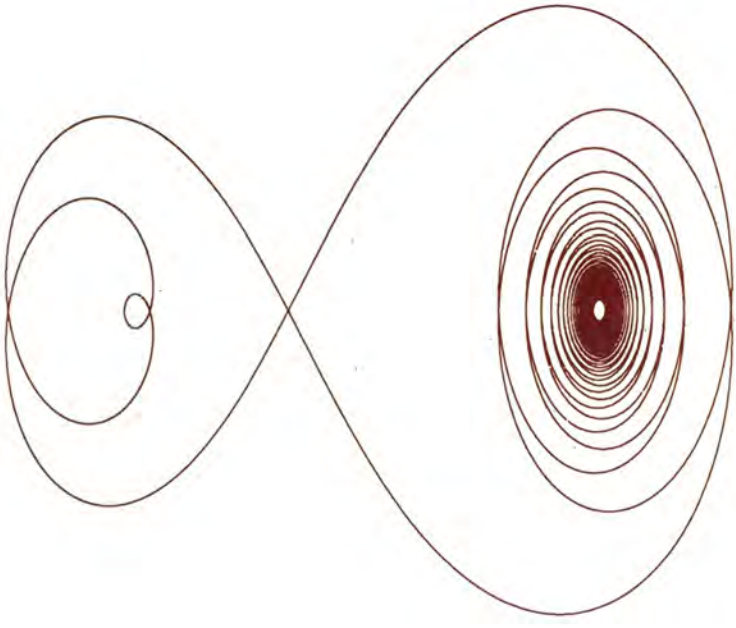
Algebraic and computational aspects of time-delay systems



Luc Habets

Algebraic and computational aspects of time-delay systems

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof.dr. J.H. van Lint, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op woensdag 1 juni 1994 om 16.00 uur

door

LUCAS CLEMENS GERARDUS JOZEF MARIA HABETS

geboren te Eindhoven

Druk: Boek- en Offsetdrukkerij Letru, Helmond, 04920-37797

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. M.L.J. Hautus
prof.dr. J.M. Schumacher

Copromotor: dr.ir. H.J.C. Huijberts

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Habets, Luc C.G.J.M.

Algebraic and computational aspects of time-delay systems / Luc C.G.J.M. Habets. - Eindhoven : Eindhoven University of Technology
Thesis Eindhoven. - With ref.
ISBN 90-386-0403-3
Subject headings: time-delay systems / linear systems over rings.

Et eunt homines mirari alta montium
et ingentes fluctus maris
et latissimos lapsus fluminum
et oceani ambitum et gyros siderum
et reliquunt se ipsos ...

Augustinus, *Conf. X, VIII, 15*

Voorwoord

Als kleine jongen boezemden de imposante gebouwen van de Technische Universiteit in mijn geboorteplaats Eindhoven mij altijd veel ontzag in. In mijn verbeelding was dit een tempel van wetenschap, waar professoren als een soort halfgoden gecompliceerde theorieën ontwikkelden, die voor gewone stervelingen totaal onbevattelijk waren. Hoe anders was de realiteit toen ik, nu bijna tien jaar geleden, aan diezelfde universiteit aan mijn studie wiskunde begon. In plaats van de halfgoden die ik verwachtte aan te treffen, bleken de meeste docenten bijzonder enthousiaste en hulpvaardige mensen. Aan hen had ik het mede te danken dat mijn studie zo vlotjes verliep.

De eerste kennismaking tijdens mijn afstuderen met echt wetenschappelijk onderzoek sprak me bijzonder aan, en toen zich hierna de mogelijkheid voordeed om in deze richting verder te gaan, was de keuze snel gemaakt. Gedurende de laatste vier jaren heb ik, financieel gesteund door SMC (Stichting Mathematisch Centrum) en NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek), gewerkt aan het promotie-onderzoek waarvan het uiteindelijke resultaat in de vorm van dit proefschrift nu voor U ligt. Aan de totstandkoming van dit proefschrift hebben velen (misschien onbewust) hun steentje bijgedragen, maar degenen die de hoekstenen hebben aangedragen wil ik hierbij speciaal bedanken.

Allereerst is dat mijn promotor, Malo Hautus. Bij vragen en problemen was zijn deur altijd open, en vond ik bij hem een welwillend oor. Op zijn steun mocht ik rekenen, ook als het zaken op niet-wetenschappelijk gebied betrof. Mijn copromotor Henri Huijberts wil ik bedanken voor zijn goede begeleiding, en voor de vele tijd die hij daarvoor heeft uitgetrokken. Op de momenten dat het nodig was, had hij een bemoedigend woord, en wanneer ik wat al te voortvarend te werk ging, zette hij me weer met beide benen op de grond. Zonder de hulp van zowel Henri als Malo had ik dit proefschrift nooit op deze manier kunnen schrijven.

Ook de andere leden van de kleine commissie wil ik bedanken omdat ik maar al te goed besef dat het lezen van dit proefschrift geen sinecure was. Naast hen verdienen nog twee mensen bijzondere vermelding. Stef van Eijndhoven wil ik bedanken voor de vele, soms pittige discussies die we samen hebben gevoerd. Vaak leidden deze discussies bij mij tot een beter inzicht in de betreffende materie, en op die manier hebben ze zeker invloed gehad op de inhoud van dit proefschrift. Ook ben ik veel dank verschuldigd aan mijn kamergenoot Anton Stoorvogel. Vele figuren in dit proefschrift waren zonder zijn hulp niet tot stand gekomen. Op hem mocht ik altijd rekenen bij mijn vragen op zowel wiskundig als computertechnisch gebied.

Tenslotte, maar zeker niet in de laatste plaats, wil ik mijn ouders bedanken voor de steun en bemoediging die ik, niet alleen in het laatste jaar, maar altijd van hun mocht ontvangen. Het is moeilijk onder woorden te brengen hoeveel jullie voor mij betekenen. Maar waarschijnlijk begrijpen jullie dat zo ook wel.

Luc Habets

Contents

Chapter 1

Introduction

The purpose of this thesis is to describe how algebraic methods and formal computations can be used in the study of systems with time-delays. In the next chapters this so-called algebraic approach to time-delay systems, based on the theory of systems over rings, will be elaborated in more detail. However, to start at the beginning, we confine ourselves in this first chapter to the main topic of the thesis itself: time-delay systems. We answer questions such as: what are time-delay systems; why are they interesting and useful to study? We also present an overview of some approaches known in the literature to describe and analyse these systems mathematically.

1.1 What are time-delay systems?

Before we can answer this question and describe what kind of systems are investigated in this thesis, another question should be answered first: what is a dynamical system? Of course this question can be answered in many different ways, but the key-idea in all these answers is the same. A dynamical system can be seen as a mathematical model for certain types of real world phenomena. These phenomena are described by some variables that are functions of time, and therefore we speak of a *dynamical* system. The variables are mutually related by laws governing the system under consideration. Some of the variables are determined by the outside world; they are called *input variables*. Throughout this thesis we often assume that we are free to choose the values of these input variables ourselves. Input variables of this special type are called *control variables*. The other variables are completely determined by the values of the input variables, and are called *output variables*. So a system can be considered as a process influenced by the outside world through inputs u, and producing outputs y to the outside world as depicted in Figure 1.1.

Inside the system, the inputs are transformed to outputs according to the laws governing the system. Depending on the types of equations describing these laws, we distinguish several classes of systems. The simplest, and probably also the most investigated class is the class of finite-dimensional linear time-invariant systems. In the continuous time case, these systems are described by the following two equations relating the m-dimensional input u to the p-dimensional output y:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{cases} \tag{1.1}$$

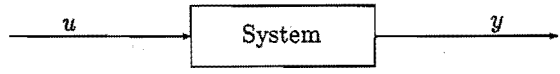


Figure 1.1: A dynamical system

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. So, except for the input and output variables u and y, there is also an evolution variable x involved. $\dot{x}(t)$, the time-derivative of this evolution variable x at time t and $y(t)$, the output y at time t, only depend on $x(t)$ and $u(t)$, the evolution variable x and the input u at time t, respectively. Given an initial value $x(0)$ of the evolution variable and an input trajectory $\{u(t) \mid t \in [0,\infty)\}$, the output trajectory is easily determined using the so-called variation-of-constants-formula

$$y(t) = Ce^{tA}x(0) + \int_0^t Ce^{(t-\tau)A}Bu(\tau)d\tau + Du(t) \qquad (t \geq 0).$$

Hence, the initial value of the evolution variable and the future inputs are the only information needed to solve the differential equation in (1.1). This observation does not only hold at $t = 0$, but for every $t \in \mathbb{R}$. So the evolution variable x has a very specific meaning: it can be seen as the memory of the system, containing all necessary information about the past of the system, that is required to compute all future outputs, when the future inputs are known. Therefore this evolution variable x has a special name: it is called the *state* of the system. (For an extensive treatment on the notion of state and the importance of this concept in system theory, we refer to e.g [48, Section 1.1] and [86, pp. 12-13]).

Linear systems with point delays, which form the class of systems that is investigated in this thesis, obey almost the same equations as (1.1). In this case, however, the time-derivative of the evolution variable x at time t, and the output y at time t, do not only depend on the values of the evolution variable and of the input at time t, but also on their values at some specific time instants in the past. As an example, consider the following linear system with delays:

$$\begin{cases} \dot{x}(t) &= A_0x(t) + A_1x(t-1) + A_2x(t-\sqrt{3}) + A_3x(t-2) + \\ &\qquad + B_0u(t) + B_1u(t-1), \\ y(t) &= C_4x(t-1-\sqrt{3}) + D_0u(t) + D_2u(t-\sqrt{3}), \end{cases} \tag{1.2}$$

where $A_i \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m}$, $C_i \in \mathbb{R}^{p\times n}$ and $D_i \in \mathbb{R}^{p\times m}$. To compute $\dot{x}$ at time t we do not only need $x(t)$ and $u(t)$, but also $x(t-1)$, $x(t-\sqrt{3})$, $x(t-2)$ and $u(t-1)$. For the computation of $y(t)$, the values of $x(t-1-\sqrt{3})$, $u(t)$ and $u(t-\sqrt{3})$ have to be available.

In general, a *time-delay system with point delays* is given by a differential-difference equation and an output equation of the form

$$\begin{cases} \dot{x}(t) &= \sum_{i=1}^{k}(A_i x(t-\tau_i) + B_i u(t-\tau_i)), \\ y(t) &= \sum_{i=1}^{k}(C_i x(t-\tau_i) + D_i u(t-\tau_i)), \end{cases} \tag{1.3}$$

where $A_i \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times m}$, $C_i \in \mathbb{R}^{p\times n}$ and $D_i \in \mathbb{R}^{p\times m}$ $(i = 1,\ldots,k)$ and $0 \leq \tau_1 < \tau_2 < \cdots < \tau_k$ are all time-delays that are involved. We speak of time-delay systems with *point delays* because for the computation of the time-derivative of the evolution variable x at time t and the output y at time t, we only need the values of x and u at some specific time instants in the past and not on a whole interval. However, to solve the differential equation in (1.3) once an input trajectory $\{u(t) \mid t \in [-\tau_k, \infty)\}$ is known, a complete initial *trajectory* $\{x(t) \mid t \in [-\tau_k, 0]\}$ of the evolution variable x is required.

From this observation it is natural to introduce the class of time-delay systems with *distributed* time-delays. For these systems, the time-derivative of the evolution variable x and the output y at time t depend on the values of x and u in a bounded time-interval in the past. In the simplest case, only the functional differential equation for the evolution variable x contains distributed time-delays:

$$\begin{cases} \dot{x}(t) &= \int_{-T}^{0}(dN(\theta))x(t+\theta) + Bu(t), \\ y(t) &= Cx(t), \end{cases} \tag{1.4}$$

with $N(\theta)$ an $n \times n$ matrix of bounded variation.

In both the point- and distributed time-delay case, $x(t)$, the value of the evolution variable at time t, is not the real state of the system. It is obvious that it is impossible to to solve the differential equation (1.3) or (1.4), if only the value of x at one time instant T, and all future inputs are known. Fortunately it is possible to generalize the classical notion of state to this more general case. In Section 1.3 this is explained in more detail. Nevertheless we shall also call the evolution variable x the state of the system. In this thesis this will turn out to be a rather convenient abuse of terminology. It reflects the idea that in both cases (1.1) and (1.3), the variable x plays the same role. It describes in what way the inputs to the system are carried over to the outputs.

Although from a differential equations point of view (1.3) and (1.4) are very similar, and the technique of finding solutions is completely the same, we confine ourselves in this thesis to systems of the form (1.3). This class of systems with point delays is still very rich, as will become apparent in the next section. Moreover, in contrast to systems with distributed time-delays, it is possible to investigate interesting properties of systems with point delays without computing a solution to the differential-difference equation (1.3) explicitly, but using algebraic tools instead. This is also the main theme of this thesis: how to apply algebraic methods to systems with point delays.

1.2 The use of time-delay systems

Time-delay systems with point delays can be used to model a large number of phenomena occurring for example in engineering, physics, biology and economy. In the mathematical modeling of a physical process there is often a trade-off between the accuracy of the model and its simplicity: one is interested in a simple model that gives a good explanation of all aspects of the process one is interested in. In a lot of cases a linear model without delays will do perfectly well as a first description of the process. However, if one wants to describe some additional features of the phenomena under consideration that are not captured by the model (1.1), a more accurate model may be necessary. The introduction of time delays might help to obtain more accurate models. In this section we give some examples of systems in engineering and biology that can be described by time-delay systems.

1.2.1 Time-delay systems in engineering

Example 1.2.1 Consider a chemical plant consisting of some reactors linked by a number of pipelines. In the plant there are several flows of different liquids from one reactor to the other. To describe the behaviour of the complete chemical plant, we first have to model the chemical processes in each of the different reactors. Clearly, the output of one reactor is the input to another one. So, once the dynamical behaviour of the first reactor is determined, the input to the second reactor is also known; one only has to take a time-delay into account: the time required for the liquid flow to cross the pipeline from the first to the second reactor. Moreover, in the reactors themselves time-delays occur too, for example in the mixing and reaction of two different liquids. Hence, for an accurate modeling of such a chemical plant time delays play an important role.

The next example is somewhat more explicit. It is taken from [60, p. 4].

Example 1.2.2 We consider a very simple model to describe ship stabilization. Assume that the ship dynamics are described by

$$I\ddot{\phi} + h\dot{\phi} = -K\psi, \qquad I > 0, K > 0, \tag{1.5}$$

where ϕ is the ship deviation angle and ψ is the turning angle of the rudder. Instead of a manual control of ψ, we apply an automatic control, described by the following helmsman rule

$$T\dot{\psi} + \psi = \alpha\xi + \beta\dot{\xi} \qquad T > 0. \tag{1.6}$$

In (1.6), ξ denotes the measured value of the ship deviation angle. In practice however, it is impossible to measure the ship deviation instantaneously. Therefore we have to assume that

$$\xi(t) = \phi(t - \tau).$$

Taking this time-delay τ into account, and combining (1.5) and (1.6), the following closed-loop system is obtained

$$TI\frac{d^3}{dt^3}\phi(t) + (Th + I)\ddot{\phi}(t) + h\dot{\phi}(t) + K\beta\dot{\phi}(t - \tau) + K\alpha\phi(t - \tau) = 0. \tag{1.7}$$

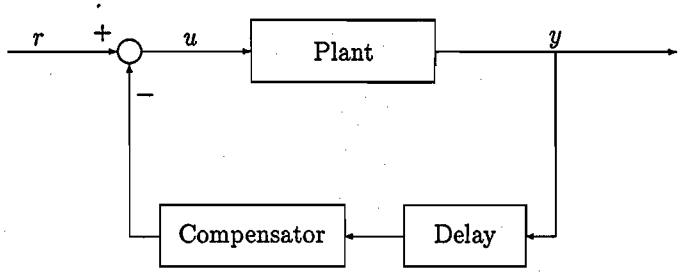


Figure 1.2: Feedback system with informational lag

The closed-loop system (1.7) can be written in a form similar to (1.3). To do so, we introduce three new variables: $x_1(t) := \phi(t)$, $x_2(t) := \dot{\phi}(t)$ and $x_3(t) := \ddot{\phi}(t)$. With this notation, (1.7) takes the following form:

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{-h}{TI} & \frac{-(Th+I)}{TI} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{-K\alpha}{TI} & \frac{-K\beta}{TI} & 0 \end{pmatrix} \begin{pmatrix} x_1(t-\tau) \\ x_2(t-\tau) \\ x_3(t-\tau) \end{pmatrix}.$$

At this moment, the parameters α and β are still free. The final choice has to guarantee the stability of the closed-loop system (1.7), which contains the time-delay τ.

The problem in Example 1.2.2 is quite common in engineering and is called *informational lag*. The general situation is depicted in Figure 1.2. To control the plant, a compensator has to be designed which takes the output y as input and is fed back to the input of the system. However, one has to take into account that the measurement of the output y takes some time τ, and that in reality $y(t-\tau)$ is the input to the compensator. Therefore the closed-loop system is a time-delay system with one point delay τ.

1.2.2 Time-delay systems in biology

Time-delay systems also occur quite naturally in biology, especially in population models. The dynamics of this kind of processes often involve large after-effects, and therefore an accurate modeling leads to a system with time-delays. To illustrate this idea, consider as a simple example an isolated population of a certain species. Let $x(t)$ denote the number of individuals in the population at time t, and assume that the life-span of every individual is a fixed constant L. The number of births per unit of time at time t is proportional to the number of individuals alive at that time instant. But then the number of deaths per unit of time is also known because

the life-span is constant. In this way the following model for the dynamics of the population is obtained:

$$\dot{x}(t) = a \cdot (x(t) - x(t - L)).$$

Of course this is a very simplified model, with rather strong assumptions, but at least it is clear that time-delays play an important role in these population models.

The next example is somewhat more realistic. It describes a predator-prey system with help of the so-called Volterra-Lotka equations, but incorporating some time-delays. Originally this model was proposed in [100].

Example 1.2.3 Consider an isolated area where two different kinds of animals live: a prey and a predator (for example goats and wolves). The predator is completely dependent on the prey for his food. The prey itself on the other hand, is herbivorous and we assume that the isolated area provides the prey animals with a constant (but bounded) amount of vegetable food per unit of time. Let $x(t)$ and $y(t)$ denote the number of individuals at time t of the prey- and predator populations, respectively. We are interested in the dynamic behaviour of both these populations.

For a moment we suppose that there are no predators present in the area and only a small number of prey animals. At first the prey population will grow exponentially because there is an abundance of food available for a relatively small group of prey animals. But at a certain moment the area gets crowded with prey animals, there is a shortage of food, diseases occur, and the population will diminish. So without predators, the dynamics of the prey population might be described by

$$\dot{x}(t) = a \cdot x(t) - b \cdot x^2(t).$$

Next suppose the predator population is present. Under the assumption that both populations are uniformly spread about the area, the number of encounters per unit of time of a predator with a prey will be proportional to $x(t) \cdot y(t)$, and in this way we obtain the following differential equation for the prey population:

$$\dot{x}(t) = a \cdot x(t) - b \cdot x^2(t) - c \cdot x(t) \cdot y(t),$$

where a, b and c are all positive constants.

This differential equation for $x(t)$ can also be interpreted in a slightly different way. Dividing both left- and right-hand side by $x(t)$, we obtain a formula for the *relative growth* $\frac{\dot{x}(t)}{x(t)}$ of the prey population:

$$\frac{\dot{x}(t)}{x(t)} = a - b \cdot x(t) - c \cdot y(t).$$

In our model, this relative growth is an affine function of the number of prey and predator animals, with constant coefficient $a > 0$, and the coefficients of the linear terms in the number of prey and predator animals ($-b$ and $-c$) both negative.

Now, consider the population of predators. If there is no prey available, the predators will starve, and their number will decline exponentially. In the presence of prey, the number of encounters per unit of time of predators with preys will again be proportional to $x(t) \cdot y(t)$. However, it takes some time for the predator to produce off-spring since the prey is not turned into a predator instantaneously.

So, in the differential-difference equation describing the dynamics of the predator population a time-delay h occurs:

$$\dot{y}(t) = -d \cdot y(t) + f \cdot x(t-h) \cdot y(t-h),$$

with the constants d and f both larger than zero.

Finally we assume that it is possible to influence the system from the outside world by shooting some individuals of the predator population. So in the last equation we introduce a control term $v(t)$, representing the number of predator animals shot per unit of time at time t. As a consequence, $v(t) \geq 0$ for all t, and we have obtained the following control system:

$$\begin{cases} \dot{x}(t) &= a \cdot x(t) - b \cdot x^2(t) - c \cdot x(t) \cdot y(t), \\ \dot{y}(t) &= -d \cdot y(t) + f \cdot x(t-h) \cdot y(t-h) - v(t), \end{cases} \tag{1.8}$$

where all variables that are involved are nonnegative.

From the ecological point of view there are now a lot of interesting questions to ask. Is the system stable without applying any control, and in what sense is it stable? Is it possible to steer the system from one equilibrium to another? Moreover, there can be different control objectives. If the predator population is of economical interest (for example because of its valuable fur), one might want to maximize the total amount of predators shot:

$$\int_0^\infty e^{-\alpha t} v(t) dt,$$

where α denotes the discount rate. This is a correction term for the fact that the economical value of future yields diminishes exponentially because of interest loss. From another point of view it is much more interesting to maintain both populations on a certain level with as little human intervention as possible. This kind of questions can all be reformulated into system-theoretic terms. To answer them, properties as reachability and stabilizability play an important role. Exactly these properties are the main topics in this thesis.

Unfortunately, the theory developed in this thesis is not directly applicable to the predator-prey model (1.8) because this model does not only contain time-delays but also nonlinearities. So first one has to linearize this model around an equilibrium to obtain a model of the form (1.3) when this is possible. Through the investigation of the linearized model it is possible to enlarge the knowledge on the original predator-prey system (1.8).

In both the examples from engineering and from ecology, the same questions on time-delay systems pop up. For example, when are these systems stable and what does stability really mean in this situation? And if these systems are not stable, how may they be stabilized by a feedback compensator? Is it possible to steer a system to a specific state and how should the input be chosen to achieve this? For linear time-invariant systems without time-delays the answers to these questions are well known. In this thesis we aim at extending the same ideas to the larger class of systems with point delays.

1.3 Two approaches to time-delay systems

Since systems with time-delays of the form (1.3) are only slight modifications of ordinary finite-dimensional linear time-invariant systems, the most obvious way to treat them is to generalize the theory to incorporate the class of time-delay systems too. This generalization can be performed in (at least) two different ways, and this gives rise to two different approaches towards time-delay systems.

In the first approach, which is most commonly used in the literature, a system with time-delays is described as an infinite-dimensional system. The key-idea behind this approach is the generalization of the notion of state for the system. Despite the fact that in this thesis with some abuse of terminology, the evolution variable x in the differential-difference equation (1.3) is called the state of the system, it is not the state of the system in the classical sense. This remark was already made in Section 1.1. The initial condition required to solve the differential-difference equation (1.3) does not consist of the value of the evolution variable x at one specific time-instant, but on a complete initial trajectory $\{x(t) \mid t \in [-\tau_k, 0]\}$, where τ_k is the largest time-delay occurring in the system (1.3). The same holds at any arbitrary time-instant T. At that time the trajectory $\{x(t) \mid t \in [T - \tau_k, T]\}$ contains all necessary information from the past that is required to compute future outputs, once future inputs are known. So not the evolution variable x itself, but time-trajectories of this evolution variable $\hat{x}(t) = \{x(\xi) \mid \xi \in [t - \tau_k, t]\}$ serve as the state of the system in the classical sense. This real state $\hat{x}$ is of course infinite-dimensional. Hence it is possible to rewrite the original system equations (1.3) in such a way that an infinite-dimensional system, i.e. a system with an infinite-dimensional state-space, is obtained. This rewriting process is mathematically rather involved and is beyond the scope of this thesis. Therefore we omit it here and instead refer to [17, pp. 48-50] for the technical details. This embedding of the class of time-delay systems with point delays in the class of infinite-dimensional systems has an important advantage: all methods and design techniques known for infinite-dimensional systems can be applied to systems with time-delays. Moreover, in this approach it does not make any difference whether we have to deal with systems with point or distributed time-delays: they both fit into the infinite-dimensional systems framework. This is one of the main advantages of this approach. But this generalization is also very straightforward from the system-theoretic point of view: the notion of the state of a system, which plays such a crucial role for finite-dimensional linear time-invariant systems, is generalized and maintains its intuitive meaning in a very clear way. Unfortunately this approach also has some shortcomings. All computations (for example for the design of a compensator) have to be carried out on operators on infinite-dimensional spaces and are therefore quite complicated. Moreover, the implementation of compensators designed in this way can become rather involved. Often the point-delay character of the system is lost after feedback because the design techniques for infinite-dimensional systems may lead to compensators with distributed time-delays. To solve these problems, another approach towards delay systems has been proposed in the literature. In the next part we give the main ideas of this approach.

To investigate the behaviour of a time-delay system like (1.3), it is not always necessary to find a complete solution of the differential-difference equation. A lot of system-theoretic properties can be studied by formal manipulation of the system defining equations themselves. These formal computations get a system-theoretic meaning when we consider a time-delay system as a so-called *system over a ring.* In the next chapter this subject is treated in more detail; we now confine ourselves to the question how delay systems fit into this algebraic framework.

First introduce a *delay operator* σ, acting on the state and input trajectories $x(t)$ and $u(t)$ respectively:

$$\sigma x(t) := x(t-1) \qquad \sigma u(t) := u(t-1).$$

From this definition it is obvious that all integer time-delays can be described by the delay operator σ because

$$\forall k \in \mathbb{N} : x(t-k) = \sigma^k x(t).$$

The use of delay operators for the description of time-delay systems is most easily explained with help of an example. For this we return to the system equations (1.2). We start introducing two delay operators working on state- and input trajectories:

$$\begin{aligned} \sigma_1 x(t) &:= x(t-1), \\ \sigma_2 x(t) &:= x(t-\sqrt{3}). \end{aligned} \tag{1.9}$$

The time-delays 1 and $\sqrt{3}$ are called *incommensurable* because it is impossible to find integers n_1 and n_2 such that $(n_1, n_2) \neq (0,0)$ and

$$n_1 + n_2 \cdot \sqrt{3} = 0.$$

Using definition (1.9), the delay system (1.2) can be rewritten as

$$\begin{aligned} \dot{x}(t) &= A_0 x(t) + A_1 \sigma_1 x(t) + A_2 \sigma_2 x(t) + A_3 \sigma_1^2 x(t) + B_0 u(t) + B_1 \sigma_1 u(t) = \\ &= (A_0 + \sigma_1 A_1 + \sigma_2 A_2 + \sigma_1^2 A_3) x(t) + (B_0 + \sigma_1 B_1) u(t), \\ y(t) &= C_4 \sigma_1 \sigma_2 x(t) + D_0 u(t) + D_2 \sigma_2 u(t) = \\ &= (\sigma_1 \sigma_2 C_4) x(t) + (D_0 + \sigma_2 D_2) u(t). \end{aligned}$$

Defining

$$\begin{aligned} \hat{A} &:= A_0 + \sigma_1 A_1 + \sigma_2 A_2 + \sigma_1^2 A_3, \\ \hat{B} &:= B_0 + \sigma_1 B_1, \\ \hat{C} &:= \sigma_1 \sigma_2 C_4, \\ \hat{D} &:= D_0 + \sigma_2 D_2, \end{aligned} \tag{1.10}$$

the formulae in (1.2) become

$$\begin{cases} \dot{x}(t) &= \hat{A} x(t) + \hat{B} u(t), \\ y(t) &= \hat{C} x(t) + \hat{D} u(t). \end{cases} \tag{1.11}$$

The equations (1.11) look very much like those of a finite-dimensional linear time-invariant system, but there is of course one big difference. The entries of the matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ in (1.11) are not real numbers, but *polynomials* in the delay operators σ_1 and σ_2 with real coefficients.

The procedure described above for example (1.2) can be used in general. Consider a system with k *incommensurable point delays* $0 < \tau_1 < \tau_2 < \cdots < \tau_k$, i.e. the existence of an n-tuple of integers $n_1, \ldots, n_k$ such that

$$n_1\tau_1 + n_2\tau_2 + \cdots + n_k\tau_k = 0,$$

implies that $n_i = 0$ $(i = 1, \ldots, k)$. Define k delay operators acting on state- and input trajectories:

$$\sigma_i x(t) := x(t - \tau_i) \qquad (i = 1, \ldots, k).$$

Then the original system equations can be rewritten in the form (1.11), where $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ are polynomial matrices in the delay operators $\sigma_1, \ldots, \sigma_k$. Next, introduce a k-tuple of indeterminates $s_1, \ldots, s_k$, and substitute in the matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ the indeterminate s_i for σ_i $(i = 1, \ldots, k)$. In this way we obtain matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ and $\tilde{D}$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$, i.e. all entries of these matrices are polynomials in the indeterminates $s_1, \ldots, s_k$ with real coefficients. However, the quadruple of matrices $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$, together with the k-tuple of time-delays $(\tau_1, \ldots, \tau_k)$, still constitutes a complete description of our original time-delay system: all information required to reconstruct the original set of equations is still available.

The quadruple of matrices $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ alone can be considered as a linear system $\Sigma = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ over the commutative ring $\mathbb{R}[s_1, \ldots, s_k]$. This kind of systems form another sort of generalization of the well-known class of linear systems over the field $\mathbb{R}$. In the literature there is a lot of theory available on systems over rings, and in Chapter 2 we study these systems in far more detail. Note already that the theory and all design methods originally developed for systems over rings are directly applicable to time-delay systems with point delays, using the construction just described. In this way several formal computation techniques, exploiting the algebraic structure of the system under consideration, become available for systems with point delays.

It is important to stress that the class of all linear systems over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ is more general than the class of systems with point delays as described in (1.3). Considering a delay system (1.3) as a system over a ring, we forget about the fact that the indeterminates $s_1, \ldots, s_k$ correspond to delay operators $\sigma_1, \ldots, \sigma_k$ with time-delays $\tau_1, \ldots, \tau_k$ respectively. So from the description of the time-delay system as a quadruple $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ of polynomial matrices and a k-tuple $(\tau_1, \ldots, \tau_k)$ of time-delays, only the first information is used: in the description as a system over a ring we forget (temporarily) about the delay character of the original system. Of course it is often possible to derive stronger results using exactly this extra information. This forgetting and remembering of the delay character of the indeterminates $s_1, \ldots, s_k$ can be seen as a leitmotiv through the whole thesis.

Finally we want to mention that also this so-called algebraic approach towards time-delay systems has its shortcomings and disadvantages. First of all it is only applicable to systems with point delays as described in (1.3); systems with distributed

time-delays do not fit into this algebraic framework. Moreover, the notion of the state of a time-delay system becomes very unclear. In the theory of systems over rings there also exists a notion of state of a system, but unfortunately this interpretation is not compatible with the classical notion of the state of a time-delay system as it was used in the infinite-dimensional systems approach. So one has to be very careful in this respect. Still, the systems over rings approach has the important advantage that computations are much easier to carry out in this framework. In this thesis we investigate how fruitful this approach can be. This does not mean that the infinite-dimensional systems approach is not interesting. On the contrary; it is the most commonly used framework for the investigation of time-delay systems and invaluable for a good understanding of time-delay systems. A lot of important results were discovered using this approach. Therefore I prefer to state it the other way around: I hope that this thesis will convince the reader that the algebraic approach towards time-delay systems is a very interesting and very useful alternative to the more classical infinite-dimensional systems framework. It is impossible to judge which approach is the best: they both have their strong and weak sides. The choice one has to make depends on the questions one wants to answer and on the applications one has in mind.

1.4 Organization of this thesis

Globally this thesis consists of two major parts. Part 1 mainly deals with the development of some theory for systems over rings in general and time-delay systems in particular. In the second part the main emphasis is on constructive methods. In this part the following problems are investigated. How can the system-theoretic properties that turned out to be of interest in Part 1 be tested explicitly? Are there constructive methods to carry out the design methods proposed in Part 1 in practice? But also within the major Parts 1 and 2 a distinction can be made, based on the forgetting and remembering of the delay character of the indeterminates. First we try to proceed as long as possible without using this additional information. So this part of the thesis treats the general case of systems over rings. Only when it is really required to make some further progress, we bring in the extra information about the delay character of the indeterminates. In these parts of course only delay systems are considered. By making this distinction as clear as possible we hope that this thesis is both valuable for people interested in delay systems and for people interested in systems over rings.

A more detailed discussion of the contents and main themes of each chapter of the thesis is given below.

Chapter 2 *Linear systems over rings*

In this chapter we give an overview of some of the main ideas behind the theory of systems over commutative rings. In the first sections we generalize some basic system theoretic properties like reachability and observability to the case of systems over rings. This part is mainly based on [85], [49] and [5].

Then we turn to the problem of stability and stabilizability of systems over rings. For this purpose a Hurwitz set framework is introduced. Although in [23] the

problem of stabilizability by dynamic state feedback was solved for the first time, we use an approach that is similar to the one in [80]. However, some of the proofs are simplified. Using the notion of detectability introduced in [44], a solution to the problem of stabilizability by dynamic output feedback is obtained. The derivation of this result for strictly proper systems was given in [58]. The presentation of Chapter 2 as a whole is based on [37].

Chapter 3 *Stabilizability of time-delay systems*

In the first part of this chapter, the results on stabilizability for linear systems over rings obtained in Chapter 2 are specialized to the case of time-delay systems. The time-delay character of the system is used explicitly to derive more easily verifiable conditions for the stabilizability of a delay system. This leads to a pointwise rank condition that can be seen as a generalization of the Hautus-test to systems with point delays. The original proof of this result in [25] is modified in such a way that with almost the same ideas more general types of stability may be treated.

The second part of Chapter 3 is devoted to the genericity of stabilizability. We use a topological approach to address this problem. First we introduce a natural topology on the parameter space of all time-delay systems with point delays that is suitable for our purposes. Next we prove that the set of stabilizable time-delay systems contains a subset that is an open and dense subset of the parameter space of all time-delay systems. This indicates that the condition of stabilizability is very weak. The proof of this result for stability in the classical sense was given in [38]. The generalization to a larger class of stability domains using inductive limit topologies is new and unpublished.

Chapter 4 *Constructive commutative algebra*

This chapter contains an overview of two methods in constructive commutative algebra: Gröbner bases and characteristic sets.

Section 4.1 is a short introduction to Gröbner bases. We explain how polynomial ideals are characterized by their Gröbner bases, and how manipulations on polynomial ideals can be carried out explicitly using Gröbner basis techniques. Although we also touch upon the computation of Gröbner bases, we mainly emphasize the applications that are useful in the sequel. Especially the computation of the variety of a polynomial ideal is important. The presentation of Section 4.1 is mainly based on [76]. However, also the books [14] and [2] are important references.

Section 4.2 is devoted to the characteristic sets method. This subject is treated in more detail. Compared to Gröbner bases, the applicability of characteristic sets is somewhat restricted because characteristic sets only characterize prime polynomial ideals. Nevertheless characteristic sets are very useful for the determination of the variety of a polynomial ideal. Besides this application we emphasize the difficulties that are caused by the two different definitions of characteristic sets that are used in literature. Using the approach developed in [35], this problem can be solved by distinguishing the two different notions of characteristic sets explicitly. Main references for this section are [79], [101] and [96].

Chapter 5 *Testing reachability and stabilizability*

In this chapter, the Gröbner basis method introduced in Chapter 4 is used to test

the reachability and stabilizability of systems over polynomial rings. We start with the introduction of some polynomial ideals that characterize the reachability and stabilizability properties of a system in a very straightforward way. Next, we derive several algorithms for the computation of the Gröbner bases of these ideals. In this way we obtain various methods to test the reachability of a system. Moreover, one of these methods can be applied to compute right-inverses of nonsquare polynomial matrices in several indeterminates. For the problem of stabilizability we confine ourselves to time-delay systems. Combining exact and numerical computations, we can find an algorithmic answer to the question of stabilizability for time-delay systems.

The main parts of Chapter 5 are new and unpublished. Some of the results were published in [40] and [39].

Chapter 6 *Stabilization of time-delay systems*

Whereas in Chapter 5 only the existence of stabilizing feedback compensators is considered, this chapter contains an overview of some existing methods described in the literature for the construction of stabilizing feedback compensators for time-delay systems. We start with the derivation of a reliable method to test the stability of a delay system, based on [36]. Next we discuss a stabilization method developed in [54], [56] and [55]. This is a constructive approach based on the notion of so-called BIBO-stability, and therefore an approximation step is required to obtain a solution within the framework of stability used in the thesis. The stability test for time-delay system is very important in this approximation step. Finally we mention some alternative stabilization methods, mainly originating from the theory of infinite-dimensional systems.

The description of the contents of the various chapters, reflects the global organization of the thesis. The first theoretical part consists of the Chapters 2 and 3; Chapters 4 to 6 form the more algorithmic part. Moreover, Chapters 2, 4 and 5 (except for Section 5.5) are dealing with systems over commutative rings in general. In Chapter 3, Section 5.5 and Chapter 6, we specialize to systems with time-delays. To obtain the results of these chapters, the delay character of the system is used explicitly.

Chapter 2

Linear systems over rings

As already mentioned in the Introduction, the theory of systems over rings is a useful tool for the investigation of algebraic properties of systems with point delays. In this chapter we give an overview of some of the results in this field that are important in the rest of this thesis. Except for time-delay systems, the theory of systems over rings has several other interesting applications. Therefore the setup of this chapter is very general: systems over rings are the central theme. Time-delay systems serve only as an illustration of the general theory.

2.1 What are systems over rings?

This first section is devoted to the definition and some applications of linear systems over rings. However, before giving the definition, we start explaining the main ideas behind this approach. Return for a moment to the class of finite-dimensional linear time-invariant systems of the form

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{cases} \tag{2.1}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. From a system-theoretic point of view there are now a lot of interesting properties to investigate. For example, is a system reachable, i.e. is it possible to steer a system in finite time from one given state to any other state using a suitable input? When is a system stable, i.e. when will the state eventually tend to zero if no input is applied? And if a system is unstable, does there exist a feedback law of the form $u(t) = Fx(t) + v(t)$ such that the closed-loop system is stable? Of course there are many other important questions to pose, but the questions above have one thing in common: they are all expressed with help of the state- and input trajectories $\{x(t) \mid t \in [0,\infty)\}$ and $\{u(t) \mid t \in [0,\infty)\}$ respectively.

The conditions under which these questions have an affirmative answer are well known: they can be found in any elementary textbook on linear systems, e.g. in [47]. Typically, these conditions are stated in terms of the system defining matrices A, B, C and D. In this way they are easy to check, e.g. the system (2.1) is stable if and only if $\sigma(A) \subset \mathbb{C}^-$. Now the following observation is very important: a finite-dimensional linear time-invariant system is completely characterized by four

matrices (A, B, C, D) of appropriate dimensions. Moreover, all properties of the system, although defined in terms of states, inputs and outputs, can also be investigated using these four matrices only. From this point of view, the notion of state is only important for our interpretation of the quadruple of matrices (A, B, C, D) as a dynamical system. Intrinsic properties of a system however, depend only on the four system defining matrices.

In the situation just described, all four matrices A, B, C and D are real: their entries are real numbers. To a certain extent, this restriction is unnecessary. The same conditions on a quadruple of matrices can be tested when A, B, C and D are matrices over a commutative ring $\mathcal{R}$, instead of matrices over the field $\mathbf{R}$ of real numbers. Therefore we come to the following definition of a system over a ring.

Definition 2.1.1 A (free) *linear system* Σ *over a commutative ring* $\mathcal{R}$ is a quadruple of matrices (A, B, C, D), where $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times m}$, $C \in \mathcal{R}^{p\times n}$ and $D \in \mathcal{R}^{p\times m}$ for some integers n, m and p, and where n is called the *rank* of the system Σ.

At first sight it seems strange that in Definition 2.1.1 no states and no dynamics occur. But linear systems over $\mathbf{R}$ still fit in this framework because they are completely characterized by four matrices of appropriate dimensions. In fact, the abstract notion of a system is useful in a far more general setting because it can be specialized to a lot of interesting situations. This idea is illustrated with a few examples.

Example 2.1.2 Let $\mathcal{R}$ be a commutative ring, and consider a linear discrete-time system over $\mathcal{R}$ defined by

$$\begin{cases} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{cases} \tag{2.2}$$

where $t \in \mathbf{Z}^+$ and $x(t) \in \mathcal{R}^n$, $u(t) \in \mathcal{R}^m$ and $y(t) \in \mathcal{R}^p$ are the state, input and output at time t respectively. The entries of the matrices A, B, C and D belong to the ring $\mathcal{R}$. So the quadruple of matrices $\Sigma = (A, B, C, D)$ is a complete description of this discrete-time system, and according to Definition 2.1.1 it can be seen as a system over the ring $\mathcal{R}$.

Example 2.1.3 Consider a system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$ and define $\sigma_1, \ldots, \sigma_k$ as the corresponding delay-operators:

$$\sigma_i x(t) = x(t - \tau_i) \qquad (i = 1, \ldots, k).$$

A time-delay system with point delays can then be written as

$$\begin{aligned} \dot{x}(t) &= A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) &= C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t), \end{aligned}$$

where A, B, C and D are polynomial matrices in the delay operators $\sigma_1, \ldots, \sigma_k$. Substituting the indeterminates $s_1, \ldots, s_k$ for $\sigma_1, \ldots, \sigma_k$, a quadruple of polynomial matrices $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ in the indeterminates $s_1, \ldots, s_k$ is obtained. Together with the k-tuple of time-delays $\tau_1, \ldots, \tau_k$ this quadruple is a complete description for the original system equations. The quadruple $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ itself can be regarded as a system over the polynomial ring $\mathbf{R}[s_1, \ldots, s_k]$.

Example 2.1.4 Consider the following linear system:

$$\begin{cases} \dot{x}(t) &= A(\alpha)x(t) + B(\alpha)u(t), \\ y(t) &= C(\alpha)x(t) + D(\alpha)u(t), \end{cases}$$

where $\alpha \in \mathbf{R}$ is a fixed parameter (of which the value is probably unknown), and $A(\alpha)$, $B(\alpha)$, $C(\alpha)$ and $D(\alpha)$ are matrices of appropriate dimension over $\mathbf{R}[\alpha]$, i.e. all their entries are polynomials in the indeterminate α with real coefficients. Hence, for each value of α another linear system is obtained. For design purposes it is interesting to investigate this whole class of systems together. This can be done by considering the quadruple of matrices $\Sigma = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ as a system over the polynomial ring $\mathbf{R}[\alpha]$.

From these three examples it is obvious that the integers m and p in Definition 2.1.1 should be interpreted as the number of inputs and outputs of the system, respectively. Moreover, these examples illustrate that the abstract notion of a system over a ring as defined in Definition 2.1.1 is a very versatile concept. It is applicable to both continuous- and discrete-time systems and a lot of interesting problems fit into this algebraic framework.

In the rest of this chapter some system-theoretic properties and design methods which are well known for systems over fields will be generalized to the systems over rings case. This is done in a rather formal way. We have just mentioned that in the classical situation of systems over the field $\mathbf{R}$, many interesting properties of a system can be reformulated as conditions on the system defining matrices. In the ring case these conditions on the matrices A, B, C and D known for systems over $\mathbf{R}$ are used to define the properties of a system. So we go exactly the other way around.

In a specific situation of an application like Example 2.1.3 or 2.1.4, there might be a discrepancy between our intuitive notion of a property, and its formal definition in the systems over rings framework. This is a price we have to pay. The constructive algebraic methods we want to use can only be applied on the quadruple of matrices $\Sigma = (A, B, C, D)$ over the ring $\mathcal{R}$. This has an enormous advantage: the design methods obtained in this way are useful for the whole range of systems that can be modeled as a system over a ring. Moreover, proceeding along this path, we obtain an elegant generalization of the theory of linear systems over fields, which gives us a better insight into the meaning of linearity for dynamical systems.

Finally we have to make a remark on the assumptions on the ring $\mathcal{R}$ we work with. In the whole chapter $\mathcal{R}$ is assumed to be a *commutative* ring. In most cases, $\mathcal{R}$ is even an integral domain, i.e. $\mathcal{R}$ is a commutative ring with identity and without zero divisors. Sometimes a specialization is made for polynomial rings $\mathcal{R} = \mathbf{R}[s_1, \ldots, s_k]$, because this case is important for systems with point delays.

2.2 Reachability

Consider a linear continuous-time system over $\mathbb{R}$ given by the following equations:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{cases} \tag{2.3}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. This system is called *reachable* if for all pairs $(x_0, \hat{x}) \in \mathbb{R}^n \times \mathbb{R}^n$ there exists a $T > 0$ and an input trajectory $\{u(t) \mid t \in [0,T]\}$, such that, starting the system in $x(0) = x_0$ and applying this input trajectory, the system reaches the state $\hat{x}$ at time T, i.e. $x(T) = \hat{x}$. It is obvious that for the formulation of the notion of reachability only the differential equation in (2.3) is used. So to verify the reachability of a system, only the matrices A and B are required. Therefore we often speak about the reachability of the matrix pair (A, B).

Theorem 2.2.1 *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Then (A, B) is reachable if and only if one of the following conditions is satisfied:*

(i) $\operatorname{rank}(B \mid AB \mid \cdots \mid A^{n-1}B) = n$,

(ii) $\forall \lambda \in \mathbb{C} : \operatorname{rank}(\lambda I - A \mid B) = n$,

(iii) $(zI - A \mid B)$ *is right-invertible over* $\mathbb{R}[z]$. ∎

The first condition in Theorem 2.2.1 is called the Kalman rank condition. It is classical and can be found in any introductory textbook (see for example [66, pp. 81-82] or [103, p.17]). Condition *(ii)* is probably the easiest one to verify in practical examples. In the literature it is known under two different names. Sometimes it is called PBH-test (= Popov-Belevitch-Hautus test) (see e.g. [47]), but in the sequel we use the name Hautus test (see [43]). Condition *(iii)* can be considered as a reformulation of the Hautus test; it follows from *(ii)* using the local-global theorem (see Appendix A). A direct proof of this condition in a more general case will be given later on in this section.

The property of reachability of a system is now generalized to the ring case by taking condition *(i)* in Theorem 2.2.1 as the new definition.

Definition 2.2.2 Let $\Sigma = (A, B, C, D)$ be a system over a commutative ring $\mathcal{R}$. Then Σ is called *reachable* if the columns of the matrix

$$(B \mid AB \mid \cdots \mid A^{n-1}B) \tag{2.4}$$

generate the free module $\mathcal{R}^n$.

In Definition 2.2.2 only the matrices A and B are involved. Therefore we often call the pair (A, B) reachable.

