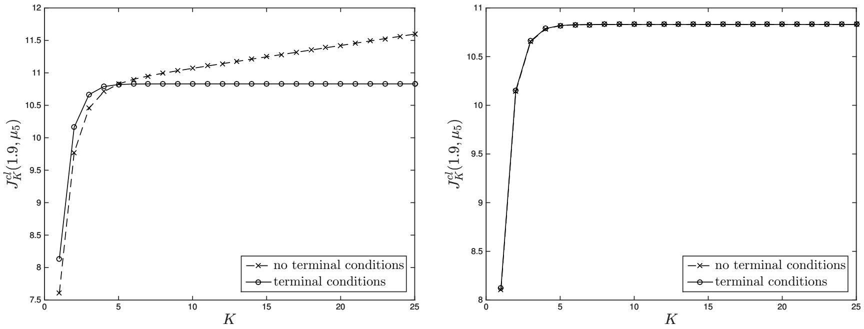


Figure 13.9: Value of $J^{cl}_K(x, \mu_N)$ for varying K, $x = 1.9$ and $N = 5$ on the left and $N = 10$ on the right, both with and without terminal conditions $\mathbb{X}_0 = \{0\}$ and $F \equiv 0$

optimal trajectories and thus the limiting values for $J^{cl}_K(x, \mu_N)$ for $N \to \infty$ are independent of the choice of $\mathbb{X}$.

(iii) Finally we observe that the main structural difference between Theorem 13.22 and 13.40 lies in the factor K in the error estimate in Theorem 13.40 without terminal conditions. This predicts a deterioration of the value $J^{cl}_K(x, \mu_N)$ for fixed N and growing K in the case without terminal conditions, which should not appear if terminal conditions are used. This effect can again be seen in numerical simulations for Example 13.1, see Fig. 13.9. Here the increase of $J^{cl}_K(x, \mu_N)$ for increasing K is clearly visible in the left figure, i.e., for $N = 5$. In the right figure, N has been increased to $N = 10$, due to which the $\delta_2(N)$-term in Theorem 13.40 becomes so small that its effect is not visible anymore for the range of K depicted in the figure.

□

Chapter 14

Analysis of stabilizing MPC schemes

In this chapter we look at the particular — but practically very relevant — special case in which the stage cost ℓ penalizes the distance from a desired equilibrium. More precisely, we consider stage costs satisfying the conditions

$$\ell(x_*, u_*) = 0 \quad \text{and} \quad \ell(x, u) \geq \alpha_3(|x|_{x_*}) \tag{14.1}$$

for all $x \in \mathbb{X}$ and a $\mathcal{K}_\infty$-function α_3. In normed spaces X and U, the simplest choice for such a function is

$$\ell(x, u) = \|x - x_*\| + \lambda\|u - u_*\|$$

for a control penalization parameter $\lambda \geq 0$.

As we have already observed in Example 13.8(i), problems of this kind are always strictly dissipative (with storage function $\lambda \equiv 0$). Hence, all results of the previous chapter apply and — under the stated conditions — we can conclude asymptotic stability for the scheme with terminal conditions and semiglobal practical asymptotic stability without terminal conditions. In practice, however, one often observes "real" asymptotic stability also in the case without terminal conditions. Also, schemes without terminal conditions are often preferred in practice, because for complex systems the design of terminal conditions satisfying Assumption 13.5 is very difficult if not impossible. Hence, in this chapter we will analyze stabilizing MPC schemes without terminal conditions.

14.1 A relaxed dynamic programming theorem

The basis for the considerations in this chapter is the following fundamental, yet simple to prove theorem.

Theorem 14.1 [Asymptotic stability and suboptimality estimate] Consider a stage cost $\ell : X \times U \to \mathbb{R}_0^+$ and a function $V : X \to \mathbb{R}_0^+$. Let $\mu : \mathbb{X} \to U$ be an admissible feedback law and let $S \subseteq \mathbb{X}$ be a forward invariant set for the closed loop system

$$x^+ = g(x) = f(x, \mu(x)). \tag{14.2}$$

Assume there exists $\alpha \in (0,1]$ such that the *relaxed dynamic programming inequality*

$$V(x) \geq \alpha\ell(x, \mu(n,x)) + V(f(x, \mu(n,x))) \tag{14.3}$$

holds for all $x \in S$. Then the *suboptimality estimate*

$$J_\infty^{cl}(x, \mu) \leq V(x)/\alpha \tag{14.4}$$

holds for all $x \in S$.

If, in addition, ℓ satisfies (14.1) and there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that the inequalities

$$\alpha_1(|x|_{x_*}) \leq V(x) \leq \alpha_2(|x|_{x_*}) \tag{14.5}$$

hold for all $x \in \mathbb{X}$, $u \in \mathbb{U}$ and an equilibrium $x_* \in \mathbb{X}$, then the closed loop system (14.2) is asymptotically stable on S in the sense of Definition 10.2.

Proof: In order to prove (14.4) consider $x \in S$ and the trajectory $x_\mu(\cdot)$ of (14.2) with $x_\mu(0) = x$. By forward invariance of the sets S this trajectory satisfies $x_\mu(k) \in S$. Hence from (14.3) for all $k \in \mathbb{N}_0$ we obtain

$$\alpha\ell(x_\mu(k), \mu(x_\mu(k))) \leq V(x_\mu(k)) - V(x_\mu(k+1)).$$

Summing over k yields for all $K \in \mathbb{N}$

$$\alpha \sum_{k=0}^{K-1} \ell(x_\mu(k), \mu(x_\mu(k))) \leq V(x_\mu(0)) - V(x_\mu(K)) \leq V(x)$$

since $V(x_\mu(K)) \geq 0$ and $x_\mu(0) = x$. Since the stage cost ℓ is nonnegative, the term on the left is monotone increasing and bounded, hence for $K \to \infty$ it converges to $\alpha J_\infty^{cl}(x, \mu)$. Since the right hand side is independent of K, this yields (14.4).