The definition of reachability completely coincides with our intuitive notion of reachability for the discrete-time interpretation of a system over a ring, as given in Example 2.1.2. Intuitively, the system

$$x(t+1) = Ax(t) + Bu(t)$$

is reachable if for all pairs $(x_0, \hat{x}) \in \mathcal{R}^n \times \mathcal{R}^n$ there exist a $T \geq 0$ and an input sequence $u(0), u(1), \cdots, u(T-1)$, such that, starting the system in $x(0) = x_0$ and applying this input sequence, it reaches the state $\hat{x}$ at time T, i.e. $x(T) = \hat{x}$. Since the Cayley-Hamilton Theorem also applies to the ring case, it is not difficult to prove that this property is satisfied if and only if the $\mathcal{R}$-module generated by all columns of $(B \mid AB \mid \cdots \mid A^{n-1}B)$ is equal to $\mathcal{R}^n$, which is exactly the condition given in Definition 2.2.2.

Unfortunately, condition (2.4) is not very practical for testing, certainly if the dimension of A is large. However, if $\mathcal{R}$ is an integral domain, it is possible to generalize condition *(iii)* in Theorem 2.2.1 to the ring case. In this way a sort of generalized Hautus test for systems over rings is obtained.

Theorem 2.2.3 *Let $\mathcal{R}$ be an integral domain, and $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Then the pair (A, B) is reachable if and only if*

$$(zI - A \mid B) \text{ is right-invertible over } \mathcal{R}[z]. \tag{2.5}$$

Proof

Assume that (A, B) is reachable. Then the columns of $(B \mid AB \mid \cdots \mid A^{n-1}B)$ generate $\mathcal{R}^n$, and therefore this matrix is right-invertible over $\mathcal{R}$. Hence, there exist matrices $Q_i \in \mathcal{R}^{m\times n}$ $(i = 0, 1, \ldots, n-1)$ such that

$$\sum_{i=0}^{n-1} A^i B Q_i = I.$$

Define for all $i \in \{0, 1, \ldots, n-1\}$ the matrices $N_i \in \mathcal{R}^{m\times n}$ as

$$N_i := Q_{n-1-i} \qquad (i = 0, 1, \ldots, n-1),$$

and the matrices $M_i \in \mathcal{R}^{n\times n}$ in the following recursive way:

$$\begin{aligned} M_0 &:= 0, \\ M_{i+1} &:= AM_i + BN_i \qquad (i = 0, 1, \ldots, n-1). \end{aligned}$$

Then $M_n = \sum_{i=0}^{n-1} A^i B N_{n-1-i} = \sum_{i=0}^{n-1} A^i B Q_i = I$. Next define

$$\begin{aligned} N(z) &:= \sum_{i=0}^{n-1} N_i z^{n-1-i}, \\ M(z) &:= \sum_{i=1}^{n-1} M_i z^{n-1-i}. \end{aligned}$$

So $N(z)$ and $M(z)$ are both matrices over $\mathcal{R}[z]$, and we have

$$\begin{aligned} &(zI - A \mid B) \cdot (-M(z)) + B \cdot N(z) = \\ &= -\sum_{i=1}^{n-1} M_i z^{n-i} + \sum_{i=1}^{n-1} AM_i z^{n-1-i} + \sum_{i=0}^{n-1} BN_i z^{n-1-i} = \\ &= (-M_1 + BN_0) z^{n-1} + \sum_{i=1}^{n-2} (-M_{i+1} + AM_i + BN_i) z^{n-1-i} + (AM_{n-1} + BN_{n-1}) = \\ &= 0 + 0 + M_n = I, \end{aligned} \tag{2.6}$$

where we used the recursive definition of M_i, and the fact that $M_n = I$. From (2.6) it follows immediately that

$$\begin{pmatrix} -M(z) \\ N(z) \end{pmatrix}$$

is a right-inverse of $(zI - A \mid B)$ over $\mathcal{R}[z]$.

Now suppose that $(zI - A \mid B)$ is right-invertible over $\mathcal{R}[z]$. Then there exist matrices $M(z) = \sum_{i=0}^{k} M_{k-i} z^i$ and $N(z) = \sum_{i=0}^{k} N_{k-i} z^i$, with for all $i \in \{0, 1, \ldots, k\}$, $M_i \in \mathcal{R}^{n \times n}$ and $N_i \in \mathcal{R}^{m \times n}$, and such that

$$(zI - A) \cdot (-M(z)) + B \cdot N(z) = I.$$

Substituting the formula for $M(z)$ and $N(z)$ we obtain:

$$\begin{aligned} I &= (zI - A)(-M(z)) + BN(z) = (zI - A) \sum_{i=0}^{k} -M_{k-i} z^i + B \sum_{i=0}^{k} N_{k-i} z^i = \\ &= -M_0 z^{k+1} + \sum_{i=1}^{k} (-M_{k-i+1} + AM_{k-i} + BN_{k-i}) z^i + (AM_k + BN_k) = \\ &= -M_0 z^{k+1} + \sum_{j=0}^{k-1} (-M_{j+1} + AM_j + BN_j) z^{k-j} + (AM_k + BN_k). \end{aligned}$$

Hence

$$\begin{cases} M_0 = 0, \\ M_{j+1} = AM_j + BN_j & (j = 0, 1, \ldots, k-1), \\ M_{k+1} = AM_k + BN_k = I, \end{cases}$$

and so

$$I = M_{k+1} = \sum_{i=0}^{k} A^i B N_{k-i}. \tag{2.7}$$

Now use the Cayley-Hamilton Theorem. Let $\chi_A(z) = \det(zI - A)$. Then $\chi_A(A) = 0$, and thus the matrix A^n can be written as an $\mathcal{R}$-linear combination of the matrices $I, A, \ldots, A^{n-1}$. Applying this fact recursively in formula (2.7), we obtain matrices $\tilde{N}_i \in \mathcal{R}^{m \times n}$ $(i = 0, 1, \ldots, n-1)$ such that

$$I = \sum_{i=0}^{n-1} A^i B \tilde{N}_{n-1-i}.$$

So $(\tilde{N}_{n-1}^T \mid \tilde{N}_{n-2}^T \mid \cdots \mid \tilde{N}_1^T \mid \tilde{N}_0^T)^T$ is a right-inverse of $(B \mid AB \mid \cdots \mid A^{n-1}B)$ over $\mathcal{R}$, and (A, B) is reachable. ∎

The proof of Theorem 2.2.3 also gives some additional information. Suppose that there exist matrices $M(z)$ and $N(z)$ over $\mathcal{R}[z]$ such that

$$(zI - A)M(z) + BN(z) = I.$$

Then (A, B) is reachable, and according to the construction in the first part of the proof, there exist matrices $\tilde{M}(z)$ and $\tilde{N}(z)$ over $\mathcal{R}[z]$ such that

$$(zI - A)\tilde{M}(z) + B\tilde{N}(z) = I,$$

and $\deg_z(\tilde{M}(z)) \leq n-2$ and $\deg_z(\tilde{N}(z)) \leq n-1$. Here $\deg_z(\tilde{M}(z))$ denotes the degree of the polynomial matrix $\tilde{M}(z)$ in the indeterminate z. It is the maximum of the degrees in z of all its entries. The bound we just found on the degrees of the matrices $\tilde{M}(z)$ and $\tilde{N}(z)$ is a direct consequence of the Cayley-Hamilton Theorem.

Condition (2.5) plays an important role in this thesis because it gives a simple algebraic description of reachability over an integral domain $\mathcal{R}$. If this integral domain is a polynomial ring, it is even possible to proceed one step further. Let $\mathcal{K}$ be a field, and $\bar{\mathcal{K}}$ be the algebraic closure of $\mathcal{K}$. Let $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$ denote the ring of all polynomials in the indeterminates $s_1, \ldots, s_k$ with coefficients in $\mathcal{K}$. In this case reachability can be tested using a pointwise rank condition.

Theorem 2.2.4 *Let $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, and $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. The pair $\Sigma = (A, B)$ is reachable over $\mathcal{R}$ if and only if*

$$\forall(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1} : \operatorname{rank}(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1 \ldots, \hat{s}_k)) = n. \qquad (2.8)$$

Proof

Assume first that $\Sigma = (A, B)$ is reachable. Then according to Theorem 2.2.3, $(zI - A \mid B)$ is right-invertible over $\mathcal{R}[z]$. Let $P(z)$ be a right-inverse of $(zI - A \mid B)$. Let $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$. Then $P(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k)$ is a right-inverse of $(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1, \ldots, \hat{s}_k))$. So $\operatorname{rank}(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1, \ldots, \hat{s}_k)) = n$. Since $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ was arbitrary, this proves (2.8).

Next, assume that (2.8) holds. Because of Theorem 2.2.3 it suffices to show that the matrix $(zI - A \mid B)$ is surjective. This will be done using the local-global theorem (see Appendix A.3).

Let $\hat{x} = (\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ and define

$$\mathcal{I}_{\hat{x}} := \{p(z, s_1, \ldots, s_k) \in \mathcal{K}[z, s_1, \ldots, s_k] \mid p(\hat{x}) = 0\}.$$

Then we know from Proposition A.3.5 that

$$\{\mathcal{I}_{\hat{x}} \mid \hat{x} \in \bar{\mathcal{K}}^{k+1}\}$$

is the set of *all* maximal ideals in $\mathcal{R}[z]$. In the proof of this claim the Hilbert Nullstellensatz (see Appendix A.2) is involved.

Now, let $\mathcal{M} = \mathcal{R}[z]^{n+m}$ and $\mathcal{N} = (zI - A \mid B)\mathcal{M}$. Then $(zI - A \mid B)$ is an $\mathcal{R}[z]$-homomorphism from the $\mathcal{R}[z]$-module $\mathcal{M}$ to the finitely generated $\mathcal{R}[z]$-module $\mathcal{N}$. Let $\hat{x} \in \bar{\mathcal{K}}^{k+1}$, and $\mathcal{I}_{\hat{x}}$ the corresponding maximal ideal. Denote by $\mathcal{M}_{\hat{x}}$ and $\mathcal{N}_{\hat{x}}$ the factor modules $\mathcal{M}/\mathcal{I}_{\hat{x}}\mathcal{M}$ and $\mathcal{N}/\mathcal{I}_{\hat{x}}\mathcal{N}$ respectively. Let $(zI - A \mid B)_{\hat{x}} : \mathcal{M}_{\hat{x}} \to \mathcal{N}_{\hat{x}}$ be the mapping between these factor modules defined by

$$(zI - A \mid B)_{\hat{x}}\overline{m} := \overline{(zI - A \mid B)m},$$

where the bar denotes the canonical projection. Since the canonical projection boils down to evaluation in $\hat{x}$ in this case, we know that $(zI - A \mid B)_{\hat{x}}$ is surjective if

and only if $(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1, \ldots, \hat{s}_k))$ is surjective. But the latter follows immediately from (2.8) for all $\hat{x} \in \bar{\mathcal{K}}^{k+1}$. So for all $\hat{x} \in \bar{\mathcal{K}}^{k+1}$, $(zI - A \mid B)_{\hat{x}}$ is surjective, and applying the local-global theorem, we conclude that $(zI - A \mid B)$ is surjective. This completes the proof. ∎

From Theorem 2.2.4 it is not immediately clear how strong the condition for reachability over a polynomial ring is. This question has been investigated by Lee and Olbrot in [67] for the case $\mathcal{K} = \mathbb{R}$. Here we state a somewhat generalized version of their result. For a proof we refer to [67].

Proposition 2.2.5 *Let $\mathcal{K}$ be a field of characteristic zero and $\mathcal{R} := \mathcal{K}[s_1, \ldots, s_k]$ be the polynomial ring in the indeterminates $s_1, \ldots, s_k$ with coefficients in $\mathcal{K}$. Consider all pairs $\Sigma = (A, B)$ with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. For this class of systems, reachability is a generic property if and only if $m > k$, i.e. if and only if the number of inputs exceeds the number of indeterminates. Here the concept of genericity is based on its interpretation in the Zariski topology (see for example [65]).* ∎

Remark 2.2.6 Theorem 2.2.4 can be seen as a Hautus test for systems with point delays, regarding them as systems over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. Moreover, Proposition 2.2.5 indicates how strong this condition is. Note however that in this particular example, the notion of reachability as stated in Definition 2.2.2 is rather formal. In the infinite-dimensional systems approach to time-delay systems there are other definitions of reachability which look intuitively more appealing. Relationships between all these notions of reachability are not very clear. It is obvious that condition (2.8) implies *spectral controllability* as it was introduced by Pandolfi in [73]. But the conditions for *approximate controllability* derived by Manitius in [68] are neither necessary nor sufficient for reachability in the systems over rings approach. This has also been noted by Yamamoto in [102]. In a later paper ([69]), Manitius argues that the property of approximate controllability is too strong in the case of systems with time-delays. As a remedy he introduces the concept of *F-controllability*. Again, the exact relationship with condition (2.8) is hard to derive, but it is obvious that there is a much closer relationship in this case. For a more detailed investigation of this problem we refer to [49, Section 3.2].

The situation described above seems rather disappointing, but is not very surprising. As we have seen in Section 1.3, the two approaches to time-delay systems are based on a completely different philosophy. In the framework of systems over rings, most definitions are just formal generalizations of well-known conditions from the theory of systems over fields. At first one has to pay a price, because these definitions are not very appealing from an intuitive point of view. Later on it becomes clear that these are exactly the conditions under which a lot of interesting design techniques carry over to the systems over rings case.

2.3 Observability

The notion of observability for systems over rings is again based on the concept of observability for systems over $\mathbb{R}$. Consider the system $\Sigma = (A, B, C, D)$ over $\mathbb{R}$, as given in (2.3). Looking at the system from the outside world, only the input and

output trajectories are observed; the state is hidden inside the system. Now a system is called *observable* if the information on the input- and output trajectory suffices to reconstruct the state trajectory. From equation (2.3) it is clear that only the matrices A and C are important for this property; the influence of the terms $Bu(t)$ and $Du(t)$ is completely known and can be eliminated easily. Therefore we often speak of observability of the matrix pair (C, A). The following theorem describes two conditions for the observability of a system over the field $\mathbf{R}$. They are the well-known dual versions of the conditions for reachability we encountered in Theorem 2.2.1 (see e.g. [47, Section 6.2]).

Theorem 2.3.1 *Let $A \in \mathbf{R}^{n\times n}$ and $C \in \mathbf{R}^{p\times n}$. The pair (C, A) is observable if and only if one of the following conditions is satisfied:*

(i) $\operatorname{rank}(C^T \mid A^TC^T \mid \cdots \mid (A^T)^{n-1}C^T)^T = n,$

(ii) $\forall \lambda \in \mathbf{C} : \operatorname{rank}\begin{pmatrix} \lambda I - A \\ C \end{pmatrix} = n.$ ∎

The definition of observability in the ring case is simply a generalization of condition *(i)* in Theorem 2.3.1. It coincides with the intuitive notion of observability for discrete-time systems over $\mathcal{R}$ as given in Example 2.1.2.

Definition 2.3.2 Let $\Sigma = (A, B, C, D)$ be a system over a commutative ring $\mathcal{R}$. Then Σ is called *observable* if the map $T : \mathcal{R}^n \to \mathcal{R}^{pn}$ defined by

$$T = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} \tag{2.9}$$

is injective.

The condition for observability can also be stated in the following alternative form: there exists no $x \in \mathcal{R}^n$, $x \neq 0$ such that $Cx = CAx = \cdots = CA^{n-1}x = 0$. But in (2.9) the analogy with condition *(i)* in Theorem 2.3.1 shows much clearer.

Next, consider the special case that $\mathcal{R}$ is an integral domain. Then the set

$$\mathcal{Q} := \{\frac{p}{q} \mid p, q \in \mathcal{R}, q \neq 0\}$$

is called the *quotient field* of $\mathcal{R}$. When we add and multiply fractions as usual, it is easily seen that $\mathcal{Q}$ is indeed a field.

Proposition 2.3.3 *Let $\mathcal{R}$ be an integral domain and $\Sigma = (A, B, C, D)$ be a system over $\mathcal{R}$. Then Σ is observable as a system over $\mathcal{R}$ if and only if Σ is observable considered as a system over the quotient field $\mathcal{Q}$ of $\mathcal{R}$.* ∎

For a proof of this result we refer to [5, p.60].

The result of Proposition 2.3.3 is rather surprising. For a system over an integral domain $\mathcal{R}$, there is no difference in the notion of observability when we regard it as

a system over the ring $\mathcal{R}$, or as a system over its quotient field $\mathcal{Q}$. In both the ring and the field case, observability is completely the same. Hence there is an important difference with the concept of reachability; the conditions for the reachability of a system over an integral domain $\mathcal{R}$ are stronger than the conditions for reachability over the corresponding quotient field $\mathcal{Q}$. So for systems over rings, there is no duality between reachability and observability in general. Nevertheless we still have:

Lemma 2.3.4 *Suppose that the pair of matrices (A, B) is reachable over the commutative ring $\mathcal{R}$ with identity. Then (B^T, A^T) is observable over $\mathcal{R}$.*

Proof

Suppose that (A, B) is reachable over $\mathcal{R}$. Then the columns of $(B \mid AB \mid \cdots A^{n-1}B)$ generate $\mathcal{R}^n$. So in particular, $(B \mid AB \mid \cdots A^{n-1}B)$ is right-invertible over $\mathcal{R}$. Let $x \in \mathcal{R}^n$ and suppose that $B^T x = B^T Ax = \cdots B^T (A^T)^{n-1} x = 0$. Then $x^T(B \mid AB \mid \cdots \mid A^{n-1}B) = 0$, and after multiplication by a right-inverse of $(B \mid AB \mid \cdots A^{n-1}B)$ over $\mathcal{R}$ we conclude that $x = 0$. ∎

Remark 2.3.5 The result of Lemma 2.3.4 is also true for other types of rings. If $\mathcal{R}$ is a commutative ring without zero divisors, but without an identity, the same implication can be proved.

Remark 2.3.6 Note that the implication in Lemma 2.3.4 only holds in one direction. If (C, A) is observable over $\mathcal{R}$, this does not imply that (A^T, C^T) is reachable over $\mathcal{R}$. This fact is illustrated in the following example.

Example 2.3.7 Let $\mathcal{R} = \mathbb{R}[s]$, i.e. $\mathcal{R}$ is the ring of all polynomials in the indeterminate s with coefficients in $\mathbb{R}$. Choose $A = 1$ and $C = s$. Then (C, A) is observable because $Cx = sx = 0$ implies that $x = 0$. So C is injective. However, $(A^T, C^T) = (1, s)$ is not reachable because the one column of $C^T = s$ does not generate $\mathbb{R}[s]$. Clearly $\text{span}(s) = \{s \cdot p(s) \mid p(s) \in \mathbb{R}[s]\}$ is a proper subset of $\mathbb{R}[s]$.

Now we have seen that for systems over rings the concepts of reachability and observability are not dual, we want to restore this duality in a certain sense. Therefore we define

Definition 2.3.8 Let $\Sigma = (A, B, C, D)$ be a system over a commutative ring $\mathcal{R}$. Then Σ is called *coreachable* if the pair (A^T, C^T) is reachable over $\mathcal{R}$.

By definition the notions of reachability and coreachability are completely dual. From Lemma 2.3.4 we know that when $\mathcal{R}$ is a commutative ring with identity, coreachability implies observability, but not the other way around. Hence coreachability is a stronger property than observability for systems over rings. This is the first difference we encounter between systems over rings and systems over fields, because in the field case these properties are the same.

2.4 Some facts about realizations

Consider a linear continuous-time system over $\mathbb{R}$ of the form (2.3). Assume that the initial state $x(0) = 0$. Then the Laplace transform with symbol z of (2.3) is given by

$$\begin{cases} z\hat{x}(z) &= A\hat{x}(z) + B\hat{u}(z), \\ \hat{y}(z) &= C\hat{x}(z) + D\hat{u}(z), \end{cases}$$

where $\hat{x}(z)$, $\hat{u}(z)$ and $\hat{y}(z)$ are the Laplace transforms of the state, input and output respectively. Eliminating $\hat{x}(z)$, we obtain

$$\hat{y}(z) = (D + C(zI - A)^{-1}B)\hat{u}(z).$$

This equation describes the direct relationship between the inputs and the outputs of the system. The state-variable is eliminated. The function relating $\hat{y}(z)$ and $\hat{u}(z)$ is called the *transfer matrix* of the system (2.3). It is a matrix over the ring of proper real rational functions, i.e. the degree of the numerator of each entry is smaller than or equal to the degree of its denominator.

The idea of transfer matrices as a description of the direct relationship between inputs and outputs, is easily generalized to the ring case. For this purpose, we first give a formal definition of the concept of proper rational functions over a ring $\mathcal{R}$.

Definition 2.4.1 Let $\mathcal{R}$ be a commutative ring with identity, and $\mathcal{R}(z)$ be the ring of all rational functions in the indeterminate z with coefficients in $\mathcal{R}$. An element $r(z) \in \mathcal{R}(z)$ is called *proper* if $r(z)$ can be written as

$$r(z) = \frac{p(z)}{q(z)},$$

where $p(z), q(z) \in \mathcal{R}[z]$ satisfy the conditions

(i) $q(z)$ is monic (i.e. the coefficient of the leading term of $q(z)$ is equal to 1),

(ii) $\deg_z(p(z)) \leq \deg_z(q(z))$.

Moreover, if $\deg_z(p(z)) < \deg_z(q(z))$, then $r(z)$ is called *strictly proper*. The set of all proper rational functions in $\mathcal{R}(z)$ is denoted by $\mathcal{R}_p(z)$.

Definition 2.4.2 Let $\Sigma = (A, B, C, D)$ be a system over an integral domain $\mathcal{R}$. Then the *transfer matrix* of Σ is the $p \times m$ matrix over $\mathcal{R}_p(z)$ defined by

$$T(z) := D + C(zI - A)^{-1}B. \tag{2.10}$$

When we are studying linear continuous-time systems over $\mathbb{R}$, these systems are often not given in state-space form (2.3), but as some differential equations involving only inputs and outputs. From these input-output equations the transfer matrix of a system is easily obtained. But a transfer matrix is often not enough; instead of it a state-space description of the system under consideration is required. Hence, we want to find real matrices A, B, C and D of appropriate dimensions such that

$T(z) = D + C(zI - A)^{-1}B$. We say that $\Sigma = (A, B, C, D)$ realizes $T(z)$ or is a *realization* of $T(z)$.

From the theory of linear systems over $\mathbf{R}$ it is well known that a proper real rational transfer matrix always has (non unique) realizations (see for example [47]). Moreover, one can always choose (A, B, C, D) in such a way that (A, B) is reachable and (C, A) is observable. Such a realization is called *canonical*. Among all realizations of $T(z)$ the canonical realizations have smallest dimension, i.e. the size n of the square matrix A is as small as possible. This is of course a very desirable property because we do not want to introduce superfluous states.

For systems over integral domains the realization problem is much more complicated. Given a transfer matrix $T(z)$, the existence of a quadruple of matrices over $\mathcal{R}$ such that $T(z) = D + C(zI - A)^{-1}B$, is not very difficult to prove. In principle, each entry can be realized separately using the same techniques as in the field case (see e.g. [47, Chapter 6]). The realization $\Sigma = (A, B, C, D)$ of the transfer matrix $T(z)$ then merely consists of a composition of all these separate parts. Although in this way a realization can be obtained, it is very unlikely that this is a minimal realization. In order to remove superfluous states, we are also interested in canonical realizations, i.e. realizations that are both reachable and observable. Unfortunately, these canonical realizations do not always exist, as is illustrated by the next example taken from [85].

Example 2.4.3 Let $\mathcal{R} = \mathbf{R}[s_1, s_2]$ be the ring of all polynomials in the indeterminates s_1 and s_2 with coefficients in $\mathbf{R}$. Consider the following 1×2 transfer matrix $T(z)$ in $\mathcal{R}_p(z)$:

$$T(z) = \left(\frac{s_1}{z-1} \mid \frac{s_2}{z-1} \right). \tag{2.11}$$

A realization of $T(z)$ is given by $\Sigma = (1, (s_1 \mid s_2), 1, 0)$. Clearly, Σ is observable, but not reachable, since the columns of $B = (s_1 \mid s_2)$ generate the ideal

$$\{s_1 \cdot p(s_1, s_2) + s_2 \cdot q(s_1, s_2) \mid p(s_1, s_2), q(s_1, s_2) \in \mathbf{R}[s_1, s_2]\}, \tag{2.12}$$

which is not equal to $\mathbf{R}[s_1, s_2]$. It can be proved that the transfer matrix $T(z)$ does not have a canonical realization. The problem is that the columns of the matrix B generate an $\mathbf{R}[s_1, s_2]$-ideal, given in (2.12), which does not have a basis, i.e., there do not exist linear independent elements in $\mathbf{R}[s_1, s_2]$ that generate the $\mathbf{R}[s_1, s_2]$-ideal given in (2.12).

Remark 2.4.4 The problem of non-existence of canonical realizations described above can be solved by a slight generalization of the definition of a system over a ring $\mathcal{R}$ (see [85]). Instead of a quadruple of matrices, a system over a ring $\mathcal{R}$ is a quintuple (X, A, B, C, D), where X (the state space) is a finitely generated $\mathcal{R}$-module, and $A : X \to X$, $B : \mathcal{R}^m \to X$, $C : X \to \mathcal{R}^p$ and $D : \mathcal{R}^m \to \mathcal{R}^p$ are $\mathcal{R}$-linear maps. The definitions for reachability and observability are then very straightforward generalizations of the Definitions 2.2.2 and 2.3.2, respectively. For example, a system is called reachable in this context if the elements $B(e_1), \ldots, B(e_m), (AB)(e_1), \ldots, (AB)(e_m), \ldots, (A^{n-1}B)(e_1), \ldots, (A^{n-1}B)(e_m)$ generate the state space module X. Here $e_1, \ldots, e_m$ denotes the standard basis in $\mathcal{R}^m$.

In this more general setting it is proven that canonical realizations always exist (see [19, Chapter 16, Section 5]).

If the state space $X = \mathcal{R}^n$, we are back in the situation of Definition 2.1.1. Such systems are called *free*. Since we are mainly interested in free systems, we forget about the generalized notion of systems over rings in the rest of the thesis, and always use our original Definition 2.1.1 instead.

The next proposition is also taken from [85] and states that canonical realizations in the classical sense do not contain superfluous states.

Proposition 2.4.5 *A canonical realization which is free (i.e. a canonical realization in the ordinary sense), has minimal rank among the free realizations.* ∎

Now there is only one question left: what are the conditions under which a transfer matrix has a (free) canonical realization? Example 2.4.3 indicates that this is a difficult problem. However, in the special case of systems over a Principal Ideal Domain (PID) there is an affirmative answer:

Proposition 2.4.6 *Systems over a PID (Principal Ideal Domain) always admit a (free) canonical realization.* ∎

For a proof of this proposition we refer to [22]. Moreover, this article contains a constructive proof of the result. So there exists an effective realization algorithm for systems over PID's.

Remark 2.4.7 Recall that systems with commensurable time-delays can be modeled as a system over the ring $\mathbf{R}[s]$. Since $\mathbf{R}[s]$ is a PID, such systems always have a (free) canonical realization. For systems with incommensurable time-delays this does not hold any more. This was already pointed out in Example 2.4.3. Proposition 2.4.6 is not applicable in this case, because a ring of polynomials in more than one indeterminate and with coefficients in $\mathbf{R}$ is not a PID.

Remark 2.4.8 It is also possible to introduce the dual concept of canonicity. A system $\Sigma = (A, B, C, D)$ is called *cocanonical* if $\Sigma^T = (A^T, C^T, B^T, D^T)$ is canonical. This means that in a cocanonical systems Σ, (C, A) is coreachable and (B^T, A^T) is observable. Note that in Example 2.4.3 the realization we gave for $T(z)$ is cocanonical, although not canonical. This illustrates the difficulties that arise for systems over rings. In contrast to the field case, canonicity and cocanonicity are not the same properties.

2.5 Stability and Hurwitz sets

At first sight, the generalization of the notion of stability to the theory of systems over rings, seems a rather troublesome exercise. At least two problems appear. First of all, the notion of stability in the classical sense, i.e. for systems over the field $\mathbf{R}$, is defined as a desirable asymptotic property of the state- and output trajectories. In the ring case, only the system defining matrices are available. Moreover, the notion of stability depends on the application one has in mind. Recalling the three

examples in Section 2.1, it is obvious that the intuitive notion of stability is different in all three cases. Hence, our general definition of stability has to be adaptable: a specialization is needed in each separate case. It is possible to incorporate all these specializations into one framework. This framework is based on the concept of so-called Hurwitz sets. The rest of this section is devoted to the elaboration of this general notion of stability.

Consider the continuous-time system (2.3) over $\mathbb{R}$ once more. For this system there are at least two different notions of stability. The system is called *internally stable* if the state x of the system eventually tends to zero if no input is applied. But not only the internal asymptotic behaviour is interesting; the external asymptotic behaviour, although somewhat more difficult to describe, is also important. A system is called *externally stable* or *BIBO-stable* (this is an acronym for Bounded Input Bounded Output) if the system, starting with a zero initial condition, and after application of a bounded input trajectory, produces a bounded output trajectory. It is well known that both properties are easily tested using only the system defining matrices A, B, C and D (see e.g. [47, pp. 175-176]):

Theorem 2.5.1 *Let $\Sigma = (A, B, C, D)$ be a continuous-time system over $\mathbb{R}$ as described in equation (2.3). Then*

(i) Σ is internally stable if and only if the characteristic polynomial $\chi_A(z) = \det(zI - A)$ of A has no zeros in $\overline{\mathbb{C}^+}$. Alternatively stated: $\sigma(A) \subset \mathbb{C}^-$.

(ii) Σ is externally stable if and only if the entries of the transfer matrix $T(z) = D + C(zI - A)^{-1}B$ have no poles in $\overline{\mathbb{C}^+}$. ∎

Both conditions *(i)* and *(ii)* in Theorem 2.5.1 can be reformulated when we introduce the set $\mathcal{D} := \{p(z) \in \mathbb{R}[z] \mid \forall \lambda \in \overline{\mathbb{C}^+} : p(\lambda) \neq 0\}$. $\mathcal{D}$ can be seen as the set of all stable polynomials; a system is internally stable if and only if its characteristic polynomial $\chi_A(z)$ is a stable polynomial: $\chi_A(z) \in \mathcal{D}$. External stability means that the denominators of the entries of the transfer matrix $T(z)$ are stable. The introduction of a set of stable polynomials has two important advantages. First of all, the same idea can be used to define stability in the systems over ring case. But what is probably more important: by changing the set of stable polynomials, the notion of stability can be changed. So, in each particular application it is possible to choose the set of stable polynomials in such a way that it coincides with the intuitive notion of stability in that particular case.

Not any arbitrary set of polynomials can serve as a set of stable polynomials. It is obvious that because of the special role these sets play, they have to satisfy certain conditions, which are naturally related to the idea of stability. A set satisfying these conditions is called a Hurwitz set. In the literature they also occur under the name stability set (see [44]) or denominator set (see [18]).

Definition 2.5.2 Let $\mathcal{R}$ be an integral domain. A subset $\mathcal{D}$ of the polynomial ring $\mathcal{R}[z]$ is called a *Hurwitz set* if it satisfies the following conditions:

(i) $\mathcal{D}$ is *multiplicative*, i.e. $1 \in \mathcal{D}$ and if $p, q \in \mathcal{D}$, then $p \cdot q \in \mathcal{D}$.

(ii) Each polynomial $p \in \mathcal{D}$ is *monic*, i.e. its leading coefficient is equal to 1, (as a consequence, $0 \notin \mathcal{D}$).

(iii) $\mathcal{D}$ is *saturated*, i.e. if $p \in \mathcal{D}$ and q is monic and divides p, then $q \in \mathcal{D}$.

(iv) There exists an $\alpha \in \mathcal{R}$ such that $(z - \alpha) \in \mathcal{D}$.

With a Hurwitz set $\mathcal{D}$ we can associate a ring of fractions, denoted by $\mathcal{R}_{\mathcal{D}}(z)$:

$$\mathcal{R}_{\mathcal{D}}(z) := \{\frac{p(z)}{q(z)} \in \mathcal{R}(z) \mid p(z) \in \mathcal{R}[z], q(z) \in \mathcal{D}\}. \tag{2.13}$$

Hence, a rational function in $\mathcal{R}(z)$ belongs to $\mathcal{R}_{\mathcal{D}}(z)$ if it has a stable denominator polynomial, i.e. the denominator is an element of $\mathcal{D}$. Adding and multiplying fractions as expected, it is obvious that $\mathcal{R}_{\mathcal{D}}(z)$ is indeed a ring. It can be considered as the ring of all stable rational functions.

It is now possible to generalize the concept of stability to systems over rings.

Definition 2.5.3 Let $\mathcal{R}$ be an integral domain and let $\Sigma = (A, B, C, D)$ be a system over $\mathcal{R}$. Let $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Then:

(i) Σ is called *internally stable* (with respect to $\mathcal{D}$) if $\chi_A(z) := \det(zI - A) \in \mathcal{D}$,

(ii) Σ is called *externally stable* (with respect to $\mathcal{D}$) if all entries of the transfer function $T(z) = D + C(zI - A)^{-1}B$ of Σ belong to $\mathcal{R}_{\mathcal{D}}(z)$.

Note that internal stability implies external stability because $(zI - A)^{-1}$ can be written as $(zI - A)^{-1} = \frac{\text{adj}(zI-A)}{\det(zI-A)}$. The converse is not true of course.

With the definition of stability as given above, the conditions imposed upon a Hurwitz set $\mathcal{D}$ become clear. First of all, the composition of two stable systems ought to be stable again, hence $\mathcal{D}$ has to be multiplicative. Clearly this works also the other way around. If a stable system can be decomposed into two completely independent subsystems, both these two subsystems have to be stable too. This clarifies condition *(iii)* of Definition 2.5.2. Since characteristic polynomials are monic, we can restrict the definition to this class of polynomials. Finally, an interpretation of condition *(iv)* in Definition 2.5.2 is difficult to give. This is only a technical condition that facilitates some of the proofs.

To illustrate the use of Hurwitz sets, and to show how the classical notion of stability can be incorporated in this general framework, we give a few examples.

Example 2.5.4 Let $\mathcal{R} = \mathbb{R}$, and consider the system $\Sigma = (A, B, C, D)$ as a continuous-time system over $\mathbb{R}$:

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{cases} \tag{2.14}$$

Let $\mathbb{C}_g$ be a subset of $\mathbb{C}$ such that $\mathbb{C}_g \cap \mathbb{R}$ is non-empty ($\mathbb{C}_g$ is a so-called stability domain). Define the Hurwitz set $\mathcal{D}$ as:

$$\mathcal{D} := \{p(z) \in \mathbb{R}[z] \mid p(z) \text{ is monic and } ([p(\alpha) = 0] \Longrightarrow [\alpha \in \mathbb{C}_g])\}. \tag{2.15}$$

By definition, the system (2.14) is internally stable w.r.t. $\mathcal{D}$ if and only if $\chi_A(z) \in \mathcal{D}$. This means that $\chi_A(\alpha) = 0$ implies that $\alpha \in \mathbb{C}_g$. Thus the spectrum of A has to be contained in $\mathbb{C}_g$: $\sigma(A) \subset \mathbb{C}_g$. So the definition of stability using the Hurwitz set $\mathcal{D}$ in (2.15) coincides with the notion of $\mathbb{C}_g$-stability. If $\mathbb{C}_g = \mathbb{C}^-$, we are back at the classical definition of stability for continuous-time linear systems over $\mathbf{R}$.

Example 2.5.5 Let $\mathcal{R} = \mathbf{R}$, and consider a discrete-time system $\Sigma = (A, B, C, D)$ over $\mathbf{R}$:

$$\begin{cases} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{cases} \tag{2.16}$$

Define the Hurwitz set $\mathcal{D}$ as

$$\mathcal{D} := \{p(z) \in \mathbf{R}[z] \mid p(z) \text{ is monic and } ([p(\alpha) = 0] \Longrightarrow [|\alpha| < 1])\}. \tag{2.17}$$

So the system (2.16) is internally stable if and only if $\chi_A(z)$ has only zeros within the open unit disc $\{z \in \mathbf{C} \mid |z| < 1\}$, i.e. all the elements of the spectrum $\sigma(A)$ of A are smaller than one in absolute value. In this way the classical notion of stability for discrete-time systems is translated into the Hurwitz set terminology.

From the point of view of systems over rings, the systems $\Sigma = (A, B, C, D)$ over $\mathbf{R}$ considered in the Examples 2.5.4 and 2.5.5, are completely the same. However, in each example the same quadruple of matrices is given another interpretation, and therefore the Hurwitz set $\mathcal{D}$ has to be adapted to this specific interpretation. Such an adaptation is also possible for time-delay systems, as is illustrated in the next example.

Example 2.5.6 Let $\mathcal{R} = \mathbf{R}[s_1, \ldots, s_k]$, and $\Sigma = (A, B, C, D)$ be a system over $\mathcal{R}$. Let $0 < \tau_1 < \tau_2 < \cdots < \tau_k$ denote a k-tuple of incommensurable time-delays and introduce the corresponding delay operators σ_i $(i = 1, \ldots, k)$, acting on the state and input trajectories:

$$\sigma_i x(t) = x(t - \tau_i) \qquad (i = 1, \ldots, k).$$

Substituting the delay operators σ_i for the indeterminates s_i $(i = 1, \ldots, k)$, the system Σ can be considered as a time-delay system:

$$\begin{cases} \dot{x}(t) &= A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) &= C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t). \end{cases} \tag{2.18}$$

This time-delay system is called internally stable if for any initial condition (which is an initial state trajectory in this case; see Sections 1.1 and 1.3), the state x tends to zero for $t \longrightarrow +\infty$ when no inputs are applied. According to Hale (see [41, p.182, Corollary 4.1]), this intuitive notion of stability is equivalent to the following condition on the matrix $A(s_1, \ldots, s_k)$: the system (2.18) is internally stable if and only if all roots of the characteristic equation

$$\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = 0, \tag{2.19}$$

lie in $\mathbf{C}^-$, the open left half plane. This condition is a rather straightforward generalization of the ordinary condition of internal stability for systems without delays because in the Laplace transform with symbol z the delay operator σ_i is transformed to $e^{-\tau_i z}$. Moreover, this definition of stability fits also into the framework of Hurwitz sets in the following way. Define $\mathcal{D}$ as

$$\mathcal{D} := \{p(z, s_1, \ldots, s_k) \in \mathbf{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } p(\lambda, e^{-\tau_1\lambda}, \ldots, e^{-\tau_k\lambda}) = 0 \Rightarrow \lambda \in \mathbf{C}^-\}. \quad (2.20)$$

Then clearly $\chi_A(z) \in \mathcal{D}$ if and only if the system (2.18) is internally stable. Note that the knowledge about the time-delay character of the system (2.18) is not evident in the system defining matrices $\Sigma = (A, B, C, D)$, but is used in the definition of the corresponding Hurwitz set $\mathcal{D}$. So definition (2.20) formalizes the intuitive notion of stability in this particular situation into the abstract setting of Hurwitz sets.

From the three examples above it is apparent that the concept of Hurwitz sets is very versatile. They can be modified in such a way that they are useful in a lot of interesting applications. Nevertheless, one of the main advantages is not mentioned yet, but will certainly become clear in the rest of this chapter. In the theory of systems over rings it is possible to work with an arbitrary Hurwitz set, so without specifying it beforehand. In this way an abstract but very general theory of stability and stabilizability can be obtained. In the rest of this chapter this abstract framework will be developed. Each particular application can be seen as a specialization of the general setup. But the main theory is the same in all these cases, and is developed only once. The specific details for each particular situation can be elaborated later on.

2.6 Pole placement and static state feedback

One of the main tools in the classical control design of ordinary linear systems over the field $\mathbf{R}$ is the use of static state feedback. It is a very straightforward technique to change the internal dynamics of the system, for example to achieve internal stability. In this section it is explained why static state feedback is not very useful in the systems over rings case. This is also the motivation for the approach which is used in the rest of this chapter, and which is based on the application of dynamic feedback.

Consider a continuous-time system over $\mathbf{R}$ given by

$$\dot{x}(t) = Ax(t) + Bu(t), \quad (2.21)$$

where $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times m}$ (in this section the output equation is not of interest). Suppose that this system is not internally stable w.r.t. a certain Hurwitz set $\mathcal{D}$. Then we may try to change the dynamics of the system to achieve stability. In order to do so, we apply the static state feedback

$$u(t) = -Fx(t) + r(t), \quad (2.22)$$

where F is an $m \times n$ matrix over $\mathbf{R}$ and $r(t)$ is the new input to the system (see Figure 2.1). In this way the following closed-loop system is obtained

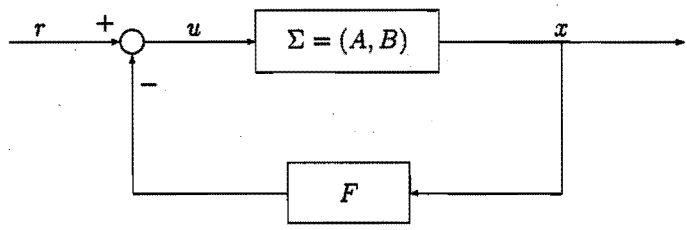


Figure 2.1: Closed-loop system with static state feedback

$$\dot{x}(t) = (A - BF)x(t) + Br(t). \tag{2.23}$$

We see that the internal dynamics of the closed-loop system (2.23) are different from the internal dynamics of the open-loop system (2.21). To stabilize the system (2.21) with respect to the Hurwitz set $\mathcal{D}$, we have to find an $F \in \mathbf{R}^{m\times n}$ such that

$$\chi_{A-BF}(z) = \det(zI - (A - BF)) \in \mathcal{D}.$$

Formally the idea of static state feedback is easily generalized to the ring case. It can be seen as a matrix $F \in \mathcal{R}^{m\times n}$ changing the system defining quadruple of matrices (A, B, C, D).

Definition 2.6.1 Let $\Sigma = (A, B, C, D)$ be a linear system over a commutative ring $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times m}$, $C \in \mathcal{R}^{p\times n}$ and $D \in \mathcal{R}^{p\times m}$. Then a matrix $F \in \mathcal{R}^{m\times n}$ is called a *static state feedback*. The feedback F transforms the *open-loop* system $\Sigma = (A, B, C, D)$ to the *closed-loop* system $\Sigma_{cl} = (A - BF, B, C - DF, D)$.

Although Definition 2.6.1 looks rather formal, it is easily seen that this is exactly the description of a static state feedback in all applications of systems over rings we have seen thus far. A static state feedback is simply a linear map from the state space to the input space. The knowledge of the state x is used to choose the input u in such a way that the internal dynamics of the system are changed in a favourable way.

In the classical situation of systems over the field $\mathbf{R}$, static state feedback is often used for pole assignment. A pole of the system (2.21) is a (complex) zero of the characteristic polynomial $\chi_A(z)$ of A. If a system $\Sigma = (A, B)$ over $\mathbf{R}$ is reachable, it is possible to assign all poles of the system to some arbitrary values by an appropriate choice of the feedback F. Moreover, reachability of (A, B) is a necessary condition for pole placement too. This so-called Pole-Shifting Theorem is very well known. For the proof we refer to [86, p. 134]; on page 185 of [86] an extensive history of the realization of this result is given. Unfortunately, the problem of pole placement is much more difficult to solve in the systems over rings case. However, the same ideas are easily generalized to this more general situation.

Definition 2.6.2 Let $\Sigma = (A, B, C, D)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times m}$, $C \in \mathcal{R}^{p\times n}$ and $D \in \mathcal{R}^{p\times m}$. Then

(i) Σ is called *coefficient assignable* if for all $\alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in \mathcal{R}$ there exists a static state feedback $F \in \mathcal{R}^{m\times n}$, such that the characteristic polynomial $\chi_{A-BF}(z)$ of the closed-loop system is equal to

$$\chi_{A-BF}(z) = z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0 = z^n + \sum_{i=0}^{n-1} \alpha_i z^i.$$

(ii) Σ is called *pole assignable* if for all $p_1, p_2, \ldots, p_n \in \mathcal{R}$ there exists a static state feedback $F \in \mathcal{R}^{m\times n}$, such that the characteristic polynomial $\chi_{A-BF}(z)$ of the closed-loop system is equal to

$$\chi_{A-BF}(z) = (z - p_1)(z - p_2)\cdots(z - p_n) = \prod_{i-1}^{n}(z - p_i).$$

From this definition it follows that coefficient assignability implies pole assignability. The implication in the opposite direction however, does not hold in general. Moreover, in the field case, pole assignability is equivalent to reachability, but for systems over rings the situation is more involved as is illustrated by the next propositions:

Proposition 2.6.3 *Let $\Sigma = (A, B, C, D)$ be a system over an integral domain $\mathcal{R}$. Suppose that Σ is pole assignable. Then Σ is reachable.* ∎

For a proof of Proposition 2.6.3 we refer to [85, p. 21] or [5, p.67].

Proposition 2.6.4 *Let $\Sigma = (A, B, C, D)$ be a single input system over an integral domain $\mathcal{R}$, i.e. $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times 1}$, $C \in \mathcal{R}^{p\times n}$ and $D \in \mathcal{R}^{p\times 1}$. Then Σ is coefficient assignable if and only if Σ is reachable.* ∎

This result may be found in [5, p. 70] or [49, Section 4.3].

Proposition 2.6.5 *Let $\Sigma = (A, B, C, D)$ be a system over a principal ideal domain $\mathcal{R}$. Then Σ is pole assignable if and only if Σ is reachable.* ∎

The proof of Proposition 2.6.5 (in the special case $\mathcal{R} = \mathbb{R}[s]$) originates from Morse ([71]). In [5, pp. 91-92] a proof for arbitrary principal ideal domains is given. An explicit algorithm to assign the poles of a reachable system over a PID to certain desired values is described by Eising in [20] and [21].

Proposition 2.6.6 *Consider the class of linear systems over the polynomial ring $\mathcal{R} = \mathbb{R}[s_1, \ldots, s_k]$, with $k > 1$. For this class of systems, reachability does not imply pole assignability.* ∎

A counterexample that proves this result is given in [89] and [90].

The result of Proposition 2.6.6 implies that reachability of a system over a ring is in general not enough to ensure its pole assignability. This indicates that static state feedback is not such a powerful tool for the control of a system over a ring.

For systems over the field $\mathbb{R}$, pole placement by static state feedback is very useful for the stabilization of a system. In this case reachability is a necessary and sufficient condition for pole assignability. Moreover, from [66] we recall that a system $\Sigma = (A, B, C, D)$ over $\mathbb{R}$ is generically reachable. So in most cases the method of pole placement can be used to stabilize the system, despite the fact that pole assignability is a sufficient condition for stabilizability but not a necessary one.

For systems over an arbitrary integral domain, the strategy for stabilization described above is simply not applicable. In general reachability of a system does not imply pole assignability, although it is a necessary condition. Moreover, the condition of reachability itself is rather restrictive. Recalling Proposition 2.2.5, we know that a system over the polynomial ring $\mathcal{K}[s_1, \ldots, s_k]$ is generically reachable if and only if the number of inputs m is strictly larger than the number of indeterminates k. So even if $\mathcal{R}$ is the principal ideal domain $\mathbb{R}[s]$, and reachability and pole assignability are equivalent, a system over $\mathcal{R} = \mathbb{R}[s]$ will only be generically reachable if it has at least two inputs.

When we recall Example 2.5.6, it is not so difficult to understand why pole placement is such a strong property for systems over (polynomial) rings. Consider a system $\Sigma = (A, B, C, D)$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$, and assume that it is pole assignable. Let $p_1, \ldots, p_n \in \mathbb{R}$. Then it is possible to find a matrix $F \in \mathbb{R}[s_1, \ldots, s_k]^{m \times n}$ such that

$$\chi_{A-BF}(z) = (z - p_1)(z - p_2) \cdots (z - p_n).$$

Note that in the characteristic polynomial of $A - BF$ the indeterminates $s_1, \ldots, s_k$ do not occur any more. Next, regard the system $\Sigma = (A, B, C, D)$ as a time-delay system in the same fashion as in (2.18). The poles of this system are the zeros of the characteristic equation

$$\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = 0.$$

Although in general equations of this type have an infinite number of solutions in the complex plane, we shall see in Chapter 3 that only a finite number of these zeros is located in $\overline{\mathbb{C}^+}$. However, after application of the feedback law

$$u(t) = F(\sigma_1, \ldots, \sigma_k)x(t) + v(t),$$

the characteristic equation becomes

$$\chi_{A-BF}(z) = (z - p_1)(z - p_2) \cdots (z - p_n) = 0,$$

and there are only a finite number of poles left. So the number of poles is reduced drastically by this static state feedback, and this extremely strong result is an immediate consequence of our assumption on the pole assignability of the system $\Sigma = (A, B, C, D)$. This explains why pole assignability is often a far too strong property to ask for.

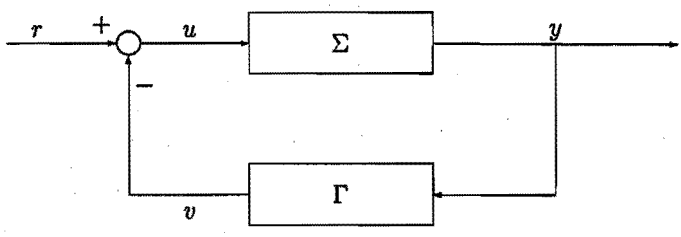


Figure 2.2: Closed-loop system with dynamic output feedback

Although pole assignability is a sufficient condition for stabilizability, it is certainly not a necessary condition. The sufficiency follows immediately from the existence of a polynomial of the form $(z-\alpha)$ (with $\alpha \in \mathcal{R}$) in $\mathcal{D}$, and the fact that $\mathcal{D}$ is multiplicative. So $(z-\alpha)^n$ is also a stable polynomial. However, to stabilize a system it is not necessary that the poles of the system can be assigned arbitrarily. The existence of a static state feedback F such that $\chi_{A-BF}(z) \in \mathcal{D}$ is enough. In the special case of time-delay systems as given above, this means that only the (finite number of) poles in $\overline{\mathbb{C}^+}$ have to be shifted into the open left half plane $\mathbb{C}^-$. But it is difficult to find necessary and sufficient conditions to test this property. Therefore we shall consider a more general concept of feedback instead. In the rest of this chapter it will turn out that the problem of stabilizability can be solved completely using this more general notion of dynamic feedback.

2.7 Dynamic feedback and stabilizability

For linear systems over the field $\mathbb{R}$ it is well known that the dynamics of a system cannot be influenced only by static but also by *dynamic* feedback. This means that the output of a system Σ is fed back to the input via another linear dynamical system Γ, called a compensator, cf. Figure 2.2. Under the condition of well posedness (which will be explained later), the closed-loop system is again a linear system of the same form as Σ and Γ.

It is obvious that the class of dynamical compensators is much larger than that of static feedbacks; in fact, all static feedbacks are contained in the class of dynamic feedback compensators. Therefore it is clear that dynamic feedback is a more general tool, that can be used to solve a larger class of control problems.

We first generalize the notion of dynamic feedback to the situation of systems over rings. This can be done in the same fashion as for static static state feedback, by first introducing the concept for systems over fields, and then generalizing the formulae to the systems over rings case. Since we have already seen this formalism several times, we now prefer to proceed in a somewhat more informal and intuitive

way. However, we still treat both discrete- and continuous-time systems together. Therefore the following set of equations is identified with a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$:

$$\begin{cases} \Delta x &= Ax + Bu, \\ y &= Cx + Du, \end{cases} \tag{2.24}$$

where Δ is an operator symbolizing differentiation with respect to time in the continuous-time case, and a right time-shift in the discrete-time case. In description (2.24) the dynamical structure of a system and the relations between the input u, the output y and the state x are much clearer than in the formal Definition 2.1.1. In the same way, a compensator $\Gamma = (F, G, H, J)$ is not only considered as a quadruple of matrices over the integral domain $\mathcal{R}$, but also regarded as a system with input y, output v and state w, evolving in time according to the equations:

$$\begin{cases} \Delta w &= Fw + Gy, \\ v &= Hw + Jy. \end{cases} \tag{2.25}$$

Given a system $\Sigma = (A, B, C, D)$ and a feedback compensator $\Gamma = (F, G, H, J)$ over $\mathcal{R}$, their feedback interconnection is called *well posed* if the closed-loop system of Σ and Γ, as depicted in Figure 2.2, constitutes a linear system again. This means that the system equations (2.24) and (2.25) together with the feedback equation

$$u = r - v \tag{2.26}$$

can be recast in such a form that a linear system $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$

$$\begin{cases} \Delta \hat{x} &= \hat{A}\hat{x} + \hat{B}r, \\ y &= \hat{C}\hat{x} + \hat{D}r, \end{cases}$$

is obtained, where r is the new external input, y the output and $\hat{x}$ the (new) state of the closed-loop system. This recasting is possible if and only if the matrix $(I + DJ)$ is invertible as a matrix over $\mathcal{R}$ (i.e. the determinant of $(I + DJ)$ is a unit of $\mathcal{R}$). The invertibility of this matrix is called the well-posedness condition.

The closed-loop system of Figure 2.2 satisfies the equations (2.24), (2.25) and (2.26). Substituting formula (2.26) for u and the second formula of (2.25) for v, formula (2.24) for y becomes

$$y = Cx + D(r - v) = Cx + Dr - DHw - DJy.$$

Hence

$$(I + DJ)y = Cx - DHw + Dr.$$

If the closed-loop system is well posed, the matrix $(I + DJ)$ is invertible over $\mathcal{R}$. Defining $E := (I + DJ)^{-1}$, the previous formula can be rewritten as

$$y = ECx - EDHw + EDr.$$

Substitution of this expression in the formulae for Δx and Δw in (2.24) and (2.25) yields:

$$\begin{aligned}\Delta x &= Ax + Bu = Ax + Br - Bv = Ax - BHw - BJy + Br = \\ &= (A - BJEC)x + (-BH + BJEDH)w + (B - BJED)r,\end{aligned}$$

$$\begin{aligned}\Delta w &= Fw + Gy = \\ &= GECx + (F - GEDH)w + GEDr.\end{aligned}$$

So the closed-loop system Σ_{cl} is determined by the equations

$$\begin{cases} \Delta \begin{pmatrix} x \\ w \end{pmatrix} = \begin{pmatrix} A - BJEC & -BH + BJEDH \\ GEC & F - GEDH \end{pmatrix} \begin{pmatrix} x \\ w \end{pmatrix} + \\ \qquad\qquad + \begin{pmatrix} B - BJED \\ GED \end{pmatrix} r, \\ y = (EC \mid -EDH) \begin{pmatrix} x \\ w \end{pmatrix} + EDr. \end{cases} \tag{2.27}$$

Finally, define

$$\hat{A} := \begin{pmatrix} A - BJEC & -BH + BJEDH \\ GEC & F - GEDH \end{pmatrix}, \tag{2.28}$$

$$\hat{B} := \begin{pmatrix} B - BJED \\ GED \end{pmatrix}, \tag{2.29}$$

$$\hat{C} := (EC \mid -EDH), \tag{2.30}$$

$$\hat{D} := ED, \tag{2.31}$$

and we conclude that the closed-loop system is characterized by the quadruple of matrices $\Sigma_{cl} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ over $\mathcal{R}$.

Note that the well-posedness condition of $(I + DJ)$ to be invertible over $\mathcal{R}$ looks rather restrictive. For a lot of rings this condition is generically not satisfied. However, if $J = 0$ or $D = 0$, then $(I + DJ) = I$, and trivially $(I + DJ)$ is invertible. In the case $J = 0$, formulae (2.28) to (2.31) become

$$\hat{A} = \begin{pmatrix} A & -BH \\ GC & F - GDH \end{pmatrix}, \hat{B} = \begin{pmatrix} B \\ GD \end{pmatrix}, \hat{C} = (C \mid -DH), \hat{D} = D. \tag{2.32}$$

So a strictly proper compensator is always well posed. Moreover, it will turn out that when a system is stabilizable by dynamic output feedback, a strictly proper compensator (i.e. a compensator $\Gamma = (F, G, H, 0)$ without a direct feedthrough) can always do the job. Hence the condition of well posedness is not at all a restriction for the stabilization problem.

From the exposition above it follows that after application of a dynamic feedback compensator Γ, the dynamics of the closed-loop system Σ_{cl} are changed drastically in comparison with the dynamics of the original open-loop system Σ. This gives the possibility to stabilize a system using dynamic output feedback.

Definition 2.7.1 Let $\mathcal{R}$ be an integral domain and $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Let $\Sigma = (A, B, C, D)$ be a system over $\mathcal{R}$. Σ is called *(internally) stabilizable by dynamic output feedback* with respect to $\mathcal{D}$ if there exists a dynamic compensator $\Gamma = (F, G, H, J)$ over $\mathcal{R}$ such that the closed-loop of Σ and Γ is well posed and the closed-loop system $\Sigma_{cl} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$, determined by (2.28) to (2.31), is internally stable with respect to $\mathcal{D}$, i.e.

$$\chi_{\hat{A}}(z) = \det(zI - \hat{A}) \in \mathcal{D}.$$

The rest of this chapter is devoted to the question of stabilizability for a linear system $\Sigma = (A, B, C, D)$. For systems over the field $\mathbf{R}$, this question can be split into two dual parts, which can be treated separately. First the problem of stabilizability using (static) *state* feedback is solved. This can be seen as a special case ($C = I$ and $D = 0$) of the original problem, so with the output equal to the state. For this system we want to find a *static* compensator (i.e. a compensator of rank zero, only consisting of a direct feedthrough term J), stabilizing the system. In the case of a general C and D, one still wants to use this stabilizing feedback, but the problem is that the state is not available for feedback. For this purpose, a so-called stable observer is built. This is a stable dynamical system taking the input u and the output y of the original system Σ as inputs, and producing an estimate $\hat{x}$ for the state x of the original system Σ as an output. The problem of finding a stable observer is called the *detectability* problem and turns out to be dual to the stabilizability problem. The idea is now to use the output of the observer, i.e. the estimated state $\hat{x}$ as the input to the original stabilizing feedback. It is even possible to combine both the observer and the static feedback into one dynamic output feedback compensator. The feedback interconnection of the system Σ with this dynamic compensator is internally stable, and in this way the stabilizability problem is solved.

For systems over rings this so-called *separation principle* still works and we can follow almost the same strategy. There are only a few differences. The stabilizability problem by state feedback has to be solved using dynamic state feedback instead of static state feedback. Moreover, the detectability problem has to be recast in a different way because it is impossible to speak of an estimate for the state x in the context of systems over rings. It is possible to find an other wording for this notion using the concept of Hurwitz sets. With this new formulation the duality of the problems of detectability and stabilizability via dynamic state feedback is preserved. It is also possible to combine both solutions to find a dynamic internally stabilizing output feedback compensator.

In the next three sections the program described above is elaborated in more detail. First the problems of stabilizability by dynamic state feedback and of detectability are solved. Finally, after the development of these tools, we put them together to solve the problem of stabilizability by dynamic output feedback.

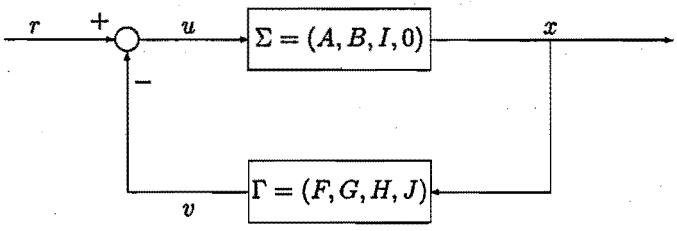


Figure 2.3: Closed-loop system with dynamic state feedback

2.8 Stabilizability by dynamic state feedback

Let $\mathcal{R}$ be an integral domain and let $\Sigma = (A, B, I, 0)$ denote a system over $\mathcal{R}$ given by the equations

$$\begin{cases} \Delta x &= Ax + Bu, \\ y &= x. \end{cases} \tag{2.33}$$

Let $\Gamma = (F, G, H, J)$ be a dynamic compensator and consider the closed-loop system depicted in Figure 2.3. From formulae (2.28) to (2.31) with $C = I$ and $D = 0$, it follows that this closed-loop system is given by the quadruple $\Sigma_{cl} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$, where

$$\hat{A} = \begin{pmatrix} A - BJ & -BH \\ G & F \end{pmatrix}, \hat{B} = \begin{pmatrix} B \\ 0 \end{pmatrix}, \hat{C} = (I \mid 0), \hat{D} = 0. \tag{2.34}$$

Since $D = 0$, we have $I + DJ = I$ for all $J \in \mathcal{R}^{m \times n}$, and thus all dynamic compensators are well posed.

Next, let $\mathcal{D}$ be a Hurwitz in $\mathcal{R}[z]$, which determines the stability of a system. Then we are interested in the existence of a compensator Γ such that the closed-loop system (2.34) is internally stable, i.e.

$$\det(zI - \hat{A}) \in \mathcal{D}.$$

To answer this existence question we need the following important lemma which also will turn out to be very useful in the sequel.

Lemma 2.8.1 *Let $\mathcal{R}$ be an integral domain, and let $A \in \mathcal{R}^{n \times n}$, $B \in \mathcal{R}^{n \times m}$ and $\varphi(z) \in \mathcal{R}[z]^n$ be given. Suppose that the equation*

$$(zI - A)\xi(z) + B\omega(z) = \varphi(z) \tag{2.35}$$

has a polynomial solution (ξ, ω), i.e. $\xi(z) \in \mathcal{R}[z]^n$ and $\omega(z) \in \mathcal{R}[z]^m$. Then there also exists a polynomial solution $(\hat{\xi}, \hat{\omega})$ with $\hat{\xi}(z) \in \mathcal{R}[z]^n$ and $\hat{\omega}(z) \in \mathcal{R}[z]^m$ which satisfies (2.35) and is such that

$$\deg_z(\hat{\omega}(z)) \leq n - 1. \tag{2.36}$$

Proof

Let $\xi(z) \in \mathcal{R}[z]^n$ and $\omega(z) \in \mathcal{R}[z]^m$ be a solution pair to equation (2.35). Since $\chi_A(z) = \det(zI - A)$ is a monic polynomial, it is possible to carry out the division algorithm with remainder on each of the entries of $\omega(z)$. So $\omega(z)$ can be written as

$$\omega(z) = \chi_A(z) \cdot \alpha(z) + \beta(z),$$

where $\alpha(z)$ and $\beta(z)$ are elements of $\mathcal{R}[z]^m$, and $\deg_z(\beta(z)) < \deg(\chi_A(z)) = n$. But then we have:

$$\begin{aligned}
\varphi(z) &= (zI - A)\xi(z) + B\omega(z) = \\
&= (zI - A)\xi(z) + B(\chi_A(z)\alpha(z) + \beta(z)) = \\
&= (zI - A)\xi(z) + \chi_A(z)B\alpha(z) + B\beta(z) = \\
&= (zI - A)\xi(z) + (zI - A) \cdot \operatorname{adj}(zI - A)B\alpha(z) + B\beta(z) = \\
&= (zI - A) \cdot [\xi(z) + \operatorname{adj}(zI - A)B\alpha(z)] + B\beta(z),
\end{aligned}$$

where we used the fact that Cramer's rule, $(zI - A) \cdot \operatorname{adj}(zI - A) = \chi_A(z) \cdot I$, also holds in the ring case. Now define

$$\begin{aligned}
\hat{\xi}(z) &:= \xi(z) + \operatorname{adj}(zI - A)B\alpha(z), \\
\hat{\omega}(z) &:= \beta(z).
\end{aligned}$$

Then clearly $\hat{\xi}(z) \in \mathcal{R}[z]^n$ and $\hat{\omega}(z) \in \mathcal{R}[z]^m$ form a solution pair for (2.35), and moreover $\deg_z(\hat{\omega}(z)) = \deg_z(\beta(z)) \leq n - 1$. This proves the claim. ∎

With help of Lemma 2.8.1 it is possible to prove the following crucial theorem which gives a necessary and sufficient condition for the solvability of the stabilizability problem by dynamic state feedback. This result was proved for the first time by Emre in [23] but the proof given below is based on the approach of Rouchaleau in [80].

Theorem 2.8.2 *Consider a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$, and assume that $C = I$ and $D = 0$. Let $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Then*

Σ is internally stabilizable with respect to the Hurwitz set $\mathcal{D}$ by dynamic state feedback,

$\Longleftrightarrow$

The matrix $(zI - A \mid B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$.

Proof

"$\Leftarrow$" Suppose that $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. Then there exist matrices $Q(z)$ and $P(z)$ over $\mathcal{R}_{\mathcal{D}}(z)$ such that

$$(zI - A)Q(z) + BP(z) = I.$$

Multiplying this equation by the least common multiple of the denominators of the entries of $Q(z)$ and $P(z)$, we obtain the equation

$$(zI - A)\hat{Q}(z) + B\hat{P}(z) = \varphi(z) \cdot I,$$

where $\hat{P}(z)$ and $\hat{Q}(z)$ are matrices over the polynomial ring $\mathcal{R}[z]$, and the least common multiple $\varphi(z)$ is an element of $\mathcal{D}$. Without loss of generality we may assume that $\deg_z(\varphi(z)) \geq n$, where n denotes the size of A. Otherwise we simply multiply this equation by a polynomial from $\mathcal{D}$ of sufficiently high degree. Because of condition *(iv)* in Definition 2.5.2 such a polynomial always exists.

Next, apply Lemma 2.8.1. We conclude that there exist polynomial matrices $\tilde{Q}(z)$ and $\tilde{P}(z)$ such that still

$$(zI - A)\tilde{Q}(z) + B\tilde{P}(z) = \varphi(z) \cdot I, \tag{2.37}$$

but also $\deg_z(\tilde{P}(z)) \leq n - 1$.

Now recall that $\varphi(z)$ is monic and of degree larger than or equal to n. Since $\deg_z(B\tilde{P}(z)) \leq n - 1$, we must have that $\deg_z((zI - A)\tilde{Q}(z)) = \deg_z(\varphi(z))$ and $(zI - A)\tilde{Q}(z)$ is monic. So $\tilde{Q}(z)$ is monic and $\deg_z(\tilde{Q}(z)) = \deg_z(\varphi(z)) - 1 \geq n - 1$. Denoting the degree of $\tilde{Q}(z)$ by k, we conclude that $\tilde{Q}(z)$ is of the form:

$$\tilde{Q}(z) = z^k \cdot I + \sum_{i=o}^{k-1} \tilde{Q}_i z^i.$$

Therefore $\det(\tilde{Q}(z)) = z^{kn} +$ lower order terms, and since $\tilde{Q}^{-1}(z) = \frac{\operatorname{adj}(\tilde{Q}(z))}{\det(\tilde{Q}(z))}$, it is clear that $\tilde{Q}(z)$ is invertible as a rational matrix (i.e. as a matrix over $\mathcal{R}(z)$).

Next consider the matrix

$$\tilde{P}(z)\tilde{Q}^{-1}(z) = \frac{1}{\det(\tilde{Q}(z))} \cdot \tilde{P}(z) \cdot \operatorname{adj}(\tilde{Q}(z)).$$

Since $\deg_z(\tilde{P}(z)) \leq n - 1$ and $\deg_z(\operatorname{adj}(\tilde{Q}(z)) \leq (n-1)k$ (this follows from the fact that an $(n-1)$-minor of $\tilde{Q}(z)$ has degree lower or equal to $(n-1)k$), we see that

$$\begin{aligned} \deg_z(\tilde{P}(z) \cdot \operatorname{adj}(\tilde{Q}(z))) &\leq (n-1) + (n-1)k = nk - k + n - 1 = \\ &= nk - (k - (n-1)) \leq nk = \deg_z(\det(\tilde{Q}(z))), \end{aligned}$$

where the last inequality follows from the fact that $k = \deg_z(\tilde{Q}(z)) \geq n - 1$. Hence $\tilde{P}(z)\tilde{Q}^{-1}(z)$ is a *proper* rational matrix.

Take a realization $\Gamma = (F, G, H, J)$ of the transfer matrix $\tilde{P}(z)\tilde{Q}^{-1}(z)$ such that $\det(zI - F) = \det(\tilde{Q}(z))$. According to the result in Appendix B, such a realization always exists. We use this dynamical system Γ as a compensator for Σ. Then the closed-loop system $\Sigma_{cl} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ is given by formula (2.34), and we have

$$\begin{aligned} \det(zI - \hat{A}) &= \det \begin{pmatrix} zI - (A - BJ) & BH \\ -G & zI - F \end{pmatrix} = \\ &= \det(zI - F) \cdot \det(zI - (A - BJ) + BH(zI - F)^{-1}G) = \\ &= \det(zI - F) \cdot \det(zI - A + B(J + H(zI - F)^{-1}G)) = \\ &= \det(zI - F) \cdot \det(zI - A + B\tilde{P}(z)\tilde{Q}^{-1}(z)) = \\ &= \frac{\det(zI - F) \cdot \det((zI - A)\tilde{Q}(z) + B\tilde{P}(z))}{\det(\tilde{Q}(z))}, \end{aligned}$$

where we took the Schur complement to make the first step. Since by construction $\det(zI-F) = \det(\tilde{Q}(z))$, and according to (2.37): $(zI-A)\tilde{Q}(z) + B\tilde{P}(z) = \varphi(z)\cdot I$, we conclude that

$$\det(zI-\hat{A}) = \det(\varphi(z)\cdot I) = (\varphi(z))^n \in \mathcal{D},$$

because $\varphi(z) \in \mathcal{D}$ and $\mathcal{D}$ is multiplicative. So the closed-loop system is internally stable.

"$\Rightarrow$" Assume that $\Sigma = (A, B, I, 0)$ is internally stabilizable by dynamic state feedback. Then there exists a compensator $\Gamma = (F, G, H, J)$ over $\mathcal{R}$ such that the closed-loop system determined by (2.34) is internally stable. So

$$\det(zI-\hat{A}) = \det\begin{pmatrix} zI-(A-BJ) & BH \\ -G & zI-F \end{pmatrix} \in \mathcal{D}.$$

Since $(zI-\hat{A})^{-1} = \frac{\operatorname{adj}(zI-\hat{A})}{\det(zI-\hat{A})}$, and $\det(zI-\hat{A}) \in \mathcal{D}$, it is easily seen that $(zI-\hat{A})$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. So there exist matrices $\hat{Q}(z)$, $\hat{R}(z)$, $\hat{P}(z)$ and $\hat{T}(z)$ over $\mathcal{R}_{\mathcal{D}}(z)$ such that

$$\begin{pmatrix} zI-(A-BJ) & BH \\ -G & zI-F \end{pmatrix} \cdot \begin{pmatrix} \hat{Q}(z) & \hat{R}(z) \\ \hat{P}(z) & \hat{T}(z) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

The equality for the (1,1)-block yields

$$(zI-(A-BJ))\hat{Q}(z) + BH\hat{P}(z) = I.$$

This formula can be rewritten as

$$(zI-A)\hat{Q}(z) + B(J\hat{Q}(z) + H\hat{P}(z)) = I.$$

Define $Q(z) := \hat{Q}(z)$ and $P(z) := J\hat{Q}(z) + H\hat{P}(z)$. Since both $\hat{Q}(z)$ and $\hat{P}(z)$ are matrices over $\mathcal{R}_{\mathcal{D}}(z)$ and J and H are matrices over $\mathcal{R}$, it is clear that also both $Q(z)$ and $P(z)$ are matrices over $\mathcal{R}_{\mathcal{D}}(z)$. Moreover, these matrices satisfy

$$(zI-A)Q(z) + BP(z) = I.$$

This completes the proof. ∎

The importance of Theorem 2.8.2 is clear. It gives a necessary and sufficient condition for a system to be stabilizable by dynamic state feedback. Therefore it is much more useful than the results on pole placement and static state feedback in Section 2.6. The conditions in that section were often too strong for stabilizability because they were sufficient but not necessary. Moreover, the proof of Theorem 2.8.2 is completely constructive. When a right-inverse of $(zI-A|B)$ over $\mathcal{R}_{\mathcal{D}}(z)$ is obtained, the proof gives a recipe for the construction of a stabilizing feedback compensator.

Theorem 2.8.2 can be seen as a rather straightforward generalization of the ordinary result on the stabilizability of systems over $\mathbb{R}$ by static state feedback.

Let $\mathcal{R} = \mathbb{R}$ and $\mathbb{C}_g \subset \mathbb{C}$ be a stability domain such that $\mathbb{C}_g \cap \mathbb{R} \neq \varnothing$. According to Example 2.5.4, the set

$$\mathcal{D} := \{p(z) \in \mathbb{R}[z] \mid p(z) \text{ is monic and } ([p(\alpha) = 0] \Longrightarrow [\alpha \in \mathbb{C}_g])\}$$

is a Hurwitz set, satisfying all conditions of Definition 2.5.2. By Theorem 2.8.2, a system $\Sigma = (A, B, I, 0)$ over $\mathbb{R}$ is stabilizable with respect to $\mathcal{D}$ if and only if $(zI - A \mid B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. Using the local-global theorem (see Appendix A.3, Theorem A.3.4), it is possible to rewrite this condition as a pointwise invertibility condition on the matrix $(\alpha I - A \mid B)$ for all $\alpha \in \mathbb{C}_g$. But a constant matrix over $\mathbb{C}$ is right-invertible if and only if it has full row rank. So a system $\Sigma = (A, B, I, 0)$ with $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ is stabilizable with respect to $\mathcal{D}$ by a dynamic state feedback compensator if and only if

$$\forall \alpha \in \mathbb{C}_g : \ \operatorname{rank}(\alpha I - A \mid B) = n. \tag{2.38}$$

But this is exactly the condition of the Hautus test for stabilizability of a system over $\mathbb{R}$ with respect to a stability domain $\mathbb{C}_g$ using *static* state feedback. Therefore Theorem 2.8.2 can be seen as a generalization of the Hautus test to the systems over rings case. There is only one important difference between these two conditions. For systems over fields *static* state feedback suffices to achieve stability, whereas in the systems over rings case we need *dynamic* state feedback to establish this goal.

Remark 2.8.3 The condition for stabilizability in Theorem 2.8.2 has one immediate consequence. Suppose that a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$ is reachable. Then $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$. Since by definition $1 \in \mathcal{D}$, this immediately implies that for each Hurwitz set $\mathcal{D}$, the matrix $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. So a reachable system is stabilizable with respect to any arbitrary Hurwitz set by dynamic state feedback.

Unfortunately, the right-invertibility condition of Theorem 2.8.2 is not always easy to check. Sometimes it is useful to state this condition in a slightly different way using polynomial ideals.

Definition 2.8.4 Let $\mathcal{R}$ be an integral domain, and $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Denote by $\alpha_0(z), \ldots, \alpha_N(z)$ all $n \times n$ minors of the matrix $(zI - A|B)$. Then $\mathcal{J}$ is defined as the ideal in $\mathcal{R}[z]$ generated by all $n \times n$ minors of $(zI - A|B)$:

$$\mathcal{J} := \langle \alpha_0(z), \ldots, \alpha_N(z) \rangle. \tag{2.39}$$

Proposition 2.8.5 *Let $\mathcal{R}$ be an integral domain, and $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Let $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$, and $\mathcal{J}$ the ideal defined in (2.39). Then*

$$(zI - A|B) \text{ is right-invertible over } \mathcal{R}_{\mathcal{D}}(z),$$

$$\Longleftrightarrow$$

$$\text{The ideal } \mathcal{J} \text{ contains an element of } \mathcal{D}, \text{ i.e. } \mathcal{D} \cap \mathcal{J} \neq \varnothing.$$

Proof

"$\Rightarrow$" Assume that $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$, and let $Q(z)$ be such a right-inverse, so

$$(zI - A|B) \cdot Q(z) = I. \tag{2.40}$$

We now use the well-known Binet-Cauchy formula (see for example [29, p. 9]), which expresses the determinant of a square $n \times n$ matrix $M = K \cdot L$ in terms of the minors of the $n \times \ell$ matrix K and the $\ell \times n$ matrix L ($\ell > n$):

$$\det \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix} = \sum_{1 \le i_1 < i_2 < \cdots < i_n \le \ell} \det \begin{pmatrix} k_{1i_1} & \cdots & k_{1i_n} \\ \vdots & & \vdots \\ k_{ni_1} & \cdots & k_{ni_n} \end{pmatrix} \cdot \det \begin{pmatrix} l_{i_1 1} & \cdots & l_{i_1 n} \\ \vdots & & \vdots \\ l_{i_n 1} & \cdots & l_{i_n n} \end{pmatrix}. \tag{2.41}$$

Application of the Binet-Cauchy formula to (2.40) yields

$$1 = \sum_{1 \le i_1 < i_2 < \cdots < i_n \le n+m} \det \begin{pmatrix} p_{1i_1} & \cdots & p_{1i_n} \\ \vdots & & \vdots \\ p_{ni_1} & \cdots & p_{ni_n} \end{pmatrix} \cdot \det \begin{pmatrix} q_{i_1 1} & \cdots & q_{i_1 n} \\ \vdots & & \vdots \\ q_{i_n 1} & \cdots & q_{i_n n} \end{pmatrix},$$

where $P(z)$ denotes the matrix $(zI - A|B)$. Now recall Definition 2.8.4 and let $\alpha_0(z), \ldots, \alpha_N(z)$ be all $n \times n$ minors of $(zI - A|B)$. Then we conclude from the last formula that there exist rational functions $q_0(z), \ldots, q_N(z)$ in $\mathcal{R}_{\mathcal{D}}(z)$ such that

$$\sum_{i=0}^{N} \alpha_i(z) \cdot q_i(z) = 1. \tag{2.42}$$

For $i = 0, 1, \ldots, N$ we know that $q_i(z) \in \mathcal{R}_{\mathcal{D}}(z)$, so $q_i(z)$ is of the form $q_i(z) = \frac{n_i(z)}{d_i(z)}$, where $n_i(z) \in \mathcal{R}[z]$ and $d_i(z) \in \mathcal{D}$. Let $d(z)$ denote the least common multiple of all the $d_i(z)$ ($i = 0, 1, \ldots, N$). Then $d(z) \in \mathcal{D}$ because $\mathcal{D}$ is a multiplicative set. Multiplication of (2.42) with $d(z)$ gives

$$d(z) = \sum_{i=0}^{N} \alpha_i(z) \cdot \tilde{n}_i(z),$$

where $\tilde{n}_i(z) = q_i(z) \cdot d(z) = \frac{n_i(z)}{d_i(z)} \cdot d(z) \in \mathcal{R}[z]$. So $d(z)$ belongs to the ideal $\mathcal{J}$ in $\mathcal{R}[z]$ generated by all $n \times n$ minors $\alpha_0(z), \ldots, \alpha_N(z)$ of $(zI - A|B)$. On the other hand we already knew that $d(z) \in \mathcal{D}$. So indeed $\mathcal{D} \cap \mathcal{J}$ is non-empty.

"$\Leftarrow$" Let $\alpha_0(z), \ldots, \alpha_N(z)$ denote all $n \times n$ minors of $(zI - A|B)$. For all $i = 0, 1, \ldots, N$, $\alpha_i(z)$ is the determinant of an $n \times n$ matrix $K_i(z)$ which consists of n columns of $(zI - A|B)$. For this square matrix we know that

$$K_i(z) \cdot \operatorname{adj}(K_i(z)) = \det(K_i(z)) \cdot I = \alpha_i(z) \cdot I.$$

Since $K_i(z)$ consists of n columns of $(zI - A|B)$, it is possible to extend $\operatorname{adj}(K_i(z))$ with zero rows on the right places to an $(n + m) \times n$ matrix $\tilde{K}_i(z)$ in such a way that

$$(zI - A|B) \cdot \tilde{K}_i(z) = \alpha_i(z) \cdot I.$$

It is obvious that the entries of the matrix $\tilde{K}_i(z)$ still belong to $\mathcal{R}[z]$.

Now, suppose that the ideal $\mathcal{J}$ generated by $\alpha_0(z), \ldots, \alpha_N(z)$ contains an element $d(z) \in \mathcal{D}$. Then there exist elements $g_0(z), g_1(z), \ldots, g_N(z)$ in $\mathcal{R}[z]$ such that

$$\sum_{i=0}^{N} \alpha_i(z) \cdot g_i(z) = d(z).$$

Define $Q(z) := \sum_{i=0}^{N} \frac{g_i(z)}{d(z)} \cdot \tilde{K}_i(z)$. Since for all $i \in \{0, 1, \ldots, N\}$ the entries of $\tilde{K}_i(z)$ and also $g_i(z)$ belong to $\mathcal{R}[z]$, and because $d(z) \in \mathcal{D}$, $Q(z)$ is a matrix over $\mathcal{R}_{\mathcal{D}}(z)$. Moreover:

$$\begin{aligned}(zI - A|B) \cdot Q(z) &= (zI - A|B) \cdot \sum_{i=0}^{N} \frac{g_i(z)}{d(z)} \cdot \tilde{K}_i(z) = \\ &= \frac{1}{d(z)} \cdot \sum_{i=0}^{N} (zI - A|B) \cdot \tilde{K}_i(z) \cdot g_i(z) = \\ &= \frac{1}{d(z)} \cdot \sum_{i=0}^{N} \alpha_i(z) g_i(z) \cdot I = \frac{1}{d(z)} d(z) \cdot I = I.\end{aligned}$$

So $Q(z)$ is a right-inverse of $(zI - A|B)$ over $\mathcal{R}_{\mathcal{D}}(z)$. ∎

The reformulation of the stabilizability condition using ideals in $\mathcal{R}[z]$ plays an important role throughout the rest of the thesis. It is an algebraically attractive restatement of the stabilizability problem: the ideal $\mathcal{J}$, completely determined by the system under consideration, must have a non-empty intersection with the Hurwitz set $\mathcal{D}$ that defines stability. In Chapter 3 this condition is used to specialize the results on stabilizability to the case of time-delay systems. But especially in Chapter 5, (polynomial) ideals play the leading role. There we shall see how system theoretic properties are translated into properties of (polynomial) ideals and how constructive methods from commutative algebra can be applied to verify these properties effectively. Proposition 2.8.5 can be considered as a little foretaste of the contents of that chapter.

2.9 Detectability

In the previous section we have seen how a system $\Sigma = (A, B, C, D)$ with $C = I$ and $D = 0$ can be stabilized by dynamic feedback. In this situation the output y is equal to the state x, so in fact the state is available for feedback to the input. This method is not applicable when we are dealing with general C and D matrices. In this case only the output y can be measured. Still we want to use the technique of dynamic feedback, developed in Section 2.8, to stabilize the system. For this purpose we have to recover information on the state x of the system from the data that are available: the input u and the output y. So for the application of the dynamic compensator of Section 2.8 in the case $C \neq I$ or $D \neq 0$, we first have to build an *observer* for the state x of the system.

In the theory of systems over $\mathbb{R}$, the same problem is encountered. Here a stable observer for the state x of a system $\Sigma = (A, B, C, D)$ is defined as a linear dynamic

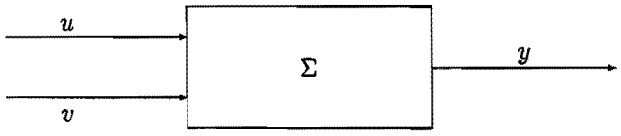


Figure 2.4: Dynamical system with disturbance input

system $\Omega = (F, G, H, J)$ which is stable, takes the input u and the output y of Σ as inputs, and produces an output $\hat{x}$, such that $\hat{x}(t) - x(t)$ tends to zero for $t \to +\infty$, irrespective of the initial conditions of the system and the observer. Conditions for the existence of such a compensator are well known (see e.g. [86, pp. 245-246]); the problem turns out to be dual to the problem of stabilization by static state feedback.

For systems over rings we have to take a somewhat different point of view. First of all, the state of a system does not play a prominent role, because formally a system is defined as a quadruple of matrices. Nevertheless a state variable can be introduced to interpret this quadruple as a dynamical system. In this way the state variable x becomes a rather formal object too, and it is difficult to incorporate initial conditions in this context. Moreover, in Example 2.1.3, where time-delay systems are modeled as systems over a ring, we have seen that the evolution variable $x(t)$ is not really the state of the system, and the initial state is in fact an initial trajectory of the evolution variable x. Therefore we do not want the initial condition to enter the theory explicitly.

These problems can be solved by replacing the concept of initial state by the addition of a "disturbance" input v as depicted in Figure 2.4. Given a system $\Sigma = (A, B, C, D)$, the disturbance input v enters only the dynamic equation and therefore it acts directly on the state x but only indirectly on the output y. So the configuration of Figure 2.4 is characterized by the set of equations:

$$\begin{cases} \Delta x &= Ax + Bu + v, \\ y &= Cx + Du. \end{cases} \tag{2.43}$$

An observer for this system has to determine the influence of the input v on the original system $\Sigma = (A, B, C, D)$. Irrespective of the disturbance input v, the output of the observer has to be an estimate for the state x of the system Σ. However, for systems over rings the notion of convergence is not defined. Instead we have to use the more abstract idea of Hurwitz sets again. In this framework the definition of a stable observer becomes:

Definition 2.9.1 Let $\mathcal{R}$ be an integral domain and $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Let $\Sigma = (A, B, C, D)$ be a system over $\mathcal{R}$ with input u and output y. A *stable observer* Ω (stable with respect to $\mathcal{D}$) is a system $\Omega = (F, (G_1|G_2), H, (J_1|J_2))$ over $\mathcal{R}$, which takes $\begin{pmatrix} u \\ y \end{pmatrix}$ as input, produces $\hat{x}$ as output, and satisfies the following conditions:

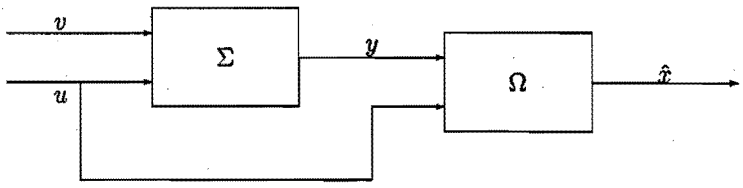


Figure 2.5: Interconnection of system and observer

(i) $\det(zI - F) \in \mathcal{D}$, i.e. Ω is internally stable with respect to $\mathcal{D}$.

(ii) The configuration of Figure 2.5, characterized by the set of equations

$$\begin{cases} \Delta x &= Ax + Bu + v, \\ y &= Cx + Du, \\ \Delta w &= Fw + (G_1|G_2)\begin{pmatrix} u \\ y \end{pmatrix}, \\ \hat{x} &= Gw + (J_1|J_2)\begin{pmatrix} u \\ y \end{pmatrix}, \end{cases} \tag{2.44}$$

gives formally rise to an equation

$$\hat{x} - x = K(z)v, \tag{2.45}$$

in which the transfer matrix $K(z)$ is stable with respect to $\mathcal{D}$. So, $\hat{x} - x$ only depends on the disturbance input v, and all entries of $K(z)$ are elements of $\mathcal{R}_{\mathcal{D}}(z)$.

Definition 2.9.1 can be seen as a generalization of the concept of stable observers to the case of systems over rings. It was suggested by Hautus and Sontag in [44]. With help of this definition, the notion of detectability can be introduced very easily.

Definition 2.9.2 A system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$ is called *detectable* (with respect to the Hurwitz set $\mathcal{D}$) if there exists a $\mathcal{D}$-stable observer for Σ.

The next theorem is also taken from [44]. It gives a necessary and sufficient condition for a system over an integral domain to be detectable.

Theorem 2.9.3 *Let $\mathcal{R}$ be an integral domain and $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Consider a system $\Sigma = (A, B, C, D)$ over $\mathcal{R}$. Then*

$$\Sigma \text{ is detectable with respect to } \mathcal{D}$$

$$\Longleftrightarrow$$

$$\begin{pmatrix} zI - A \\ C \end{pmatrix} \text{ is left-invertible over } \mathcal{R}_{\mathcal{D}}(z) \tag{2.46}$$

Moreover, if (2.46) holds, then there exists a strictly proper $\mathcal{D}$-stable observer $\Omega = (F, G, H, 0)$ for Σ.

Proof

"$\Leftarrow$" Suppose that $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. Then there exist matrices $M(z)$ and $N(z)$ over $\mathcal{R}_{\mathcal{D}}(z)$ such that

$$M(z)(zI - A) + N(z)C = I.$$

Multiplying this equation with the least common multiple $\varphi(z)$ of the denominators of the entries of $M(z)$ and $N(z)$ we obtain

$$\hat{M}(z)(zI - A) + \hat{N}(z)C = \varphi(z) \cdot I,$$

where $\hat{M}(z)$ and $\hat{N}(z)$ are matrices over $\mathcal{R}[z]$ and $\varphi(z) \in \mathcal{D}$. Without loss of generality we assume that $\deg_z(\varphi(z)) \geq n$; otherwise we multiply the last equation with an element of $\mathcal{D}$ of sufficiently high degree. The existence of such a stable polynomial is guaranteed by condition *(iv)* of Definition 2.5.2.

Since A is an $n \times n$ matrix and C is $p \times n$, it follows from Lemma 2.8.1 (or rather from a transposed version of this lemma), that there exist polynomial matrices $\tilde{M}(z)$ and $\tilde{N}(z)$ such that $\deg_z(\tilde{N}(z)) \leq n - 1$, and still

$$\tilde{M}(z)(zI - A) + \tilde{N}(z)C = \varphi(z) \cdot I.$$

$\varphi(z)$ is monic and of degree greater than or equal to n, and because $\tilde{N}(z)C$ is of degree less than n, we conclude that $\tilde{M}(z)$ has degree $\deg_z(\tilde{M}(z)) = \deg_z(\varphi(z)) - 1$. Define

$$(\bar{M}(z)|\bar{N}(z)) := \frac{1}{\varphi(z)} \cdot (\tilde{M}(z)|\tilde{N}(z)) = (\tilde{M}(z)|\tilde{N}(z))(\varphi(z) \cdot I)^{-1}. \tag{2.47}$$

It is obvious that both $\bar{M}(z)$ and $\bar{N}(z)$ are strictly proper matrices over $\mathcal{R}_{\mathcal{D}}(z)$ and still

$$\bar{M}(z)(zI - A) + \bar{N}(z)C = I. \tag{2.48}$$

Moreover, with the second description of $(\bar{M}(z)|\bar{N}(z))$ in (2.47), it follows from appendix B that there exists a realization $(F, (G_1|G_2), H, 0)$ of the transfer matrix $(\bar{M}(z)|\bar{N}(z))$, satisfying $\det(zI - F) = \det(\varphi(z) \cdot I) \in \mathcal{D}$.

Next, define the observer Ω as

$$\Omega \left\{ \begin{array}{rcl} \Delta w & = & Fw + (G_1 B - G_2 D|G_2) \begin{pmatrix} u \\ y \end{pmatrix}, \\ \hat{x} & = & Hw. \end{array} \right. \tag{2.49}$$

Clearly, Ω is internally stable, so we only have to check that from the systems equations (2.43) together with the observer equations (2.49) a $\mathcal{D}$-stable transfer matrix $K(z)$ from v to $\hat{x} - x$ is obtained.

Substitution of the output equation of (2.43) into (2.49) gives

$$\begin{aligned}\Delta w &= Fw + (G_1B - G_2D)u + G_2y = \\ &= Fw + G_1Bu - G_2Du + G_2Cx + G_2Du = \\ &= Fw + G_1Bu + G_2Cx.\end{aligned}$$

Combining this equation with the output equation $\hat{x} = Hw$ of the observer, we obtain the transfer matrix from u and x to $\hat{x}$:

$$\hat{x} = H(zI - F)^{-1}G_1Bu + H(zI - F)^{-1}G_2Cx = \bar{M}(z)Bu + \bar{N}(z)Cx,$$

where we used the fact that $(F, (G_1|G_2), H, 0)$ is a realization of $(\bar{M}(z)|\bar{N}(z))$. The transfer matrix from u and v to x is easily derived from the first equation of (2.43):

$$x = (zI - A)^{-1}Bu + (zI - A)^{-1}v.$$

Substitution of this formula for x in the one obtained for $\hat{x}$ yields

$$\begin{aligned}\hat{x} &= \bar{M}(z)Bu + \bar{N}(z)C(zI - A)^{-1}Bu + \bar{N}(z)C(zI - A)^{-1}v = \\ &= (\bar{M}(z)(zI - A) + \bar{N}(z)C)(zI - A)^{-1}Bu + \bar{N}(z)C(zI - A)^{-1}v = \\ &= (zI - A)^{-1}Bu + \bar{N}(z)C(zI - A)^{-1}v,\end{aligned}$$

where in the last step equation (2.48) is used. Subtracting x from $\hat{x}$ gives

$$\begin{aligned}\hat{x} - x &= \bar{N}(z)C(zI - A)^{-1}v - (zI - A)^{-1}v = (\bar{N}(z)C - I)(zI - A)^{-1}v = \\ &= (\bar{N}(z)C - \bar{N}(z)C - \bar{M}(z)(zI - A))(zI - A)^{-1}v = -\bar{M}(z)v.\end{aligned}$$

So $\hat{x} - x$ only depends on v, and the transfer matrix $K(z)$ from v to $\hat{x} - x$ is given by $K(z) = -\bar{M}(z)$. Since $\bar{M}(z)$ is a matrix over $\mathcal{R}_{\mathcal{D}}(z)$, this proves that Ω is a stable observer for Σ, so Σ is detectable. Moreover, because the observer Ω defined in (2.49) is strictly proper, we have also proven the assertion that a strictly proper observer can do the job.