The stability assertion now immediately follows by observing that V satisfies all assumptions of Theorem 10.5 with $\alpha_V = \alpha\,\alpha_3$. □

14.2 Bounds on V_N

The central assumption we will use in order to ensure asymptotic stability and performance bounds imposes upper bounds on the optimal value functions V_N. These bounds are formulated relative to the stage cost ℓ. To this end, we define

$$\ell^*(x) := \inf_{u \in U} \ell(x, u). \tag{14.6}$$

With this notation, we can formulate our central assumption.

Assumption 14.2 [Bound on V_N] Consider the optimal control problem ($\mathrm{OCP_N}$). We assume that there exist functions $B_K \in \mathcal{K}_\infty$, $K \in \mathbb{N}$ such that for each $x \in \mathbb{X}$ the inequality

$$V_K(x) \leq B_K(\ell^*(x)) \tag{14.7}$$

holds for all $K \in \mathbb{N}$. □

We observe that $V_K(x) \geq \ell(x, u^\star(0)) \geq \ell^*(x)$ implies $B_K(r) \geq r$.

Before we state consequences from this assumption in the next section, we discuss a sufficient controllability condition which ensures Assumption 14.2. To this end, we first slightly enlarge the class of $\mathcal{KL}$-functions introduced in Definition 10.1.

Definition 14.3 We say that a continuous function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{KL}_0$ if for each $r > 0$ we have $\lim_{t\to\infty} \beta(r,t) = 0$ and for each $t \geq 0$ we either have $\beta(\cdot, t) \in \mathcal{K}_\infty$ or $\beta(\cdot, t) \equiv 0$. □

Compared to the class $\mathcal{KL}$, here we do not assume monotonicity in the second argument and we allow for $\beta(\cdot, t)$ being identically zero for some t. This allows for tighter bounds for the actual controllability behavior of the system. It is, however, easy to see that each $\beta \in \mathcal{KL}_0$ can be overbounded by a $\tilde\beta \in \mathcal{KL}$, e.g., by setting $\tilde\beta(r,t) = \max_{\tau \geq t} \beta(r, \tau) + e^{-t} r$. Using the $\mathcal{KL}_0$ functions we now formulate our controllability assumption.

Assumption 14.4 [Asymptotic controllability wrt. ℓ] Consider the optimal control problem ($\mathrm{OCP_N}$). We assume that the system is *asymptotically controllable with respect to ℓ with rate $\beta \in \mathcal{KL}_0$*, i.e., for each $x \in \mathbb{X}$ and each $N \in \mathbb{N}$ there exists an admissible control sequence $u_x \in \mathbb{U}^N(x)$ satisfying

$$\ell(x_{u_x}(n,x), u_x(n)) \leq \beta(\ell^*(x), n)$$

for all $n \in \{0, \ldots, N-1\}$. □

An important special case for $\beta \in \mathcal{KL}_0$ is

$$\beta(r,n) = C\sigma^n r \tag{14.8}$$

for real constants $C \geq 1$ and $\sigma \in (0,1)$, i.e., *exponential controllability*.

The following lemma links Assumptions 14.4 and 14.2.

Lemma 14.5 If Assumption 14.4 holds then Assumption 14.2 holds. More precisely, for each $K \in \mathbb{N}$ and each $x \in \mathbb{X}$ the inequality

$$V_K(x) \leq J_K(x, u_x) \leq B_K(\ell^*(x)) \tag{14.9}$$

holds for u_x from Assumption 14.4 and

$$B_K(r) := \sum_{n=0}^{K-1} \beta(r,n). \tag{14.10}$$

Proof: The inequality follows immediately from

$$V_K(x) \leq J_K(x, u_x) \ = \ \sum_{n=0}^{K-1} \ell(x(n, u_x), u_x(n))$$

$$\leq \sum_{n=0}^{K-1} \beta(\ell^*(x), n) \;=\; B_K(\ell^*(x)).$$

□ □

In the special case (14.8) the values B_K, $K \in \mathbb{N}$, evaluate to

$$B_K(r) = C\frac{1-\sigma^K}{1-\sigma}r.$$

It is easily seen that if the state trajectories itself are exponentially controllable to some equilibrium x_* then exponential controllability, i.e., Assumption 14.4 with β from (14.8), holds if ℓ has polynomial growth. In particular, this covers the usual linear-quadratic setting for stabilizable systems.

However, even if the system itself is not exponentially controllable, exponential controllability in the sense of Assumption 14.4 can be achieved by proper choice of ℓ, as the following example shows.

Example 14.6 Consider the control system

$$x^+ = x + ux^3$$

with $\mathbb{X} = [-1,1]$ and $U = [-1,1]$. The system is controllable to $x_* = 0$, which can be seen by choosing $u = -1$. This results in the system $x^+ = x - x^3$ whose solutions approach $x_* = 0$ monotonically for $x_0 \in \mathbb{X}$.

However, the system it is not exponentially controllable to 0: exponential controllability would mean that there exist constants $C > 0$, $\sigma \in (0,1)$ such that for each $x \in \mathbb{X}$ there is $u_x \in \mathbb{U}^\infty(x)$ with

$$|x_{u_x}(n,x)| \leq C\sigma^n|x|.$$

This implies that by choosing $n^* > 0$ so large such that $C\sigma^{n^*} \leq 1/2$ holds the inequality

$$|x_{u_x}(n^*,x)| \leq |x|/2 \tag{14.11}$$

must hold for each $x \in \mathbb{X}$. However, for each $x \geq 0$ the restriction $u \in [-1,1]$ implies $x^+ \geq x - x^3 = (1-x^2)x$ which by induction yields

$$x_u(n^*,x) \geq (1-x^2)^{n^*}x$$

for all $u \in \mathbb{U}^\infty(x)$ which contradicts (14.11) for $x < 1 - 2^{-1/n^*}$.