"$\Rightarrow$" Suppose that Σ is detectable. Then there exists a $\mathcal{D}$-stable observer Ω for Σ. Let $(L(z)|N(z))$ denote the transfer matrix of Ω from $\begin{pmatrix} u \\ y \end{pmatrix}$ to $\hat{x}$. So $L(z)$ and $N(z)$ are transfer matrices over $\mathcal{R}_{\mathcal{D}}(z)$ and $\hat{x} = L(z)u + N(z)y$. Substituting the output equation $y = Cx + Du$ of the system Σ in this equation we obtain

$$\hat{x} = (L(z) + N(z)D)u + N(z)Cx.$$

Recall that the transfer matrix from u and v to x is given by

$$x = (zI - A)^{-1}Bu + (zI - A)^{-1}v.$$

Subtraction of the equations for $\hat{x}$ and x yields

$$\begin{aligned}\hat{x} - x &= (L(z) + N(z)D)u + (N(z)C - I)x = \\ &= (L(z) + N(z)D)u + (N(z)C - I)((zI - A)^{-1}Bu + (zI - A)^{-1}v) = \\ &= (L(z) + N(z)(D + C(zI - A)^{-1}B) - (zI - A)^{-1}B)u + \\ &\qquad + ((N(z)C - I)(zI - A)^{-1})v.\end{aligned}$$

Now recall that Ω is a stable observer for Σ. So the interconnection of Figure 2.5 leads to an equation of the form $\hat{x} - x = K(z)v$, with $K(z)$ a matrix over $\mathcal{R}_{\mathcal{D}}(z)$. Therefore we conclude that $(L(z) + N(z)(D + C(zI - A)^{-1}B) - (zI - A)^{-1}B) = 0$ and that $M(z) := (N(z)C - I)(zI - A)^{-1}$ is a stable matrix. By definition we have $M(z)(zI - A) = N(z)C - I$. So

$$-M(z)(zI - A) + N(z)C = I,$$

and $(-M(z)|N(z))$ is a left-inverse of $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ over $\mathcal{R}_{\mathcal{D}}(z)$.

This completes the proof. ∎

Again, the proof of Theorem 2.9.3 is completely constructive. When a left-inverse of $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is known, a stable observer can be obtained by carrying out the construction method described in the proof.

From condition (2.46) in Theorem 2.9.3 it is clear that for the detectability of a system $\Sigma = (A, B, C, D)$ only the matrices A and C are important. The input matrix B does not play any role. This is similar to the condition for stabilizability by dynamic state feedback in Theorem 2.8.2, which is completely determined by the matrices A and B. Moreover, we see that the conditions for detectability and stabilizability are dual. This is stated more formally in the next

Corollary 2.9.4 *Let $\Sigma = (A, B, C, D)$ be a system over an integral domain $\mathcal{R}$. Define $\Sigma^T := (A^T, C^T, B^T, D^T)$. Then*

(i) Σ is stabilizable by dynamic state feedback $\iff$ Σ^T is detectable,

(ii) Σ is detectable $\iff$ Σ^T is stabilizable by dynamic state feedback. ∎

Hence, analogous to systems over fields, stabilizability and detectability are dual concepts, also in the ring case. There is only one difference. For a system over a field a stabilizing feedback and a stable observer can be obtained with static state feedback and pole placement techniques. For systems over rings there are some dynamics involved. So we conclude that the duality between reachability and observability that was lost in the case of systems over rings, can be restored for stabilizability and detectability when dynamic compensators are used to define and to achieve these properties, instead of static compensators.

2.10 Stabilizability by dynamic output feedback

In the two previous sections we have developed all tools that are required to tackle the problem of stabilization by dynamic output feedback. According to the separation principle, this problem can be split into two parts. Now that both parts have been solved, we want to combine the construction methods for a stabilizing state feedback compensator and a stable observer in order to find a stabilizing output feedback compensator. However, we start with a rather naive (but effective) method to stabilize a system.

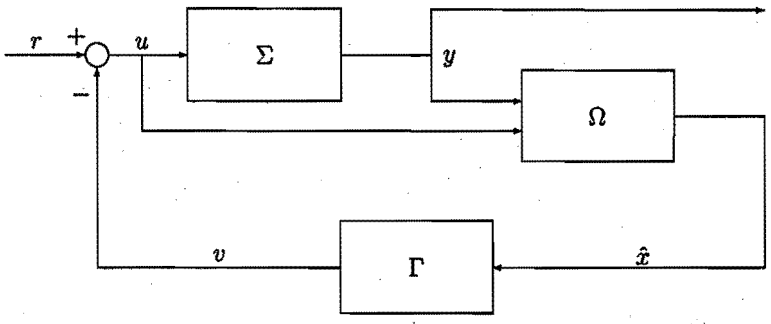


Figure 2.6: Closed-loop system with observer and feedback compensator

Consider a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$, and suppose that Σ is both stabilizable and detectable with respect to a Hurwitz set $\mathcal{D}$. Construct a stabilizing dynamic state feedback compensator Γ according to the method of Theorem 2.8.2, and a stable observer Ω with the method of Theorem 2.9.3. The idea is now to estimate the state x of the system Σ with help of the observer Ω, and to use this estimated state $\hat{x}$ as the input to the compensator Γ. So we are interested in the configuration of Figure 2.6, and show that this system with input r and output y is internally stable.

Since Σ is stabilizable by dynamic state feedback, the matrix $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$, and according to Theorem 2.8.2 and its proof, there exists a polynomial $\varphi(z) \in \mathcal{D}$ and matrices $Q(z)$ and $P(z)$ over $\mathcal{R}[z]$ such that

(i) $(zI - A)Q(z) + BP(z) = \varphi(z) \cdot I$,

(ii) $P(z)Q^{-1}(z)$ exists as a rational matrix and is proper,

(iii) $P(z)Q^{-1}(z)$ has a realization $\Gamma = (F, G, H, J)$ such that $\det(zI - F) = \det(Q(z))$.

Since Σ is detectable, the matrix $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$, and it follows from Theorem 2.9.3 and its proof, that there exist rational strictly proper stable matrices $M(z)$ and $N(z)$ over $\mathcal{R}_D(z)$ such that

$$M(z)(zI - A) + N(z)C = I.$$

Moreover, $(M(z)|N(z))$ has a realization $(R, (T_1|T_2), V, 0)$ with $\det(zI - R) \in \mathcal{D}$.

Let Ω denote the stable observer $\Omega = (R, (T_1B - T_2D|T_2), V, 0)$, and Γ the stabilizing state feedback compensator $\Gamma = (F, G, H, J)$. Then the configuration of

Figure 2.6 is described by the following set of equations

$$\begin{cases} \Delta x &= Ax + Bu, \\ y &= Cx + Du, \\ \Delta w &= Rw + (T_1B - T_2D)u + T_2y, \\ \hat{x} &= Vw, \\ \Delta z &= Fz + G\hat{x}, \\ v &= Hz + J\hat{x}, \\ u &= r - v. \end{cases}$$

We now show that the resulting linear dynamical system with input r and output y is internally stable.

First substitute the output equations for y, $\hat{x}$ and v, and the feedback equation $u = r - v$ into the dynamic equations for Δx, Δw and Δz. In this way we obtain:

$$\begin{aligned} \Delta x &= Ax + Bu = Ax + Br - Bv = Ax - BJVw - BHz + Br, \\ \Delta w &= Rw + (T_1B - T_2D)u + T_2y = Rw + T_2Cx + T_1Br - T_1Bv = \\ &= T_2Cx + (R - T_1BJV)w - T_1BHz + T_1Br, \\ \Delta z &= Fz + G\hat{x} = GVw + Fz. \end{aligned}$$

And the output y equals

$$y = Cx + Du = Cx + Dr - Dv = Cx - DJVw - DHz + Dr.$$

So the closed-loop system of Figure 2.6 can be written as

$$\begin{cases} \Delta \begin{pmatrix} x \\ w \\ z \end{pmatrix} &= \begin{pmatrix} A & -BJV & -BH \\ T_2C & R - T_1BJV & -T_1BH \\ 0 & GV & F \end{pmatrix} \begin{pmatrix} x \\ w \\ z \end{pmatrix} + \begin{pmatrix} B \\ T_1B \\ 0 \end{pmatrix} r, \\ y &= (C \ -DJV \ -DH) \begin{pmatrix} x \\ w \\ z \end{pmatrix} + Dr. \end{cases} \tag{2.50}$$

Defining

$$\hat{A} := \begin{pmatrix} A & -BJV & -BH \\ T_2C & R - T_1BJV & -T_1BH \\ 0 & GV & F \end{pmatrix},$$

this system is internally stable if and only if

$$\det(zI - \hat{A}) = \det \begin{pmatrix} zI - A & BJV & BH \\ -T_2C & zI - R + T_1BJV & T_1BH \\ 0 & -GV & zI - F \end{pmatrix} \in \mathcal{D}. \tag{2.51}$$

Adding the first block row, multiplied by $-T_1$, to the second row gives

$$\det(zI - \hat{A}) = \det \begin{pmatrix} zI - A & BJV & BH \\ -T_1(zI - A) - T_2C & zI - R & 0 \\ 0 & -GV & zI - F \end{pmatrix}.$$

Multiplying the third block column with $(zI - F)^{-1}GV$ and adding this to the second block column yields

$$\begin{aligned}\det(zI - \hat{A}) &= \det\begin{pmatrix} zI - A & BJV + BH(zI - F)^{-1}GV & BH \\ -T_1(zI - A) - T_2C & zI - R & 0 \\ 0 & 0 & zI - F \end{pmatrix} \\ &= \det(zI - F) \cdot \det\begin{pmatrix} zI - A & B(J + H(zI - F)^{-1}G)V \\ -T_1(zI - A) - T_2C & zI - R \end{pmatrix} \\ &= \det(zI - F) \cdot \det\begin{pmatrix} zI - A & BP(z)Q^{-1}(z)V \\ -T_1(zI - A) - T_2C & zI - R \end{pmatrix},\end{aligned}$$

where we used the fact that $\Gamma = (F, G, H, J)$ is a realization of the transfer matrix $P(z)Q^{-1}(z)$. Finally, taking the Schur complement and recalling that the system $(R, (T_1|T_2), V, 0)$ is a realization of $(M(z)|N(z))$, we arrive at

$$\begin{aligned}\det(zI - \hat{A}) &= \det(zI - F) \cdot \det(zI - R) \cdot \\ &\quad \cdot \det(zI - A - BP(z)Q^{-1}(z)V(zI - R)^{-1}(-T_1(zI - A) - T_2C)) = \\ &= \det(zI - F) \cdot \det(zI - R) \cdot \\ &\quad \cdot \det(zI - A + BP(z)Q^{-1}(z)(M(z)|N(z)) \begin{pmatrix} zI - A \\ C \end{pmatrix}).\end{aligned}$$

Now, $M(z)$ and $N(z)$ are constructed in such a way that $M(z)(zI - A) + N(z)C = I$. Moreover, $Q(z)$ and $P(z)$ satisfy the conditions *(i)* to *(iii)*. Thus we have

$$\begin{aligned}\det(zI - \hat{A}) &= \det(zI - F) \cdot \det(zI - R) \cdot \det(zI - A + BPQ^{-1}(z)) = \\ &= \frac{\det(zI - F) \cdot \det(zI - R) \cdot \det((zI - A)Q(z) + BP(z))}{\det(Q(z))} = \\ &= \det(zI - R) \cdot \det(\varphi(z) \cdot I) = \det(zI - R) \cdot \varphi^n(z). \qquad (2.52)\end{aligned}$$

Since both $\varphi(z) \in \mathcal{D}$ and $\det(zI - R) \in \mathcal{D}$, and since $\mathcal{D}$ is multiplicative, it follows that $\det(zI - \hat{A}) \in \mathcal{D}$. So the closed-loop system of Figure 2.6 is internally stable.

We conclude that the configuration of Figure 2.6 gives us a method to stabilize a system $\Sigma = (A, B, C, D)$ under the condition that Σ is both stabilizable by dynamic state feedback and detectable, using only the input and output of Σ. It is shown how a dynamic feedback compensator Γ and an observer Ω can be combined to achieve internal stability. But the closed-loop system of Figure 2.6 is not really an output feedback system, because the observer Ω requires both the input and the output of the system Σ as inputs.

With a slight modification of the configuration of Figure 2.6 however, we can make it into a dynamic output feedback. Instead of taking u (the input to the system Σ) as one of the inputs to the observer Ω, we only feed the output v of the compensator Γ, multiplied by -1, back to the observer Ω as depicted in Figure 2.7. In this way the interconnection of Ω and Γ (i.e. the dashed box in Figure 2.7) with input y and output v, can be considered as an output feedback compensator.

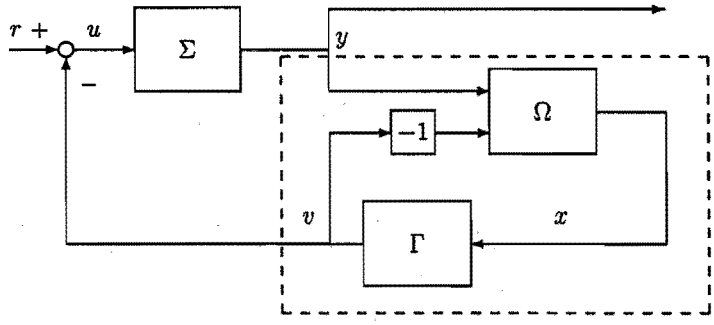


Figure 2.7: Closed-loop system with output feedback

The dynamic equations (2.50) are only slightly changed by this modification. Only one of the inputs to the observer Ω is altered: in the equation for Δw the input u has to be replaced by $-v$. In this way we obtain

$$\begin{aligned}\Delta w &= Rw - (T_1B - T_2D)v + T_2y = \\ &= Rw - T_1Bv + T_2Dv + T_2Cx + T_2Du = \\ &= T_2Cx + (R - T_1BJV)w - T_1BHz + T_2Dr,\end{aligned}$$

and the closed-loop system of Figure 2.7 can be written as

$$\begin{cases} \Delta \begin{pmatrix} x \\ w \\ z \end{pmatrix} = \begin{pmatrix} A & -BJV & -BH \\ T_2C & R - T_1BJV & -T_1BH \\ 0 & GV & F \end{pmatrix} \begin{pmatrix} x \\ w \\ z \end{pmatrix} + \begin{pmatrix} B \\ T_2D \\ 0 \end{pmatrix} r, \\ y = (C \ -DJV \ -DH) \begin{pmatrix} x \\ w \\ z \end{pmatrix} + Dr. \end{cases} \tag{2.53}$$

In comparison with (2.50), only the input matrix is changed. The matrix $\hat{A}$ however, is still the same, and therefore also the system configuration of Figure 2.7 is internally stable.

The configuration in the dashed box, with input y and output v satisfies the equations

$$\begin{cases} \Delta w = Rw - (T_1B - T_2D)v + T_2y, \\ \hat{x} = Vw, \\ \Delta z = Fz + G\hat{x}, \\ v = Hz + J\hat{x}. \end{cases}$$

Substitution of the expressions for v and $\hat{x}$ in the formulae for Δw and Δz yields

$$\begin{aligned}\Delta w &= (R - T_1BJV + T_2DJV)w + (-T_1BH + T_2DH)z + T_2y, \\ \Delta z &= GVw + Fz.\end{aligned}$$

We conclude that the dashed box can be considered as a linear system with input y and output v, governed by the system equations

$$\begin{cases} \Delta \begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} R - T_1BJV + T_2DJV & -T_1BH + T_2DH \\ GV & F \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} + \begin{pmatrix} T_2 \\ 0 \end{pmatrix} y, \\ v = (JV|H) \begin{pmatrix} w \\ z \end{pmatrix}. \end{cases}$$

This linear system over the integral domain $\mathcal{R}$ is a dynamic output feedback stabilizing compensator for the system Σ. Note that this compensator is strictly proper (there is no direct feedthrough term), so the closed-loop system is well posed. Therefore we have proven the sufficiency of the following conditions for stabilizability by dynamic output feedback.

Theorem 2.10.1 *Consider a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$, and let $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Then*

Σ is internally stabilizable with respect to $\mathcal{D}$ by dynamic output feedback,

$\Longleftrightarrow$

(i) $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$, and

(ii) $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$.

Moreover, if (i) and (ii) hold, this dynamic output feedback can be chosen strictly proper.

Proof (of the necessity)

Assume that $\Sigma = (A, B, C, D)$ is stabilizable by dynamic output feedback. Then there exists a feedback compensator $\Gamma = (F, G, H, J)$ such that the closed-loop system of Figure 2.2 is well posed and internally stable. So $(I + DJ)$ is invertible as a matrix over $\mathcal{R}$, with inverse $E := (I + DJ)^{-1}$, and the matrix $\hat{A}$, given in (2.28),

$$\hat{A} = \begin{pmatrix} A - BJEC & -BH + BJEDH \\ GEC & F - GEDH \end{pmatrix},$$

is a stable matrix, i.e. $\det(zI - \hat{A}) \in \mathcal{D}$. Then $(zI - \hat{A})$ is invertible over $\mathcal{R}_{\mathcal{D}}(z)$ because

$$(zI - \hat{A})^{-1} = \frac{1}{\det(zI - \hat{A})} \cdot \operatorname{adj}(zI - \hat{A}).$$

Let $Q(z)$, $P(z)$, $R(z)$ and $T(z)$ be matrices over $\mathcal{R}_D(z)$ such that

$$\begin{pmatrix} Q(z) & R(z) \\ P(z) & T(z) \end{pmatrix}$$

is an inverse of $(zI - \hat{A})$ over $\mathcal{R}_{\mathcal{D}}(z)$. So

$$\begin{pmatrix} zI - A + BJEC & BH - BJEDH \\ -GEC & zI - F + GEDH \end{pmatrix} \begin{pmatrix} Q(z) & R(z) \\ P(z) & T(z) \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

The (1,1)-block of this equation yields:

$$(zI - A)Q(z) + BJECQ(z) + BHP(z) - BJEDHP(z) = I.$$

Defining $\hat{Q}(z) := Q(z)$ and $\hat{P}(z) := JECQ(z) + HP(z) - JEDHP(z)$, both $\hat{Q}(z)$ and $\hat{P}(z)$ are matrices over $\mathcal{R}_{\mathcal{D}}(z)$ and moreover

$$(zI - A)\hat{Q}(z) + B\hat{P}(z) = I.$$

Hence, $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$.

On the other hand, the matrix $\begin{pmatrix} Q(z) & R(z) \\ P(z) & T(z) \end{pmatrix}$ is of course also a left-inverse of $(zI - \hat{A})$. Multiplying the matrix $(zI - \hat{A})$ from the left with its inverse, and taking the (1,1)-block again, we obtain:

$$Q(z)(zI - A) + Q(z)BJEC - R(z)GEC = I.$$

Define $M(z) := Q(z)$ and $N(z) := Q(z)BJE - R(z)GE$. It is clear that $M(z)$ and $N(z)$ are both matrices over $\mathcal{R}_{\mathcal{D}}(z)$ satisfying

$$M(z)(zI - A) + N(z)C = I.$$

So $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. This completes the proof. ∎

Corollary 2.10.2 *Consider a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$. Assume that Σ is internally stabilizable with respect to a Hurwitz set $\mathcal{D}$ by dynamic output feedback. Then this system is also stabilizable by a strictly proper dynamic output feedback compensator.*

Proof

If $\Sigma = (A, B, C, D)$ is stabilizable by dynamic output feedback, it follows from Theorem 2.10.1 that $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$ and $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ is left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. Again applying Theorem 2.10.1 yields that a strictly proper compensator can do the job. ∎

Theorem 2.10.1 was first stated and proved for strictly proper systems by Khargonekar and Sontag in [58]. The idea of the proof for proper but not strictly proper systems is based on a proof in [24]. Unfortunately, the proof in this article is not completely correct because the existence of a strictly proper observer is not guaranteed. We have solved this problem using Lemma 2.8.1. This lemma is very important because it enables us to design strictly proper observers and output feedback compensators, and in this way it is possible to prove Theorem 2.10.1 in a constructive way.

The conditions in Theorem 2.10.1 for the existence of a dynamic output feedback compensator look very similar to the conditions known for linear systems over the field $\mathbb{R}$. The separation principle still works and in both cases stabilizability by state feedback and detectability are necessary and sufficient to guarantee stabilizability

by dynamic output feedback. The only difference is that the dynamics have to be incorporated one step earlier. Note that the conditions of detectability and stabilizability by dynamic state feedback are dual, so we only have to develop a test for one of these conditions. A test for the other property is then obtained by dualization. Stabilizability by dynamic output feedback is checked by successively carrying out both (dual) tests.

The invertibility conditions of Theorem 2.10.1 are not always easy to check. Of course this depends on the application one has in mind, so on the choice of the Hurwitz set $\mathcal{D}$. For time-delay systems however, this problem can be facilitated a lot. The next chapter is devoted to this subject. There it is shown that this invertibility condition can be replaced by a pointwise rank-condition, and in this way a generalization of the Hautus test to time-delay systems is obtained.

Chapter 3

Stabilizability of time-delay systems

After the introduction in Chapter 2 of a rather general framework for the investigation of stability for linear systems over integral domains, we now return to the class of time-delay systems with point delays. For this class of systems it is possible to characterize the structure of the Hurwitz sets describing stability more explicitly, because we are allowed to use the delay-character of the system. With this additional information the right-invertibility condition on the matrix $(zI - A|B)$ can be replaced by a pointwise rank condition, which facilitates the testing of stabilizability considerably . Moreover, the same test can be applied to verify the detectability of a system because this concept is dual to the problem of stabilizability by dynamic state feedback, as we have seen in Chapter 2.

Besides the derivation of the pointwise rank condition mentioned above, this chapter also has another goal. When we investigate the stabilizability condition more carefully, it seems not very restrictive. In fact, we prove that this condition is generically satisfied. To do so, we have to introduce a topology on the space of all time-delay systems with point delays. In this topology, the set of stabilizable time-delay systems contains an open and dense subset of the space of all time-delay systems. This indicates that the condition of stabilizability is very weak; it is satisfied for almost all time-delay systems.

3.1 Stability of systems with time-delays

Consider a linear system with k incommensurable time-delays $0 < \tau_1 < \cdots < \tau_k$. Let σ_i $(i = 1, \ldots, k)$ denote the delay operator corresponding to the time-delay τ_i,

$$\sigma_i x(t) = x(t - \tau_i) \qquad (i = 1, \ldots, k).$$

Then the time-delay system with k incommensurable time-delays can be written as

$$\begin{cases} \dot{x}(t) &= A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) &= C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t), \end{cases} \tag{3.1}$$

where A, B, C and D are polynomial matrices in the delay operators $\sigma_1, \ldots, \sigma_k$ of appropriate dimensions. After substitution of the indeterminates $s_1, \ldots, s_k$ for the

delay operators $\sigma_1, \ldots, \sigma_k$, we obtain a quadruple of matrices over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. This quadruple can be seen as a linear system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$, and together with the k-tuple $(\tau_1, \ldots, \tau_k)$ of time-delays it is still a complete description of the time-delay system (3.1).

To study stabilizability of time-delay systems in the framework of Chapter 2, we have to find out first how stability of time-delay systems is defined originally. The choice of our Hurwitz set has to correspond to this classical notion of stability of time-delay systems. Although this subject has been under discussion in Example 2.5.6, we treat it in more detail in this section.

Consider the autonomous delay system

$$\dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t). \tag{3.2}$$

Classically this system is called stable if for any arbitrary initial state trajectory the corresponding state x tends to zero when t tends to infinity. Sometimes however, one goes one step further, and demands that the state x tends to zero with a certain exponential decay rate α. Both conditions can be investigated with help of the characteristic equation of the polynomial matrix $A(s_1, \ldots, s_k)$, thanks to the following result.

Proposition 3.1.1 *Let $A(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]^{n \times n}$ and $(\tau_1, \ldots, \tau_k)$ be k-tuple of incommensurable time-delays. Let $\sigma_1, \ldots, \sigma_k$ denote the delay operators corresponding to $\tau_1, \ldots, \tau_k$. Then for any arbitrary initial state trajectory the corresponding solution of the differential-difference equation*

$$\dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t),$$

tends to zero with an exponential decay rate greater than α if and only if all roots of the characteristic equation

$$\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = 0,$$

are contained in the half-plane

$$\mathbb{C}_{-\alpha} = \{z \in \mathbb{C} \mid \operatorname{Re} z < -\alpha\}.$$

■

For a proof of this result we refer to [41, Chapter 7, Section 4]. The case $\alpha = 0$, i.e. the case of stability in the classical sense, can also be found in [3, p. 190].

The statement in Proposition 3.1.1 can be proved with aid of the Laplace transformation with symbol z. For example, consider the differential-difference equation (3.2) with initial condition

$$x(t) = \begin{cases} 0 & \text{for } t < 0, \\ x_0 & \text{for } t = 0. \end{cases}$$

Let $\hat{x}(z)$ denote the Laplace transform (with symbol z) of $x(t)$. Then we have

$$z\hat{x}(z) - x_0 = A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\hat{x}(z),$$

hence

$$\hat{x}(z) = (zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))^{-1}x_0. \tag{3.3}$$

From equation (3.3) it is obvious that the zeros of the characteristic equation $\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = 0$ are exactly the eigenvalues (or poles) of the original autonomous system. If these poles are contained in the half plane

$$\mathbf{C}_{-\alpha} = \{z \in \mathbf{C} | \operatorname{Re} z < -\alpha\},$$

the state of the system tends to zero with an exponential decay rate greater than α.

The poles of a system also play an important role for the performance of a system. Although a system is stable if all its poles are contained in $\mathbf{C}^-$, the presence of a pole with a small but negative real part, and a very large imaginary part, can have a very unwanted effect. It may lead to a solution which is highly oscillating but poorly damped. This is a very unsatisfactory behaviour, and therefore we should like to restrict the location of the poles to a smaller subset than the left half plane. The Hurwitz set framework allows us to do so: we are free to choose the set of favourable pole locations first, and adapt the definition of our Hurwitz set to this specific situation thereafter.

However, to be able to simplify the right-invertibility condition for stabilizability to a pointwise rank condition, the Hurwitz set $\mathcal{D}$ has to satisfy some regularity conditions that correspond to the set of favourable pole locations.

Definition 3.1.2 Consider the complex plane $\mathbf{C}$, decomposed disjointly into the regions $\mathbf{C}_g$ and $\mathbf{C}_b$ and the Jordan curve J; so J is the boundary of both $\mathbf{C}_g$ and $\mathbf{C}_b$. Suppose that this decomposition of $\mathbf{C}$ satisfies the following conditions:

(i) $\mathbf{C}_g$ and $\mathbf{C}_b$ are symmetric w.r.t. the real axis,

(ii) $\mathbf{C}_g$ is connected and $\mathbf{C}_b$ is simply connected,

(iii) $\mathbf{C}_g$ and $\mathbf{C}_b$ are both unbounded,

(iv) $\exists \alpha \in \mathbf{R} : \mathbf{C}_\alpha = \{z \in \mathbf{C} | \operatorname{Re} z < \alpha\} \subset \mathbf{C}_g$.

Then $\mathbf{C}_g$ is called a *stability domain.*

Note that some of the conditions in Definition 3.1.2 may be omitted without changing the definition. Condition *(iv)* implies that $\mathbf{C}_g$ is unbounded, and since $\mathbf{C}_b$ is simply connected, it follows that $\mathbf{C}_g$ is connected.

It is obvious that the definition of a stability domain formalizes the idea of a set of favourable pole locations. This is illustrated in the next example.

Example 3.1.3 Let $\beta < \alpha \leq 0$ and $\omega_2 > \omega_1 > 0$ be given, and consider the stability domain $\mathbf{C}_g$ depicted in Figure 3.1. So $\mathbf{C}_g$ is defined by

$$\begin{aligned} \mathbf{C}_g = \{z \in \mathbf{C} \mid & \operatorname{Re} z < \beta \text{ or} \\ & (\beta \leq \operatorname{Re} z < \alpha \text{ and } |\operatorname{Im} z| < \tfrac{\omega_2 - \omega_1}{\beta - \alpha}(\operatorname{Re} z - \alpha) + \omega_1)\}, \end{aligned} \tag{3.4}$$

$\mathbf{C}_b$ is the interior of $\mathbf{C} \backslash \mathbf{C}_g$, and the Jordan curve J is the boundary of $\mathbf{C} \backslash \mathbf{C}_g$. It is obvious that this decomposition of $\mathbf{C}$ satisfies all conditions of Definition 3.1.2.

This stability domain is also of practical interest. A system is stable with respect to $\mathbf{C}_g$ if its exponential decay rate is at least $(-\beta)$. However, a lower exponential

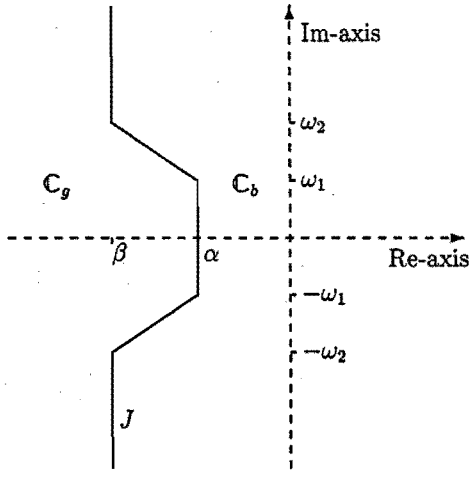


Figure 3.1: The stability domain $\mathbb{C}_g$

decay rate (up to $(-\alpha)$) is allowed if the corresponding solutions only involve a low frequency oscillatory behaviour at frequencies less than a value between ω_1 and ω_2. In this way, desirable performance criteria can be formalized by an appropriate choice of the stability domain $\mathbb{C}_g$.

When a particular stability domain $\mathbb{C}_g$ is fixed, the corresponding Hurwitz set $\mathcal{D}_g$ is defined as follows.

$$\mathcal{D}_g := \{p(z, s_1, \ldots, s_k) \in \mathbb{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } \forall z \in \mathbb{C} : p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) = 0 \Rightarrow z \in \mathbb{C}_g\}. \quad (3.5)$$

A polynomial $p(z, s_1, \ldots, s_k) \in \mathbb{R}[z, s_1, \ldots, s_k]$ is called stable if and only if the analytic function $p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$ that is obtained by substitution of $e^{-\tau_i z}$ for the indeterminate s_i $(i = 1, \ldots, k)$, has all its zeros in $\mathbb{C}_g$. Note that the Hurwitz set (3.5) does not only depend on the choice of the stability domain $\mathbb{C}_g$, but also on the incommensurable time-delays $\tau_1, \ldots, \tau_k$ occurring in the system. In this way, the Hurwitz set framework for the stability of linear systems over rings is specialized to the case of systems with incommensurable time-delays. Therefore the design methods developed in Chapter 2 are also applicable to this class of systems. Unfortunately, the conditions under which these methods work are still rather difficult to check. However, in the time-delay case and with Hurwitz sets of the form (3.5), the testing of these conditions can be facilitated a lot. This is the subject of the next section.

3.2 Stabilizability conditions for time-delay systems

In this section it is shown that for time-delay systems, the right-invertibility condition for stabilizability is equivalent with a pointwise rank condition similar to the

Hautus test. For the proof of this result we need some results from the theory of analytic functions. In literature, these theorems from complex analysis are stated in such a way that they are not directly applicable for our purpose. Therefore, we devote the first subsection to a modification of these results so that we can use them later on.

3.2.1 Some auxiliary results on analytic functions

We start with the definition of a commutative (Banach) algebra of analytic functions (see for example [13, p. 191]) that plays an important role throughout this section.

Definition 3.2.1 Let Ω be a bounded simply connected region in the complex plane, and let $\overline{\Omega}$ denote the closure of Ω. Assume that the boundary J of $\overline{\Omega}$ is a Jordan curve. Then $\mathcal{A}(\Omega)$ is defined as the algebra of all functions f which are analytic in Ω and are continuous on $\overline{\Omega}$. Moreover, if we equip $\mathcal{A}(\Omega)$ with the norm

$$\|f\|_\Omega := \sup\{|f(z)| \mid z \in \overline{\Omega}\}, \tag{3.6}$$

$\mathcal{A}(\Omega)$ becomes a commutative Banach algebra.

Instead of taking Ω an arbitrary bounded simply connected region in the complex plane, we may also assume that $\Omega = \mathcal{U} := \{z \in \mathbb{C} \mid |z| < 1\}$ (so $\mathcal{U}$ is the unit disc). This is only a small restriction because a lot of results for the special case $\Omega = \mathcal{U}$ remain invariant under conformal mappings, and therefore these results are easily generalized to arbitrary bounded simply connected regions, using the Riemann Mapping Theorem. In the sequel we only need the following rather restricted version of this more general result.

Theorem 3.2.2 *Let Ω be a bounded simply connected region in the complex plane and assume that the boundary of Ω is a Jordan curve. Then every conformal mapping from Ω onto $\mathcal{U}$ extends to a homeomorphism from $\overline{\Omega}$ onto $\overline{\mathcal{U}}$.* ∎

An extensive treatment on conformal mappings is given in [83, Section 14]. The result stated above is mainly based on [83, Theorem 14.19] and [83, Remark 14.20].

For the algebra $\mathcal{A}(\Omega)$ there exists a sort of adapted version of the Hilbert Nullstellensatz. This result is stated in the next theorem.

Theorem 3.2.3 *Let $f_1, \ldots, f_n$ be functions in $\mathcal{A}(\Omega)$, and assume that $f_1, \ldots, f_n$ do not have a common zero in $\overline{\Omega}$. Then there exist functions $g_1, \ldots, g_n \in \mathcal{A}(\Omega)$ such that*

$$\forall z \in \overline{\Omega} : \sum_{i=1}^{n} f_i(z) \cdot g_i(z) = 1. \tag{3.7}$$

Proof

The case $\Omega = \mathcal{U}$ can be found in [46, p. 88]. It is based on some results of Rudin in [82].

If Ω is an arbitrary bounded simply connected region such that the boundary of Ω is a Jordan curve, we take a conformal mapping from Ω onto $\mathcal{U}$, and extend it

to a homeomorphism from $\overline{\Omega}$ onto $\overline{\mathcal{U}}$. In this way a function φ is obtained that is analytic in Ω and continuous on $\overline{\Omega}$. Moreover, φ^{-1} exists and is analytic in $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$.

Define for $i = 1, \ldots, n$:

$$\tilde{f}_i : \quad \overline{\mathcal{U}} \longrightarrow \mathbf{C} : \quad \tilde{f}_i := f_i \circ \varphi^{-1}.$$

Clearly $\tilde{f}_i \in \mathcal{A}(\mathcal{U})$, and $\tilde{f}_1, \ldots, \tilde{f}_n$ do not have a common zero in $\overline{\mathcal{U}}$. We now apply the theorem for $\Omega = \mathcal{U}$, and in this way we find functions $\tilde{g}_i \in \mathcal{A}(\mathcal{U})$ $(i = 1, \ldots, n)$ such that

$$\forall z \in \overline{\mathcal{U}} : \sum_{i=1}^{n} \tilde{f}_i(z) \cdot \tilde{g}_i(z) = 1.$$

Finally, define for $i = 1, \ldots, n$ the functions $g_i \in \mathcal{A}(\Omega)$ by:

$$g_i : \quad \overline{\Omega} \longrightarrow \mathbf{C} : \quad g_i := \tilde{g}_i \circ \varphi.$$

Let $z \in \overline{\Omega}$ and define $s := \varphi(z)$. Then we have

$$\sum_{i=1}^{n} f_i(z) \cdot g_i(z) = \sum_{i=1}^{n} f_i(\varphi^{-1}(s)) \cdot \tilde{g}_i(\varphi(z)) = \sum_{i=1}^{n} \tilde{f}_i(s) \cdot \tilde{g}_i(s) = 1,$$

and thus the functions $g_1, \ldots, g_n \in \mathcal{A}(\Omega)$ satisfy the claim. ∎

The next theorem indicates that functions that are continuous on a compact set K, satisfying some regularity conditions, and that are analytic in the interior of K, can be uniformly approximated by polynomials.

Theorem 3.2.4 (Mergelyan's Theorem) *Let K be a compact set in the complex plane whose complement is connected, and let f be a function that is continuous on K and analytic in the interior of K. Let $\varepsilon > 0$. Then there exists a polynomial $p \in \mathbf{C}[z]$ such that*

$$\forall z \in K : |f(z) - p(z)| < \varepsilon. \tag{3.8}$$

∎

For a proof of Mergelyan's Theorem we refer to e.g [83, pp. 390-394]. In Section 6.2, a special case of this result is elaborated in more detail.

Combining Theorem 3.2.3 and Theorem 3.2.4, we obtain the following result.

Corollary 3.2.5 *Let Ω be a bounded simply connected region in the complex plane such that the complement of $\overline{\Omega}$ is connected. Let $f_1, \ldots, f_n \in \mathcal{A}(\Omega)$, and assume that $f_1, \ldots, f_n$ do not have a common zero in $\overline{\Omega}$. Then for every $\varepsilon > 0$ there exist polynomials $p_1, \ldots, p_n \in \mathbf{C}[z]$ such that*

$$\forall z \in \overline{\Omega} : \left| \sum_{i=1}^{n} f_i(z) p_i(z) - 1 \right| < \varepsilon. \tag{3.9}$$

Proof

Since $f_1, \ldots, f_n$ are functions in $\mathcal{A}(\Omega)$ without a common zero in $\overline{\Omega}$, there exist according to Theorem 3.2.3 functions $g_1, \ldots, g_n \in \mathcal{A}(\Omega)$ such that (3.7) is satisfied. By definition the functions $f_1, \ldots, f_n$ are continuous on the compact set $\overline{\Omega}$, and thus there exists an $M \in \mathbf{R}$ such that

$$\forall i \in \{1, \ldots, n\}\ \forall z \in \overline{\Omega}: \ |f_i(z)| < M.$$

Let $\varepsilon > 0$, and apply Theorem 3.2.4 on the functions $g_1, \ldots, g_n$. Then for every $i \in \{1, \ldots, n\}$ we may find a polynomial $p_i \in \mathbf{C}[z]$ such that

$$\forall z \in \overline{\Omega}: \ |g_i(z) - p_i(z)| < \frac{\varepsilon}{n \cdot M}.$$

These polynomials $p_1, \ldots, p_n$ satisfy (3.9) because for every $z \in \overline{\Omega}$ we have

$$\begin{aligned} \left| \sum_{i=1}^{n} f_i(z) p_i(z) - 1 \right| &\leq \left| \sum_{i=1}^{n} f_i(z)(p_i(z) - g_i(z)) \right| \leq \\ &\leq \sum_{i=1}^{n} |f_i(z)| \cdot |p_i(z) - g_i(z)| < \sum_{i=1}^{n} M \frac{\varepsilon}{nM} = \varepsilon. \end{aligned}$$

■

Corollary 3.2.5 is only valid for functions defined on a compact subset $\overline{\Omega}$ of the complex plane. So, in particular, $\overline{\Omega}$ is bounded. However, the stability of a time-delay system is defined with help of a stability domain $\mathbf{C}_g$ and its complement $\overline{\mathbf{C}_b} = \mathbf{C} \backslash \mathbf{C}_g$. So for our purpose we are interested in functions that are analytic in $\mathbf{C}_b$ and continuous on $\overline{\mathbf{C}_b}$ and at infinity, because such functions are considered to be stable.

Definition 3.2.6 Let $\mathbf{C}_g$ be a stability domain satisfying the conditions of Definition 3.1.2 and let $\mathbf{C}_b$ denote the interior of $\mathbf{C} \backslash \mathbf{C}_g$. The algebra $\mathcal{A}_0(\mathbf{C}_b)$ associated with $\mathbf{C}_g$ is defined as the set of all functions f that are analytic in $\mathbf{C}_b$, continuous on $\overline{\mathbf{C}_b}$ and satisfy

$$\exists L \in \mathbf{C}: \lim_{|z| \to \infty,\, z \in \overline{\mathbf{C}_b}} |f(z) - L| = 0. \tag{3.10}$$

(Condition (3.10) simply says that f can be extended continuously to infinity).

To generalize Corollary 3.2.5 to the case $\Omega = \mathbf{C}_b$ we have to solve the problem that arises because $\overline{\mathbf{C}_b}$ is not a compact set. This can be done by transforming $\overline{\mathbf{C}_b}$ to a compact set using a so-called Möbius transformation. In this way we obtain the following proposition which plays a crucial role in the proof of the main result of the next subsection.

Proposition 3.2.7 *Let $f_1, \ldots, f_n$ be functions in $\mathcal{A}_0(\mathbf{C}_b)$. Assume that $f_1, \ldots, f_n$ do not have a common zero in $\overline{\mathbf{C}_b}$ and that*

$$\exists i \in \{1, \ldots, n\}: \lim_{|z| \to \infty,\, z \in \overline{\mathbf{C}_b}} |f_i(z)| \neq 0. \tag{3.11}$$

Let $\varepsilon > 0$. Then there exist proper stable rational functions $r_1, \ldots, r_n \in \mathbf{C}(z)$ (i.e. for all $i = 1, \ldots, n$ we have $r_i \in \mathbf{C}(z) \cap \mathcal{A}_0(\mathbf{C}_b)$) such that

$$\forall z \in \overline{\mathbf{C}_b} : \left| \sum_{i=1}^{n} f_i(z) \cdot r_i(z) - 1 \right| < \varepsilon. \tag{3.12}$$

Moreover, if for all $i \in \{1, \ldots, n\}$ the equality $\overline{f_i(\bar{z})} = f_i(z)$ is satisfied for all $z \in \overline{\mathbf{C}_b}$, then these proper stable rational functions $r_1, \ldots, r_n$ can be chosen to be real rational functions, i.e.

$$\forall i \in \{1, \ldots, n\} : r_i \in \mathcal{A}_0(\mathbf{C}_b) \cap \mathbf{R}(z).$$

Proof

Choose a $\beta \in \mathbf{C}_g \cap \mathbf{R}$. Such a β exists because of condition *(iv)* in Definition 3.1.2. Define the Möbius transformation

$$T : \mathbf{C}\backslash\{\beta\} \longrightarrow \mathbf{C}\backslash\{1\} : \qquad T(z) := \frac{z+\beta}{z-\beta},$$

with inverse T^{-1} :

$$T^{-1} : \mathbf{C}\backslash\{1\} \longrightarrow \mathbf{C}\backslash\{\beta\} : \qquad T^{-1}(s) := \beta \cdot \frac{s+1}{s-1}.$$

Note that the Möbius transformation T maps the point β to infinity. Since there exists a neighbourhood of β that does not contain points of $\mathbf{C}_b$, the image $\Omega := T(\mathbf{C}_b)$ of $\mathbf{C}_b$ under T stays away from infinity, and is therefore bounded. So it is obvious that $\overline{\Omega} = T(\overline{\mathbf{C}_b}) \cup \{1\}$ is a compact set. Moreover, because of condition *(ii)* in Definition 3.1.2, $\mathbf{C}_b$ is simply connected, and therefore Ω is also simply connected. The same reasoning is applicable to $\mathbf{C}_g$: since $\mathbf{C}_g$ is an open and connected set, also $\mathbf{C}\backslash\overline{\Omega} = T(\mathbf{C}_g\backslash\{\beta\})$ is open and connected.

Define functions $\tilde{f}_i : \overline{\Omega} \longrightarrow \mathbf{C}$ $(i = 1, \ldots, n)$ in the following way:

$$\tilde{f}_i(s) := \begin{cases} f_i(T^{-1}(s)) & \text{for} \quad s \neq 1, \\ \lim_{|s|\to\infty,\, z\in\overline{\mathbf{C}_b}} f_i(s) & \text{for} \quad s = 1. \end{cases}$$

Since T^{-1} is analytic in $\mathbf{C}\backslash\{1\}$, all $\tilde{f}_i$ are analytic in Ω and continuous on $\overline{\Omega}\backslash\{1\}$. The continuity of f_i at infinity (condition (3.10)) implies that for all $i \in \{1, \ldots, n\}$, the function $\tilde{f}_i$ is also continuous in 1. Hence $\tilde{f}_i \in \mathcal{A}(\Omega)$ for $i = 1, \ldots, n$. Moreover, since $f_1, \ldots, f_n$ have a common zero neither in $\overline{\mathbf{C}_b}$, nor at infinity (condition (3.11)), the functions $\tilde{f}_1, \ldots, \tilde{f}_n$ do not have a common zero in $\overline{\Omega}$.

Let $\varepsilon > 0$, and apply Corollary 3.2.5 on the functions $\tilde{f}_1, \ldots, \tilde{f}_n$. Then we find polynomials $p_i \in \mathbf{C}[s]$ $(i = 1, \ldots, n)$ such that for all $s \in \overline{\Omega}$:

$$\left| \sum_{i=1}^{n} \tilde{f}_i(s) p_i(s) - 1 \right| < \varepsilon.$$

Define for $i = 1, \ldots, n$:

$$r_i : \overline{\mathbf{C}_b} \longrightarrow \mathbf{C} : \qquad r_i(z) := p_i(T(z)).$$

Clearly, all r_i are rational functions which are proper and have poles in β only. Hence, all r_i are stable: $r_i \in \mathbb{C}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$ $(i = 1, \ldots, n)$. Let $z \in \overline{\mathbb{C}_b}$ and define $s := T(z)$. Then $s \in \overline{\Omega}$ and $z = T^{-1}(s)$, and we have

$$\left|\sum_{i=1}^{n} f_i(z) r_i(z) - 1\right| = \left|\sum_{i=1}^{n} f_i(T^{-1}(s)) p_i(T(z)) - 1\right| = \left|\sum_{i=1}^{n} \tilde{f}_i(s) p_i(s) - 1\right| < \varepsilon.$$

Finally, if for all $i \in \{1, \ldots, n\}$ and $z \in \overline{\mathbb{C}_b}$ the equality $\overline{f_i(\bar{z})} = f_i(z)$ is satisfied, we define $q_i(z) := \frac{1}{2}(r_i(z) + \overline{r_i(\bar{z})})$. Since $\overline{\mathbb{C}_b}$ is symmetric w.r.t. the real axis (condition *(i)* of Definition 3.1.2), all q_i are well defined. Moreover, the functions q_i $(i = 1, \ldots, n)$ remain proper stable rational functions, and since $\overline{q_i(\bar{z})} = q_i(z)$, they are also real: all q_i $(i = 1, \ldots, n)$ belong to $\mathbb{R}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$. Let again $z \in \overline{\mathbb{C}_b}$. Then we have

$$\begin{aligned}\left|\sum_{i=1}^{n} f_i(z) q_i(z) - 1\right| &= \left|\frac{1}{2}\sum_{i=1}^{n} f_i(z) r_i(z) + \frac{1}{2}\sum_{i=1}^{n} f_i(z)\overline{r_i(\bar{z})} - 1\right| \leq \\ &\leq \frac{1}{2}\left|\sum_{i=1}^{n} f_i(z) r_i(z) - 1\right| + \frac{1}{2}\left|\sum_{i=1}^{n} \overline{f_i(\bar{z}) r_i(\bar{z})} - 1\right| < \varepsilon.\end{aligned}$$

This completes the proof. ∎

At this point, the importance of Proposition 3.2.7 is not very clear. In the next subsection, this result is used to construct a stable polynomial in the ideal $\mathcal{J}$ described in Definition 2.8.4. For this purpose, the proper stable rational functions $r_1, \ldots, r_n$ are used explicitly. According to the proof of Proposition 2.8.5, this enables us to construct a right-inverse of the matrix $(zI - A|B)$ over the ring $\mathcal{R}_{\mathcal{D}_g}(z)$. Therefore the results of this subsection play a crucial role in the reformulation of the stabilizability conditions for time-delay systems.

3.2.2 A pointwise rank condition for stabilizability

In this subsection, we derive a condition for the stabilizability of a time-delay system that is more easily verifiable than the right-invertibility condition of Theorem 2.8.2. We follow the same approach as Emre and Knowles in [25]. They were the first ones to present such a stabilizability condition, both for time-delay and neutral systems, in the case that the stability domain $\mathbb{C}_g$ is an arbitrary open left half plane $\mathbb{C}_\alpha = \{z \in \mathbb{C} | \operatorname{Re} z < \alpha\}$. We confine ourselves to time-delay systems, but allow a more general class of stability domains. In this more abstract setting, we obtain, besides a more general result, also somewhat more insight in the ideas behind the proof.

Theorem 3.2.8 *Let $\mathcal{R} = \mathbb{R}[s_1, \ldots, s_k]$ and $A = A(s_1, \ldots, s_k) \in \mathcal{R}^{n \times n}$ and $B = B(s_1, \ldots, s_k) \in \mathcal{R}^{n \times m}$. Let $(\tau_1, \ldots, \tau_k)$ be a k-tuple of incommensurable time-delays. Let $\mathbb{C}_g$ be a stability domain as described in Definition 3.1.2 and $\mathcal{D}_g$ the Hurwitz set corresponding to $\mathbb{C}_g$ as defined in (3.5). Then*

$$(zI - A|B) \text{ is right-invertible over } \mathcal{R}_{\mathcal{D}_g}(z), \tag{3.13}$$

$\Longleftrightarrow$

$$\forall z \in \mathbb{C} \backslash \mathbb{C}_g : \operatorname{rank}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) | B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = n. \tag{3.14}$$

Proof

"$\Rightarrow$" Assume that $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}_g}(z)$. Then there exists a matrix $G(z, s_1, \ldots, s_k)$ over $\mathcal{R}_{\mathcal{D}_g}(z)$ such that

$$(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k)) \cdot G(z, s_1, \ldots, s_k) = I. \tag{3.15}$$

Let $\hat{z} \in \mathbb{C}\backslash\mathbb{C}_g$, and substitute $z = \hat{z}$ and $s_j = e^{-\tau_j \hat{z}}$ $(j = 1, \ldots, k)$ in (3.15). Since $G(z, s_1, \ldots, s_k)$ is a matrix over $\mathcal{R}_{\mathcal{D}_g}(z)$, the matrix $G(\hat{z}, e^{-\tau_1 \hat{z}}, \ldots, e^{-\tau_k \hat{z}})$ is well defined, and it is a right-inverse of

$$(\hat{z}I - A(e^{-\tau_1 \hat{z}}, \ldots, e^{-\tau_k \hat{z}})|B(e^{-\tau_1 \hat{z}}, \ldots, e^{-\tau_k \hat{z}})). \tag{3.16}$$

So, in particular, the matrix (3.16) is of full row rank. Since $\hat{z} \in \mathbb{C}\backslash\mathbb{C}_g$ was arbitrary, this proves (3.14).

"$\Leftarrow$" The implication in the other direction is much more involved. Denote the $n \times n$ minors of the matrix $(zI - A|B)$ by $\alpha_0(z, s_1, \ldots, s_k), \ldots, \alpha_N(z, s_1, \ldots, s_k)$, and assume that $\alpha_0(z, s_1, \ldots, s_k) = \det(zI - A(s_1, \ldots, s_k))$. Let $\mathcal{J}$ be the ideal in $\mathcal{R}[z]$ generated by all these $n \times n$ minors. Suppose that (3.14) is satisfied. We shall prove that this implies that the ideal $\mathcal{J}$ contains an element of $\mathcal{D}_g$.

Let $\beta \in \mathbb{C}_g \cap \mathbb{R}$ (according to condition *(iv)* of Definition 3.1.2 such a β exists), and define for $i = 0, 1, \ldots, N$ the functions:

$$a_i : \overline{\mathbb{C}_b} \longrightarrow \mathbb{C} : \quad a_i(z) := \frac{\alpha_i(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z})}{(z - \beta)^n}.$$

Claim 1: The functions $a_0, \ldots, a_N$ belong to $\mathcal{A}_0(\mathbb{C}_b)$, have no common zeros in $\overline{\mathbb{C}_b}$, and satisfy condition (3.11) of Proposition 3.2.7.

Proof of Claim 1. It is apparent that all functions a_i $(i = 0, 1, \ldots, N)$ are analytic on $\mathbb{C}_b$ and continuous on $\overline{\mathbb{C}_b}$, so to prove that they belong to the algebra $\mathcal{A}_0(\mathbb{C}_b)$, we have to show that they can be extended continuously to infinity.

Because of condition *(iv)* in Definition 3.1.2 there exists a $\gamma \in \mathbb{R}$ such that $\mathbb{C}_b \subset \{z \in \mathbb{C} | \operatorname{Re} z \geq \gamma\}$. So for all $\tau > 0$ and $z \in \overline{\mathbb{C}_b}$ we have

$$|e^{-\tau z}| \leq e^{-\tau \gamma}.$$

The minors $\alpha_0(z, s_1, \ldots, s_k), \ldots, \alpha_N(z, s_1, \ldots, s_k)$ are of the form

$$\alpha_i(z, s_1, \ldots, s_k) = \sum_{j=0}^{\ell} q_j(s_1, \ldots, s_k) z^j \qquad (i = 0, 1, \ldots, N),$$

where $q_i(s_1, \ldots, s_k)$ is a polynomial in $\mathbb{R}[s_1, \ldots, s_k]$, and $\ell \leq n$. Since $|e^{-\tau z}|$ is bounded in $\overline{\mathbb{C}_b}$, this implies that there exist $M_0, \ldots, M_\ell \in \mathbb{R}$ such that

$$\forall j \in \{0, 1, \ldots, \ell\} \forall z \in \overline{\mathbb{C}_b} : |q_j(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})| \leq M_j.$$

Now, for $\alpha_1, \ldots, \alpha_N$ we even know that $\ell < n$. Hence for $z \in \overline{\mathbb{C}_b}$ and $|z| \to \infty$ we have

$$|a_i(z)| = \left| \frac{\alpha_i(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z})}{(z - \beta)^n} \right| \leq \frac{M_\ell |z|^\ell + \cdots + M_1 |z| + M_0}{|z - \beta|^n} \longrightarrow 0.$$

So $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} a_i(z) = 0$ for all $i = 1, \ldots, N$. On the other hand, $\alpha_0(z, s_1, \ldots, s_k)$ is of the form

$$\alpha_0(z, s_1, \ldots, s_k) = z^n + \sum_{j=0}^{n-1} q_j(s_1, \ldots, s_k) z^j.$$

So in this case we have

$$\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} a_0(z) = \lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} \frac{z^n}{(z-\beta)^n} + \lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} \frac{\sum_{j=0}^{n-1} q_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \cdot z^j}{(z-\beta)^n} = 1.$$

Hence $a_i \in \mathcal{A}_0(\mathbb{C}_b)$ for all $i = 0, 1, \ldots, N$, and, moreover, condition (3.11) of Proposition 3.2.7 is satisfied. Finally, assumption (3.14) guarantees that $a_0, \ldots, a_N$ do not have a common zero in $\overline{\mathbb{C}_b}$. □

Claim 2: $\forall \varepsilon > 0\ \exists r_0, \ldots, r_N \in \mathbb{R}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$ such that:

(i) $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} r_0(z) = 1,$

(ii) $\forall z \in \overline{\mathbb{C}_b} : \left| \sum_{i=0}^{N} a_i(z) r_i(z) - 1 \right| < \varepsilon.$

Proof of Claim 2. Let $\varepsilon > 0$, and define $\hat{M} := \max\{|a_0(z)| \mid z \in \overline{\mathbb{C}_b}\}$. This maximum exists because a_0 is continuous on the closed set $\overline{\mathbb{C}_b}$ and can be extended continuously to infinity (recall $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} a_0(z) = 1$). Choose $0 < \varepsilon_1 < \min(\frac{\varepsilon}{3\hat{M}}, \frac{\varepsilon}{3})$ and apply Proposition 3.2.7 on the functions $a_0, \ldots, a_N$. Since all functions a_i satisfy the condition

$$\forall i \in \{0, 1, \ldots, N\} \forall z \in \overline{\mathbb{C}_b} : \quad \overline{a_i(\bar{z})} = a_i(z),$$

we conclude that there exist proper real rational functions $r_0, \ldots, r_N \in \mathbb{R}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$ such that

$$\forall z \in \overline{\mathbb{C}_b} : \left| \sum_{i=0}^{N} a_i(z) r_i(z) - 1 \right| < \varepsilon_1. \tag{3.17}$$

Since all r_i are proper, $\lim_{|z|\to\infty} r_i(z)$ exists for every $i \in \{0, 1, \ldots, N\}$. Recall that $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} a_i(z) = 0$ for $i = 1, \ldots, N$ and $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} a_0(z) = 1$. Combining these observations we obtain

$$\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} \sum_{i=0}^{N} a_i(z) r_i(z) = \lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} r_0(z),$$

and because of (3.17) this implies that $\left|\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} r_0(z) - 1\right| \leq \varepsilon_1$.

Define

$$\tilde{r}_0 : \overline{\mathbb{C}_b} \longrightarrow \mathbb{C} : \quad \tilde{r}_0(z) := r_0(z) + 1 - \lim_{|w|\to\infty, w\in\overline{\mathbb{C}_b}} r_0(w).$$

Then $\tilde{r}_0$ is still a proper element in $\mathbb{R}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$, and moreover $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} \tilde{r}_0(z) = 1$. Replacing $r_0(z)$ by $\tilde{r}_0(z)$ yields for an arbitrary $z \in \overline{\mathbb{C}_b}$:

$$\left|a_0(z)\tilde{r}_0(z) + \sum_{i=1}^{N} a_i(z)r_i(z) - 1\right| = \left|\sum_{i=0}^{N} a_i(z)r_i(z) - 1 + a_0(z)\cdot(\tilde{r}_0(z) - r_0(z))\right| \leq$$

$$\leq \left|\sum_{i=0}^{N} a_i(z)r_i(z) - 1\right| + |a_0(z)| \cdot \left|\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} r_0(z) - 1\right| \leq$$

$$\leq \varepsilon_1 + \hat{M}\cdot\varepsilon_1 < \frac{\varepsilon}{3} + \frac{\varepsilon\hat{M}}{3\hat{M}} < \varepsilon.$$

So we conclude that $\tilde{r}_0, r_1, \ldots, r_N$ satisfy both *(i)* and *(ii)*. □.

With help of Claim 2 we are able to construct a polynomial in $\mathcal{D}_g \cap \mathcal{J}$. Let $0 < \varepsilon < 1$, and choose rational functions $r_0, \ldots, r_N$ in $\mathbb{R}(z) \cap \mathcal{A}_0(\mathbb{C}_b)$ such that both condition *(i)* and condition *(ii)* of Claim 2 are satisfied. Without loss of generality we may assume that the denominator polynomials of the rational functions $r_0, \ldots, r_N$ are monic. (Otherwise we simply divide by the leading coefficient unequal to 1). Now, $r_i(z)$ $(i = 0, 1, \ldots, N)$ may be written as $r_i(z) = \frac{n_i(z)}{d_i(z)}$, with $n_i(z) \in \mathbb{R}[z]$ and $d_i(z) \in \mathcal{D}_g$. The denominators $d_i(z)$ $(i = 0, 1, \ldots, N)$ belong to $\mathcal{D}_g$ because all their zeros are contained in $\mathbb{C}_g$. Define $\psi(z)$ as the least common multiple of all polynomials $d_i(z)$ $(i = 0, 1, \ldots, N)$, and $g_i(z) \in \mathbb{R}[z]$ $(i = 0, 1, \ldots, N)$ by

$$g_i(z) := r_i(z)\cdot\psi(z) \qquad (i = 0, 1, \ldots, N).$$

Claim 3: The polynomial

$$\alpha(z, s_1, \ldots, s_k) := \sum_{i=0}^{N} \alpha_i(z, s_1, \ldots, s_k) g_i(z) \tag{3.18}$$

is an element of $\mathcal{D}_g \cap \mathcal{J}$.

Proof of Claim 3. It is immediately clear that $\alpha(z, s_1, \ldots, s_k)$ is an element of the ideal $\mathcal{J}$, so it suffices to prove that $\alpha(z, s_1, \ldots, s_k) \in \mathcal{D}_g$.

First we show that $\alpha(z, s_1, \ldots, s_k)$ is monic in z. Recall that the rational function $r_0(z)$, constructed in Claim 2, was proper, but not strictly proper, while the other rational functions $r_i(z)$ $(i = 1, \ldots, N)$ were strictly proper. This implies that

$$\forall i \in \{1, \ldots, N\} : \deg_z(g_i(z)) < \deg_z(g_0(z)).$$

Moreover, from the fact that $\lim_{|z|\to\infty, z\in\overline{\mathbb{C}_b}} r_0(z) = 1$ we conclude that $\deg(n_0(z)) = \deg(d_0(z))$ and that $n_0(z)$ is a monic polynomial. Since $\psi(z)$ is monic too, this implies that $g_0(z)$ is a monic polynomial. Recalling our convention on the numbering of the $n \times n$ minors of $(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k))$, we know that $\alpha_0(z, s_1, \ldots, s_k)$ is monic in z and

$$\forall i \in \{1, \ldots, N\} : \deg_z(\alpha_i(z, s_1, \ldots, s_k)) < \deg_z(\alpha_0(z, s_1, \ldots, s_k)) = n.$$

Combining both results, we conclude that the polynomial $\alpha_0(z, s_1, \ldots, s_k)g_0(z)$ is monic in z and that

$$\forall i \in \{1, \ldots, N\} : \deg_z(\alpha_i(z, s_1, \ldots, s_k)g_i(z)) < \deg_z(\alpha_0(z, s_1, \ldots, s_k)g_0(z)).$$

This immediately implies that $\alpha(z, s_1, \ldots, s_k)$ is monic in z.

Next, substituting $s_j = e^{-\tau_j z}$ $(j = 1, \ldots, k)$, we obtain for all $z \in \overline{\mathbb{C}_b}$:

$$\left| \frac{\alpha(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z})}{\psi(z)(z-\beta)^n} - 1 \right| =$$

$$= \left| \frac{1}{\psi(z)(z-\beta)^n} \cdot \sum_{i=0}^{N} \alpha_i(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) r_i(z) \psi(z) - 1 \right| =$$

$$= \left| \sum_{i=0}^{N} a_i(z) r_i(z) - 1 \right| < \varepsilon.$$

So in particular:

$$\forall z \in \overline{\mathbb{C}_b} : \left| \frac{\alpha(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z})}{\psi(z)(z-\beta)^n} \right| > 1 - \varepsilon > 0.$$

Since $\psi(z) \in \mathcal{D}_g$, it has no zeros in $\overline{\mathbb{C}_b}$, and pole-zero cancellations cannot occur. Therefore we conclude

$$\forall z \in \overline{\mathbb{C}_b} : \alpha(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \neq 0,$$

hence $\alpha(z, s_1, \ldots, s_k) \in \mathcal{D}_g$. □

The polynomial $\alpha(z, s_1, \ldots, s_k)$ constructed in Claim 3 is an element of $\mathcal{D}_g \cap \mathcal{J}$. According to Proposition 2.8.5 this implies that $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}_g}(z)$.

This completes the proof. ■

Combining Theorem 3.2.8 and the results on stabilizability and detectability for linear systems over rings in Chapter 2, we obtain the following explicit conditions for the stabilizability of a time-delay system.

Corollary 3.2.9 *Let $\Sigma = (A, B, C, D)$ be a linear system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ of rank n, describing a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$. For $i = 1, \ldots, k$ the indeterminate s_i corresponds to the delay operator σ_i with time-delay τ_i. Let $\mathbb{C}_g$ be a stability domain. Then*

(i) $\Sigma = (A, B, C, D)$ is internally stabilizable by dynamic state feedback if and only if

$$\forall z \in \mathbb{C} \backslash \mathbb{C}_g : \operatorname{rank}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) | B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = n, \quad (3.19)$$

(ii) $\Sigma = (A, B, C, D)$ is detectable if and only if

$$\forall z \in \mathbb{C} \backslash \mathbb{C}_g : \operatorname{rank} \begin{pmatrix} zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \\ C(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \end{pmatrix} = n, \quad (3.20)$$

(iii) $\Sigma = (A, B, C, D)$ is stabilizable by dynamic output feedback if and only if both (3.19) and (3.20) are satisfied. ■

Corollary 3.2.9 gives a pointwise rank condition to verify the stabilizability of a time-delay system. This condition can be seen as a rather straightforward generalization of the well-known Hautus test for stabilizability to the case of systems with time delays.

The stabilizability conditions (3.19) and (3.20) are also very satisfactory when we compare them to similar results coming from the infinite-dimensional systems approach, mentioned in Chapter 1. Using the latter approach, Pandolfi (see [73]) derived the same condition for the stabilizability of a time-delay system. However, he does not need dynamic feedback to achieve stability; static feedback turns out to be sufficient in his approach. Unfortunately, in this static feedback, distributed time-delays occur, even if the original system only has point delays. Therefore, Pandolfi's result is not directly applicable in the systems over rings approach to time-delay systems. On the other hand, the stabilizing dynamic compensator obtained with Corollary 3.2.9 contains only point delays, and therefore it fits better in our algebraic framework.

Remark 3.2.10 The conditions for stabilizability by dynamic feedback enable us to give another interpretation to the algebraic definition of reachability (i.e. of Definition 2.2.2) in the case of systems with point delays. According to Remark 2.8.3, reachability of a system implies stabilizability with respect to any Hurwitz set $\mathcal{D}$. Now, consider exponential stabilizability with a guaranteed decay rate α. So look at Hurwitz sets of the form

$$\mathcal{D}_{-\alpha} := \{p(z, s_1, \ldots, s_k) \in \mathbb{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } \forall z \in \mathbb{C} : p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) = 0 \Rightarrow \operatorname{Re} z < -\alpha\}.$$

According to the arguments above, a reachable system is always stabilizable with respect to $\mathcal{D}_{-\alpha}$ independent of the actual value of α. This means that for any arbitrary exponential decay rate α, a dynamic compensator can be found such that the closed-loop system is stable with exponential decay rate α. So in a time-delay system that is reachable in the algebraic sense, the state x can be steered to zero arbitrarily fast. In a way, this property resembles null-controllability of linear systems without delays.

3.2.3 Pointwise stabilizability

One of the main problems in the stabilizability test of Corollary 3.2.9 is the fact that the time-delays $\tau_1, \ldots, \tau_k$ have to be known exactly. In the second part of this chapter it is shown that the stabilizability condition is not very sensitive to small perturbations of the lengths $\tau_1, \ldots, \tau_k$ of the time-delays occurring in the system. However, in a lot of cases a good estimate of the length of a time-delay is already difficult to obtain. In this situation, the stabilizability condition (3.19) is not very helpful. Instead, we want to have a stabilizability condition that can be used without any knowledge of the time-delays $\tau_1, \ldots, \tau_k$. This idea of stability independent of delay was introduced by Kamen, who investigated this subject in several articles (see [50], [51] and [52]).

Definition 3.2.11 Let $A(s_1,\ldots,s_k) \in \mathbb{R}[s_1,\ldots,s_k]^{n\times n}$ and consider a k-tuple of delay operators $\sigma_1,\ldots,\sigma_k$ corresponding to k incommensurable time-delays $\tau_1,\ldots,\tau_k$. Then the time-delay system

$$\dot{x}(t) = A(\sigma_1,\ldots,\sigma_k)x(t), \tag{3.21}$$

is called *stable independent of delay* if for all k-tuples $(\tau_1,\ldots,\tau_k) \in (\mathbb{R}^+)^k$ and for any arbitrary initial state trajectory, the state $x(t)$ tends to zero when t tends to infinity.

Proposition 3.2.12 *The system (3.21) is stable independent of delay if and only if*

$$\forall z \in \overline{\mathbb{C}^+}\ \forall (h_1,\ldots,h_k) \in (\mathbb{R}^+)^k :\ \det(zI - A(e^{-h_1 z},\ldots,e^{-h_k z})) \neq 0. \tag{3.22}$$

■

A proof of this result may be found in [15].

The next step is to define a Hurwitz set $\mathcal{D}$ which translates the notion of stability independent of delay to the abstract framework of stability developed in Chapter 2. This is not a very difficult task, but unfortunately this Hurwitz set will not lead to easily verifiable conditions for stabilizability independent of delay. However, this problem is partly solvable if we use the following Hurwitz set instead:

$$\begin{aligned}\mathcal{D}_p := \{p(z,s_1,\ldots,s_k) \in \mathbb{R}[z,s_1,\ldots,s_k] \ \mid\ & p(z,s_1,\ldots,s_k) \text{ is monic in } z \\ & \text{and } \forall z \in \overline{\mathbb{C}^+}\ \forall (s_1,\ldots,s_k) \in \overline{\mathcal{U}}^k :\ p(z,s_1,\ldots,s_k) \neq 0\},\end{aligned} \tag{3.23}$$

where $\mathcal{U}$ denotes the open unit disc $\{s \in \mathbb{C} \mid |s| < 1\}$.

Definition 3.2.13 Let $\Sigma = (A,B,C,D)$ be a linear system over the polynomial ring $\mathcal{R} := \mathbb{R}[s_1,\ldots,s_k]$. Substitute delay operators $\sigma_1,\ldots,\sigma_k$ for the indeterminates $s_1,\ldots,s_k$, and consider Σ as a time-delay system with k incommensurable delays. $\Sigma = (A,B,C,D)$ is called *pointwise stable* (stabilizable) if and only if it is stable (stabilizable) w.r.t. the Hurwitz set $\mathcal{D}_p$.

It follows from Definition 3.2.13 and Proposition 3.2.12 that a delay system which is pointwise stable is certainly stable independent of delay. But pointwise stability and stability independent of delay are not completely equivalent; in [53] it was pointed out that pointwise stability is a slightly stronger property. Although the intuitive meaning of pointwise stability remains a little bit obscure, it has one important advantage: in the same way as in Subsection 3.2.2, pointwise stabilizability can be tested with a rank condition that is very similar to the Hautus test. Since pointwise stability implies stability independent of delay, the practical importance of this result is obvious.

Proposition 3.2.14 *Let $\mathcal{R} = \mathbb{R}[s_1,\ldots,s_k]$ and $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Then*

$(zI - A|B)$ *is right-invertible over* $\mathcal{R}_{\mathcal{D}_p}(z)$

$\Longleftrightarrow$

$$\forall z \in \overline{\mathbb{C}^+}\ \forall (s_1,\ldots,s_k) \in \overline{\mathcal{U}}^k :\ \operatorname{rank}(zI - A(s_1,\ldots,s_k)|B(s_1,\ldots,s_k)) = n. \ ■$$

The proof of Proposition 3.2.14 is a modification (or even a simplification) of the proof of Theorem 3.2.8. The details can be found in [53].

Corollary 3.2.15 *Let $\Sigma = (A, B, C, D)$ be a linear system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ of rank n, describing a time-delay system with k incommensurable time-delays. Then*

(i) $\Sigma = (A, B, C, D)$ is internally pointwise stabilizable by dynamic state feedback if and only if

$$\forall z \in \overline{\mathbb{C}^+}\, \forall (s_1, \ldots, s_k) \in \overline{\mathcal{U}}^k : \operatorname{rank}(zI - A(s_1, \ldots, s_k) | B(s_1, \ldots, s_k)) = n, \quad (3.24)$$

(ii) $\Sigma = (A, B, C, D)$ is pointwise detectable if and only if

$$\forall z \in \overline{\mathbb{C}^+}\, \forall (s_1, \ldots, s_k) \in \overline{\mathcal{U}}^k : \operatorname{rank} \begin{pmatrix} zI - A(s_1, \ldots, s_k) \\ C(s_1, \ldots, s_k) \end{pmatrix} = n, \quad (3.25)$$

(iii) $\Sigma = (A, B, C, D)$ is pointwise stabilizable by dynamic output feedback if and only if both (3.24) and (3.25) are satisfied. ∎

The results of this section are a good illustration of the versatility of the Hurwitz set framework for stability. Several notions of stability can be considered within one framework. The main theorems of Chapter 2 remain valid in all special cases. Only the explicit rank conditions for stabilizability change according to the Hurwitz set under consideration.

3.3 On the genericity of stabilizability

In this section we return to the ordinary notion of stability for time-delay systems as described in Section 3.1, and consider the question how restrictive the stabilizability conditions of Corollary 3.2.9 are. This gives an indication how large the class of stabilizable delay systems is, and whether or not we can expect an arbitrary time-delay system to belong to this class. Or put differently: is the property of stabilizability generic? The answer to this question is affirmative, but involves a lot of technicalities. The rest of this section is devoted to the proof of this result, but before we start with this, we present a more intuitive explanation. This explanation is not a proof; it only contains the main idea.

First, consider a linear system $\Sigma = (A, B, C, D)$ over $\mathbb{R}$. The pair (A, B) is reachable if and only if

$$\forall z \in \mathbb{C} : \operatorname{rank}(zI - A | B) = n, \quad (3.26)$$

where n is the size of A. In [66, p. 100] it was shown that this condition is generically satisfied. In this case, genericity is considered in the normed parameter-space of matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. Although in [66] the proof is based on considerations different from those presented here, the result is not difficult to understand. First of all, square $n \times n$ matrices over $\mathbb{R}$ generically have n distinct eigenvalues and

thus they are generically diagonilizable. This implies that the matrix $(zI - A)$ is singular for only n different values of z, and that in these points the matrix $(zI - A)$ has rank $n - 1$. It is obvious that outside these singularity points (3.26) is certainly valid, while in a singularity point, the matrix B will generically prevent a rank drop of the complete matrix $(zI - A|B)$. Of course this argument is not very precise in this form. Nevertheless, the strategy proposed to solve this problem is used later on to prove the genericity of stabilizability for time-delay systems. The crucial point is that the characteristic function $p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ of a delay system can only have a finite number of zeros in any arbitrary right half plane $\{z \in \mathbb{C} \mid \operatorname{Re} z \geq \alpha\}$, and therefore the same ideas apply.

Next, we recall the condition for the reachability of a system $\Sigma = (A, B, C, D)$ over a polynomial ring $K[s_1, \ldots, s_k]$. According to Theorem 2.2.4, this condition can be restated as a rank condition in almost the same way as for systems over fields:

$$\forall(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{K}^{k+1} : \operatorname{rank}(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1 \ldots, \hat{s}_k)) = n,$$

where n is the size of $A(s_1, \ldots, s_k)$. In Proposition 2.2.5 we quoted from [67] a genericity result on this condition: a system over a polynomial ring is generically reachable if and only if the number of inputs m to the system is strictly larger than the number k of indeterminates in the polynomial ring. The proof of this result is based on a completely algebraic approach. The number of polynomial equations that have to be satisfied is compared with the number of unknowns, and after application of some results from algebraic geometry one can prove the result: except on some hypersurfaces in the system-parameter space, the reachability condition is always satisfied. So in this case the concept of genericity is considered within the so-called Zariski topology.

At first sight, this approach also looks very promising for solving the genericity problem of stabilizability for time-delay systems. In this case each indeterminate s_i in the polynomial ring corresponds to an exponential function $e^{-\tau_i z}$. Therefore some additional (exponential) equations are obtained which are probably sufficient to remove the condition on the number of inputs. However, this method fails because we are now dealing with both polynomial and exponential equations, which do not fit into the algebro-geometric framework.

In the rest of this section, we give a proof of the genericity of stabilizability based on the first strategy we mentioned. So we shall not use the Zariski topology. Instead, we follow the approach developed in [38]. We consider the set of all polynomial matrices of dimension $p \times q$ as a linear space, and construct a topological framework from a functional analytic point of view. Note however that this setup has nothing to do with the functional analytic approach to time-delay systems mentioned in Chapter 1. The topology only has to formalize our intuitive ideas on the question: when are the parametrizations of two time-delay systems said to be close to each other? In Subsection 3.3.1 this topological framework is treated in detail.

Having fixed a topology for time-delay systems, a property is called generic if the set of all systems satisfying this property contains a subset which is both an open and dense subset of the space of all parametrizations of time-delay systems. In Subsection 3.3.2 we show that in our topology the set of all delay systems which are stabilizable w.r.t. an arbitrary open left half plane is open. The proof of denseness is more involved. We start in Subsection 3.3.3 with some preliminary results

on matrices over the ring of analytic functions. These results are used in Subsection 3.3.4 to show that the subset of stabilizable delay systems is a dense subset of the parameter-space describing all time-delay systems.

Finally we have to make an important remark. Throughout the whole section it is tacitly assumed that we are dealing with time-delay systems with commensurable delays. This implies that there is only one delay operator σ needed to describe the system equations (3.1). This situation seems much simpler than the incommensurable delay case, but this distinction does not make any difference for the approach we take to the problem. All results are easily generalized to the case of incommensurable time-delays, because the assumption of the presence of only one delay operator is never used explicitly. This assumption is only made for notational convenience. In Subsection 3.3.5 we return to this subject briefly, and explain why the methods developed in this section are also applicable to systems with incommensurable time-delays.

3.3.1 A topological framework for time-delay systems

As we have seen in Section 1.3, a time-delay system with commensurable delays is completely characterized by a system $\Sigma = (A, B, C, D)$ over the polynomial ring $\mathbf{R}[s]$ and the length τ of the time-delay occurring in the system. We are interested in the question how strong the conditions of stabilizability by dynamic output feedback for these systems are. We know that this problem can be split into two dual parts: the problem of stabilizability by dynamic state feedback and the detectability problem. In the rest of this section we concentrate on the first problem. When this is solved, the same results can be proved for detectability by dualization. According to Corollary 3.2.9 only the matrices A and B and the time-delay τ are involved in the condition for stabilizability by dynamic state feedback. When these three data are given, condition (3.19) may be verified.

Consider a triple $\Sigma = (A(s), B(s), \tau)$, with $A(s) \in \mathbf{R}[s]^{n\times n}$, $B(s) \in \mathbf{R}[s]^{n\times m}$ and $\tau \in \mathbf{R}^+$. After substitution of the delay operator σ with time-delay τ for the indeterminate s, such a triple is a complete description of the time delay system:

$$\begin{cases} \dot{x}(t) &= A(\sigma)x(t) + B(\sigma)u(t), \\ \sigma x(t) &= x(t-\tau), \qquad \sigma u(t) = u(t-\tau). \end{cases} \tag{3.27}$$

On the other hand, the triple $\Sigma = (A(s), B(s), \tau)$ can be seen as a point in the parameter-space

$$\mathcal{W} := \{(A(s), B(s), \tau) \mid A(s) \in \mathbf{R}[s]^{n\times n}, B(s) \in \mathbf{R}[s]^{n\times m}, \tau \in \mathbf{R}^+\}. \tag{3.28}$$

Each element of $\mathcal{W}$ corresponds to a time-delay system as defined in (3.27), for which the property of stabilizability can be tested. So, to study the concept of genericity, we need a topology on this parameter-space $\mathcal{W}$. In this subsection we introduce two different topologies on $\mathcal{W}$. The simplest one turns $\mathcal{W}$ into a metric space, and is used to prove the genericity of stabilizability in the classical sense, i.e. for stabilizability with respect to the stability domain $\mathbf{C}^-$. Unfortunately, this topology has some unwanted consequences. First of all it does not capture some of the features of the parameter-space describing all time-delay systems. Moreover, in this topology the proof of genericity for an arbitrary stability domain $\mathbf{C}_g$ is very troublesome. To

solve both problems a more sophisticated topology is introduced: an inductive limit topology. This topology is really adapted to our specific situation of time-delay systems. In the framework of the inductive limit topology, the genericity result for the stability domain $\mathbb{C}^-$ remains true, and the generalization to arbitrary stability domains is relatively easy.

We start with the introduction of a topology on the space of polynomial matrices in $\mathbb{R}[s]^{p\times q}$. Although the degree of each element of $\mathbb{R}[s]^{p\times q}$ is bounded, it can become arbitrarily large. Let us first look at the situation in which the degree is bounded. Let $\ell \in \mathbb{N}\cup\{0\}$, and define

$$\mathcal{V}_\ell := \{P(s) \in \mathbb{R}[s]^{p\times q} \mid \deg_s(P(s)) \leq \ell\}. \tag{3.29}$$

For each element $P(s) \in \mathcal{V}_\ell$ there exist matrices $P_0, P_1, \ldots, P_\ell \in \mathbb{R}^{p\times q}$ such that

$$P(s) = \sum_{i=0}^{\ell} P_i s^i.$$

From this observation it is clear that $\mathcal{V}_\ell$ is a finite-dimensional linear space. Defining the norm of $P(s) \in \mathcal{V}_\ell$ by

$$\|P(s)\|_\ell := \sum_{i=0}^{\ell} \|P_i\|, \tag{3.30}$$

where $\|P_i\|$ denotes the operator induced matrix norm, $\mathcal{V}_\ell$ becomes a *normed* linear space.

This exercise may be carried out for each $\ell \in \mathbb{N}\cup\{0\}$ separately. Note that the spaces $\mathcal{V}_\ell$ are strongly related because for all $\ell \in \mathbb{N}\cup\{0\}$, $\mathcal{V}_\ell$ is a closed subspace of $\mathcal{V}_{\ell+1}$, and if $P(s) \in \mathcal{V}_\ell$, then $\|P(s)\|_{\ell+1} = \|P(s)\|_\ell$. Define

$$\mathcal{V} := \bigcup_{\ell=0}^{\infty} \mathcal{V}_\ell. \tag{3.31}$$

In spite of the fact that $\mathcal{V} = \mathbb{R}[s]^{p\times q}$, the choice of a topology on $\mathcal{V}$ is not so obvious.

For each element $P(s) \in \mathbb{R}[s]^{p\times q}$, there exists an $\ell \in \mathbb{N}$ such that $P(s) \in \mathcal{V}_\ell$, and thus $P(s)$ can be written as $P(s) = \sum_{i=0}^{\ell} P_i s^i$. Defining $P_i := 0$ for $i > \ell$, we can map the polynomial matrix $P(s)$ to the sequence $(P_i)_{i=0}^{\infty}$ of real matrices. In this way we obtain an explicit description of $P(s)$ in terms of its parameters. In fact, there is a 1-1 correspondence between polynomial matrices and the space $\ell_0(\mathbb{R}^{p\times q})$ consisting of all real matrix sequences with only a finite number of nonzero elements (i.e matrices with at least one nonzero entry), via the bijection:

$$\psi : \ell_0(\mathbb{R}^{p\times q}) \to \mathbb{R}[s]^{p\times q} : \psi((P_i)_{i=0}^{\infty}) = \sum_{i=0}^{\infty} P_i s^i.$$

The space $\ell_0(\mathbb{R}^{p\times q})$ is easily turned into a normed space by defining the norm of $(P_i)_{i=0}^{\infty}$ by

$$\|(P_i)_{i=0}^{\infty}\| = \sum_{i=0}^{\infty} \|P_i\|.$$

It is evident that the same norm can also be used for polynomial matrices:

Definition 3.3.1 Let $P(s)$ be a $p \times q$ matrix over $\mathbf{R}[s]$ and $(P_i)_{i=0}^{\infty} \in \ell_0(\mathbf{R}^{p\times q})$ be such that

$$P(s) = \sum_{i=0}^{\infty} P_i s^i.$$

Then the *norm* of $P(s)$ is defined as

$$\|P(s)\|_{pm} := \sum_{i=0}^{\infty} \|P_i\|, \tag{3.32}$$

where $\|P_i\|$ denotes the operator induced matrix norm of P_i for all $i \in \mathbf{N} \cup \{0\}$.

In the case of square polynomial matrices, so if $p = q$, the norm of Definition 3.3.1 turns $\mathbf{R}[s]^{p\times p}$ even into a (non-commutative) *normed ring*, because

$$\|P(s) \cdot Q(s)\|_{pm} = \sum_{i=0}^{\infty} \| \sum_{j=0}^{i} P_j Q_{i-j} \| \leq \sum_{i=0}^{\infty} \sum_{j=0}^{i} \|P_j\| \cdot \|Q_{i-j}\| \leq \|P(s)\|_{pm} \cdot \|Q(s)\|_{pm}.$$

This norm for polynomial matrices also has another very important property. Recall that in the stabilizability condition of Corollary 3.2.9 the indeterminate s is replaced by $e^{-\tau z}$. Because of Definition 3.3.1, the norm of $P(s)$ is a uniform upper bound for the norm of $P(e^{-\tau z})$ in the closed right half plane:

Lemma 3.3.2 *Let $P(s) \in \mathbf{R}[s]^{p\times q}$. Then*

$$\forall \tau > 0 \, \forall z \in \overline{\mathbf{C}^+} : \|P(e^{-\tau z})\| \leq \|P(s)\|_{pm}. \tag{3.33}$$

Proof

There exists an $\ell \in \mathbf{N}$ such that $P(s) \in \mathcal{V}_\ell$ and thus $P(s)$ can be written as

$$P(s) = \sum_{i=0}^{\ell} P_i s^i,$$

with $P_0, P_1, \ldots, P_\ell \in \mathbf{R}^{p\times q}$. Let $\tau > 0$, $z \in \overline{\mathbf{C}^+}$. Then $|e^{-\tau z}| \leq 1$ and we have

$$\|P(e^{-\tau z})\| = \| \sum_{i=0}^{\ell} P_i e^{-i\tau z} \| \leq \sum_{i=0}^{\ell} \|P_i\| \cdot |e^{-i\tau z}| \leq \sum_{i=0}^{\ell} \|P_i\| = \|P(s)\|_{pm}.$$

■

With condition (3.19) in mind, Lemma 3.3.2 has a very interesting consequence for square polynomial matrices.

Corollary 3.3.3 *Let $A(s) \in \mathbf{R}[s]^{n\times n}$. Then*

$$\forall \tau > 0 \, \forall z \in \overline{\mathbf{C}^+} \, \forall w \in \mathbf{C} \text{ s.t. } |w| > \|A(s)\|_{pm} : \operatorname{rank}(wI - A(e^{-\tau z})) = n. \tag{3.34}$$

Proof

Let $\tau > 0$ and $z \in \overline{\mathbf{C}^+}$. Let $w \in \mathbf{C}$ be such that $|w| > \|A(s)\|_{pm}$. According to Lemma 3.3.2 we have that

$$\|\frac{1}{w}A(e^{-\tau z})\| \leq \frac{1}{|w|} \cdot \|A(s)\|_{pm} < 1.$$

Therefore $I - \frac{1}{w}A(e^{-\tau z})$ is invertible (by the corresponding Neumann series). Hence $wI - A(e^{-\tau z})$ is invertible. ∎

Unfortunately, the norm of Definition 3.3.1 also has an important shortcoming. The space $\mathcal{V} = \mathbf{R}[s]^{p\times q}$ we are considering basically consists of sequences of real matrices with a finite number of nonzero elements. However, the norm we imposed on this space is a sort of ℓ_1-norm. This norm does not distinguish between sequences with a finite and an infinite number of nonzero elements. Therefore it is easy to construct a Cauchy-sequence that does not converge. Hence the normed linear space we constructed is not complete.

The problem noticed above also has some practical implications. The number of delays occurring in a time-delay system is always finite. In the normed linear space obtained with Definition 3.3.1 it is possible to find a sequence of polynomial matrices that converges to a power series in the norm $\|\cdot\|_{pm}$. Such a power series does not correspond to a time-delay system any more because it involves an infinite number of time-delays. Intuitively speaking, the space $\mathbf{R}[s]^{p\times q}$ contains too many convergent sequences when we impose the norm $\|\cdot\|_{pm}$ on it.

The problem just mentioned may be solved by introducing the inductive limit topology for $\mathcal{V} = \mathbf{R}[s]^{p\times q}$. For this, we have to return to the linear finite-dimensional spaces $\mathcal{V}_\ell$ ($\ell \in \mathbf{N} \cup \{0\}$). With its norm $\|\cdot\|_\ell$ defined in (3.30) each $\mathcal{V}_\ell$ is a normed linear space. The sequence $(\mathcal{V}_\ell)_{\ell=0}^{\infty}$ satisfies the following properties:

(i) $\mathcal{V}_\ell \subset \mathcal{V}_{\ell+1}$,

(ii) $\forall P(s) \in \mathcal{V}_\ell\colon\ \|P(s)\|_{\ell+1} = \|P(s)\|_\ell$,

(iii) $\mathcal{V}_\ell$ is a closed subspace of $\mathcal{V}_{\ell+1}$ with respect to the norm $\|\cdot\|_{\ell+1}$.

So, according to [13, Chapter IV, Definitions 5.1 and 5.12], the pair $(\mathcal{V}, (\mathcal{V}_\ell)_{\ell\in\mathbf{N}\cup\{0\}})$ is a (strict) *inductive system.* Therefore we can apply the same approach as in [13, Chapter IV, Section 5] to construct a topology on $\mathcal{V}$, using the topologies on $\mathcal{V}_\ell$ generated by their respective norms $\|\cdot\|_\ell$. This topology is called the (strict) *inductive limit topology.*

Definition 3.3.4 Let $\mathcal{B}$ denote the collection of all subsets $B \subset \mathcal{V}$ that satisfy the following conditions:

(i) $0 \in B$,

(ii) $\forall \ell \in \mathbf{N} \cup \{0\}\colon\ B \cap \mathcal{V}_\ell$ is open w.r.t. the norm $\|\cdot\|_\ell$ on $\mathcal{V}_\ell$,

(iii) B is convex and symmetric around 0.

Define $\mathcal{T}$ to be the collection of subsets of $\mathcal{V}$ consisting of all sets $\mathcal{O} \subset \mathcal{V}$ with the property

$$\forall P(s) \in \mathcal{O}\ \exists B \in \mathcal{B}:\ P(s) + B \subset \mathcal{O}. \tag{3.35}$$

Then the pair $(\mathcal{V}, \mathcal{T})$ is a locally convex space. $\mathcal{T}$ is called the (strict) *inductive limit topology*, and $(\mathcal{V}, \mathcal{T})$ is said to be the (strict) *inductive limit* of $(\mathcal{V}_\ell)_{\ell \in \mathbf{N} \cup \{0\}}$.

The claim that the pair $(\mathcal{V}, \mathcal{T})$ constitutes a locally convex space is not trivial. A proof of this fact can be found in [13, p. 120].

At first sight, the definition of the inductive limit topology looks very abstract. It is completely determined by the collection $\mathcal{T}$ containing all the open subsets of $\mathcal{V}$. However, the relationship between $\mathcal{T}$ and the topology generated by the norm $\|\cdot\|_{pm}$ of Definition 3.3.1 is not difficult to discover.

Lemma 3.3.5 *Let $\mathcal{O} \subset \mathcal{V}$ be an open set in $\mathcal{V}$ w.r.t. the norm $\|\cdot\|_{pm}$ on $\mathcal{V}$, i.e.*

$$\forall P_0(s) \in \mathcal{O}\ \exists \varepsilon > 0:\ \{P(s) \in \mathcal{V} \mid \|P(s) - P_0(s)\|_{pm} < \varepsilon\} \subset \mathcal{O}. \tag{3.36}$$

Then $\mathcal{O} \in \mathcal{T}$.

Proof

Let $P_0(s) \in \mathcal{O}$. Then there exists an $\ell_1 \in \mathbf{N}$ such that $P_0(s) \in \mathcal{V}_{\ell_1}$. Choose $\varepsilon > 0$ such that the inclusion of (3.36) is satisfied. Define

$$B := \bigcup_{\ell=0}^{\infty} \{P(s) \in \mathcal{V}_\ell \mid \|P(s)\|_\ell < \varepsilon\}.$$

It is obvious that $B \in \mathcal{B}$. Let $P(s) \in B$. Then there exists an $\ell_2 \in \mathbf{N}$ such that $P(s) \in \mathcal{V}_{\ell_2}$ and $\|P(s)\|_{\ell_2} < \varepsilon$. Define $Q(s) := P_0(s) + P(s)$. Clearly $Q(s) \in \mathcal{V}_{\ell_1+\ell_2}$ and we have

$$\|Q(s) - P_0(s)\|_{pm} = \|Q(s) - P_0(s)\|_{\ell_1+\ell_2} = \|P(s)\|_{\ell_1+\ell_2} = \|P(s)\|_{\ell_2} < \varepsilon.$$

So according to (3.36), $Q(s) = P_0(s) + P(s) \in \mathcal{O}$. Since $P(s) \in B$ was arbitrary, we conclude that $P_0(s) + B \subset \mathcal{O}$. ∎

Let $\mathcal{T}_{pm}$ denote the topology generated by the norm $\|\cdot\|_{pm}$ of Definition 3.3.1. Then Lemma 3.3.5 indicates that

$$\mathcal{T}_{pm} \subsetneqq \mathcal{T},$$

i.e. every subset of $\mathcal{V}$ that belongs to $\mathcal{T}_{pm}$ is also an element of $\mathcal{T}$. The next example illustrates that the topology $\mathcal{T}$ is really stronger than $\mathcal{T}_{pm}$, and contains open subsets of $\mathcal{V}$ that are not open in the topology $\mathcal{T}_{pm}$.

Example 3.3.6 Consider the following subset B of $\mathcal{V}$:

$$B := \bigcup_{\ell=0}^{\infty} \{P(s) \in \mathcal{V}_\ell \mid \|P(s)\|_\ell < \frac{1}{\ell+1}\}.$$

It is obvious that $B \in \mathcal{B}$. Then it follows from [13, p. 120, Lemma 5.5] that $B \in \mathcal{T}$. However, when we consider B in the topology generated by the norm $\|\cdot\|_{pm}$, B is

not open any more. This may be seen as follows. It is clear that $0 \in B$. Let $\varepsilon > 0$. Then there exists an $\ell \in \mathbf{N}$ such that $\frac{1}{\ell+1} < \varepsilon$. Let $P_0 \in \mathbf{R}^{p\times q}$ be such that $\|P_0\| = 1$ and define

$$P(s) := \frac{2}{2\ell+2} \cdot P_0 \cdot s^{2\ell+2}.$$

Then $\|P(s)\|_{pm} = \frac{2}{2\ell+2} \cdot \|P_0\| = \frac{1}{\ell+1} < \varepsilon$. But, on the other hand, $P(s) \in \mathcal{V}_{2\ell+2}$ and

$$\|P(s)\|_{2\ell+2} = \frac{2}{2\ell+2} > \frac{1}{2\ell+2}.$$

So $P(s) \notin B$, and thus 0 is not an internal point of B w.r.t. the norm $\|\cdot\|_{pm}$.

Corollary 3.3.7 *The inductive limit topology $\mathcal{T}$ on $\mathcal{V}$ is stronger than the topology $\mathcal{T}_{pm}$ generated by the norm $\|\cdot\|_{pm}$, i.e.*

$$\mathcal{T}_{pm} \subsetneq \mathcal{T}$$

■

Corollary 3.3.7 implies that the notion of convergence in the inductive limit topology $\mathcal{T}$ is stronger than in $\mathcal{T}_{pm}$.