On the other hand, since $|x| \leq 1$ we obtain $(1-x^2)^2(2x^2+1) = 1 + 2x^6 - 3x^4 \leq 1$ which implies

$$\frac{1}{(1-x^2)^2} \geq 2x^2 + 1 \;\;\Rightarrow\;\; -\frac{1}{2x^2(1-x^2)^2} \leq -\frac{2x^2+1}{2x^2} = -1 - \frac{1}{2x^2}.$$

Hence, choosing

$$\ell(x,u) = \ell(x) = e^{-\frac{1}{2x^2}},$$

for $u \equiv -1$ we obtain

$$\ell(x^+) = \ell(x - x^3) = e^{-\frac{1}{2x^2(1-x^2)^2}} = e^{-\frac{1}{2x^2(1-x^2)^2}} \leq e^{-1}e^{-\frac{1}{2x^2}} = e^{-1}\ell(x).$$

By induction this implies Assumption 14.4 with β from (14.8) with $C = 1$ and $\sigma = e^{-1}$.

□

For certain results it will be useful that β in Assumption 14.4 has the property

$$\beta(r, n+m) \leq \beta(\beta(r,n), m) \quad \text{for all } r \geq 0, n, m \in \mathbb{N}_0. \tag{14.12}$$

Inequality (14.12), often referred to as *submultiplicativity*, ensures that any sequence of the form $b_n = \beta(r,n)$, $r > 0$, also fulfills $b_{n+m} \leq \beta(b_n, m)$. It is, for instance, always satisfied in the exponential case (14.8). If needed, this property can be assumed without loss of generality, because by Sontag's $\mathcal{KL}$-Lemma [18, Proposition 7] the function β in Assumption 14.4 can be replaced by a β of the form $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$ for $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$. Then, (14.12) is easily verified if $\alpha_2 \circ \alpha_1(r) \geq r$, which is equivalent to $\alpha_1 \circ \alpha_2(r) \geq r$, which in turn is a necessary condition for Assumption 14.4 to hold for $n = 0$ and $\beta(r,t) = \alpha_1(\alpha_2(r)e^{-t})$.

14.3 Implications of the bounds on V_N

In this section we will use the bound on the V_N from Assumption 14.2 in order to establish two lemmas which yield bounds for optimal value functions and functionals along pieces of optimal trajectories. In the subsequent section, these bounds will then be used for the calculation of α in (14.3).

In order to be able to calculate α in (14.3), we will need an upper bound for $V_N(f(x, \mu_N(x)))$. To this end, recall from Step (3) of Algorithm 11.1 that $\mu_N(x_0)$ is the first element of the optimal control sequence $u^\star(\cdot)$ for ($\mathrm{OCP_N}$) with initial value x_0. In particular, this implies $f(x_0, \mu_N(x_0)) = x_{u^\star}(1, x_0)$. Hence, if we want to derive an upper bound for $V_N(f(x_0, \mu_N(x_0)))$ then we can alternatively derive an upper bound for $V_N(x_{u^\star}(1, x_0))$. This will be done in the following lemma.

Lemma 14.7 Suppose Assumption 14.2 holds and consider $x_0 \in \mathbb{X}$ and an optimal control $u^\star \in \mathbb{U}^N(x_0)$ for ($\mathrm{OCP_N}$). Then for each $j = 0, \ldots, N-2$ the inequality

$$V_N(x_{u^\star}(1, x_0)) \leq J_j(x_{u^\star}(1, x_0)), u^\star(1 + \cdot)) + B_{N-j}(\ell^*(x_{u^\star}(1 + j, x_0)))$$

holds for B_K from (14.7).

Proof: We define the control sequence

$$\tilde{u}(n) = \begin{cases} u^\star(1+n), & n \in \{0, \ldots, j-1\} \\ u_x(n-j), & n \in \{j, \ldots, N-1\}, \end{cases}$$

where u_x is an optimal control for initial value $x = x_{u^\star}(1+j, x_0)$ and $N = N - j$. By construction, this control sequence is admissible for $x_{u^\star}(1, x_0)$ and we obtain

$$\begin{aligned} V_N(x_{u^\star}(1,x_0)) &\le J(x_{u^\star}(1,x_0), \tilde{u}) \\ &= J_j(x_{u^\star}(1,x_0), u^\star(1+\cdot)) + J_{N-j}(x_{u^\star}(1+j,x_0), u_x) \\ &\le J_j(x_{u^\star}(1,x_0), u^\star(1+\cdot)) + B_{N-j}(\ell^*(x_{u^\star}(1+j,x_0))) \end{aligned}$$

where we used $J_{N-j}(x_{u^\star}(1+j, x_0), u_x) = V_{N-j}(x_{u^\star}(1+j, x_0))$ and Assumption 14.2 in the last step. This is the desired inequality. □

In words, the idea of this proof is as follows. The upper bound for each $j \in \{0, \dots, N-2\}$ is obtained from a specific trajectory. We follow the optimal trajectory for initial value x_0 and horizon N for j steps and for the point x reached this way we use the optimal control sequence for initial value x and horizon $N - j$ for another $N - j$ steps.

In the next lemma we derive upper bounds for the J_k-terms along tails of the optimal trajectory $x_{u^\star}$, which will later be used in order to bound the right hand side of the inequality from Lemma 14.7. To this end we use that these tails are optimal trajectories themselves.

Lemma 14.8 Suppose Assumption 14.2 holds and consider $x_0 \in \mathbb{X}$ and an optimal control $u^\star \in \mathbb{U}^N(x_0)$ for $(\mathrm{OCP_N})$. Then for each $k = 0, \dots, N-1$ the inequality

$$J_{N-k}(x_{u^\star}(k,x_0), u^\star(k+\cdot)) \le B_{N-k}(\ell^*(x_{u^\star}(k,x_0)))$$

holds for B_K from (14.7).