Proposition 3.3.8 *Let $(P_n(s))_{n\in\mathbf{N}}$ be a sequence in $\mathcal{V}$ that converges to $P(s)$ in the inductive limit topology $\mathcal{T}$. Then*

(i) $\exists \ell \in \mathbf{N}: \ [P(s) \in \mathcal{V}_\ell \text{ and } \forall n \in \mathbf{N}: \ P_n(s) \in \mathcal{V}_\ell]$,

(ii) $\lim_{n\to\infty} \|P_n(s) - P(s)\|_\ell = 0.$

■

The proof of Proposition 3.3.8 follows from the fact that a convergent sequence is bounded; subsequent application of Proposition 5.16 in [13, Chapter IV, Section 5, p. 123] yields the desired result.

The statement of Proposition 3.3.8 is the main motivation for the introduction of the inductive limit topology: in this way we exactly obtain the notion of convergence we are interested in. A sequence of polynomial matrices can only converge if the whole sequence has a fixed bounded degree. This implies that a convergent sequence of delay systems can only converge to a delay system containing a finite number of time-delays. The undesired behaviour that is possible in the topology $\mathcal{T}_{pm}$ generated by the norm $\|\cdot\|_{pm}$ of Definition 3.3.1 cannot occur. Moreover, using the same argument as in the proof of Proposition 3.3.8, it is easily seen that in the inductive limit topology, every Cauchy-sequence in $\mathcal{V}$ converges; hence $\mathcal{V}$ is complete in this topology. So we have found a topology on $\mathcal{V}$ that really fits out purposes.

We conclude this introduction of the inductive limit topology with the following proposition on the continuity of linear mappings.

Proposition 3.3.9 *Let $T : \mathcal{V} \longrightarrow \mathcal{V}$ be a linear transformation. Then T is continuous w.r.t. the topology $\mathcal{T}$ if and only if T is sequentially continuous w.r.t. $\mathcal{T}$.*

■

For a proof we refer to [13, p. 123, Corollary 5.18]. Note that in general the equivalence of continuity and sequential continuity, which is well known for Banach spaces, does not hold for locally convex spaces.

With the topologies on polynomial matrices described above, also the parameter-space $\mathcal{W}$ may be equipped with a suitable topology. Of course this can be done in two different ways.

Definition 3.3.10 Let $\Sigma_1 = (A_1(s), B_1(s), \tau_1)$ and $\Sigma_2 = (A_2(s), B_2(s), \tau_2)$ be two elements of the parameter-space $\mathcal{W}$. Then the *distance* between Σ_1 and Σ_2 is defined as

$$d_{\mathcal{W}}(\Sigma_1, \Sigma_2) := \|A_1(s) - A_2(s)\|_{pm} + \|B_1(s) - B_2(s)\|_{pm} + |\tau_1 - \tau_2|. \tag{3.37}$$

With this distance function $d_{\mathcal{W}}$, $\mathcal{W}$ becomes a metric space.

But also the inductive limit topology may be applied:

Definition 3.3.11 Consider the parameter-space

$$\mathcal{W} = \{(A(s), B(s), \tau) \mid A(s) \in \mathbb{R}[s]^{n\times n}, B(s) \in \mathbb{R}[s]^{n\times m}, \tau \in \mathbb{R}^+\},$$

and define by $\mathcal{T}_A$ the inductive limit topology on $\mathcal{V}_A = \mathbb{R}[s]^{n\times n}$, and by $\mathcal{T}_B$ the inductive limit topology on $\mathcal{V}_B = \mathbb{R}[s]^{n\times m}$. The topology on $\mathbb{R}^+$ generated by the norm $|\cdot|$ is denoted by $\mathcal{T}_{|\cdot|}$. It is clear that

$$\mathcal{W} = \mathcal{V}_A \times \mathcal{V}_B \times \mathbb{R}^+.$$

The topology $\mathcal{T}$ on $\mathcal{W}$ is defined as the product topology of the three factors:

$$\mathcal{T} := \mathcal{T}_A \times \mathcal{T}_B \times \mathcal{T}_{|\cdot|}.$$

With abuse of terminology, $\mathcal{T}$ is called the *inductive limit topology* on $\mathcal{W}$. A subset $\mathcal{O} \subset \mathcal{W}$ belongs to $\mathcal{T}$ if and only if $\mathcal{O}$ is of the form $\mathcal{O} = \mathcal{O}_A \times \mathcal{O}_B \times \mathcal{O}_{|\cdot|}$, where $\mathcal{O}_A \in \mathcal{T}_A$, $\mathcal{O}_B \in \mathcal{T}_B$ and $\mathcal{O}_{|\cdot|} \in \mathcal{T}_{|\cdot|}$.

Our final task is to show that the topology on $\mathcal{W}$ generated by the metric $d_{\mathcal{W}}$ and the inductive limit topology $\mathcal{T}$, both reflect our intuitive ideas on genericity. For each triple $\Sigma = (A(s), B(s), \tau)$ in $\mathcal{W}$, it is possible to check the stabilizability of the corresponding time-delay system with help of Corollary 3.2.9. Let $\mathbb{C}_g$ be a stability domain, and denote the set of all systems in $\mathcal{W}$ which are stabilizable w.r.t. $\mathbb{C}_g$ by S_g:

$$S_g := \{(A(s), B(s), \tau) \in \mathcal{W} \mid \forall z \in \mathbb{C}\backslash\mathbb{C}_g : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n\}.$$

Now, stabilizability w.r.t. $\mathbb{C}_g$ is called *generic* if S_g contains a subset S that is an open and dense subset of the parameter-space $\mathcal{W}$. In the topologies defined above this implies that the set S, and thus certainly the set S_g, covers almost the whole space $\mathcal{W}$:

(i) S is open: Every system Σ in S is contained in an open neighbourhood of Σ that completely belongs to S,

(ii) S is a dense subset of $\mathcal{W}$: Every element $\Sigma \in \mathcal{W}$ can be approximated arbitrarily close by a sequence of systems in S.

We conclude that both topologies lead to a formal description of genericity on the parameter-space $\mathcal{W}$, which looks very natural and is completely in accordance with our intuitive ideas on this concept. Note that the condition that S is open is somewhat stronger in the topology generated by the metric $d_{\mathcal{W}}$ since this topology does not contain as many open sets as $\mathcal{T}$. On the other hand, the condition that S must be dense is stronger in the inductive limit topology because there the notion of convergence is much stronger.

Remark 3.3.12 It is important to note that we have defined topologies on the parameter-space $\mathcal{W}$ of all time-delay systems, and not on the time-delay systems itself. For a delay system with a specified input-output behaviour there exist many different parametrizations. In the metric $d_{\mathcal{W}}$ the distance between two different parametrizations of the same input-output behaviour is greater than zero, while also the inductive limit topology does not reflect that two different parametrizations characterize essentially the same system. This might seem strange, but it is no problem at all because the stabilizability condition of Corollary 3.2.9 is a rank condition on the *parametrization* of a system. To show that time-delay systems are generically stabilizable, it suffices to show that this rank condition is generically satisfied on the parameter-space $\mathcal{W}$ describing all time-delay systems. Input-output behaviours are not directly involved in this question.

In almost the same way as for polynomial matrices, it is possible to regard the polynomial ring $\mathbf{R}[s,z]$ as a linear space, and to define a topology on this space. Also in this situation it is possible to introduce the inductive limit topology. However, in the sequel this topology is not used explicitly, and therefore we confine ourselves to the definition of a norm.

Definition 3.3.13 Let $p(s,z) \in \mathbf{R}[s,z]$, and write $p(s,z)$ as

$$p(s,z) = \sum_{i=0}^{\ell} \sum_{j=0}^{k} p_{ij} s^j z^i. \tag{3.38}$$

Then the norm of $p(s,z)$ is defined as

$$\|p(s,z)\|_p := \sum_{i=0}^{\ell} \sum_{j=0}^{k} |p_{ij}|. \tag{3.39}$$

With this norm, $\mathbf{R}[s,z]$ becomes a normed ring. For all $q \in \mathbf{R}$ and $i,j \in \mathbf{N} \cup \{0\}$, we have

$$\|p(s,z) \cdot q s^j z^i\|_p = |q| \cdot \|p(s,z)\|_p.$$

So, when $q(s,z) = \sum_{i=0}^{m} \sum_{j=0}^{n} q_{ij} s^j z^i$, the triangle inequality implies that

$$\|p(s,z) \cdot q(s,z)\|_p \leq \|p(s,z)\|_p \cdot \|q(s,z)\|_p. \tag{3.40}$$

Besides the topology generated by this norm, Definition 3.3.13 has several other interesting consequences. Analogously to the polynomial matrix case, there exists a 1-1 correspondence between polynomials $p(s,z)$ in two indeterminates and analytic functions $p(e^{-\tau z},z)$ that are obtained after substitution of $e^{-\tau z}$ for the indeterminate s. Moreover, in Section 3.1 we have seen that characteristic polynomials of this form determine the stability of a time-delay system. In the same way as in Lemma 3.3.2, the norm $\|p(s,z)\|_p$ contains information on some interesting properties of the function $p(e^{-\tau z},z)$ in the right half plane. For example, it is a good measure for the magnitude of $|p(e^{-\tau z},z)|$ in a bounded part of $\overline{\mathbb{C}^+}$:

Lemma 3.3.14 *Let $p(s,z) \in \mathbb{R}[s,z]$, and assume that the degree of p in z is n, i.e.*

$$p(s,z) = \sum_{i=0}^{n}\sum_{j=0}^{k} p_{ij}s^j z^i.$$

Furthermore assume that there is a $j \in \{0,\ldots,k\}$ such that $p_{nj} \neq 0$. Let $M > 0$ and $\varepsilon > 0$. If

$$\|p(s,z)\|_p < \varepsilon \cdot \frac{M-1}{M^{n+1}-1}, \tag{3.41}$$

then

$$\forall \tau > 0\ \forall z \in \overline{\mathbb{C}^+}\ s.t.\ |z| \leq M : |p(e^{-\tau z},z)| < \varepsilon. \tag{3.42}$$

Proof

Assume that (3.41) is satisfied. Let $\tau > 0$ and $z \in \overline{\mathbb{C}^+}$ be such that $|z| \leq M$. Then $|e^{-\tau z}| \leq 1$ and thus

$$|p(e^{-\tau z},z)| = \left|\sum_{i=0}^{n}\sum_{j=0}^{k} p_{ij}e^{-j\tau z}z^i\right| \leq \sum_{i=0}^{n}\sum_{j=0}^{k}|p_{ij}|\cdot|e^{-j\tau z}|\cdot|z|^i \leq$$

$$\leq \sum_{i=0}^{n}(\sum_{j=0}^{k}|p_{ij}|)\cdot|z|^i < \sum_{i=0}^{n}\varepsilon\frac{M-1}{M^{n+1}-1}\cdot|z|^i \leq \varepsilon\frac{M-1}{M^{n+1}-1}\cdot\sum_{i=0}^{n}M^i = \varepsilon.$$

■

Finally, we elaborate on the relationship between polynomial matrices in one indeterminate on the one hand, and 2-D polynomials on the other hand. Here the characteristic polynomial plays the leading part. In fact we prove that the map χ:

$$\chi : \mathbb{R}[s]^{n\times n} \longrightarrow \mathbb{R}[s,z] : \chi(A(s)) = \det(zI - A(s)), \tag{3.43}$$

is continuous with respect to the norms $\|\cdot\|_{pm}$ on $\mathbb{R}[s]^{n\times n}$ and $\|\cdot\|_p$ on $\mathbb{R}[s,z]$ as defined in (3.32) and (3.39), respectively.

Proposition 3.3.15 *Let $A(s) \in \mathbb{R}[s]^{n\times n}$. Then*

$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall B(s) \in \mathbb{R}[s]^{n\times n} :$$

$$\|A(s) - B(s)\|_{pm} < \delta \Longrightarrow \|\det(zI - A(s)) - \det(zI - B(s))\|_p < \varepsilon. \tag{3.44}$$

Proof

First we consider the map from the $n \times n$ matrices over $\mathbf{R}[s,z]$ to their determinants. Recall from formula (3.40) that the norm of Definition 3.3.13 on the space of 2-D polynomials is sub-multiplicative. Regarding $\mathbf{R}[s,z]$ as a normed linear space, this implies that the multiplication of two polynomials in $\mathbf{R}[s,z]$ defines a continuous mapping from $\mathbf{R}[s,z] \times \mathbf{R}[s,z]$ to $\mathbf{R}[s,z]$. Moreover, the definition of a norm immediately implies that addition is a continuous mapping too. Now the determinant of a matrix is simply a sum of products of entries of the matrix. So, if we define a norm on the $n \times n$ matrices over $\mathbf{R}[s,z]$ entry-wise, the mapping from a square 2-D polynomial matrix to its determinant is continuous. Formally, this may be stated in the following way. Let $P(s,z) \in \mathbf{R}[s,z]^{n\times n}$. Then for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all matrices $Q(s,z) \in \mathbf{R}[s,z]^{n\times n}$ satisfying

$$\forall i,j \in \{1,\ldots,n\} : \|p_{ij}(s,z) - q_{ij}(s,z)\|_p < \delta, \tag{3.45}$$

(where $p_{ij}(s,z)$ and $q_{ij}(s,z)$ denote the $(i,j)^{\text{th}}$ entry of $P(s,z)$ and $Q(s,z)$, respectively), we have

$$\|\det(P(s,z)) - \det(Q(s,z))\|_p < \varepsilon. \tag{3.46}$$

To prove the claim, we consider $P(s,z) := (zI - A(s))$. Let $\varepsilon > 0$. Choose $\delta > 0$ such that (3.45) implies (3.46). Let now $B(s) \in \mathbf{R}[s]^{n\times n}$ be such that $\|A(s) - B(s)\|_{pm} < \delta$. There exists an $\ell \in \mathbf{N}$ such that $A(s)$ and $B(s)$ can be written as

$$A(s) = \sum_{i=0}^{\ell} A_i s^i, \qquad B(s) = \sum_{i=0}^{\ell} B_i s^i,$$

with A_i and $B_i \in \mathbf{R}^{n\times n}$ $(i = 0,\ldots,\ell)$. Let $A_{jk}(s)$ and A_{ijk} denote the $(j,k)^{\text{th}}$ entry of $A(s)$ and A_i respectively. The same notation is used for $B(s)$ and B_i.

Let $j,k \in \{1,\ldots,n\}$. Then

$$\begin{aligned}
\|(zI - A(s))_{jk} - (zI - B(s))_{jk}\|_p &= \|B(s)_{jk} - A(s)_{jk}\|_p = \|(B(s) - A(s))_{jk}\|_p = \\
&= \|\sum_{i=0}^{\ell}(B_{ijk} - A_{ijk}) \cdot s^i\|_p = \sum_{i=0}^{\ell} |B_{ijk} - A_{ijk}| \leq \\
&\leq \sum_{i=0}^{\ell} \|B_i - A_i\| = \|A(s) - B(s)\|_{pm} < \delta,
\end{aligned}$$

where we used the fact that the absolute value of every entry of a constant matrix is smaller than or equal to the operator induced norm of that matrix.

Now apply (3.46) with $(zI - B(s))$ playing the role of $Q(s,z)$. Then we obtain:

$$\|\det(zI - A(s)) - \det(zI - B(s))\|_p < \varepsilon.$$

This completes the proof. ∎

3.3.2 On the robustness of the property of stabilizability

In this subsection the first part of our genericity result is proved. Initially, we only consider stability domains of the form $\mathbb{C}_g = \mathbb{C}_{-\alpha} = \{z \in \mathbb{C} \mid \operatorname{Re} z < -\alpha\}$. We prove that the set $S_{-\alpha}$ consisting of all elements in $\mathcal{W}$ that are stabilizable w.r.t. $\mathbb{C}_{-\alpha}$ is itself an open subset of $\mathcal{W}$. This implies that stabilizability w.r.t $\mathbb{C}_{-\alpha}$ is a robust property: it is preserved under small perturbations of the parameters describing the system. Later on it turns out that this result is also enough to prove the genericity of stabilizability w.r.t. arbitrary stability domains.

We start considering stabilizability w.r.t. the stability domain $\mathbb{C}^-$, and use the topology on $\mathcal{W}$ generated by the metric $d_{\mathcal{W}}$. Given a nominal stabilizable system, an upper bound is derived for the distance between this nominal system, and all perturbed systems that are allowed: if the distance between a perturbed system and the nominal system is smaller than this upper bound, the perturbed system is still stabilizable. Since this upper bound is always larger than zero, this immediately implies that the set S_0 consisting of all delay systems in $\mathcal{W}$ that are stabilizable w.r.t. $\mathbb{C}^-$ is an open subset of $\mathcal{W}$.

Then we switch over to the inductive limit topology. Since this topology is stronger, the robustness result for stabilizability w.r.t. the stability domain $\mathbb{C}^-$ still holds. But in this setting this result can be generalized to arbitrary open left half planes. Suppose that $\alpha \in \mathbb{R}$ is given. In the inductive limit topology $\mathcal{T}$ the set $S_{-\alpha}$ describing all time-delay systems that are stabilizable with respect to the stability domain $\mathbb{C}_{-\alpha} = \{z \in \mathbb{C} \mid \operatorname{Re} z < -\alpha\}$ is an open subset of $\mathcal{W}$, or in the terminology of Definition 3.3.11: $S_{-\alpha} \in \mathcal{T}$.

We start this program with the following well-known result on linear operators (see e.g. [61]) that is also valid in the far more general context of Banach algebras. In the proof of one of the main results of this subsection we only need this restricted version.

Lemma 3.3.16 *Let $A_0 \in \mathbb{C}^{p\times q}$, with $q > p$. Assume that A_0 is of full row rank, so A_0 is right-invertible, with right-inverse B_0. Then*

$$\forall A \in \mathbb{C}^{p\times q} : \|A - A_0\| < \frac{1}{\|B_0\|} \Longrightarrow A \text{ is right-invertible.} \tag{3.47}$$

■

Recall from Corollary 3.2.9 that the stabilizability condition for time-delay systems is a full rank condition on a matrix in the variable z, that has to be satisfied for all $z \in \mathbb{C}\backslash\mathbb{C}_g$. Therefore it is apparent that Lemma 3.3.16 is helpful to prove the robustness of this condition. We first study the special case in which the stability domain $\mathbb{C}_g$ equals the open left half plane $\mathbb{C}^-$.

Theorem 3.3.17 *Let $\Sigma_0 = (A_0(s), B_0(s), \tau_0)$ be a point in $\mathcal{W}$, and assume that the time-delay system (3.27) corresponding to Σ_0 is stabilizable w.r.t. the stability domain $\mathbb{C}^-$, i.e.*

$$\forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A_0(e^{-\tau_0 z}) | B_0(e^{-\tau_0 z})) = n.$$

Then there exists a $\rho > 0$ such that in the topology generated by the metric $d_{\mathcal{W}}$ all systems Σ in the ball around Σ_0 with radius ρ,

$$\text{Ball}(\Sigma_0, \rho) := \{\Sigma \in \mathcal{W} \mid d_{\mathcal{W}}(\Sigma, \Sigma_0) < \rho\},$$

are stabilizable w.r.t. $\mathbb{C}^-$.

Proof

First of all, there exists an $\ell \in \mathbb{N}$ such that $A_0(s)$ and $B_0(s)$ can be written as

$$A_0(s) = \sum_{i=0}^{\ell} A_i s^i, \qquad B_0(s) = \sum_{i=0}^{\ell} B_i s^i.$$

Next, define G as

$$G := \{z \in \mathbb{C} \mid \text{Re}\, z \geq 0 \text{ and } |z| \leq \|A_0(s)\|_{pm} + 1\}. \tag{3.48}$$

From Theorem 3.2.8 it follows that the matrix $(zI - A_0(s)|B_0(s))$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$, so that $(zI - A_0(e^{-\tau_0 z})|B_0(e^{-\tau_0 z}))$ has a right-inverse $T(z)$ which is analytic on $\overline{\mathbb{C}^+}$. Since G is a compact subset of $\overline{\mathbb{C}^+}$, $T(z)$ is bounded on G and

$$K := \max\{\|T(z)\| \mid z \in G\} \tag{3.49}$$

is well-defined.

Choose

$$\rho := \min\left(1, \frac{1}{4K}, \frac{1}{\|A_0(s)\|_{pm} + 1} \cdot \frac{1}{4K\ell} \cdot \min\left(\frac{1}{\|A_0(s)\|_{pm}}, \frac{1}{\|B_0(s)\|_{pm}}\right)\right). \tag{3.50}$$

Then clearly $\rho > 0$. We show that all systems in $\text{Ball}(\Sigma_0, \rho)$ are stabilizable.

Let $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$ be such that $d_{\mathcal{W}}(\Sigma, \Sigma_0) < \rho$. The proof that Σ is stabilizable w.r.t. $\mathbb{C}^-$ is divided into two parts: the case $|z| > \|A_0(s)\|_{pm} + 1$, and the case $|z| \leq \|A_0(s)\|_{pm} + 1$.

Let $z \in \overline{\mathbb{C}^+}$, and assume that $|z| > \|A_0(s)\|_{pm} + 1$. Since $d_{\mathcal{W}}(\Sigma, \Sigma_0) < \rho$ we have

$$\|A(s)\|_{pm} \leq \|A_0(s)\|_{pm} + \|A(s) - A_0(s)\|_{pm} < \|A_0(s)\|_{pm} + \rho.$$

Using (3.50) it follows that

$$|z| > \|A_0(s)\|_{pm} + 1 \geq \|A_0(s)\|_{pm} + \rho > \|A(s)\|_{pm}$$

According to Corollary 3.3.3 (with $w = z$) this implies that

$$\text{rank}(zI - A(e^{-\tau z})) = n,$$

and thus certainly $\text{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n$.

The second case is more involved. Let $z \in \overline{\mathbb{C}^+}$, $|z| \leq \|A_0(s)\|_{pm} + 1$. We start by proving that

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\| < \frac{1}{K}. \tag{3.51}$$

First note that

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\| \leq$$
$$\leq \|A_0(e^{-\tau_0 z}) - A(e^{-\tau z})\| + \|B(e^{-\tau z}) - B_0(e^{-\tau_0 z})\|. \tag{3.52}$$

Now clearly

$$\|A_0(e^{-\tau_0 z}) - A(e^{-\tau z})\| \leq \|A_0(e^{-\tau_0 z}) - A_0(e^{-\tau z})\| + \|A_0(e^{-\tau z}) - A(e^{-\tau z})\|. \tag{3.53}$$

Since $\|A(s) - A_0(s)\|_{pm} \leq d_{\mathcal{W}}(\Sigma, \Sigma_0) < \rho$ and $\rho \leq \frac{1}{4K}$, it follows from Lemma 3.3.2 that the second term in (3.53) is bounded from above:

$$\|A_0(e^{-\tau z}) - A(e^{-\tau z})\| \leq \|A_0(s) - A(s)\|_{pm} < \rho \leq \frac{1}{4K}. \tag{3.54}$$

To estimate the other term, we apply the Mean Value Theorem:

$$\|A_0(e^{-\tau_0 z}) - A_0(e^{-\tau z})\| = \|\sum_{i=0}^{\ell} A_i \cdot (e^{-i\tau_0 z} - e^{-i\tau z})\| = \|\sum_{i=0}^{\ell} A_i \cdot iz \int_{\tau_0}^{\tau} e^{-i\xi z} d\xi\|$$
$$\leq \sum_{i=0}^{\ell} \|A_i\| \cdot i|z| \int_{\tau_0}^{\tau} |e^{-i\xi z}| d\xi \leq \|A_0(s)\|_{pm} \cdot \ell \cdot (\|A_0(s)\|_{pm} + 1) \cdot |\tau - \tau_0| \leq$$
$$\leq \|A_0(s)\|_{pm} \cdot \ell \cdot (\|A_0(s)\|_{pm} + 1) \cdot \rho < \frac{1}{4K}, \tag{3.55}$$

where in the last inequality (3.50) was used. (Note that in the derivation of formula (3.55) i is a summation variable and not the complex number i). In a completely analogous way we may prove that

$$\|B(e^{-\tau z}) - B_0(e^{-\tau_0 z})\| \leq \|B(e^{-\tau z}) - B_0(e^{-\tau z})\| + \|B_0(e^{-\tau z}) - B_0(e^{-\tau_0 z})\| < \frac{1}{2K}.$$

Using the previous inequality together with (3.54) and (3.55) in (3.52), we obtain (3.51).

Next recall that $(zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))$ is right-invertible, with right-inverse $T(z)$. Moreover $\|T(z)\| \leq K$. So

$$\|(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) - (zI - A_0(e^{-\tau_0 z}) \mid B_0(e^{-\tau_0 z}))\| < \frac{1}{K} \leq \frac{1}{\|T(z)\|}.$$

After application of Lemma 3.3.16 we immediately see that $(zI - A(e^{-\tau z}) \mid B(e^{-\tau z}))$ is right-invertible, hence

$$\operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n.$$

This completes the proof. ∎

According to Theorem 3.3.17, the set of all parametrizations of time-delay systems that are stabilizable w.r.t. $\mathbb{C}^-$, is an open subset of $\mathcal{W}$ in the topology generated by the metric $d_{\mathcal{W}}$. This immediately implies that the same result holds in the inductive limit topology, because we have seen in Lemma 3.3.5 that this topology is stronger.

Corollary 3.3.18 *The set of all parametrizations of time-delay systems that are stabilizable w.r.t. the stability domain $\mathbb{C}^-$, is an open subset of $\mathcal{W}$ in the inductive limit topology $\mathcal{T}$ of Definition 3.3.11:*

$$\{(A(s), B(s), \tau) \in \mathcal{W} \mid \forall z \in \overline{\mathbb{C}^+} : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n\} \in \mathcal{T}. \quad \blacksquare$$

Our next step is to generalize this result on stabilizability w.r.t. the stability domain $\mathbb{C}^-$ to a result on the stabilizability w.r.t. an arbitrary open left half plane $\mathbb{C}_{-\alpha}$. For this purpose we introduce the following operator.

Definition 3.3.19 Let $\alpha \in \mathbb{R}$. Then the operator $H_\alpha : \mathcal{W} \to \mathcal{W}$ is defined by

$$H_\alpha(A(s), B(s), \tau) = (\alpha I + A(e^{\tau\alpha}s), B(e^{\tau\alpha}s), \tau). \tag{3.56}$$

Denote the set of all $n \times n$ and $n \times m$ matrices over $\mathbb{R}[s]$ of degree smaller than or equal to ℓ by $\mathcal{V}_{n\times n,\ell}$ and $\mathcal{V}_{n\times m,\ell}$, respectively. If $A(s) \in \mathcal{V}_{n\times n,\ell}$ and $B(s) \in \mathcal{V}_{n\times m,\ell}$, i.e. if $A(s)$ and $B(s)$ are of the form

$$A(s) = \sum_{i=0}^{\ell} A_i s^i, \qquad B(s) = \sum_{i=0}^{\ell} B_i s^i,$$

and if $\tau > 0$ is given, then the operator H_α maps $A(s)$ and $B(s)$ to

$$\alpha I + \sum_{i=0}^{\ell} A_i e^{i\tau\alpha} s^i \quad \text{and} \quad \sum_{i=0}^{\ell} B_i e^{i\tau\alpha} s^i,$$

respectively. It is obvious that for all $\alpha \in \mathbb{R}$, the operator H_α is *affine*. Moreover, H_α is invertible with inverse $H_{-\alpha}$. The main reason for our interest in the operator H_α is the following result.

Lemma 3.3.20 *Let $\alpha \in \mathbb{R}$, and consider a delay system $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$. Then we have:*

$$(A(s), B(s), \tau) \text{ is stabilizable w.r.t. the stability domain } \mathbb{C}_{-\alpha}, \tag{3.57}$$

$\Longleftrightarrow$

$$H_\alpha(A(s), B(s), \tau) \text{ is stabilizable w.r.t. the stability domain } \mathbb{C}^-. \tag{3.58}$$

Proof

"$\Rightarrow$" Assume that $\Sigma = (A(s), B(s), \tau)$ is stabilizable w.r.t. $\mathbb{C}_{-\alpha}$, i.e.

$$\forall z \in \mathbb{C}\backslash\mathbb{C}_{-\alpha} : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n.$$

Let $z \in \overline{\mathbb{C}^+}$ and define $w := z - \alpha$. Then $w \in \mathbb{C}\backslash\mathbb{C}_{-\alpha}$, and thus

$$\operatorname{rank}(wI - A(e^{-\tau w}) \mid B(e^{-\tau w})) = n.$$

Substitution of $w = z - \alpha$ in the last formula yields

$$\operatorname{rank}(zI - (\alpha I + A(e^{\tau\alpha}e^{-\tau z})) \mid B(e^{\tau\alpha}e^{-\tau z})) = n.$$

Since $z \in \overline{\mathbb{C}^+}$ was arbitrary, the definition of H_α implies that $H_\alpha(A(s), B(s), \tau)$ is stabilizable w.r.t. $\mathbb{C}^-$.

"$\Leftarrow$" Assume that $H_\alpha(A(s), B(s), \tau)$ is stabilizable w.r.t. $\mathbb{C}^-$, so

$$\forall z \in \overline{\mathbb{C}^+}: \ \operatorname{rank}(zI - (\alpha I + A(e^{\tau\alpha}e^{-\tau z})) \mid B(e^{\tau\alpha}e^{-\tau z})) = n.$$

Let $z \in \mathbb{C}\backslash\mathbb{C}_{-\alpha}$ and define $w := z + \alpha$. Then $w \in \overline{\mathbb{C}^+}$ and substituting $w = z + \alpha$ in the last formula, we obtain

$$\operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n.$$

This holds for all $z \in \mathbb{C}\backslash\mathbb{C}_{-\alpha}$, so $(A(s), B(s), \tau)$ is stabilizable w.r.t. $\mathbb{C}_{-\alpha}$. ∎

According to Lemma 3.3.20, it is possible to translate the problem of stabilizability of a system $\Sigma = (A(s), B(s), \tau)$ w.r.t. an open left half plane $\mathbb{C}_{-\alpha}$ to the problem of stabilizability of the transformed system $H_\alpha(A(s), B(s), \tau)$ w.r.t. the half plane $\mathbb{C}^-$. Now the crucial point is that in the inductive limit topology this transformation operator H_α is continuous.

Proposition 3.3.21 *Let $\alpha \in \mathbb{R}$ and consider the operator $H_\alpha : \mathcal{W} \to \mathcal{W}$, as defined in (3.56). In the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$, the operator H_α is continuous.*

Proof

First note that H_α is an affine operator; without the term αI in the first component, the operator H_α would have been linear. In this situation, Proposition 3.3.9 is still valid, and it suffices to show that H_α is sequentially continuous.

Let $(A_0(s), B_0(s), \tau_0)$ be an element of $\mathcal{W}$, and let $(A_j(s), B_j(s), \tau_j)_{j\in\mathbb{N}}$ be a sequence in $\mathcal{W}$ converging to $(A_0(s), B_0(s), \tau_0)$. Then $A_j(s)$ converges to $A_0(s)$, and according to Proposition 3.3.8 there exists an $\ell_1 \in \mathbb{N}$ such that

(i) $A_0(s) \in \mathcal{V}_{n\times n,\ell_1}$ and $\forall j \in \mathbb{N}: \ A_j(s) \in \mathcal{V}_{n\times n,\ell_1}$,

(ii) $\lim_{j\to\infty} \|A_j(s) - A_0(s)\|_{\ell_1} = 0$.

In the same way there exists an $\ell_2 \in \mathbb{N}$ such that

(iii) $B_0(s) \in \mathcal{V}_{n\times m,\ell_2}$ and $\forall j \in \mathbb{N}: \ B_j(s) \in \mathcal{V}_{n\times m,\ell_2}$,

(iv) $\lim_{j\to\infty} \|B_j(s) - B_0(s)\|_{\ell_2} = 0$,

and finally

(v) $\lim_{j\to\infty} |\tau_j - \tau_0| = 0$.

Define $\ell := \max(\ell_1, \ell_2)$ and denote for $j \in \mathbb{N}\cup\{0\}$ the point $H_\alpha(A_j(s), B_j(s), \tau_j)$ in $\mathcal{W}$ by $(\bar{A}_j(s), \bar{B}_j(s), \tau_j)$. Then $\bar{A}_j(s) \in \mathcal{V}_{n\times n,\ell}$ and $\bar{B}_j(s) \in \mathcal{V}_{n\times m,\ell}$ and we only have to show that

$$\lim_{j\to\infty} \|\bar{A}_j(s) - \bar{A}_0(s)\|_\ell = 0, \tag{3.59}$$

$$\lim_{j\to\infty} \|\bar{B}_j(s) - \bar{B}_0(s)\|_\ell = 0. \tag{3.60}$$

For all $j \in \mathbb{N} \cup \{0\}$ the matrix $A_j(s)$ may be written as

$$A_j(s) = \sum_{i=0}^{\ell} A_{ji} s^i.$$

So

$$\begin{aligned} \bar{A}_j(s) - \bar{A}_0(s) &= \sum_{i=0}^{\ell} (A_{ji} e^{i\tau_j \alpha} - A_{0i} e^{i\tau_0 \alpha}) \cdot s^i = \\ &= \sum_{i=0}^{\ell} [(A_{ji} - A_{0i}) e^{i\tau_j \alpha} + A_{0i}(e^{i\tau_j \alpha} - e^{i\tau_0 \alpha})] \cdot s^i. \end{aligned} \tag{3.61}$$

Let $K := \sup(\{e^{\ell \tau_j \alpha} \mid j \in \mathbb{N}\} \cup \{1\})$. Since $(\tau_j)_{j=1}^{\infty}$ is a convergent sequence, it is bounded from above and thus K exists. Next, define for $j \in \mathbb{N}$: $a_j := \max\{|e^{i\tau_j \alpha} - e^{i\tau_0 \alpha}| \mid i = 0, 1, \ldots, \ell\}$. Then $\lim_{j \to \infty} a_j = 0$ because τ_j converges to τ_0 when j tends to infinity. Using both the definitions of K and $(a_j)_{j \in \mathbb{N}}$ in (3.61) we obtain

$$\|\bar{A}_j(s) - \bar{A}_0(s)\|_\ell \leq K \cdot \|A_j(s) - A_0(s)\|_\ell + a_j \cdot \|A_0(s)\|_\ell \longrightarrow 0 \quad (j \to \infty),$$

and indeed (3.59) is satisfied. In a completely analogous way (3.60) is proved. After application of Proposition 3.3.8, we conclude that in the inductive limit topology $\mathcal{T}$

$$H_\alpha(A_j(s), B_j(s), \tau_j) \longrightarrow H_\alpha(A_0(s), B_0(s), \tau_0) \quad (j \to \infty).$$

Thus H_α is continuous. ∎

The proof of Proposition 3.3.21 shows another advantage of the inductive limit topology. In this topology the proof of the continuity of H_α is relatively easy. In the topology on $\mathcal{W}$ generated by the metric $d_{\mathcal{W}}$ the derivation of a similar result seems very troublesome.

Using the previous results on the operator H_α, it is possible to generalize Corollary 3.3.18 to arbitrary open left half planes.

Corollary 3.3.22 *Let $\alpha \in \mathbb{R}$, and let $\mathbb{C}_{-\alpha}$ denote the stability domain $\{z \in \mathbb{C} \mid \operatorname{Re} z < -\alpha\}$. In the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$, the set of all parametrizations of time-delay systems that are stabilizable w.r.t. $\mathbb{C}_{-\alpha}$ is open:*

$$\{(A(s), B(s), \tau) \in \mathcal{W} \mid \forall z \in \mathbb{C} \backslash \mathbb{C}_{-\alpha} : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n\} \in \mathcal{T}.$$

Proof

Let S_0 and $S_{-\alpha}$ denote the sets of all parametrizations of delay systems in $\mathcal{W}$ that are stabilizable w.r.t. $\mathbb{C}^-$ and $\mathbb{C}_{-\alpha}$, respectively. According to Corollary 3.3.18, S_0 is an open subset of $\mathcal{W}$. From Lemma 3.3.20 it follows that

$$S_{-\alpha} = H_\alpha^{\leftarrow}(S_0),$$

i.e. $S_{-\alpha}$ is the inverse image of S_0 under the mapping H_α. Since H_α is a continuous mapping and S_0 is open, this implies that $S_{-\alpha}$ is also an open subset of $\mathcal{W}$. In other words $S_{-\alpha} \in \mathcal{T}$. ∎

Remark 3.3.23 In this subsection, only stabilizability w.r.t. an arbitrary open left half plane was considered. This specialization looks rather restrictive, but it is sufficient to prove our genericity result for arbitrary stability domains $\mathbf{C}_g$. Recall from Definition 3.1.2 that for every stability domain $\mathbf{C}_g$ there exists an $\hat{\alpha} \in \mathbb{R}$ such that the left half plane $\mathbf{C}_{-\hat{\alpha}}$ is contained in $\mathbf{C}_g$. It is our objective to prove that the set $S_{-\hat{\alpha}}$ consisting of all delay systems in $\mathcal{W}$ that are stabilizable w.r.t. $\mathbf{C}_{-\hat{\alpha}}$ is an open and dense subset of $\mathcal{W}$. Since $\mathbf{C}_{-\hat{\alpha}} \subset \mathbf{C}_g$, the set S_g of all parametrizations of time-delay systems that are stabilizable w.r.t. $\mathbf{C}_g$, contains the set $S_{-\hat{\alpha}}$. So, after proving the genericity of stabilizability w.r.t. arbitrary open left half planes, the same result for general stability domains as described in Definition 3.1.2 follows immediately.

Remark 3.3.24 At first sight, Theorem 3.3.17 and Corollaries 3.3.18 and 3.3.22 seem to have very much in common with the result of Pandolfi in [74, Section 5]. However, there are several differences. First of all, Pandolfi's result is obtained within the framework of distributed parameter systems. As already noted in Chapter 1, this is a more general class of systems. Moreover, in the setting of Pandolfi, perturbations are described in a completely different way. In [74], systems are described with help of linear operators acting on the (infinite-dimensional) state space and the (finite-dimensional) input space. Perturbations of these systems are considered as perturbations of these operators and they are measured in the operator norm induced by the norms on the input and state space. In this context, the robustness of stabilizability against these perturbations is studied. In our approach, the state space does not play any role. We describe perturbations within a topology on the space $\mathcal{W}$ of all parametrizations of time-delay systems. Pandolfi's result certainly holds in a much more general setting, but our result is more suitable to capture the concept of genericity for the time-delay systems under consideration. So the results of this subsection are not an immediate consequence of the work of Pandolfi. Although these results look very similar, there is not a clear relationship, and the differences are more conspicuous.

3.3.3 Some results on matrices of analytic functions

In this subsection, we make some preparations for the second part of the proof of our genericity result. In this proof (which is given in the next subsection) we need some properties of matrices of analytic functions. Especially the relationship between the rank of these matrices and their determinants is studied. This relationship is clarified using projection matrices. Since these results are also interesting in themselves, we isolate them from the rest, and devote this subsection to this subject.

The first lemma describes how projections can be useful for the computation of the determinant of a matrix.

Lemma 3.3.25 *Let A_1 and A_2 be two arbitrary square matrices, and E a projection. Define $\rho(E) := \operatorname{rank}(E)$. Let α be an indeterminate. Then*

$$\det(\alpha E A_1 + (I - E)A_2) = \alpha^{\rho(E)} \cdot \det(EA_1 + (I - E)A_2). \tag{3.62}$$

Proof

Choose a basis $(x_1, \ldots, x_n)$ such that $\text{range}(E) = \langle x_1, \ldots, x_{\rho(E)} \rangle$ and $\text{range}(I - E) = \langle x_{\rho(E)+1}, \ldots, x_n \rangle$. Let $B = \begin{pmatrix} B_1 \\ \hline B_2 \end{pmatrix}$ denote the matrix of $EA_1 + (I - E)A_2$ with respect to this new basis, where B_1 consists of the first $\rho(E)$ rows of B, and B_2 of the last $n - \rho(E)$ rows. Then $\begin{pmatrix} \alpha B_1 \\ \hline B_2 \end{pmatrix}$ is the matrix of $\alpha EA_1 + (I - E)A_2$ with respect to this basis. Hence

$$\begin{aligned} \det(\alpha EA_1 + (I - E)A_2) &= \det \begin{pmatrix} \alpha B_1 \\ \hline B_2 \end{pmatrix} = \alpha^{\rho(E)} \cdot \det \begin{pmatrix} B_1 \\ \hline B_2 \end{pmatrix} \\ &= \alpha^{\rho(E)} \cdot \det(EA_1 + (I - E)A_2). \end{aligned}$$

■

Let $Q(z)$ be an $n \times n$ matrix over the ring of analytic functions on $\mathbb{C}$, i.e. all entries of $Q(z)$ are analytic functions in z. Define

$$p(z) := \det(Q(z)). \tag{3.63}$$

It is clear that in a point $\lambda \in \mathbb{C}$, the matrix $Q(\lambda)$ is of full rank if and only if $p(\lambda) \neq 0$. Using a suitable projection E, it is also possible to obtain more precise information on the rank of $Q(\lambda)$ from the determinant function $p(z)$ when $p(\lambda) = 0$.

Proposition 3.3.26 *Let $Q(z)$ be an $n \times n$ matrix of analytic functions, and $p(z) = \det(Q(z))$. Assume that for a certain $\lambda \in \mathbb{C}$: $p(\lambda) = 0$. Define the matrix of analytic functions $Q_1(z)$ as*

$$Q_1(z) := \frac{Q(z) - Q(\lambda)}{z - \lambda} = \sum_{j=1}^{\infty} \frac{1}{j!} Q^{(j)}(\lambda)(z - \lambda)^{j-1}.$$

Let E be a projection such that $EQ(\lambda) = 0$. Then

$$p(z) = (z - \lambda)^{\rho(E)} \cdot \det(EQ_1(z) + (I - E)Q(z)). \tag{3.64}$$

Moreover, if $\rho(E) = k$, then

$$\begin{cases} p^{(j)}(\lambda) = 0 \text{ for } j = 1, \ldots, k-1, \\ p^{(k)}(\lambda) = k! \cdot \det(EQ'(\lambda) + (I - E)Q(\lambda)). \end{cases} \tag{3.65}$$

Proof

$Q(z)$ may be written as $Q(z) = Q(\lambda) + (z - \lambda)Q_1(z)$. Therefore,

$$\begin{aligned} p(z) &= \det(EQ(z) + (I - E)Q(z)) = \det((z - \lambda)EQ_1(z) + (I - E)Q(z)) = \\ &= (z - \lambda)^{\rho(E)} \cdot \det(EQ_1(z) + (I - E)Q(z)), \end{aligned}$$

where in the last step Lemma 3.3.25 is used. The result on the derivatives of $p(z)$ in λ when $\rho(E) = k$, is an easy consequence of formula (3.64) and the definition of $Q_1(z)$. ■

Corollary 3.3.27 *Let $Q(z)$ be an $n \times n$ matrix of analytic functions, and $p(z) := \det(Q(z))$. Then*

$$\forall \lambda \in \mathbb{C} : \left[\begin{array}{rcl} p(\lambda) & = & 0 \\ p'(\lambda) & \neq & 0 \end{array} \right\} \Longrightarrow \operatorname{rank}(Q(\lambda)) = n-1 \Bigg]. \tag{3.66}$$

Proof

Let $\lambda \in \mathbb{C}$ be such that $p(\lambda) = 0$ and $p'(\lambda) \neq 0$. Choose a projection E with $\operatorname{range}(Q(\lambda)) = \ker(E)$. Since $Q(\lambda)$ is singular, $\rho(E) = \operatorname{rank}(E) \geq 1$. According to Proposition 3.3.26 we have

$$p(z) = (z-\lambda)^{\rho(E)} \cdot \det(EQ_1(z) + (I-E)Q(z)).$$

Suppose that $\rho(E) > 1$. Then $p'(\lambda) = 0$. This contradicts our assumption, and therefore $\rho(E) = 1$. This implies that $\dim(\operatorname{range}(Q(\lambda))) = n-1$. ∎

Remark 3.3.28 From Proposition 3.3.26 it follows that ($p(\lambda) = 0$ and $p'(\lambda) \neq 0$) is a sufficient condition for $Q(\lambda)$ to have rank $n-1$. However, it is not a necessary condition because it is also possible that $\operatorname{rank}(Q(\lambda)) = n-1$ while $p'(\lambda) = 0$. In that case the matrix $EQ'(\lambda) + (I-E)Q(\lambda)$ is singular.

In Subsection 3.3.4, it turns out that for our purposes matrices $Q(z)$ of analytic functions, for which the determinant $p(z)$ has only simple zeros, are of special interest. According to Corollary 3.3.27 this type of matrices has the property: if $p(\lambda) = 0$ then $\operatorname{rank}(Q(\lambda)) = n-1$.

Let $Q(z)$ be given, and assume that $\lambda \in \mathbb{C}$ is such that $p(\lambda) = \det(Q(\lambda)) = 0$ and also $p'(\lambda) = 0$. Then it is possible to perturb $Q(z)$ in such a way that λ becomes a simple zero of $p(z)$. However, before we can prove this result, we first need a lemma that describes how a constant matrix can be perturbed in order to increase its rank.

Lemma 3.3.29 *Let A be an $n \times n$ matrix over $\mathbb{C}$, and assume that $\operatorname{rank}(A) = \ell$. For each $j \in \{1, \ldots, n-\ell\}$, there exists a matrix $B \in \mathbb{R}^{n \times n}$ satisfying the following properties:*

(i) $\|B\| = 1$ *and* $\operatorname{rank}(B) = j$.

(ii) $\forall \alpha, \beta \neq 0 :$ $\operatorname{range}(\alpha A + \beta B) = \operatorname{range}(A) \oplus \operatorname{range}(B)$.

Proof

Let $e_1, \ldots, e_n$ denote the standard basis in $\mathbb{C}^n$. Then there exists a permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that

$$\operatorname{range}(A) = \langle Ae_{\pi(1)}, \ldots, Ae_{\pi(\ell)} \rangle. \tag{3.67}$$

Choose vectors $e_{i_1}, \ldots, e_{i_{n-\ell}}$ from the standard basis satisfying

$$\langle e_{i_1}, \ldots, e_{i_{n-\ell}} \rangle \oplus \operatorname{range}(A) = \mathbb{C}^n \tag{3.68}$$

Let $j \in \{1, \ldots, n-\ell\}$, and define B as

$$\left\{ \begin{array}{rcll} Be_{\pi(k)} & = & 0 & \text{for } k = 1, \ldots, \ell, \ell+j+1, \ldots, n, \\ Be_{\pi(k)} & = & e_{i_{k-\ell}} & \text{for } k = \ell+1, \ldots, \ell+j. \end{array} \right.$$

With this choice of B, it is obvious that *(i)* is satisfied.

From the construction of B it is immediately clear that $\text{range}(A) \cap \text{range}(B) = \{0\}$. Moreover, the inclusion $\text{range}(\alpha A + \beta B) \subset \text{range}(A) + \text{range}(B)$ is trivial. So, to prove *(ii)*, we only have to show the correctness of the other inclusion.