Proof: Corollary 12.3 implies $J_{N-k}(x_{u^\star}(k, x_0), u^\star(k+\cdot)) = V_{N-k}(x_{u^\star}(k, x_0))$. Hence the assertion follows immediately from Assumption 14.2. □

Remark 14.9 Since $u^\star \in \mathbb{U}^N(x_0)$ we obtain $x_{u^\star}(k, x_0) \in \mathbb{X}$ for $k = 0, \dots, N$. For $k = 0, \dots, N-1$ this property is crucial for the proof of Lemma 14.7 because it ensures that an optimal control for initial value $x = x_{u^\star}(1+j, x_0)$ exists. Note, however, that we do not need $x_{u^\star}(N, x_0) \in \mathbb{X}$. In fact, all results in this and the ensuing sections remain true if we remove the state constraint on $x_{u^\star}(N, x_0) \in \mathbb{X}$ from the definition of $\mathbb{U}^N(x_0)$ or replace it by some weaker constraint. □

14.4 Computation of α

We will now use the inequalities derived in the previous section in order to compute α for which (14.3) holds for all $x \in \mathbb{X}$. When trying to put together these inequalities in order to bound $V_N(x_{u^\star}(1, x_0))$ from above, one notices that the functionals in Lemma 14.7 and 14.8 are not exactly the same. Hence, in order to combine these results into a closed form which is suitable for computing α we need to look at the single terms of the stage cost ℓ contained in these functionals.

To this end, let $u^\star$ be an optimal control for $(\mathrm{OCP_N})$ with initial value $x_0 = x$. Then from the definition of V_N and μ_N it follows that (14.3) is equivalent to

$$\sum_{k=0}^{N-1} \ell(x_{u^\star}(k,x), u^\star(k)) \geq \alpha\ell(x, u^\star(0)) + V_N(x_{u^\star}(1,x)). \tag{14.13}$$

Thus, in order to compute α for which (14.3) holds for all $x \in \mathbb{X}$ we can equivalently compute α for which (14.13) holds for all optimal trajectories $x_{u^\star}(\cdot, x)$ with initial values $x \in \mathbb{X}$.

For this purpose we now consider arbitrary real values $\lambda_0, \ldots, \lambda_{N-1}, \nu \geq 0$ and start by deriving necessary conditions which hold if these values coincide with the cost along an optimal trajectory $\ell(x_{u^\star}(k,x), u^\star(k))$ and an optimal value $V_N(x_{u^\star}(1,x))$, respectively.

Proposition 14.10 Suppose Assumption 14.2 holds and consider $N \geq 1$, values $\lambda_n \geq 0$, $n = 0, \ldots, N-1$, and a value $\nu \geq 0$. Consider $x \in \mathbb{X}$ and assume that there exists an optimal control sequence $u^\star \in \mathbb{U}^N(x)$ for $(\mathrm{OCP_N})$ such that

$$\lambda_k = \ell(x_{u^\star}(k,x), u^\star(k)), \quad k = 0, \ldots, N-1$$

holds. Then

$$\sum_{n=k}^{N-1} \lambda_n \leq B_{N-k}(\lambda_k), \quad k = 0, \ldots, N-2 \tag{14.14}$$

holds. If, furthermore,

$$\nu = V_N(x_{u^\star}(1,x))$$

holds then

$$\nu \leq \sum_{n=0}^{j-1} \lambda_{n+1} + B_{N-j}(\lambda_{j+1}), \quad j = 0, \ldots, N-2 \tag{14.15}$$

holds. □

Proof: If the stated conditions hold, then λ_n and ν must meet the inequalities given in Lemmas 14.7 and 14.8, which is exactly (14.15) and (14.14). □

Using this proposition we can give a sufficient condition for (14.13) and thus for (14.3). The idea behind the following proposition is to express the terms in inequality (14.13) using the values $\lambda_0, \ldots, \lambda_{N-1}$ and ν introduced above.

Proposition 14.11 Consider $N \geq 1$ and $B_K \in \mathcal{K}_\infty$, $K = 2, \ldots, N$ and assume that all values $\lambda_n \geq 0$, $n = 0, \ldots, N-1$ and $\nu \geq 0$ fulfilling (14.14) and (14.15) satisfy the inequality

$$\sum_{n=0}^{N-1} \lambda_n - \nu \geq \alpha\lambda_0 \tag{14.16}$$

for some $\alpha \in (0, 1]$. Then for this α and each optimal control problem $(\mathrm{OCP_N})$ satisfying Assumption 14.2 inequality (14.3) holds for μ_N from Algorithm 11.1 and all $x \in \mathbb{X}$. □

Proof: Consider a control system satisfying Assumption 14.2 and an optimal control sequence $u^\star \in \mathbb{U}^N(x)$ for initial value $x \in \mathbb{X}$. Then by Proposition 14.10 the values $\lambda_k = \ell(x_{u^\star}(k,x), u^\star(k))$ and $\nu = V_N(x_{u^\star}(1,x))$ satisfy (14.14) and (14.15), hence by assumption also (14.16). Thus, using $\ell(x, u^\star(0)) = \ell(x_{u^\star}(0,x), u^\star(0)) = \lambda_0$ we obtain

$$V_N(x_{u^\star}(1,x)) + \alpha\ell(x, u^\star(0)) = \nu + \alpha\lambda_0 \quad \leq \quad \sum_{k=0}^{N-1} \lambda_k = \sum_{k=0}^{N-1} \ell(x_{u^\star}(k,x), u^\star(k)).$$

This proves (14.13) and thus also (14.3). □

Proposition 14.11 is the basis for computing α as specified in the following theorem.

Theorem 14.12 [Abstract optimization problem] Consider $N \geq 1$ and $B_K \in \mathcal{K}_\infty$, $K = 2, \ldots, N$ and assume that the optimization problem

$$\begin{aligned} &\alpha := \inf_{\lambda_0,\ldots,\lambda_{N-1},\nu} \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\lambda_0} \\ &\text{subject to the constraints (14.14), (14.15), and} \\ &\lambda_0 > 0, \lambda_1, \ldots, \lambda_{N-1}, \nu \geq 0 \end{aligned} \tag{14.17}$$

has an optimal value $\alpha \in (0,1]$. Then for this α and each optimal control problem (OCP$_\mathrm{N}$) satisfying Assumption 14.2 inequality (14.3) holds for μ_N from Algorithm 11.1 and all $x \in \mathbb{X}$.

Proof: Consider arbitrary values $\lambda_0, \ldots, \lambda_{N-1}, \nu \geq 0$ satisfying (14.14) and (14.15).

If $\lambda_0 > 0$ then the definition of Problem (14.17) immediately implies (14.16).

If $\lambda_0 = 0$, then inequality (14.14) for $k = 0$ together with $B_K(0) = 0$ implies $\lambda_1, \ldots, \lambda_{N-1} = 0$. Thus, (14.15) for $j = 1$ yields $\nu = 0$ and again (14.16) holds.

Hence, (14.16) holds in both cases and Proposition 14.11 yields the assertion. □

Remark 14.13 (i) Theorem 14.12 shows Inequality (14.3) for all $x \in \mathbb{X}$ if Assumption 14.2 or, alternatively, Assumption 14.4 holds for all $x \in \mathbb{X}$ and $K = 2, \ldots, N$.

If we want to establish Inequality (14.3) only for states $x_0 \in Y$ for a subset $Y \subset \mathbb{X}$, then from the proofs of the Lemmas 14.7 and 14.8 it follows that Proposition 14.10 holds for all $x_0 \in Y$ (instead of for all $x_0 \in \mathbb{X}$) under the following condition:

$$\begin{aligned} &\text{(14.7) holds for } x = x_{u^\star}(k, x_0) \text{ for all } k = 0, \ldots, N-1, \text{ all } x_0 \in Y \\ &\text{and all } K = 2, \ldots, N, \text{ where } u^\star \text{ is the optimal control for } J_N(x_0, u). \end{aligned} \tag{14.18}$$

This implies that under condition (14.18) Theorem 14.12 holds for all $x_0 \in Y$ and consequently (14.3) holds for all $x_0 \in Y$.

(ii) A further relaxation of the assumptions of Theorem 14.12 can be obtained by observing that if we are interested in establishing Inequality (14.3) only for states $x_0 \in Y$, then in (14.17) we only need to optimize over those λ_i which correspond to optimal trajectories starting in Y. In particular, if we know that $\inf_{x_0 \in Y} \ell^*(x_0) \geq \zeta$ for some $\zeta > 0$, then the constraint $\lambda_0 > 0$ can be tightened to $\lambda_0 \geq \zeta$. □

The following lemma shows that the optimization problem (14.17) specializes to a linear program if the functions $B_K(r)$ are linear in r.

Lemma 14.14 If the functions $B_K(r)$ from (14.7) in the constraints (14.14), (14.15) are linear in r, then α from Problem (14.17) coincides with

$$\begin{aligned}&\alpha := \min_{\lambda_0,\dots,\lambda_{N-1},\nu} \sum_{n=0}^{N-1} \lambda_n - \nu \\ &\text{subject to the (now linear) constraints (14.14), (14.15), and} \\ &\lambda_0 = 1, \lambda_1, \dots, \lambda_{N-1}, \nu \geq 0.\end{aligned} \tag{14.19}$$

In particular, this holds if Assumption 14.4 holds with functions $\beta(r,t)$ being linear in r.

Proof: Due to the linearity, all sequences $\bar\lambda_0, \dots, \bar\lambda_{N-1}, \bar\nu$ satisfying the constraints in (14.17) can be written as $\gamma\lambda_0, \dots, \gamma\lambda_{N-1}, \gamma\nu$ for some $\lambda_0, \dots, \lambda_{N-1}, \nu$ satisfying the constraints in (14.19), where $\gamma = \bar\lambda_0$. Since

$$\frac{\sum_{n=0}^{N-1} \bar\lambda_n - \bar\nu}{\bar\lambda_0} = \frac{\sum_{n=0}^{N-1} \gamma\lambda_n - \gamma\nu}{\gamma\lambda_0} = \frac{\sum_{n=0}^{N-1} \lambda_n - \nu}{\lambda_0} = \sum_{n=0}^{N-1} \lambda_n - \nu,$$

the values α in Problems (14.17) and (14.19) coincide. □

The next result gives an explicit bound for Problem (14.19) and thus also (14.17) if the functions B_K are linear.

Proposition 14.15 If the functions $B_K(r)$ from (14.7) in the constraints (14.14), (14.15) are linear in r, then the solution of Problems (14.17) and (14.19) satisfies the inequality

$$\alpha \geq \tilde\alpha_N \tag{14.20}$$

for

$$\tilde\alpha_N := 1 - (\gamma_2 - 1)(\gamma_N - 1) \prod_{k=2}^{N-1} \left(\frac{\gamma_k - 1}{\gamma_k}\right) \quad \text{with } \gamma_k = B_k(r)/r. \tag{14.21}$$

□

Proof: We prove the theorem by showing the inequality

$$\lambda_{N-1} \leq (\gamma_N - 1) \prod_{k=2}^{N-1} \left(\frac{\gamma_k - 1}{\gamma_k}\right) \lambda_0 \tag{14.22}$$

for all feasible $\lambda_0, \dots, \lambda_{N-1}$. From this (14.20) follows since (14.15) with $j = N-2$ implies

$$\nu \leq \sum_{n=1}^{N-2} \lambda_n + \gamma_2 \lambda_{N-1}$$

and thus (14.22), $\gamma_2 \geq 1$ and $\lambda_0 = 1$ yield

$$\sum_{n=0}^{N-1} \lambda_n - \nu \geq \lambda_0 + (1-\gamma_2)\lambda_{N-1} \geq \lambda_0 - (\gamma_2 - 1)(\gamma_N - 1)\prod_{k=2}^{N-1}\left(\frac{\gamma_k - 1}{\gamma_k}\right)\lambda_0 = \tilde{\alpha}_N$$

for all feasible $\lambda_1, \ldots, \lambda_{N-1}$ and ν, which yields $\alpha \geq \tilde{\alpha}_N$.