Let $x_1 \in \text{range}(A)$. Then there exists an $y_1 \in \langle e_{\pi(1)}, \ldots, e_{\pi(\ell)} \rangle$ such that $x_1 = Ay_1$. But clearly $By_1 = 0$. Hence

$$(\alpha A + \beta B)(\frac{1}{\alpha} y_1) = Ay_1 + \frac{\beta}{\alpha} By_1 = x_1,$$

and $x_1 \in \text{range}(\alpha A + \beta B)$.

Let $x_2 \in \text{range}(B)$. Then there exists an $y_2 \in \langle e_{\pi(\ell+1)}, \ldots, e_{\pi(\ell+j)} \rangle$ such that $By_2 = x_2$. Since $Ay_2 \in \text{range}(A)$, there exists an $y_3 \in \langle e_{\pi(1)}, \ldots, e_{\pi(\ell)} \rangle$ such that $Ay_2 = Ay_3$. Now

$$(\alpha A + \beta B) \cdot \frac{1}{\beta}(y_2 - y_3) = \frac{\alpha}{\beta}(Ay_2 - Ay_3) + By_2 - By_3 = By_2 = x_2,$$

and $x_2 \in \text{range}(\alpha A + \beta B)$. This completes the proof of *(ii)*. ∎

At this stage all ingredients to prove the main result of this subsection are available. The result describes how a matrix of analytic functions may be perturbed in order to reduce the multiplicity of one of the zeros of its determinant to 1.

Proposition 3.3.30 *Let $Q(z)$ be an $n \times n$ matrix of analytic functions, and define $p(z) = \det(Q(z))$. Assume that $\lambda \in \mathbf{C}$ satisfies $p(\lambda) = 0$. Let $g(z)$ be an analytic function such that $g'(\lambda) \neq 0$.*

Then for each $\varepsilon > 0$ there exists an $n \times n$ polynomial matrix $\Delta(s)$ over $\mathbf{R}[s]$, that satisfies the following properties (where $\tilde{Q}(z) := Q(z) + \Delta(g(z))$ and $\tilde{p}(z) := \det(\tilde{Q}(z))$):

(i) $\|\Delta(s)\|_{pm} < \varepsilon$,

(ii) $\deg(\Delta(s)) \leq 1$ *if* $g(\lambda)$ *is real, and* $\deg(\Delta(s)) \leq 2$ *if* $g(\lambda)$ *is complex,*

(iii) $\tilde{p}(\lambda) = 0$ *and* $\tilde{p}'(\lambda) \neq 0$.

Proof

If $p'(\lambda) \neq 0$, the proof is trivial: take $\Delta(s) = 0$.

Assume $p'(\lambda) = 0$. Let $\varepsilon > 0$. If $\text{rank}(Q(\lambda)) = n - 1$, define $B_1 := 0$. Otherwise, choose a matrix B_1 according to Lemma 3.3.29, with $\|B_1\| = 1$ and $\text{rank}(B_1) = n - 1 - \text{rank}(Q(\lambda))$ in such a way that

$$\forall \alpha \neq 0 : \ \text{range}(Q(\lambda) + \alpha B_1) = \text{range}(Q(\lambda)) \oplus \text{range}(B_1).$$

This implies that for all $\alpha \neq 0$: $\text{rank}(Q(\lambda) + \alpha B_1) = n - 1$.

Fix $\alpha := \frac{1}{3}\varepsilon$ and apply Lemma 3.3.29 again, but now to the matrix $Q(\lambda) + \alpha B_1$. In this way we find a matrix B_2 (possibly depending on α), satisfying $\|B_2\| = 1$, $\text{rank}(B_2) = 1$, and

$$\forall \beta \neq 0 : \ \text{range}(Q(\lambda) + \alpha B_1 + \beta B_2) = \text{range}(Q(\lambda)) \oplus \text{range}(B_1) \oplus \text{range}(B_2).$$

So for every $\beta \neq 0$, the matrix $(Q(\lambda) + \alpha B_1 + \beta B_2)$ has rank n.

Let E denote the projection on range(B_2) along range$(Q(\lambda) + \alpha B_1)$, so that $E(Q(\lambda) + \alpha B_1) = 0$ and $EB_2 = B_2$. Then $\rho(E) = \text{rank}(E) = 1$. Define $Q_\alpha(z) := Q(z) + \alpha B_1$, and $p_\alpha(z) := \det(Q_\alpha(z))$. So $p_\alpha(\lambda) = \det(Q(\lambda) + \alpha B_1) = 0$, and using formula (3.65) from Proposition 3.3.26 we obtain:

$$p'_\alpha(\lambda) = \det(EQ'_\alpha(\lambda) + (I - E)Q_\alpha(\lambda)) = \det(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)).$$

Next we show that

$$\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) \subset \ker(Q_\alpha(\lambda)).$$

Let $x \in \ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda))$. Then $EQ'(\lambda)x = 0$ and $(I - E)Q_\alpha(\lambda)x = 0$. By construction $EQ_\alpha(\lambda) = 0$, hence the last equality implies that $Q_\alpha(\lambda)x = 0$, and we conclude that $x \in \ker(Q_{alpha}(\lambda))$.

Since $\dim(\ker(Q_\alpha(\lambda))) = 1$, we now divide the problem into two different cases:

Case 1: $\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) = \{0\}$.

Then $p'_\alpha(\lambda) = \det(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) \neq 0$, and $\Delta(s) := \alpha B_1$ satisfies both *(ii)* and *(iii)* and also *(i)* because $\|\Delta(s)\|_{pm} = \|\alpha B_1\| \leq \frac{1}{3}\varepsilon$.

Case 2: $\ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda)) = \ker(Q_\alpha(\lambda))$.

If $g(\lambda)$ is real, define for all $\beta \in \mathbb{R}\backslash\{0\}$

$$\Delta_\beta(s) := \alpha B_1 + \beta(s - g(\lambda))B_2;$$

if $g(\lambda)$ is complex, define for all $\beta \in \mathbb{R}\backslash\{0\}$

$$\Delta_\beta(s) := \alpha B_1 + \beta(s - g(\lambda))(s - \overline{g(\lambda)})B_2.$$

Then in each case $\Delta_\beta(s) \in \mathbb{R}[s]^{n\times n}$, and moreover *(ii)* is satisfied.

Let $\beta \in \mathbb{R}\backslash\{0\}$ and define $\tilde{Q}(z) := Q(z) + \Delta_\beta(g(z))$. Then $\tilde{Q}(\lambda) = Q(\lambda) + \alpha B_1$, and in both the real and the complex case there exist a $\gamma \neq 0$ such that $\tilde{Q}'(\lambda) = Q'(\lambda) + \gamma B_2$. Since $\tilde{Q}(\lambda) = Q(\lambda) + \alpha B_1$ is singular, we still have that $\tilde{p}(\lambda) = 0$, and according to Proposition 3.3.26,

$$\tilde{p}'(\lambda) = \det(E\tilde{Q}'(\lambda) + (I - E)\tilde{Q}(\lambda)) = \det(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda)).$$

Assume that $x \in \ker(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda))$. Then $x \in \ker(Q_\alpha(\lambda))$. So by assumption $x \in \ker(EQ'(\lambda) + (I - E)Q_\alpha(\lambda))$. Moreover we have that $EB_2 = B_2$, and thus we obtain

$$\gamma B_2 x = (E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda))x - (EQ'(\lambda) + (I - E)Q_\alpha(\lambda))x = 0.$$

So $(Q_\alpha(\lambda) + \gamma B_2)x = 0$. By construction $Q_\alpha(\lambda) + \gamma B_2 = Q(\lambda) + \alpha B_1 + \gamma B_2$ has full rank, and thus $x = 0$. This implies that $\text{rank}(E(Q'(\lambda) + \gamma B_2) + (I - E)Q_\alpha(\lambda)) = n$. Therefore $\tilde{p}'(\lambda) \neq 0$, and $\tilde{Q}(z)$ satisfies condition *(iii)* for all $\beta \neq 0$.

In order to satisfy *(i)*, we choose

$$\beta := \frac{1}{4}\varepsilon \cdot \min(\frac{1}{g(\lambda)}, 1),$$

when $g(\lambda)$ is real, and

$$\beta := \frac{1}{8}\varepsilon \cdot \min\left(\frac{1}{g(\lambda)+\overline{g(\lambda)}}, \frac{1}{g(\lambda)\cdot\overline{g(\lambda)}}, 1\right),$$

when $g(\lambda)$ is complex. Then it is easily verified that $\|\Delta_\beta(s)\|_{pm} < \varepsilon$. This completes the proof. ∎

Remark 3.3.31 If the matrix $Q(z)$ of analytic functions has the property that $\overline{Q(\overline{z})} = Q(z)$, also its determinant $p(z)$ has this property. This implies that λ is a zero of $p(z)$ of multiplicity k, if and only if $\overline{\lambda}$ is a zero of $p(z)$ of the same multiplicity. Note that if also $\overline{g(\overline{z})} = g(z)$, and $g(\lambda)$ is complex, the reduction process described in the proof of Proposition 3.3.30 reduces the multiplicity of both the zeros λ and $\overline{\lambda}$ to 1 in only one step. Although in general a perturbation matrix $\Delta(s)$ of degree 2 is needed to reduce the multiplicity, this matrix handles both the zeros λ and $\overline{\lambda}$ at the same time.

Remark 3.3.32 Corollary 3.3.27 and Proposition 3.3.30 are formulated in a very general context of matrices over analytic functions, but in the next subsection they are only used for a very special case. It is clear that for the stabilizability properties of a system Σ corresponding to a point $(A(s), B(s), \tau) \in \mathcal{W}$, the matrix $(zI - A(e^{-\tau z}))$ is very important. Therefore the results of this subsection are applied to the case $Q(z) = (zI - A(e^{-\tau z}))$ and $g(z) = e^{-\tau z}$. Then clearly $g'(\lambda) = -\tau e^{-\tau\lambda} \neq 0$ for all $\lambda \in \mathbb{C}$. In this perspective Proposition 3.3.30 describes how the matrix $A(s) \in \mathbb{R}[s]^{n\times n}$ has to be perturbed in such a way that $(zI - (A(e^{-\tau z}) + \Delta(e^{-\tau z})))$ satisfies the condition of Corollary 3.3.27. Note that the degree of the perturbation matrix $\Delta(s)$ is bounded above by 2. Therefore it does not matter whether perturbations are considered within the norm $\|\cdot\|_{pm}$ on $\mathbb{R}[s]^{n\times n}$ or in the inductive limit topology: the result holds in both cases.

3.3.4 Approximation by stabilizable time-delay systems

In this subsection, the second and final part of our genericity result is proved. We show that the subset S_g of $\mathcal{W}$, consisting of all parametrizations of time-delay systems that are stabilizable w.r.t. the stability domain $\mathbb{C}_g$, is a dense subset of $\mathcal{W}$. This means that in every open neighbourhood of a non-stabilizable system $\Sigma \in \mathcal{W}$, there exists a point $\tilde{\Sigma} \in \mathcal{W}$ corresponding to a time-delay system that is stabilizable w.r.t. $\mathbb{C}_g$. In other words, there exists a sequence of stabilizable systems $(\Sigma_j)_{j\in\mathbb{N}}$ in S_g that converges to Σ. In this subsection, such an approximation by stabilizable time-delay systems is constructed explicitly.

The strategy to prove this result is the same as in Subsection 3.3.2. First we consider the problem for the special case of stabilizability w.r.t. the stability domain $\mathbb{C}^-$, and use the topology on $\mathcal{W}$ generated by the metric $d_{\mathcal{W}}$. When this is proved, the generalization to the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$ is very straightforward. Finally, this result for the stability domain $\mathbb{C}^-$, can be shifted to any other open left half plane with help of the operator H_α of Definition 3.3.19. Since every stability domain $\mathbb{C}_g$ contains an open left half plane, this immediately implies that approximation by $\mathbb{C}_g$-stabilizable delay systems is also possible.

The main idea of the proof is as follows. Suppose that $\Sigma = (A(s), B(s), \tau)$ is a point in $\mathcal{W}$, such that the corresponding time-delay system is not stabilizable w.r.t. $\mathbb{C}^-$. Using Corollary 3.3.3, it is easy to show that for all matrices $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$, the analytic function $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$. Using Rouché's Theorem, and Corollary 3.3.27 and Proposition 3.3.30 of the previous subsection, it is possible to prove that for every $\varepsilon > 0$, there exists a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n\times n}$ such that $\|A(s) - A_\varepsilon(s)\|_{pm} < \frac{1}{2}\varepsilon$ and

$$\forall z \in \overline{\mathbb{C}^+} : \left[\operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) < n \Longrightarrow \operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) = n - 1 \right].$$

So, in all points $z \in \overline{\mathbb{C}^+}$ where the matrix $(zI - A_\varepsilon(e^{-\tau z}))$ loses rank, it only loses rank 1. This loss of rank has to be compensated by the matrix $B(s)$. Therefore this matrix is perturbed in such a way that the perturbed version $B_\varepsilon(s)$ satisfies the inequality $\|B_\varepsilon(s) - B(s)\|_{pm} < \frac{1}{2}\varepsilon$ and is such that

$$\forall z \in \overline{\mathbb{C}^+} : \left[\operatorname{rank}(zI - A_\varepsilon(e^{-\tau z})) < n \Longrightarrow \operatorname{rank}(zI - A_\varepsilon(e^{-\tau z}) \mid B_\varepsilon(e^{-\tau z})) = n \right].$$

Since the analytic function $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ only has a finite number of zeros in the closed right half plane, it is possible to satisfy this condition. In this way we find a stabilizable time-delay system $\Sigma_\varepsilon = (A_\varepsilon(s), B_\varepsilon(s), \tau)$ such that $d_{\mathcal{W}}(\Sigma, \Sigma_\varepsilon) < \varepsilon$, and the proof is complete.

The rest of this subsection only consists of a detailed elaboration of the scheme of the proof given above. The first lemma describes the location of the zeros of the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ corresponding to the square polynomial matrix $A(s)$.

Lemma 3.3.33 *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Then the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ only has a finite number of zeros in the closed right half plane $\overline{\mathbb{C}^+}$. Moreover, all zeros of $p(z)$ in $\overline{\mathbb{C}^+}$ are located within the semi-disc*

$$D := \{z \in \mathbb{C} \mid \operatorname{Re} z \geq 0 \text{ and } |z| \leq \|A(s)\|_{pm}\}. \tag{3.69}$$

Proof

From Corollary 3.3.3 (with w equal to z), we know that

$$\forall z \in \overline{\mathbb{C}^+}, |z| > \|A(s)\|_{pm} : \operatorname{rank}(zI - A(e^{-\tau z})) = n.$$

This implies that $p(z)$ has no zeros in $\overline{\mathbb{C}^+} \backslash D$. So all zeros of $p(z)$ in $\overline{\mathbb{C}^+}$ are contained in D. Since D is a compact set, the analytic function $p(z)$ only has a finite number of zeros inside D. This proves the claim. ∎

Corollary 3.3.34 *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let $\ell \in \mathbb{N}$ be such that $A(s)$ can be written as*

$$A(s) = \sum_{j=0}^{\ell} A_j s^j.$$

Let $b \in \mathbf{R}$, and define the polynomial matrix $A_b(s)$ by

$$A_b(s) := -b \cdot I + \sum_{j=0}^{\ell} A_j e^{-j\tau b} s^j.$$

Then the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ only has a finite number of zeros in the half plane $\{z \in \mathbf{C} \mid \operatorname{Re} z \geq b\}$. Moreover, all zeros of $p(z)$ in this half plane are located within the semi-disc

$$D_b := \{z \in \mathbf{C} \mid \operatorname{Re} z \geq b \text{ and } |z - b| \leq \|A_b(s)\|_{pm}\}.$$

Proof

Define the analytic function $p_b(z) := \det(zI - A_b(e^{-\tau z}))$. It is easily seen that $p_b(z)$ is just a shifted version of $p(z)$:

$$p_b(z) = \det(zI - A_b(e^{-\tau z})) = \det((z+b)I - A(e^{-\tau(z+b)})) = p(z+b).$$

Therefore λ is a zero of $p(z)$ if and only if $(\lambda - b)$ is a zero of $p_b(z)$. So the zeros of $p(z)$ in the half plane $\{z \in \mathbf{C} \mid \operatorname{Re} z \geq b\}$ correspond to the zeros of $p_b(z)$ in $\overline{\mathbf{C}^+}$. Application of Lemma 3.3.33 to the matrix $A_b(s)$ yields the desired result. ∎

Remark 3.3.35 If the polynomial matrix $A(s)$ in Corollary 3.3.34 is completed to a system $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$, then $A_b(s)$ is simply the first component of $H_{-b}(A(s), B(s), \tau)$.

In the proof of one of the main results of this subsection, we have to assume that the analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on the imaginary axis. Moreover, we are interested in suitable perturbations of the corresponding polynomial matrix $A(s)$. The next proposition describes how a polynomial matrix $A(s)$ can be perturbed in such a way that the corresponding analytic function $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on the imaginary axis. Combining the results of Lemma 3.3.33 and Corollary 3.3.34, it is shown that an arbitrary small perturbation of $A(s)$ is enough to satisfy this condition.

Proposition 3.3.36 *Let $A(s) \in \mathbf{R}[s]^{n \times n}$ and $\tau > 0$ be given. Let $\varepsilon > 0$. Then there exists a polynomial matrix $A_1(s) \in \mathbf{R}[s]^{n \times n}$ satisfying the following properties:*

(i) $\|A(s) - A_1(s)\|_{pm} < \varepsilon$,

(ii) $\deg(A_1(s)) = \deg(A(s))$,

(iii) $p_1(z) := \det(zI - A_1(e^{-\tau z}))$ *has no zeros on the imaginary axis.*

Proof

We construct the matrix $A_1(s)$ in the following way. From Corollary 3.3.34 it follows that the analytic function $p(z)$ only has a finite number of zeros in any arbitrary right half plane. So there exists a $b < 0$ such that $p(z)$ has no zeros in the strip

$$\{z \in \mathbf{C} \mid b < \operatorname{Re} z < 0\}.$$

There exists an $\ell \in \mathbb{N}$ such that $A(s)$ can be written as $A(s) = \sum_{j=0}^{\ell} A_j s^j$. Define for all $\delta \in (b, 0)$:

$$A_\delta(s) := -\delta I + \sum_{j=0}^{\ell} A_j e^{-j\tau\delta} s^j.$$

Then $\|A_\delta(s) - A(s)\|_{pm} \leq \delta + \sum_{j=0}^{\ell} \|A_j\| \cdot |e^{-j\tau\delta} - 1|$, and it is easily verified that there exists a $\hat{\delta} \in (b, 0)$ such that $\|A_{\hat{\delta}}(s) - A(s)\|_{pm} < \varepsilon$. Define $A_1(s) := A_{\hat{\delta}}(s)$. Then clearly *(i)* and *(ii)* hold. To prove *(iii)*, recall that $p(z)$ has no zeros on the line $\{z \in \mathbb{C} \mid \operatorname{Re} z = \hat{\delta}\}$. Let $\omega \in \mathbb{R}$. With exactly the same argument on the shifting of zeros as in Corollary 3.3.34, we have

$$p_1(\imath\omega) = \det(\imath\omega I - A_1(e^{-\imath\tau\omega})) = \det((\imath\omega + \hat{\delta}) \cdot I - A(e^{-\tau(\imath\omega+\hat{\delta})})) = p(\imath\omega + \hat{\delta}) \neq 0,$$

and thus *(iii)* is satisfied too. ∎

The next theorem is a restatement of a well-known result from complex analysis. It plays a crucial role in the rest of this subsection because it describes in what way small perturbations of an analytic function influence the location of its zeros. For a proof of this result we refer to e.g. [83, Theorem 10.43].

Theorem 3.3.37 (Rouché's Theorem) *Let f and g be two functions that are analytic inside and on a bounded Jordan curve J. Suppose that f and g have no zeros on J. Denote by N_f and N_g the total number of zeros of f and g inside J, also counting multiplicities. Then*

$$[\, \forall z \in J : |f(z) - g(z)| < |f(z)| \,] \implies N_g = N_f. \tag{3.70}$$

∎

It is evident that under the same conditions as in Theorem 3.3.37, and after the definition of $\delta := \min\{|f(z)| \mid z \in J\}$, the condition $|f(z) - g(z)| < \delta$ implies that f and g have the same number of zeros inside J. This observation is exploited in the next lemma.

Lemma 3.3.38 *Let $A(s) \in \mathbb{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let J be a bounded Jordan curve in $\overline{\mathbb{C}^+}$ such that $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on J. Then there exists an $\bar{\varepsilon} > 0$ such that for all polynomial matrices $\tilde{A}(s) \in \mathbb{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$, the characteristic function $\tilde{p}(z) := \det(zI - \tilde{A}(e^{-\tau z}))$ corresponding to $\tilde{A}(s)$, has the same number of zeros within J as $p(z)$ (counting multiplicities), and no zeros on J.*

Proof

Define $p_c(s, z) := \det(zI - A(s))$. Then $p_c(s, z) \in \mathbb{R}[s, z]$, and the degree of $p_c(s, z)$ in z is n. Define

$$\delta := \min\{|p(z)| \mid z \in J\}, \tag{3.71}$$

and $M := 1+\max\{|z| \mid z \in J\}$. Now apply Proposition 3.3.15. Choose an $\bar{\varepsilon} > 0$ such that for all matrices $\tilde{A}(s) \in \mathbf{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$, the following inequality holds:

$$\|p_c(s,z) - \tilde{p}_c(s,z)\|_p < \delta \frac{M-1}{M^{n+1}-1}. \tag{3.72}$$

Here $\tilde{p}_c(s,z)$ denotes the characteristic polynomial $\det(zI - \tilde{A}(s))$ of $\tilde{A}(s)$, which is also of degree n in z. We show that for $\bar{\varepsilon}$, the claim of Lemma 3.3.38 holds.

Let $\tilde{A}(s) \in \mathbf{R}[s]^{n\times n}$ be such that $\|A(s) - \tilde{A}(s)\|_{pm} < \bar{\varepsilon}$. First apply Lemma 3.3.14 to $r(s,z) := p_c(s,z) - \tilde{p}_c(s,z)$ and use inequality (3.72). In this way we obtain:

$$\forall z \in \overline{\mathbb{C}^+}, |z| \leq M : \ |p(z) - \tilde{p}(z)| < \delta. \tag{3.73}$$

So in particular $|p(z) - \tilde{p}(z)| < \delta$ for all $z \in J$. Apparently $\tilde{p}(z)$ has no zeros on J. (Otherwise there would be a $\lambda \in J$ such that $|p(\lambda)| < \delta$, which contradicts definition (3.71)). Finally, because both $p(z)$ and $\tilde{p}(z)$ are analytic functions without zeros on J, Rouché's Theorem and formulae (3.71) and (3.73) imply that $p(z)$ and $\tilde{p}(z)$ have the same number of zeros inside the Jordan curve J (counting multiplicities). ∎

Lemma 3.3.38 indicates that small perturbations of the matrix $A(s)$ affect the zeros of $p(z)$ only slightly: they cannot cross the Jordan curve J. The idea is now to perturb $A(s)$ in such a way that the multiple zeros of $p(z)$ inside J become simple, without changing the total number of zeros inside J. In this approach, Rouché's Theorem (in the disguised form of Lemma 3.3.38), plays an important role.

Proposition 3.3.39 *Let $A(s) \in \mathbf{R}[s]^{n\times n}$ and $\tau > 0$ be given. Let J be a bounded Jordan curve in $\overline{\mathbb{C}^+}$, and assume that $p(z) = \det(zI - A(e^{-\tau z}))$ has no zeros on J. Choose $\bar{\varepsilon} > 0$ such that Lemma 3.3.38 is satisfied. Let N_p denote the total number of zeros of $p(z)$ within J, counting multiplicities. Then:*

$\forall \varepsilon \in (0, \bar{\varepsilon})\ \exists \tilde{A}(s) \in \mathbf{R}[s]^{n\times n}$ *such that*

(i) $\|A(s) - \tilde{A}(s)\|_{pm} \leq \varepsilon$,

(ii) $\deg(\tilde{A}(s)) \leq \max(\deg(A(s)), 2)$,

(iii) The analytic function $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ has N_p zeros within J, and all these zeros are simple.

Proof

Let $\varepsilon \in (0, \bar{\varepsilon})$. Then it follows from Lemma 3.3.38 that for all $\tilde{A}(s) \in \mathbf{R}[s]^{n\times n}$ satisfying $\|A(s) - \tilde{A}(s)\|_{pm} < \varepsilon < \bar{\varepsilon}$, the number of zeros of $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ inside J equals N_p. Let L_p denote the number of simple zeros of $p(z)$ within J. The proposition is proved with the following induction argument:

$\forall i \in \{0, 1, \ldots, N_p - L_p\}\ \exists A_i(s) \in \mathbf{R}[s]^{n\times n}$ such that

(1) $\|A(s) - A_i(s)\|_{pm} \leq \frac{2^i - 1}{2^i} \cdot \varepsilon$,

(2) $\deg(A_i(s)) \leq \max(\deg(A(s)), 2)$.

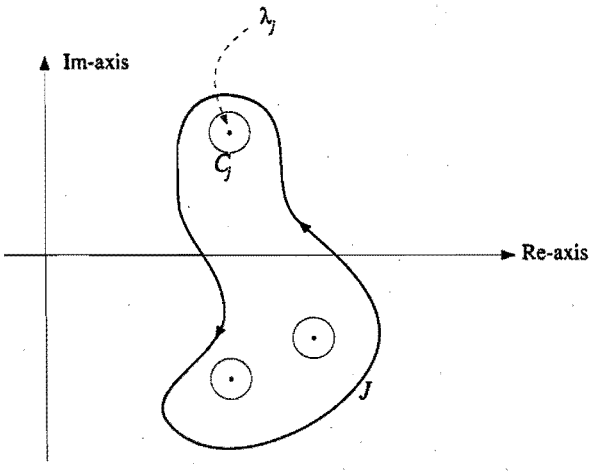


Figure 3.2: Location of the zeros inside a Jordan curve J

(3) The analytic function $p_i(z) = \det(zI - A_i(e^{-\tau z}))$ has at least $L_p + i$ simple zeros within J, i.e. $L_{p_i} \geq L_p + i$, where L_{p_i} denotes the number of simple zeros of $p_i(z)$ enclosed by J.

$\underline{i = 0}$: This is trivial: choose $A_0(s) = A(s)$.

Induction step: Suppose that for certain $i \in \{0, 1, \ldots, N_p - L_p - 1\}$ we have found a matrix $A_i(s)$, satisfying *(1)*, *(2)* and *(3)*. If $L_{p_i} \geq L_p + i + 1$, choose $A_{i+1}(s) = A_i(s)$, and we are ready.

Next assume that $L_{p_i} = L_p + i$. Since $i < N_p - L_p$, we know that at least one of the N_p zeros of $p_i(z)$ inside J is a multiple zero. Let λ_j, $j \in \{1, \ldots, \ell\}$ denote all distinct zeros of $p_i(z)$. Then there exists a $\rho > 0$ such that the circles C_j defined by

$$C_j = \{z \in \mathbb{C} \mid |z - \lambda_j| = \rho\}, \tag{3.74}$$

neither intersect one another nor the Jordan curve J (see Figure 3.2). Apply Lemma 3.3.38 to each of these circles C_j. Then for all $j = 1, \ldots, \ell$ we find an $\bar{\varepsilon}_j > 0$, such that for all $\hat{A}(s) \in \mathbb{R}[s]^{n \times n}$, the inequality $\|A(s) - \hat{A}(s)\|_{pm} < \bar{\varepsilon}_j$ implies that $p_i(z)$ and $\hat{p}(z) = \det(zI - \hat{A}(e^{-\tau z}))$ have the same number of zeros within C_j, and no zeros on C_j. Define $\hat{\varepsilon} := \min\{\bar{\varepsilon}_j \mid j = 1, \ldots, \ell\}$.

Assume without loss of generality that λ_1 is a multiple zero of $p_i(z)$. Apply Proposition 3.3.30 to $Q(z) := zI - A_i(e^{-\tau z})$, with $g(z) = e^{-\tau z}$ and $\lambda = \lambda_1$. Clearly $g'(\lambda_1) = -\tau e^{-\tau \lambda_1} \neq 0$, and so there exists a polynomial matrix $\Delta(s) \in \mathbb{R}[s]^{n \times n}$, with $\deg(\Delta(s)) \leq 2$, in norm bounded by

$$\|\Delta(s)\|_{pm} < \min(\hat{\varepsilon}, \frac{1}{2^{i+1}} \cdot \varepsilon),$$

and such that $\tilde{p}(z) = \det(Q(z) + \Delta(e^{-\tau z}))$ only has a simple zero in $z = \lambda_1$. We show that $A_{i+1}(s) := A_i(s) - \Delta(s)$ satisfies the requirements *(1)*, *(2)* and *(3)*, with i replaced by $i + 1$.

(1) and *(2)* are very straightforward:

$$\begin{aligned}\|A(s) - A_{i+1}(s)\|_{pm} &\leq \|A(s) - A_i(s)\|_{pm} + \|A_i(s) - A_{i+1}(s)\|_{pm} \leq \\ &\leq \frac{2^i - 1}{2^i} \cdot \varepsilon + \frac{1}{2^{i+1}} \cdot \varepsilon = \frac{2^{i+1} - 1}{2^{i+1}} \cdot \varepsilon,\end{aligned}$$

and $\deg(A_{i+1}(s)) \leq \max(\deg(A_i(s)), 2) \leq \max(\deg(A(s)), 2)$.

(3) Since $\|A_{i+1}(s) - A_i(s)\|_{pm} < \hat{\varepsilon}$, we can apply Lemma 3.3.38 to each of the circles $\mathcal{C}_j$ defined in (3.74), separately. In this way we obtain that for all $j \in \{1, \ldots, \ell\}$, the number of zeros of $p_{i+1}(z)$ within $\mathcal{C}_j$ is equal to the number of zeros of $p_i(z)$ within $\mathcal{C}_j$ (counting multiplicities). This implies that the L_{p_i} circles containing a simple zero of $p_i(z)$, also contain exactly one (simple) zero of $p_{i+1}(z)$. Moreover, the multiple zero λ_1 has become simple by construction, and thus

$$L_{p_{i+1}} \geq L_{p_i} + 1 = L_p + i + 1.$$

This completes the proof of the induction argument. The correctness of Proposition 3.3.39 follows immediately by taking $\tilde{A}(s) = A_{N_p - L_p}(s)$. ∎

Proposition 3.3.39 shows that the matrix perturbations introduced in Proposition 3.3.30 can be used successively to reduce the multiplicity of zeros to 1. Rouché's Theorem guarantees not only that the total number of zeros within the Jordan curve J remains constant, but also that simple zeros remain simple. Combining Propositions 3.3.36 and 3.3.39, together with the results of the previous subsection, we can finish the first part of the proof as indicated in the introduction of this subsection, by an appropriate choice of the Jordan curve J.

Theorem 3.3.40 *Let $A(s) \in \mathbb{R}[s]^{n \times n}$ and $\tau > 0$ be given. Then for all $\varepsilon > 0$ there exists a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n \times n}$ such that*

(i) $\|A(s) - A_\varepsilon(s)\|_{pm} < \varepsilon$,

(ii) $\deg(A_\varepsilon(s)) \leq \max(\deg(A(s)), 2)$,

(iii) $\forall \lambda \in \overline{\mathbb{C}^+}: \left[\operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) < n \Longrightarrow \operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) = n - 1\right]$.

Proof

Let $\varepsilon > 0$. Choose, according to Proposition 3.3.36, a matrix $A_1(s) \in \mathbb{R}[s]^{n \times n}$ of the same degree as $(A(s))$, satisfying $\|A(s) - A_1(s)\|_{pm} < \frac{1}{2}\varepsilon$, and such that $p_1(z) := \det(zI - A_1(e^{-\tau z}))$ has no zeros on the imaginary axis.

Define $R := \|A_1(s)\|_{pm} + 1$, and the Jordan curve J, as depicted in Figure 3.3 by

$$J := \{z \in \mathbf{C} \mid (\operatorname{Re} z = 0 \text{ and } |z| < R) \text{ or } (\operatorname{Re} z \geq 0 \text{ and } |z| = R)\}. \tag{3.75}$$

So, according to Lemma 3.3.33, all zeros of $p_1(z)$ in $\overline{\mathbb{C}^+}$ are located inside the Jordan curve J. Let N_{p_1} denote the number of zeros of $p_1(z)$ enclosed by J (counting multiplicities). We choose $\bar{\varepsilon} > 0$ according to Lemma 3.3.38, and apply Proposition 3.3.39 with $\bar{\varepsilon} := \frac{1}{2} \cdot \min(1, \varepsilon, \bar{\varepsilon})$. Then we find a matrix $A_\varepsilon(s) \in \mathbb{R}[s]^{n \times n}$ such that

(1) $\|A_1(s) - A_\varepsilon(s)\|_{pm} \leq \frac{1}{2} \cdot \min(1, \varepsilon, \bar{\varepsilon}) \leq \frac{1}{2}\varepsilon$,

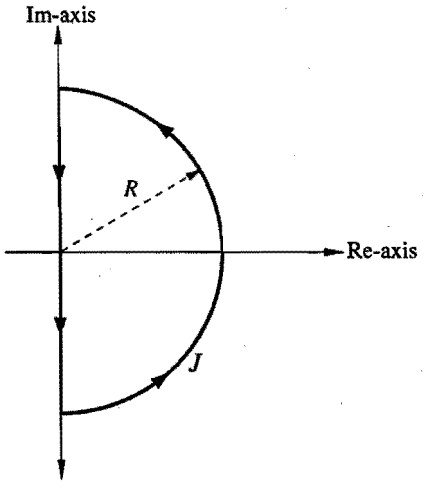


Figure 3.3: The Jordan curve J

(2) $\deg(A_\varepsilon(s)) \leq \max(\deg(A_1(s)), 2)$,

(3) $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ has N_{p_1} zeros within J, that are all simple.

This matrix $A_\varepsilon(s)$ clearly satisfies *(i)* and *(ii)*:

$$\|A(s) - A_\varepsilon(s)\|_{pm} \leq \|A(s) - A_1(s)\|_{pm} + \|A_1(s) - A_\varepsilon(s)\|_{pm} < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon,$$

and $\deg(A_\varepsilon(s)) \leq \max(\deg(A_1(s)), 2) \leq \max(\deg(A(s)), 2)$.

Next we prove *(iii)*. Since $\|A_\varepsilon(s)\|_{pm} < \|A_1(s)\|_{pm} + \frac{1}{2}$, Lemma 3.3.33 implies that $p_\varepsilon(z)$ has no zeros in $\overline{\mathbb{C}^+}$ outside J. Moreover, because $\|A_1(s) - A_\varepsilon(s)\|_{pm} < \bar{\varepsilon}$, we know from Lemma 3.3.38 that $p_\varepsilon(z)$ has no zeros on J. Therefore all zeros of $p_\varepsilon(z)$ in $\overline{\mathbb{C}^+}$ are located within J. According to (3), all these zeros are simple, and thus we have:

$$\forall z \in \overline{\mathbb{C}^+} : \; [p_\varepsilon(z) = 0 \Longrightarrow p'_\varepsilon(z) \neq 0]\,. \tag{3.76}$$

Now, let $\lambda \in \overline{\mathbb{C}^+}$, and assume that $\operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) < n$. Define $Q(z) = (zI - A_\varepsilon(e^{-\tau z}))$. Then $p_\varepsilon(z) = \det(Q(z))$, and we know that $p_\varepsilon(\lambda) = 0$. Moreover, according to (3.76), $p'_\varepsilon(\lambda) \neq 0$, and thus it follows from Corollary 3.3.27 that

$$\operatorname{rank}(Q(\lambda)) = \operatorname{rank}(\lambda I - A_\varepsilon(e^{-\tau\lambda})) = n - 1.$$

This completes the proof. ∎

In the next part of this subsection we are concerned with perturbations of the matrix $B(s)$. Suppose that a point $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$ is given. First perturb $A(s)$ in such a way that for $A_\varepsilon(s)$ conditions *(i)*, *(ii)* and *(iii)* of Theorem 3.3.40 are satisfied. Then it follows from Lemma 3.3.33 that the analytic function $p_\varepsilon(z) = \det(zI - A_\varepsilon(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$, say $\lambda_1, \ldots, \lambda_k$. We

know that for each $i \in \{1,\ldots,k\}$, $\text{rank}(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i})) = n-1$, and therefore the left-kernel of the matrix $(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i}))$, i.e. the linear subspace of $\mathbb{C}^n$ consisting of all row vectors x^T, such that $x^T \cdot (\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i})) = 0$, is one-dimensional. So for each $i \in \{1,\ldots,k\}$, this left-kernel is spanned by one row vector $v_i^T \in \mathbb{C}^n$. Now $(\lambda_i I - A_\varepsilon(e^{-\tau\lambda_i})) \mid B(e^{-\tau\lambda_i}))$ has rank n if and only if

$$v_i^T \cdot B(e^{-\tau\lambda_i}) \neq 0. \tag{3.77}$$

So, in order to achieve stabilizability, we have to perturb $B(s)$ in such a way that for the perturbed version $B_\varepsilon(s)$ the following holds:

$$\forall i \in \{1,\ldots,k\} : \; v_i^T \cdot B_\varepsilon(e^{-\tau\lambda_i}) \neq 0. \tag{3.78}$$

To find such a perturbation of $B(s)$, we first look for a vector b that is not perpendicular to a given finite set of vectors.

Lemma 3.3.41 *Let the column vectors $v_1,\ldots,v_k \in \mathbb{C}^n$ be given, and assume that they are all nonzero. Then there exists a vector $b \in \mathbb{R}^n$ such that*

$$\forall i \in \{1,\ldots,k\} : v_i^T \cdot b \neq 0.$$

Proof

First define for all $i = 1,\ldots,k$ the linear spaces

$$V_i := \{x \in \mathbb{R}^n \mid v_i^T \cdot x = 0\}.$$

Since all vectors v_i are nonzero, the sets V_i are linear subspaces of $\mathbb{R}^n$, with dimension smaller than or equal to $n-1$. This implies that each V_i is a nowhere dense subset of $\mathbb{R}^n$. Application of Baire's Category Theorem (see for example [83, Theorem 5.6 and Remark 5.7]) yields

$$\mathbb{R}^n \neq \bigcup_{i=1}^{k} V_i.$$

■

Intuitively, the result of Lemma 3.3.41 is clear. The vectors $v_1,\ldots,v_k$ correspond to linear subspaces $V_1,\ldots,V_k$ in $\mathbb{R}^n$ of dimension smaller than or equal to $n-1$. Now one simply has to pick a vector $b \in \mathbb{R}^n$, that is not an element of one of these subspaces $V_1,\ldots,V_k$. Since we only consider a finite number of subspaces, this is a rather easy task.

Lemma 3.3.41 makes it possible to find a perturbation of the matrix $B(s)$ that is suitable for our purpose. This result is stated in the next lemma.

Lemma 3.3.42 *Let the vectors $v_1,\ldots,v_k \in \mathbb{C}^n$ and $b_1,\ldots,b_k \in \mathbb{C}^n$ be given. Assume that for all $i \in \{1,\ldots,k\} : \|v_i\| = 1$. Then*

$$\forall \varepsilon > 0 \; \exists \beta \in \mathbb{R}^n \; : \; (i) \; \|\beta\| < \varepsilon,$$
$$(ii) \; \forall i \in \{1,\ldots,k\} : \; v_i^T \cdot (b_i + \beta) \neq 0.$$

Proof

Let $\varepsilon > 0$. Choose, according to Lemma 3.3.41, a vector $\gamma \in \mathbb{R}^n$ such that $v_i^T \cdot \gamma \neq 0$ for all $i \in \{1, \ldots, k\}$. If for all $i \in \{1, \ldots, k\}$ we have $v_i^T \cdot b_i = 0$, then $\beta = \frac{1}{2}\varepsilon \cdot \gamma$ satisfies the claim. Otherwise, choose a $\rho \in (0, \min\{|v_i^T \cdot b_i| \mid v_i^T \cdot b_i \neq 0,\ i = 1, \ldots, k\})$, and define

$$\beta := \frac{1}{2} \cdot \min(\varepsilon, \rho) \cdot \frac{1}{\|\gamma\|} \cdot \gamma.$$

Then *(i)* is clear: $\|\beta\| \leq \frac{1}{2} \cdot \varepsilon \cdot 1 < \varepsilon$. To prove *(ii)*, let $i \in \{1, \ldots, k\}$. If $v_i^T \cdot b_i = 0$, then

$$v_i^T \cdot (b_i + \beta) = v_i^T \cdot \beta = \frac{1}{2} \cdot (v_i^T \gamma) \cdot \frac{1}{\|\gamma\|} \cdot \min(\varepsilon, \rho) \neq 0.$$

On the other hand, if $v_i^T \cdot b_i \neq 0$, then

$$|v_i^T \cdot (b_i + \beta)| = |v_i^T b_i + v_i^T \beta| \geq |v_i^T b_i| - |v_i^T \beta| \geq \rho - \|v_i\| \cdot \|\beta\| \geq \rho - \frac{1}{2}\rho > 0.$$

So, in either case $v_i^T \cdot (b_i + \beta) \neq 0$. ∎

At this point, the proof for the stability domain $\mathbb{C}^-$, outlined in the introduction of this subsection, is almost complete. We only have to state and prove the main result.

Theorem 3.3.43 *Let $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$ be given. For all $\varepsilon > 0$ there exists a point $\tilde{\Sigma} = (\tilde{A}(s), \tilde{B}(s), \tilde{\tau}) \in \mathcal{W}$ such that*

(i) $d_{\mathcal{W}}(\Sigma, \tilde{\Sigma}) < \varepsilon$,

(ii) $\deg(\tilde{A}(s)) \leq \max(\deg(A(s)), 2)$ *and* $\deg(\tilde{B}(s)) = \deg(B(s))$,

(iii) The time-delay system corresponding to $\tilde{\Sigma}$ is stabilizable w.r.t. $\mathbb{C}^-$, i.e.

$$\forall z \in \overline{\mathbb{C}^+} : \ \operatorname{rank}(zI - \tilde{A}(e^{-\tilde{\tau}z}) \mid \tilde{B}(e^{-\tilde{\tau}z})) = n.$$

Proof

Let $\varepsilon > 0$. First apply Theorem 3.3.40 to $A(s)$, and choose a matrix $\tilde{A}(s) \in \mathbb{R}[s]^{n \times n}$ such that

(1) $\|A(s) - \tilde{A}(s)\|_{pm} < \frac{1}{2}\varepsilon$,

(2) $\deg(\tilde{A}(s)) \leq \max(\deg(A(s)), 2)$,

(3) $\forall z \in \overline{\mathbb{C}^+} : \ \left[\operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) < n \Longrightarrow \operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) = n - 1\right]$.

According to Lemma 3.3.33, the function $\tilde{p}(z) = \det(zI - \tilde{A}(e^{-\tau z}))$ has only a finite number of zeros in $\overline{\mathbb{C}^+}$, say $\lambda_1, \ldots, \lambda_k$. Only in these points $(zI - \tilde{A}(e^{-\tau z}))$ loses rank, but then still $\operatorname{rank}(zI - \tilde{A}(e^{-\tau z})) = n - 1$. So the left-kernel of $(zI - \tilde{A}(e^{-\tau z}))$

is one-dimensional for all $z \in \{\lambda_1, \ldots, \lambda_k\}$. Choose vectors $v_1, \ldots, v_k$ of norm 1 in $\mathbb{C}^n$, spanning these left-kernels:

$$\forall i \in \{1, \ldots, k\} : \ \mathrm{span}\{v_i\} = \{x \in \mathbb{C}^n \mid x^T \cdot (\lambda_i I - \tilde{A}(e^{-\tau\lambda_i})) = 0\}.$$

Denote for all $i \in \{1, \ldots, k\}$ the first column of $B(e^{-\tau\lambda_i})$ by b_i. According to Lemma 3.3.42 there exists a $\beta \in \mathbb{R}^n$ such that $\|\beta\| < \frac{1}{2}\varepsilon$ and $v_i^T \cdot (b_i + \beta) \neq 0$ for all $i = 1, \ldots, k$.

Define $\tilde{B}(s)$ as the sum of $B(s)$ and the $n \times m$ matrix $(\beta \mid 0)$, consisting of the column β, completed with zeros:

$$\tilde{B}(s) := B(s) + (\beta \mid 0).$$

Then obviously *(ii)* holds, and we only need to show that $\tilde{\Sigma} = (\tilde{A}(s), \tilde{B}(s), \tau)$ satisfies both *(i)* and *(iii)*.

$$\begin{aligned} d_{\mathcal{W}}(\Sigma, \tilde{\Sigma}) &= \|A(s) - \tilde{A}(s)\|_{pm} + \|B(s) - \tilde{B}(s)\|_{pm} + |\tau - \tau| < \\ &< \frac{1}{2}\varepsilon + \|(\beta \mid 0)\| = \frac{1}{2}\varepsilon + \|\beta\| < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon. \end{aligned}$$

To prove *(iii)*, let $z \in \overline{\mathbb{C}^+}$. If $z \notin \{\lambda_1, \ldots, \lambda_k\}$, then $\mathrm{rank}(zI - \tilde{A}(e^{-\tau z})) = n$, so certainly $\mathrm{rank}(zI - \tilde{A}(e^{-\tau z}) \mid \tilde{B}(e^{-\tau z})) = n$.

Otherwise, suppose that $z = \lambda_i$ for certain $i \in \{1, \ldots, k\}$. Let $x \in \mathbb{C}^n$ be such that

$$x^T \cdot (\lambda_i I - \tilde{A}(e^{-\tau\lambda_i}) \mid \tilde{B}(e^{-\tau\lambda_i})) = 0. \tag{3.79}$$

Hence, x^T is an element of the left-kernel of $(\lambda_i I - \tilde{A}(e^{-\tau\lambda_i}))$, and there exists an $\alpha \in \mathbb{C}$ such that $x = \alpha \cdot v_i$. Now the first column of $\tilde{B}(e^{-\tau\lambda_i})$ is $b_i + \beta$, and

$$0 = x^T \cdot (b_i + \beta) = \alpha v_i^T \cdot (b_i + \beta) = \alpha \cdot [v_i^T \cdot (b_i + \beta)].$$

We conclude that $\alpha = 0$. This completes the proof. ∎

From Theorem 3.3.43 it follows directly that the subset of $\mathcal{W}$ consisting of all parametrizations of time-delay systems that are stabilizable w.r.t. $\mathbb{C}^-$, is a dense subset of $\mathcal{W}$. Note that the conditions on the degrees of $\tilde{A}(s)$ and $\tilde{B}(s)$ are essential. Construction of a sequence of stabilizable systems converging to Σ, but with an increasing degree in s, is of no use for our genericity result because this would require systems with time-delays of constantly increasing length. However, since Theorem 3.3.43 enables us to construct a sequence of stabilizable systems in $\mathcal{W}$ of bounded degree in s and converging to Σ, the same result holds in the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$.