In order to prove (14.22), we start by observing that (14.14) with $j = p$ implies

$$\sum_{k=p+1}^{N-1} \lambda_k \leq (\gamma_{N-p} - 1)\lambda_p \tag{14.23}$$

for $p = 0, \ldots, N-2$. From this we can conclude

$$\lambda_p + \sum_{k=p+1}^{N-1} \lambda_k \geq \frac{\sum_{k=p+1}^{N-1} \lambda_k}{\gamma_{N-p} - 1} + \sum_{k=p+1}^{N-1} \lambda_k = \frac{\gamma_{N-p}}{\gamma_{N-p} - 1}\sum_{k=p+1}^{N-1} \lambda_k.$$

Using this inequality inductively for $p = 1, \ldots, N-2$ yields

$$\sum_{k=1}^{N-1} \lambda_k \geq \prod_{k=1}^{N-2}\left(\frac{\gamma_{N-k}}{\gamma_{N-k} - 1}\right)\lambda_{N-1} = \prod_{k=2}^{N-1}\left(\frac{\gamma_k}{\gamma_k - 1}\right)\lambda_{N-1}.$$

Using (14.23) for $p = 0$ we then obtain

$$(\gamma_N - 1)\lambda_0 \geq \sum_{k=1}^{N-1} \lambda_k \geq \prod_{k=2}^{N-1}\left(\frac{\gamma_k}{\gamma_k - 1}\right)\lambda_{N-1}$$

which implies (14.22). □

A much more complicated proof (see [8, Proposition 6.18]) shows that the optimal α_N is given by

$$\alpha_N := 1 - \frac{(\gamma_N - 1)\prod_{k=2}^{N}(\gamma_k - 1)}{\prod_{k=2}^{N}\gamma_k - \prod_{k=2}^{N}(\gamma_k - 1)} \quad \text{with } \gamma_k = B_k(r)/r, \tag{14.24}$$

A comparison of the two formulas (14.24) and (14.20) can be found in Remark 14.17, below.

14.5 Main Stability and Performance Results

We are now ready to state our main result on stability and performance of stabilizing MPC without terminal conditions.

Theorem 14.16 [Stability without terminal conditions] Consider the MPC Algorithm 11.1 with optimization horizon $N \in \mathbb{N}$ and stage cost ℓ satisfying $\alpha_3(|x|_{x_*}) \leq \ell^*(x) \leq \alpha_4(|x|_{x_*})$ for suitable $\alpha_3, \alpha_4 \in \mathcal{K}_\infty$. Suppose that Assumption 14.2 holds and that $\alpha = \alpha_N$

from Formula (14.24) or $\alpha = \tilde{\alpha}_N$ from Formula (14.20) satisfies $\alpha \in (0,1]$. Then the MPC closed loop system (11.2) with MPC-feedback law μ_N is asymptotically stable on $\mathbb{X}$.

In addition, the inequality

$$J^{cl}_\infty(x,\mu_N) \le V_N(x)/\alpha \le V_\infty(x)/\alpha$$

holds for each $x \in \mathbb{X}$.

Proof: First note that $V_N \le V_\infty$ follows immediately from $\ell \ge 0$. Hence, the assertion follows readily from Theorem 14.1 if we prove the inequalities (14.3) and (14.5). Inequality (14.3) follows directly from Theorem 14.12 and Proposition 14.15 or [8, Proposition 6.18].

Regarding (14.5), observe that the inequality for ℓ follows immediately from our assumptions. From the definition of V_N we get

$$V_N(x) = \inf_{u\in\mathbb{U}^N(x)} J_N(x,u) \ge \inf_{u\in\mathbb{U}^N(x)} \ell(x,u(0)) = \ell^*(x) \ge \alpha_3(|x|_{x_*}),$$

thus the lower inequality for V_N follows with $\alpha_1 = \alpha_3$. The upper inequality in (14.5) follows from Assumption 14.2 and the upper bound on ℓ^* via

$$V_N(x) \le B_N(\ell^*(x)) \le B_N(\alpha_4(|x|_{x_*})),$$

i.e., for $\alpha_2 = B_N \circ \alpha_4$. □

Remark 14.17 Let us compare the two different bounds on α given by $\tilde{\alpha}_N$ from (14.20) and α_N from (14.24). In order to illustrate that the criterion $\tilde{\alpha}_N > 0$ is more conservative than the criterion $\alpha_N > 0$, we consider the case where $\gamma_k = \gamma$ for all k, i.e., the γ_k are independent of k, and compute the minimal N for which $\tilde{\alpha}_N > 0$ and $\alpha_N > 0$, respectively, hold. For $\gamma_k = \gamma$ the expressions simplify to

$$\tilde{\alpha}_N = 1 - \frac{(\gamma-1)^N}{\gamma^{N-2}} \quad \text{and} \quad \alpha_N = 1 - \frac{(\gamma-1)^N}{\gamma^{N-1} - (\gamma-1)^{N-1}}.$$

Thus, an optimization horizon N for which $\tilde{\alpha}_N > 0$ must satisfy

$$N > 2 + 2\frac{\ln\gamma}{\ln\gamma - \ln(\gamma-1)}$$

while the same condition for $\tilde{\alpha}_N > 0$ is given by

$$N > 2 + \frac{\ln(\gamma-1)}{\ln\gamma - \ln(\gamma-1)}.$$

This means that the estimate for the minimal stabilizing horizon based on $\tilde{\alpha}_N$ is about twice as large as the estimate based on α_N.

In this context, it is interesting to look at the asymptotic behavior of the bounds on N for $\gamma \to \infty$. For large γ the denominator is approximately $1/\gamma$. This implies that asymptotically for $\gamma \to \infty$ the first estimate for N behaves like $2\gamma\ln\gamma$ while the second behaves like $\gamma\ln\gamma$. □

The class of systems which is covered by Theorem 14.16 is quite large, since, e.g., exponential controllability holds on compact sets $\mathbb{X}$ whenever the linearization of f in x_* is stabilizable and ℓ is quadratic.

The following simple example illustrates the use of Theorem 14.16 for the case of a nonexponentially controllable system.

Example 14.18 We reconsider Example 14.6, i.e.,

$$x^+ = x + ux^3 \quad \text{with} \quad \ell(x,u) = e^{-\frac{1}{2x^2}}.$$

As shown in Example 14.6, Assumption 14.4 holds with $\beta(r,k) = C\sigma^k r$ with $C = 1$ and $\sigma = e^{-1}$. The bounds in Assumption 14.2 resulting from this β according to (14.10) are

$$B_K(r) = C\frac{1-\sigma^K}{1-\sigma}r = C\frac{1-e^{-K}}{1-e^{-1}}r,$$

thus Theorem 14.16 is applicable and we obtain $\alpha \geq \alpha_N$ with α_N from Formula (14.24). The γ_k in Formula (14.24) are given by

$$\gamma_k = C\frac{1-e^{-k}}{1-e^{-1}}.$$

A straightforward computation reveals that for these values Formula (14.24) simplifies to

$$1 - \frac{(\gamma_N - 1)\prod_{k=2}^{N}(\gamma_k - 1)}{\prod_{k=2}^{N}\gamma_k - \prod_{k=2}^{N}(\gamma_k - 1)} = 1 - e^{-N}.$$

Hence, for $N = 2$ we obtain $\alpha = 1 - e^{-2} \approx 0.865$ and for $N = 3$ we get $\alpha \geq 1 - e^{-3} \approx 0.95$. Hence, Theorem 14.16 ensures asymptotic stability for all $N \geq 2$ and — since $1/0.95 \approx 1.053$ — for $N = 3$ the performance of the MPC controller is at most about 5.3% worse than the infinite horizon controller. □

While in this simple example the computation of α via Formula (14.24) is possible, in many practical examples this will not be the case. However, Formula (14.24) can still be used to obtain valuable information for the design of MPC schemes. This aspect will be discussed at the end of this section.

Although the main benefit of the approach developed in this chapter compared to other approaches is that we can get rather precise quantitative estimates, it is nevertheless good to know that our approach also guarantees asymptotic stability for sufficiently large optimization horizons N under suitable assumptions. This is the statement of our final stability result.

Theorem 14.19 [Stability for sufficiently large N] Consider the MPC Algorithm 11.1 with optimization horizon $N \in \mathbb{N}$ and stage cost ℓ satisfying $\alpha_3(|x|_{x_*}) \leq \ell^*(x) \leq \alpha_4(|x|_{x_*})$

for suitable $\alpha_3, \alpha_4 \in \mathcal{K}_\infty$. Suppose that Assumption 14.2 holds for linear $B_K \in \mathcal{K}_\infty$ of the form $B_K(r) = \gamma_K r$ with $\gamma_\infty := \sup_{k\in\mathbb{N}} \gamma_k < \infty$.

Then the MPC closed loop system (11.2) with MPC-feedback law μ_N is asymptotically stable on $\mathbb{X}$ provided N is sufficiently large.

Furthermore, for each $C > 1$ there exists $N_C > 0$ such that

$$J_\infty^{cl}(x, \mu_N) \le CV_N(x) \le CV_\infty(x)$$

holds for each $x \in \mathbb{X}$ and each $N \ge N_C$.

Proof: The assertion follows immediately from Theorem 14.16 if we show that $\tilde{\alpha}_N \to 1$ holds in (14.20) as $N \to \infty$. Since all factors in (14.20) are monotone increasing in γ_k and the product has a negative sign, we obtain

$$\tilde{\alpha}_N \ge 1 - (\gamma_\infty - 1)^2 \left(\frac{\gamma_\infty - 1}{\gamma_\infty}\right)^{N-2}.$$

Since $(\gamma_\infty - 1)/\gamma_\infty < 1$ we obtain that

$$\left(\frac{\gamma_\infty - 1}{\gamma_\infty}\right)^{N-2} \to 0$$

as $N \to \infty$ and thus $\tilde{\alpha}_N \to 1$. □

Remark 14.20 For B_K of the form (14.10), a sufficient condition for the γ_k being bounded by γ_∞ is that Assumption 14.4 holds for a $\beta \in \mathcal{KL}_0$ which is linear in its first argument and is summable, i.e.,

$$\sum_{k=0}^{\infty} \beta(r, k) < \infty \quad \text{for all } r > 0.$$

□

Theorem 14.19 justifies what is often done in practice: we set up an MPC scheme using a reasonable stage cost ℓ and increase N until the closed loop system becomes stable.

Of course, Theorem 14.19 immediately leads to the question how large the optimization horizon N needs to be for achieving stability or a certain performance. As the computational cost grows with the length of a horizon, this is also important for the practical implementability of the MPC scheme. We investigate this question for the case that the asymptotic controllability condition from Assumption 14.4 holds with the exponential functions $\beta(r, n) = C\sigma^N r$ from (14.8). To this end, we look at the minimal horizon N for which α_N is larger than a certain threshold depending on the parameters C and σ. This dependence is illustrated in Figure 14.5 for thresholds 0 and 0.5.

As we see, the two parameters C and σ play a very different role. While for fixed $\sigma > 0$ it is always possible to reduce the necessary horizon to $N = 2$, i.e., to the shortest possible horizon, by making C smaller, this is not possible for fixed C by reducing σ. Hence, the constant C plays a more important role for obtaining stability and performance with small optimization horizon N. Particularly, any tuning of the stage cost ℓ which leads to a reduction of C is likely also to reduce the necessary optimization horizon. In the lecture, it will be shown how this observation can explain the parameter dependence of the stability behavior of the third example in Section 9.1.

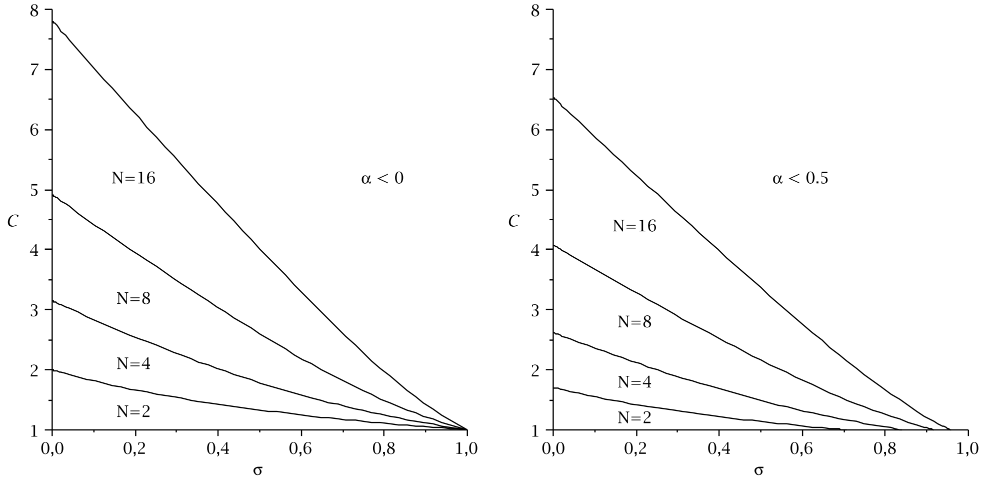


Figure 14.1: Suboptimality regions for different optimization horizons N depending on C and σ in (14.8) for $\alpha_N > 0$ (left) and $\alpha_N > 0.5$ (right)

Bibliography

[1] D. Angeli, R. Amrit, and J. B. Rawlings, *On average performance and stability of economic model predictive control*, IEEE Trans. Autom. Control, 57 (2012), pp. 1615–1626.

[2] F. Colonius, *Einführung in die Steuerungstheorie.* Vorlesungsskript, Universität Augsburg, 1992, eine aktuelle Version ist erhältlich unter dem Link "Lehre" auf `scicomp.math.uni-augsburg.de/~colonius/`.

[3] J. Doležal, *Existence of optimal solutions in general discrete systems*, Kybernetika, 11 (1975), pp. 301–312.

[4] T. Faulwasser and D. Bonvin, *On the design of economic NMPC based on approximate turnpike properties*, in Proceedings of the 54th IEEE Conference on Decision and Control — CDC 2015, 2015, pp. 4964–4970.

[5] L. Grüne, *Stabilität und Stabilisierung linearer Systeme.* Vorlesungsskript, Universität Bayreuth, 2003, `www.math.uni-bayreuth.de/~lgruene/linstab0203/`.

[6] L. Grüne, *Economic receding horizon control without terminal constraints*, Automatica, 49 (2013), pp. 725–734.

[7] L. Grüne and O. Junge, *Gewöhnliche Differentialgleichungen. Eine Einführung aus der Perspektive der Dynamischen Systeme*, Springer Spektrum, 2. aktualisierte auflage ed., 2016.

[8] L. Grüne and J. Pannek, *Nonlinear Model Predictive Control. Theory and Algorithms*, Springer-Verlag, London, 2nd ed., 2017.

[9] L. Grüne and M. Stieler, *Asymptotic stability and transient optimality of economic MPC without terminal conditions*, J. Proc. Control, 24 (2014), pp. 1187–1196.

[10] W. Hahn, *Stability of Motion*, Springer–Verlag Berlin, Heidelberg, 1967.

[11] D. Hinrichsen and A. J. Pritchard, *Mathematical systems theory I*, vol. 48 of Texts in Applied Mathematics, Springer, Heidelberg, 2010. Modelling, state space analysis, stability and robustness, Corrected reprint [of MR2116013].

[12] S. S. Keerthi and E. G. Gilbert, *An existence theorem for discrete-time infinite horizon optimal control problems*, IEEE Trans. Automat. Contr., 30 (1985), pp. 907–909.

[13] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.

[14] J. Lunze, *Regelungstechnik 1*, Springer, 10 ed., 2010.

[15] M. A. Müller, D. Angeli, and F. Allgöwer, *On necessity and robustness of dissipativity in economic model predictive control*, IEEE Trans. Autom. Control, 60 (2015), pp. 1671–1676.

[16] M. A. Müller and L. Grüne, *Economic model predictive control without terminal constraints for optimal periodic behavior*, Automatica, 70 (2016), pp. 128–139.

[17] E. D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[18] E. D. Sontag, *Comments on integral variants of ISS*, Syst. Control Lett., 34 (1998), pp. 93–100.

[19] ——, *Mathematical Control Theory*, Springer Verlag, New York, 2nd ed., 1998.