Corollary 3.3.44 *Let $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$ be given, and assume that Σ is not stabilizable w.r.t. $\mathbb{C}^-$. Then there exists a sequence $\Sigma_j = (A_j(s), B_j(s), \tau)_{j \in \mathbf{N}}$ in $\mathcal{W}$ such that all systems Σ_j are stabilizable w.r.t. $\mathbb{C}^-$. Moreover, in the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$, the sequence Σ_j converges to Σ for $j \to \infty$.*

Proof

Define $\ell_1 := \deg(A(s))$, $\ell_2 := \deg(B(s))$ and $\ell := \ell_1 + \ell_2 + 2$. According to Theorem 3.3.43, for each $j \in \mathbf{N}$ there exist matrices $A_j(s) \in \mathcal{V}_{n \times n, \ell}$ and $B_j(s) \in \mathcal{V}_{n \times m, \ell}$ such that

(i) $\|A(s) - A_j(s)\|_\ell < \frac{1}{j}$ and $\|B(s) - B_j(s)\|_\ell < \frac{1}{j}$,

(ii) $\Sigma_j := (A_j(s), B_j(s), \tau)$ is stabilizable w.r.t. $\mathbf{C}^-$.

From *(i)* and Proposition 3.3.8 we conclude that in the inductive limit topology the sequence $(\Sigma_j)_{j\in\mathbf{N}}$ converges to Σ when j tends to infinity. Together with *(ii)*, this proves the claim. ∎

In the inductive limit topology, Corollary 3.3.44 may be generalized to an arbitrary stability domain $\mathbf{C}_g$.

Corollary 3.3.45 *Let $\mathbf{C}_g$ be a stability domain, and $\Sigma = (A(s), B(s), \tau) \in \mathcal{W}$. Assume that Σ is not stabilizable w.r.t. $\mathbf{C}_g$. Then there exists a sequence $\Sigma_j = (A_j(s), B_j(s), \tau)_{j\in\mathbf{N}}$ in $\mathcal{W}$ such that all systems Σ_j are stabilizable w.r.t. $\mathbf{C}_g$. Moreover, in the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$, the sequence Σ_j converges to Σ for $j \to \infty$.*

Proof

According to Definition 3.1.2, there exists an $\alpha \in \mathbf{R}$ such that $\mathbf{C}_{-\alpha} \subset \mathbf{C}_g$. Apply Corollary 3.3.44 to the system $H_\alpha(A(s), B(s), \tau)$, where H_α is the operator defined in (3.56). In this way we obtain a sequence $\Sigma_j = (A_j(s), B_j(s), \tau)_{j\in\mathbf{N}}$ of systems that are stabilizable w.r.t. $\mathbf{C}^-$, and converging to $H_\alpha(A(s), B(s), \tau)$. Both the operator H_α and its inverse $H_{-\alpha}$ are (sequentially) continuous, because of Proposition 3.3.21. So the sequence $(H_{-\alpha}(A_j(s), B_j(s), \tau))_{j\in\mathbf{N}}$ converges to $H_{-\alpha}H_\alpha(A(s), B(s), \tau) = (A(s), B(s), \tau)$ when $j \to \infty$. Moreover, all systems $\Sigma_j = (A_j(s), B_j(s), \tau)$ are stabilizable w.r.t. $\mathbf{C}^-$, so for all $j \in \mathbf{N}$ $H_{-\alpha}(A_j(s), B_j(s), \tau)$ is stabilizable w.r.t. $\mathbf{C}_{-\alpha}$. Since $\mathbf{C}_{-\alpha} \subset \mathbf{C}_g$, we conclude that the sequence $(H_{-\alpha}(A_j(s), B_j(s), \tau))_{j\in\mathbf{N}}$ satisfies the claim. ∎

After all these preparations, the main result on genericity is stated in the next theorem.

Theorem 3.3.46 *Let $\mathbf{C}_g$ be a stability domain. Then time-delay systems of the form (3.27) are generically stabilizable w.r.t $\mathbf{C}_g$ by dynamic state feedback. In other words:*

In the inductive limit topology $\mathcal{T}$ on $\mathcal{W}$, the subset S_g of the parameter-space $\mathcal{W}$, consisting of all parametrizations $\Sigma = (A(s), B(s), \tau)$ of time-delay systems satisfying

$$\forall z \in \mathbf{C}\backslash\mathbf{C}_g : \operatorname{rank}(zI - A(e^{-\tau z}) \mid B(e^{-\tau z})) = n,$$

contains an open and dense subset of the space $\mathcal{W}$.

Proof

There exists an $\alpha \in \mathbf{R}$ such that $\mathbf{C}_{-\alpha} = \{z \in \mathbf{C} \mid \operatorname{Re} z < -\alpha\} \subset \mathbf{C}_g$. Let $S_{-\alpha}$ and S_g denote the subsets of $\mathcal{W}$ consisting of all delay systems that are stabilizable w.r.t. $\mathbf{C}_{-\alpha}$ and $\mathbf{C}_g$ respectively. Then $S_{-\alpha} \subset S_g$. Combining Corollaries 3.3.22 and 3.3.45, we know that $S_{-\alpha}$ is an open and dense subset of $\mathcal{W}$. Hence S_g contains an open and dense subset of $\mathcal{W}$, and thus stabilizability w.r.t. $\mathbf{C}_g$ is a generic property.

∎

3.3.5 Generalization to the case of incommensurable time-delays

In Subsections 3.3.1 to 3.3.4, a derivation of our genericity result is given for systems with commensurable time-delays. This restriction was only made for notational convenience; the incommensurable delay case is not significantly more difficult. In this subsection we explain that our genericity result on stabilizability also holds for the more general class of systems with incommensurable time-delays. We point out that exactly the same arguments as in Subsections 3.3.1 to 3.3.4 may be used to prove this result.

In the algebraic terminology, a time-delay system with k incommensurable delays $\tau_1, \ldots, \tau_k$ is modeled as a system over the ring $\mathbf{R}[s_1, \ldots, s_k]$, where the indeterminate s_i corresponds to the delay operator σ_i with time-delay τ_i. To apply a topological approach to our genericity problem, first a parameter-space $\mathcal{W}_k$ (the incommensurable version of $\mathcal{W}$) has to be introduced. Denoting $\mathbf{R}[s_1, \ldots, s_k]$ by $\mathcal{R}$, $\mathcal{W}_k$ is defined as:

$$\mathcal{W}_k := \{\Sigma = (A, B, (\tau_1, \ldots, \tau_k)) \quad | \quad A \in \mathcal{R}^{n \times n},\ B \in \mathcal{R}^{n \times m},\ \tau_i \in \mathbf{R}^+ (i = 1, \ldots, k)\}. \tag{3.80}$$

In the same way as in the commensurable delay case, a matrix over $\mathbf{R}[s_1, \ldots, s_k]^{p \times q}$ can be seen as a k-dimensional sequence of $p \times q$ matrices over $\mathbf{R}$, with only a finite number of nonzero elements. So definition of an ℓ_1-norm is possible, and in this way the polynomial matrix norm $\|\cdot\|_{pm}$ of Definition 3.3.1 can be generalized to the case of incommensurable delays. Also the inductive limit topology (see Definition 3.3.4) is applicable in this more general context, but the definition becomes technically more involved. Finally, the norm $\|\cdot\|_p$ for polynomials in two indeterminates, introduced in Definition 3.3.13, is easily extended to polynomials in more than two indeterminates.

With these generalized definitions of the norms and topologies, the results of Subsection 3.3.1 remain valid. Most of these results rely either on the structure of inductive limits, or on the fact that for all $z \in \overline{\mathbf{C}^+}$: $|e^{-\tau z}| \leq 1$. Since all time-delays τ_i are strictly greater than zero, we still have:

$$\forall i \in \{1, \ldots, k\}\ \forall \tau_i > 0\ \forall z \in \overline{\mathbf{C}^+} :\ |e^{-\tau_i z}| \leq 1, \tag{3.81}$$

and the same proofs can be applied. The only difficulty left is the result of Proposition 3.3.15 on the continuity of the map χ from a polynomial matrix to its characteristic polynomial. Here exponentials do not play a role. However, in the proof of this result the number of indeterminates is not significant. Therefore this result also holds in the case of incommensurable time-delays.

The results of Subsection 3.3.2 are easily generalized, as far as the perturbations of the matrices $A(s_1, \ldots, s_k)$ and $B(s_1, \ldots, s_k)$ are concerned. However, in Theorem 3.3.17 perturbations of the lengths of the time-delays lead to a more complicated situation. But generalization to the incommensurable delay case is still possible. Because of (3.81), all perturbations of the time-delays can be treated successively. In each step i, $(i = 1, \ldots, k)$, the exponentials $e^{-\tau_1 z}, \ldots, e^{-\tau_{i-1} z}$ and $e^{-\tau_{i+1} z}, \ldots, e^{-\tau_k z}$,

corresponding to all time delays except τ_i, are bounded above by 1 in absolute value because we assume that $z \in \overline{\mathbb{C}^+}$. Therefore exactly the same techniques as in formula (3.55) may be applied successively, for each τ_i separately, to arrive at the desired result.

To explain this idea more clearly, consider the case of two incommensurable time-delays. Let $A_0(s_1, s_2) = \sum_{i=0}^{k} \sum_{j=0}^{\ell} A_{ij} s_1^i s_2^j$, and consider a $z \in \overline{\mathbb{C}^+}$ such that $|z| \leq \|A_0(s_1, s_2)\|_{pm} + 1$. Then

$$\|A_0(e^{-\tau_1 z}, e^{-\tau_2 z}) - A_0(e^{-\hat{\tau}_1 z}, e^{-\hat{\tau}_2 z})\| \leq \sum_{i=0}^{k} \sum_{j=0}^{\ell} \|A_{ij}\| \cdot |e^{-i\tau_1 z} e^{-j\tau_2 z} - e^{-i\hat{\tau}_1 z} e^{-j\hat{\tau}_2 z}| \leq$$

$$\leq \sum_{i=0}^{k} \sum_{j=0}^{\ell} \|A_{ij}\| \cdot (|e^{-i\tau_1 z} - e^{-i\hat{\tau}_1 z}| + |e^{-j\tau_2 z} - e^{-j\hat{\tau}_2 z}|) \leq$$

$$\leq \sum_{i=0}^{k} \sum_{j=0}^{\ell} \|A_{ij}\| \cdot (|iz \int_{\tau_1}^{\hat{\tau}_1} e^{-i\xi z} d\xi| + |jz \int_{\tau_2}^{\hat{\tau}_2} e^{-j\xi z} d\xi|) \leq$$

$$\leq \sum_{i=0}^{k} \sum_{j=0}^{\ell} \|A_{ij}\| \cdot (|i| \cdot |z| \cdot |\hat{\tau}_1 - \tau_1| + |j| \cdot |z| \cdot |\hat{\tau}_2 - \tau_2|) \leq$$

$$\leq (\|A_0(s_1, s_2)\|_{pm} + 1) \cdot \|A_0(s_1, s_2)\|_{pm} \cdot (k|\hat{\tau}_1 - \tau_1| + \ell|\hat{\tau}_2 - \tau_2|).$$

Taking $\hat{\tau}_1$ and $\hat{\tau}_2$ close enough to τ_1 and τ_2 respectively, this last expression can be made arbitrarily small.

Subsection 3.3.3 is already put in a general context, so here nothing has to be done. Note however that in Proposition 3.3.30 only one time-delay is needed to achieve an appropriate perturbation of the matrix $Q(z)$, and that $g(z) = e^{-\tau_i z}$ satisfies the condition $g'(\lambda) \neq 0$ for all $\lambda \in \mathbb{C}$.

In the first part of Subsection 3.3.4 we are now dealing with analytic functions of the form

$$p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})).$$

The assumption on the absence of zeros on the imaginary axis can be removed in almost the same way as it was done in Proposition 3.3.36. The proof becomes somewhat unclear because of notational difficulties, but the same ideas still apply. Of course Rouché's Theorem is still valid, and it is also easily seen that all zeros of $p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ in $\overline{\mathbb{C}^+}$ are contained in a compact subset of $\overline{\mathbb{C}^+}$. Therefore, Lemma 3.3.38 still holds, and the same process of successively reducing the order of the zeros to 1 can be used. Again, Rouché's Theorem guarantees that the total number of zeros in $\overline{\mathbb{C}^+}$ remains constant, and that simple zeros remain simple. Moreover, the results of Subsection 3.3.3 imply that the condition on the degree of $A(s_1, \ldots, s_k)$ is satisfied. Also perturbations of the matrix $B(s_1, \ldots, s_k)$ are easily found. For a suitable perturbation β, one has to substitute the right half plane zeros $\{\lambda_1, \ldots, \lambda_h\}$ of $p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ in $B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$. Since there is only a finite number of zeros of $p(z)$ in $\overline{\mathbb{C}^+}$, the method of Lemma

3.3.42 is still applicable. Therefore also Theorem 3.3.43 can be generalized to the case of incommensurable time-delays. Finally, the generalization of the genericity result for stabilizability w.r.t. $\mathbb{C}^-$ to arbitrary stability domains $\mathbb{C}_g$ is completely analogous to the commensurable delay case. Only the shift operator H_α of Definition 3.3.19 has to be adapted a little.

Summarizing, we conclude that our genericity result for the stabilizability of time-delay systems with commensurable time-delays, is also true for systems with incommensurable time-delays.

Theorem 3.3.47 *Let $\mathbb{C}_g$ be a stability domain satisfying the conditions of Definition 3.1.2. Time-delay systems with incommensurable time-delays of the form*

$$\dot{x}(t) = A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t),$$

where σ_i $(i = 1, \ldots, k)$ denotes the delay operator corresponding to a time-delay τ_i, are generically stabilizable w.r.t. $\mathbb{C}_g$ by dynamic state feedback in the following sense:

The subset of the parameter-space

$$\mathcal{W}_k = \{(A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), (\tau_1, \ldots, \tau_k)) \mid A(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]^{n \times n},$$
$$B(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]^{n \times m} \wedge \tau_i > 0 \ (i = 1, \ldots, k)\},$$

consisting of all parametrizations $\Sigma = (A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), (\tau_1, \ldots, \tau_k))$, of time-delay systems satisfying

$$\forall z \in \mathbb{C} \backslash \mathbb{C}_g : \ \operatorname{rank}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \mid B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) = n,$$

contains an open and dense subset of the space $\mathcal{W}_k$. ∎

Since stabilizability by dynamic state feedback and detectability are dual concepts, application of Corollary 3.2.9 *(iii)* yields the following result.

Corollary 3.3.48 *Let $\mathbb{C}_g$ be a stability domain satisfying the conditions of Definition 3.1.2. Consider all time-delay systems with incommensurable time-delays of the form*

$$\begin{aligned} \dot{x}(t) &= A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) &= C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t), \end{aligned}$$

where σ_i $(i = 1, \ldots, k)$ denotes the delay operator corresponding to a time-delay τ_i. For this class of systems both detectability and stabilizability by dynamic output feedback w.r.t. the stability domain $\mathbb{C}_g$ are generic properties. ∎

Theorem 3.3.47 and Corollary 3.3.48 are rather strong results: when a stability domain $\mathbb{C}_g$ satisfying the conditions of Definition 3.1.2 is fixed, the property of stabilizability w.r.t. $\mathbb{C}_g$ by dynamic output feedback is generic, both for systems with commensurable and incommensurable time-delays. This indicates that the condition for stabilizability is very weak; it is satisfied for almost all time-delay systems. On the other hand, the problem of finding an internally stabilizing feedback compensator turns out to be very difficult. We return to this question in later chapters.

Chapter 4

Constructive commutative algebra

In Chapter 2 we have seen that the framework of linear systems over rings provides a very versatile and powerful method for the description of a large class of systems and also for the design of feedback compensators. Especially systems over polynomial rings turned out to be important, certainly for the application to time-delay systems. In this algebraic framework, a lot of interesting properties of a system can be reformulated as properties of polynomial ideals. A little foretaste of this approach was given in Proposition 2.8.5, but in the next chapter we elaborate more on this subject. However, the translation of a problem into terms of polynomial ideals does not yield a final answer to our question; it is only a restatement of the same problem, in the expectation that the new question is somewhat easier to answer. For polynomial ideals, this is indeed the case: there exist several constructive methods in commutative algebra to verify properties of polynomial ideals algorithmically. This chapter gives an introduction to two of these methods: Gröbner bases and characteristic sets. In the next chapter, we shall study the application of these methods to solve some problems for linear systems over polynomial rings.

Although a lot of problems in the field of constructive commutative algebra may be considered as classical (e.g. the membership problem for a polynomial ideal), the interest in this field has increased rapidly in the last two decades. This "revival" has two main causes. First of all, the Gröbner-basis method, invented by Buchberger in 1965 and refined in the years after, made it possible to find algorithmic solutions to a lot of problems in commutative algebra. The only problem was (and in some sense still is) that the computations were very time and memory consuming. With the increasing speed and memory capacity of new generations of computers, the Gröbner basis method became also practically applicable. Nowadays most computer algebra packages contain standard software for the computation of Gröbner bases.

This wide availability of powerful algorithms also opens the way for new applications. Formerly, the restatement of a question into a problem of commutative algebra was not useful, because this new problem was not solvable anyway. Nowadays this situation has changed drastically. Because of the existence of techniques like the Gröbner-basis method, it becomes interesting to find reformulations of a problem into algebraic terms, and to find solutions using the powerful tools in this

field. Our approach to systems over polynomial rings as sketched in the first lines of this chapter, should also be seen in this perspective. It is motivated by the powerful algorithmic tools nowadays offered by computer algebra.

In the methods from constructive commutative algebra considered in this chapter, polynomial ideals play the leading role. Manipulations on the elements of a polynomial ideal are carried out with the objective to find a set of generators with a number of useful properties. With this new set of generators a lot of problems are solvable, e.g. the membership problem or the problem of computation modulo a polynomial ideal. The same techniques may also be used to eliminate some variables from a system of polynomial equations. This enables us to find a solution to such a system of equations. There are several methods to carry out these simplifications, for example Gröbner bases, characteristic sets and resultants (see e.g. [32, Section 122]). In this chapter we only consider the first two methods.

In Section 4.1 we give an introduction to the Gröbner basis method. It only contains the main ideas of the algorithm, and some of its applications. Within the scope of this thesis we are only able to give a little of the flavour of this method. For a more thorough treatment we refer to the literature. Two recent books on this subject are [14] and [2].

Section 4.2 is devoted to the method of characteristic sets. We shall treat this subject in much more detail. This is motivated by the fact that this method is not as well known as Gröbner bases, and not as widespread. Compared to Gröbner bases, there exists only very little literature on characteristic sets, and therefore it is probably more useful to dwell somewhat longer on this subject. Finally we make a comparison of both methods in Section 4.3. It turns out that characteristic sets are not as well developed as Gröbner bases yet, and that Gröbner bases form a more powerful and versatile tool for the problems we are interested in. Therefore mainly the method of Gröbner bases is applied in the rest of this thesis.

4.1 Gröbner bases

As already mentioned in the introduction of this chapter, the Gröbner basis method is nowadays probably the most powerful method in constructive commutative algebra. It was invented in 1965 by Buchberger, who named the method after his thesis advisor Prof. W. Gröbner. In fact, it was Hironaka who introduced the concept of Gröbner bases under the name "standard bases" in 1964 (see [45]), but unfortunately his proof of existence was not constructive. Instead, Buchberger obtained an algorithm for the computation of these bases. An implementation of this algorithm is nowadays available in most computer algebra packages like Mathematica, Maple and Reduce.

The purpose of this section is to give a short introduction to the Gröbner basis method: how does it work, and what can you do with it? For a more detailed study and for refinements of the algorithms we refer to the vast literature on this subject. We mainly follow the same approach as in [76], because this treatment is very compact, but we shall also give some extensions from other references. Other short introductions are given in [7] and [27, Chapter 2].

4.1.1 The Euclidean algorithm

The Gröbner basis method can be seen as a generalization of the well-known Euclidean division algorithm. To illustrate the resemblances, we start with a short glimpse at this algorithm.

Let $\mathcal{K}$ be a field and consider the Euclidean ring $\mathcal{K}[x]$ consisting of all polynomials in the indeterminate x with coefficients in $\mathcal{K}$. Every ideal $\mathcal{I}$ in $\mathcal{K}[x]$ is principal and its monic generator is uniquely determined. This is the monic polynomial in $\mathcal{I}$ of lowest degree; it can be seen as the simplest nonzero polynomial contained in the ideal $\mathcal{I}$. When the generator g of an ideal $\mathcal{I}$ is known, a lot of questions on the ideal $\mathcal{I}$ are easily answered. For example, a polynomial $p \in \mathcal{K}[x]$ belongs to $\mathcal{I}$ if and only if g divides p. The variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$, i.e. the set of all common zeros of the polynomials in $\mathcal{I}$, is simply the set of zeros of g. Two ideals are equal if and only if their monic generators are equal to each other.

Often, a finite number of polynomials $p_1, \ldots, p_k \in \mathcal{K}[x]$ are given, generating together the ideal $\mathcal{I} = \langle p_1, \ldots, p_k \rangle$. In this case, questions like those posed above, are not easily solved directly. A monic generator of $\mathcal{I}$ has to be computed first. This can be done with the Euclidean division algorithm.

Definition 4.1.1 Let p_1 and p_2 be polynomials in $\mathcal{K}[x]$, and assume that p_1 is nonzero. Then there exist two uniquely determined polynomials $q \in \mathcal{K}[x]$ and $r \in \mathcal{K}[x]$, with $\deg(r) < \deg(p_1)$ such that p_2 can be written as

$$p_2 = q \cdot p_1 + r.$$

r is called a *remainder* of p_2 after division by p_1, and is denoted by $r = \mathrm{rem}(p_2, p_1)$.

Proposition 4.1.2 (Euclidean algorithm) *Let p_1 and p_2 be two nonzero polynomials in $\mathcal{K}[x]$, and assume that $\deg(p_1) \leq \deg(p_2)$. Apply the following algorithm:*

```
g := p2; r := p1;
while r ≠ 0 do
    remainder := rem(g, r);
    g := r;
    r := remainder;
od;
```

After termination of the algorithm g is a generator of the ideal $\langle p_1, p_2 \rangle$. Moreover, g is a greatest common divisor of the polynomials p_1 and p_2. ∎

A proof of the termination of this algorithm, and of the correctness of the result can be found in for example [14, pp. 41-42] or [93, pp. 53-55].

The generator of an ideal spanned by more than two polynomials can be computed by successive application of the Euclidean division algorithm.

The algorithm of Proposition 4.1.2 has two important features. First of all, the degree of a polynomial plays an important role. It can be seen as a measure for the complexity of a polynomial. In the algorithm a sequence of polynomials with strictly decreasing degree is obtained. Since such a sequence must be finite, this

yields a polynomial of minimal degree that is contained in the original ideal. By definition, this has to be a generator of the ideal. For the construction of the sequence of polynomials with decreasing degree, the concept of division with remainder as described in Definition 4.1.1 is used.

The Gröbner basis algorithm can be seen as a generalization of this idea to rings of polynomials with more than one indeterminate; it is based on the same principles. Given a polynomial ideal, we want to obtain a set of generators of low complexity. So we first need a measure for the complexity of a polynomial in more than one indeterminate: a generalized notion of degree. For this an ordering on monomials (i.e. on polynomials consisting of only one term) is required. Moreover, a generalized division algorithm is necessary to make the computation of polynomials of lower (generalized) degree possible. The Gröbner basis algorithm is a method that guarantees that with this division algorithm, and in a finite number of steps, a set of generators is obtained that have some very useful properties.

4.1.2 Term orderings

Let $\mathcal{K}$ be an arbitrary field, and let $\mathcal{R} := \mathcal{K}[x_1, \ldots, x_n]$ denote the ring of polynomials in the indeterminates $x_1, \ldots, x_n$ with coefficients in $\mathcal{K}$. In this subsection an ordering on the monomials in $\mathcal{R}$ is defined. In this way the notion of degree is generalized to polynomials in more than one indeterminate.

First we need a ranking of the indeterminates of the polynomial ring. This ranking indicates what indeterminate is considered to be more important than the others, and induces a multidegree on the monomials in $\mathcal{R}$.

Definition 4.1.3 A *ranking* of the indeterminates $x_1, \ldots, x_n$ is a permutation π of the index set $\{1, \ldots, n\}$. The indeterminates are ordered according to the sequence $x_{\pi(1)}, \ldots, x_{\pi(n)}$. If $1 \leq i < j \leq n$, then $x_{\pi(i)}$ is said to be of *higher rank* than $x_{\pi(j)}$, and we write $x_{\pi(i)} \succ x_{\pi(j)}$.

Definition 4.1.4 Let $m = \beta x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ be a monomial in $\mathcal{R}$, and suppose that a ranking π of the indeterminates $x_1, \ldots, x_n$ is fixed. Then the *multidegree* of m, denoted by $\mathrm{mdeg}(m)$, is defined as the n-tuple

$$\mathrm{mdeg}(m) := (\deg_{x_{\pi(1)}}(m), \deg_{x_{\pi(2)}}(m), \ldots, \deg_{x_{\pi(n)}}(m)). \tag{4.1}$$

Example 4.1.5 Let $\mathcal{R} = \mathbb{R}[x_1, x_2, x_3]$ and fix the ranking $x_2 \succ x_3 \succ x_1$. Then the multidegree of $x_1^2 x_2^7 x_3$ is $(7, 1, 2)$.

We conclude that a ranking of indeterminates specifies a bijective map from the set of all monomials in $\mathcal{R}$ to the set $\mathbb{N}_0^n$, where $\mathbb{N}_0$ denotes the set $\mathbb{N} \cup \{0\}$ consisting of all non-negative integers. So, if a total ordering on the n-tuples in $\mathbb{N}_0^n$ is fixed, this automatically leads to an ordering of monomials in $\mathcal{R}$. However, to maintain the properties of the notion of degree, this ordering on $\mathbb{N}_0^n$ has to satisfy some regularity conditions.

Definition 4.1.6 A *term ordering* on the monomials in the polynomial ring $\mathcal{R}$ consists of a ranking π, and an ordering $>$ on $\mathbb{N}_0^n$ satisfying the following properties:

(i) $>$ is a total ordering on $\mathbf{N}_0^n$,

(ii) If $\alpha, \beta, \gamma \in \mathbf{N}_0^n$ are such that $\alpha > \beta$, then $\alpha + \gamma > \beta + \gamma$,

(iii) $>$ is a well-ordering. This means that every non-empty subset of $\mathbf{N}_0^n$ has a smallest element under $>$,

(iv) $(1, 0, \ldots, 0) > (0, 1, 0, \ldots, 0) > \cdots > (0, \ldots, 0, 1)$.

A monomial m_1 is said to be smaller than a monomial m_2 if $\mathrm{mdeg}(m_1) < \mathrm{mdeg}(m_2)$.

Conditions *(ii)* and *(iii)* imply that $(0, \ldots, 0)$ is the smallest element of $\mathbf{N}_0^n$, so a constant term is always smaller than a monomial containing at least one indeterminate. Moreover, condition *(ii)* indicates that if a pair of monomials m_1 and m_2 satisfies $m_1 > m_2$, and we multiply both by another monomial a, then also $am_1 > am_2$. Condition *(iv)* ensures that the ordering of the indeterminates in the term ordering $>$ is the same as in the ranking π. When an ordering on $\mathbf{N}_0^n$ (and thus on the monomials in $\mathcal{R}$) is fixed, notions like maximum, minimum etc. are also well defined. Some important orderings satisfying the conditions of Definition 4.1.6 are the following:

Definition 4.1.7 Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n) \in \mathbf{N}_0^n$.

(i) The *(strict) lexicographic ordering* (or pure lexicographic ordering) $>_{\mathrm{lex}}$ is defined by

$$\alpha >_{\mathrm{lex}} \beta \iff \text{there exists a } j \in \{1, \ldots, n\} \text{ such that } \alpha_i = \beta_i \text{ for all } i < j \text{ and } \alpha_j > \beta_j.$$

(ii) The *graded lexicographic ordering* (or total-degree ordering) $>_{\mathrm{tdeg}}$ is defined by

$$\alpha >_{\mathrm{tdeg}} \beta \iff \Big(\sum_{i=1}^{n} \alpha_i > \sum_{i=1}^{n} \beta_i\Big) \text{ or } \Big(\Big(\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \beta_i\Big) \text{ and } \alpha >_{\mathrm{lex}} \beta\Big).$$

(iii) The *graded inverse lexicographic ordering* $>_{\mathrm{grevlex}}$ is defined by

$$\alpha >_{\mathrm{grevlex}} \beta \iff \Big(\sum_{i=1}^{n} \alpha_i > \sum_{i=1}^{n} \beta_i\Big) \text{ or } \Big(\Big(\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \beta_i\Big) \text{ and there exists a } j \in \{1, \ldots, n\} \text{ such that } \alpha_i = \beta_i \text{ for all } i > j \text{ and } \alpha_j < \beta_j\Big).$$

Example 4.1.8 Let $\mathcal{R}$ be the polynomial ring $\mathbf{R}[x_1, x_2, x_3]$ with ranking $x_2 \succ x_1 \succ x_3$. Consider the monomials

$$\begin{aligned} m_1(x_1, x_2, x_3) &= x_1^3 x_2 x_3; & \mathrm{mdeg}(m_1) &= (1, 3, 1), \\ m_2(x_1, x_2, x_3) &= x_1 x_2^2 x_3; & \mathrm{mdeg}(m_2) &= (2, 1, 1), \\ m_3(x_1, x_2, x_3) &= x_1 x_2^2 x_3^2; & \mathrm{mdeg}(m_3) &= (2, 1, 2). \end{aligned}$$

Then it is easily verified that

$$\begin{aligned}
&(2,1,2) >_{\text{lex}} (2,1,1) >_{\text{lex}} (1,3,1) \quad \text{so} \quad m_3 >_{\text{lex}} m_2 >_{\text{lex}} m_1,\\
&(2,1,2) >_{\text{tdeg}} (1,3,1) >_{\text{tdeg}} (2,1,1) \quad \text{so} \quad m_3 >_{\text{tdeg}} m_1 >_{\text{tdeg}} m_2,\\
&(1,3,1) >_{\text{grevlex}} (2,1,2) >_{\text{grevlex}} (2,1,1) \quad \text{so} \quad m_1 >_{\text{grevlex}} m_3 >_{\text{grevlex}} m_2.
\end{aligned}$$

Next we adopt the convention to write a monomial in $\mathcal{R}$ with multidegree $\alpha = (\alpha_1, \ldots, \alpha_n)$ as x^α. A polynomial in $\mathcal{R}$ is simply a finite sum of monomials. We will use the notation

$$p(x) = \sum_{\alpha \in \mathbb{N}_0^n} c_\alpha x^\alpha,$$

always tacitly assuming that only finitely many of the coefficients c_α are nonzero.

Definition 4.1.9 Let $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ and assume that a ranking of the indeterminates $x_1, \ldots, x_n$, and an ordering on $\mathbb{N}_0^n$ are fixed. Let $p(x) = \sum_{\alpha \in \mathbb{N}_0^n} c_\alpha x^\alpha$ be a nonzero polynomial in $\mathcal{R}$. Then

(i) The *degree* of p is defined as the maximum of the multidegrees of the terms in p:

$$\deg(p) := \max\{\alpha \in \mathbb{N}_0^n \mid c_\alpha \neq 0\}. \tag{4.2}$$

(ii) The *leading coefficient* of p is

$$\operatorname{lc}(p) := c_{\deg(p)}. \tag{4.3}$$

(iii) The *initial term* of p is defined as

$$\operatorname{in}(p) := \operatorname{lc}(p) \cdot x^{\deg(p)}. \tag{4.4}$$

The definition of degrees and initial terms is easily generalized to subsets F of the polynomial ring $\mathcal{R}$:

$$\deg(F) := \{\deg(p) \mid p \in F \backslash \{0\}\}, \tag{4.5}$$

$$\operatorname{in}(F) := \{\operatorname{in}(p) \mid p \in F \backslash \{0\}\}. \tag{4.6}$$

The initial terms play an important role in the definition of Gröbner bases. However, before giving this definition in Subsection 4.1.4, we first generalize the other important ingredient of the Euclidean algorithm to the case of polynomial rings in more than one indeterminate: the division algorithm.

4.1.3 Generalized division

Let F be a finite subset of the polynomial ring $\mathcal{R}$, and let $\mathcal{I} = \langle F \rangle$ be the ideal generated by the polynomials in F. In analogy with Proposition 4.1.2, a division algorithm has a double purpose. On the one hand it is used to manipulate the polynomials in F in such a way that new polynomials in $\mathcal{I}$ are obtained that are of lower degree than the polynomials in F. On the other hand we want to use the same algorithm to verify whether or not a given polynomial belongs to the ideal $\mathcal{I}$. The algorithm presented in this subsection is based on these two ideas.

Definition 4.1.10 Let F be a finite subset of the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. A polynomial $q \in \mathcal{R}$ is called an *admissible combination of* F if either $q = 0$ or q can be written as an expression of the form

$$q = \sum_{\gamma \in \mathbf{N}_0^n; p \in F} c(\gamma, p) \cdot x^\gamma \cdot p, \tag{4.7}$$

with $c(\gamma, p) \in \mathcal{K}$, and satisfying the condition

$$\deg(q) = \max\{\deg(x^\gamma \cdot p) \mid c(\gamma, p) \neq 0\}.$$

The case $q = 0$ may be considered as a special case of (4.7), when we regard the polynomial $q = 0$ as the polynomial generated by the empty sum. If $q \neq 0$ is an admissible combination of F, q can be generated in such a way that the terms of highest degree do not cancel out.

Example 4.1.11 If p_1 and p_2 are polynomials in $\mathcal{R}$, and $\alpha_1, \alpha_2 \in \mathbf{N}_0^n$, then $x^{\alpha_1} \cdot p_1 - x^{\alpha_2} \cdot p_2$ is an admissible combination of $\{p_1, p_2\}$ if and only if $x^{\alpha_1} \cdot \mathrm{in}(p_1) \neq x^{\alpha_2} \cdot \mathrm{in}(p_2)$.

Lemma 4.1.12 *Let F be a finite subset of the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$, and denote the ideal generated by the initial terms of F by $\langle \mathrm{in}(F) \rangle$. For every element $p \in \langle \mathrm{in}(F) \rangle$ there exists an admissible combination q of F such that*

$$\mathrm{in}(q) = \mathrm{in}(p). \tag{4.8}$$

Proof

Let $p \in \langle \mathrm{in}(F) \rangle$. Then there exists a polynomial $f \in F$ such that $\mathrm{in}(f)$ divides $\mathrm{in}(p)$. Take $q = \frac{\mathrm{in}(p)}{\mathrm{in}(f)} \cdot f$. ∎

In the next proposition, the notion of admissible combination plays the same role as division with remainder in the Euclidean division algorithm (compare Definition 4.1.1 and Proposition 4.1.2).

Proposition 4.1.13 (Generalized division algorithm) *Let F be a finite subset of the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. For every polynomial $p \in \mathcal{R}$ there exists a decomposition $p = q + \bar{p}$ satisfying the following properties:*

(i) q is an admissible combination of F,

(ii) $\bar{p} = 0$ or $\mathrm{in}(\bar{p}) \notin \langle \mathrm{in}(F) \rangle$.

Proof
Let $p \in \mathcal{R}$ and apply the following algorithm

$\bar{p} := p$; $q := 0$;
while ($\bar{p} \neq 0$ and $\mathrm{in}(\bar{p}) \in \langle \mathrm{in}(F) \rangle$) do
  $m :=$ admissible combination of F such that $\mathrm{in}(m) = \mathrm{in}(\bar{p})$;
  $q := q + m$;
  $\bar{p} := \bar{p} - m$;
od;

Since in every step of the algorithm, the degree of the admissible combination m strictly decreases, it follows that the desired result is obtained, if the algorithm terminates. To prove termination, let $p_0 = p$, and p_k be the value $\bar{p}$ after k loops. Suppose that $p_k \neq 0$ and $\mathrm{in}(p_k) \in \langle \mathrm{in}(F) \rangle$. Then $\deg(p_{k+1}) < \deg(p_k)$. So $(p_0, p_1, \ldots)$ is a sequence of polynomials of strictly descending degree. Since $>$ is a well-ordering (condition *(iii)* of Definition 4.1.6), every strictly decreasing sequence is finite. Hence the algorithm terminates. ∎

Note that if a nonzero polynomial p, with $\mathrm{in}(p) \in \langle \mathrm{in}(F) \rangle$, is decomposed into $p = q + \bar{p}$ according to Proposition 4.1.13, then $\mathrm{in}(q) = \mathrm{in}(p)$.

Remark 4.1.14 The algorithm in the proof of Proposition 4.1.13 is called *division by F*. The polynomial $\bar{p}$ that is obtained after termination of this algorithm is called a *remainder of p after division by F*. This polynomial plays the same role as the remainder in the Euclidean division algorithm.

Remark 4.1.15 Since there is a lot of freedom in the choice of an admissible combination $m \in F$ such that $\mathrm{in}(m) = \mathrm{in}(\bar{p})$, the remainder of a polynomial after division by F is not unique.

Remark 4.1.16 If there exists a remainder of the polynomial p after division by F that is zero, then it is obvious that $p \in \langle F \rangle$. However, in general it is possible that there are polynomials in the ideal $\langle F \rangle$ for which there exists a remainder after division by F that is nonzero. In the next subsection we shall see that exactly the property

$$p \in \langle F \rangle \iff \text{each remainder of } p \text{ after division by } F \text{ is zero},$$

characterizes a Gröbner basis.

4.1.4 The definition of Gröbner bases

This subsection is devoted to the formal definition of the concept of Gröbner bases. Moreover, the relationship between Gröbner bases and the generalized division algorithm is elaborated.

Definition 4.1.17 Let $\mathcal{I}$ be an ideal in the polynomial ring $\mathcal{R} = K[x_1, \ldots, x_n]$ with term ordering $>$, and assume that $\mathcal{I}$ is nonzero. A finite subset G of $\mathcal{I}$ is called a *Gröbner basis* of $\mathcal{I}$ if and only if $\mathrm{in}(G)$ generates the ideal $\langle \mathrm{in}(\mathcal{I}) \rangle$, i.e.

$$\langle \mathrm{in}(G) \rangle = \langle \mathrm{in}(\mathcal{I}) \rangle \tag{4.9}$$

Remark 4.1.18 Every nonzero polynomial ideal $\mathcal{I}$ has a Gröbner basis. Namely, consider the ideal $\langle \text{in}(\mathcal{I}) \rangle$ and let M be a finite subset of $\text{in}(\mathcal{I})$ generating the ideal $\langle \text{in}(\mathcal{I}) \rangle$. Then any finite subset G of $\mathcal{I}$ such that $M \subseteq \text{in}(G)$ is a Gröbner basis of $\mathcal{I}$. This implies that a Gröbner basis is not unique. In fact, every finite subset of $\mathcal{I} \backslash \{0\}$ containing a Gröbner basis of $\mathcal{I}$, is itself a Gröbner basis of $\mathcal{I}$.

Remark 4.1.19 A finite set of monomials is a Gröbner basis for the ideal generated by these monomials.

Definition 4.1.17 indicates that for a finite subset F of an ideal $\mathcal{I}$ the property of being a Gröbner basis of $\mathcal{I}$ is completely determined by the initial terms of F and $\mathcal{I}$. The definition can be seen as a generalization of the one-indeterminate case. In that situation, the generator g of an ideal $\mathcal{I}$ is a nonzero polynomial of minimal degree. Hence $\langle \text{in}(g) \rangle = \langle \text{in}(\mathcal{I}) \rangle$. In this perspective, Definition 4.1.17 is simply a generalization of this characterization to polynomial rings in more than one indeterminate. There are even more resemblances: also the relationship between the generator of an ideal and the Euclidean division algorithm is preserved in some sense.

Proposition 4.1.20 *Let $\mathcal{I}$ be an ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. Let F be a finite subset of $\mathcal{I}$. Then the following conditions are equivalent:*

(i) F is a Gröbner basis of $\mathcal{I}$.

(ii) For every polynomial $p \in \mathcal{I}$, each remainder of p after division by F is zero.

(iii) For every polynomial $p \in \mathcal{I}$, there exists a remainder of p after division by F that is zero.

Proof

(i) $\Rightarrow$ *(ii)* Assume that F is a Gröbner basis of $\mathcal{I}$, and let $p \in \mathcal{I}$. After division by F, p can be written as $p = q + \bar{p}$, where q is an admissible combination of F, and either $\bar{p} = 0$ or $\text{in}(\bar{p}) \notin \langle \text{in}(F) \rangle$. Clearly $q \in \mathcal{I}$, and thus also $\bar{p} = p - q \in \mathcal{I}$. Since F is a Gröbner basis of $\mathcal{I}$, we have

$$\text{in}(\bar{p}) \in \langle \text{in}(\mathcal{I}) \rangle = \langle \text{in}(F) \rangle.$$

Hence $\bar{p} = 0$.

(ii) $\Rightarrow$ *(iii)* Trivial.

(iii) $\Rightarrow$ *(i)* Assume that for every $p \in \mathcal{I}$ there exists a remainder of p after division by F that is zero. Let $p \in \mathcal{I}$. Then there exists an admissible combination q of F such that $p = q$. So $\text{in}(p) \in \langle \text{in}(F) \rangle$. Since $p \in \mathcal{I}$ was arbitrary we conclude that $\langle \text{in}(\mathcal{I}) \rangle = \langle \text{in}(F) \rangle$, and F is a Gröbner basis of $\mathcal{I}$.

■

Corollary 4.1.21 *Let $\mathcal{I}$ be an ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$, and let G be a Gröbner basis of $\mathcal{I}$. Then*

(i) G generates the ideal $\mathcal{I}$: $\langle G \rangle = \mathcal{I}$.

(ii) Let $p \in \mathcal{R}$. Then $p \in \mathcal{I}$ if and only if each remainder of p after division by G is zero. ∎

Corollary 4.1.21 states some very interesting properties of Gröbner bases. First of all they are generating sets for the corresponding ideal, but moreover, when a Gröbner basis of an ideal is obtained, the membership problem can be solved algorithmically using the generalized division algorithm of Proposition 4.1.13. Therefore a Gröbner basis of an ideal can be seen as a generating subset with some very useful properties. Note that the terminology "basis" for a generating subset is a little bit old-fashioned because the elements of a Gröbner basis are not linearly independent in general.

Unfortunately not every generating subset of an ideal is a Gröbner basis.

Example 4.1.22 Let $\mathcal{R} = \mathbb{R}[x_1, x_2]$, and choose the ranking $x_1 \succ x_2$ and the pure lexicographic ordering of Definition 4.1.7 *(i)*. Define

$$\begin{aligned} f_1 &:= x_1 + x_2^2 - 1, & \deg(f_1) &= (1,0); \quad \mathrm{in}(f_1) = x_1. \\ f_2 &:= x_1x_2 + x_2 + 3, & \deg(f_2) &= (1,1); \quad \mathrm{in}(f_2) = x_1x_2. \end{aligned}$$

Let $F := \{f_1, f_2\}$ and $\mathcal{I} := \langle F \rangle$. Then F is <u>not</u> a Gröbner basis of $\mathcal{I}$. This due to the fact that

$$g := -x_2^3 + 2x_2 + 3 = -x_2 \cdot f_1 + f_2 \in \mathcal{I},$$

and $\mathrm{in}(g) = -x_2^3$ is not an element of the ideal $\langle x_1, x_1x_2 \rangle = \langle \mathrm{in}(F) \rangle$. With the results of the next subsection it is easy to show that $\{f_1, g\}$ is a Gröbner basis of $\mathcal{I}$, but at this moment this statement is difficult to verify.

We conclude that there is still one problem left. Given a finite set of generators of a polynomial ideal, we want to construct a Gröbner basis of that ideal. This is the subject of the next subsection.

4.1.5 Computation of Gröbner bases

The importance of the Gröbner basis method does not only originate from the nice structure of these generating sets. The interest in this method was mainly initiated by the fact that Gröbner bases can be constructed algorithmically. This makes it possible to apply this technique for the solution of several problems concerning polynomial ideals.

The idea behind the construction algorithm is straightforward. Given a finite set F of polynomials, we take a polynomial p in the ideal $\langle F \rangle$, and compute the remainder of p after division by F. This remainder has to be an element of $\langle F \rangle$. If it is nonzero, we add the remainder to the set F, and continue with this new extended set of generators. Now a few questions arise. How do you choose a polynomial $p \in \langle F \rangle$? Is there any guarantee that the algorithm terminates, in other words, is a Gröbner basis obtained after a finite number of steps, and how is this verified? The main point of the Gröbner basis algorithm is that when we consider only a special kind of polynomials in the ideal $\langle F \rangle$ (namely the so-called S-polynomials), all these questions are answered automatically.

Definition 4.1.23 Let p_1, p_2 be elements of the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. Define $\alpha := \deg(p_1)$ and $\beta := \deg(p_2)$. Let γ be the n-tuple in $\mathbb{N}_0^n$ such that $\gamma_i = \max(\alpha_i, \beta_i)$, $(i = 1, \ldots, n)$. The monomial x^γ is called a *least common multiple* of $\mathrm{in}(p_1)$ and $\mathrm{in}(p_2)$. Now the *S-polynomial* of p_1 and p_2 is defined as

$$S(p_1, p_2) := \frac{x^\gamma}{\mathrm{in}(p_1)} \cdot p_1 - \frac{x^\gamma}{\mathrm{in}(p_2)} \cdot p_2. \tag{4.10}$$

The name S-polynomial is an abbreviation of subtraction polynomial. They are constructed in such a way that the initial terms of the two components p_1 and p_2 cancel out.

Example 4.1.24 The polynomial g in Example 4.1.22 is the S-polynomial of f_2 and f_1.

S-polynomials have the following important property:

Proposition 4.1.25 *Let $\mathcal{I}$ be an ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. Assume that $\mathcal{I}$ is generated by a finite subset $F \subset \mathcal{R}\backslash\{0\}$. Then*

$$\begin{array}{ll} & F \text{ is a Gröbner basis of } \mathcal{I}, \\ \Longleftrightarrow & \\ & \forall p, q \in F: \text{ there exists a remainder of } S(p,q) \text{ after division by } F \\ & \qquad \text{that is zero.} \end{array}$$

■

For a proof we refer to [76, p. 226], [7, p.191] or [14, p.84].

Proposition 4.1.25 implies that we only have to consider remainders of S-polynomials to verify whether a generating set is a Gröbner basis of an ideal. This observation leads quite naturally to the following construction method for Gröbner bases.

Theorem 4.1.26 (Gröbner basis algorithm) *Let $\mathcal{I}$ be a nonzero ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$. Let F be a finite set in $\mathcal{R}$ which generates $\mathcal{I}$. Then a Gröbner basis of $\mathcal{I}$ can be constructed with the following algorithm in a finite number of steps:*

$$\begin{aligned} F_0 &:= F; \\ F_{i+1} &:= F_i \cup (\{\overline{S(p,q)}_i \mid p, q \in F_i\}\backslash\{0\}); \end{aligned}$$

where $\overline{S(p,q)}_i$ denotes a remainder of the S-polynomial of p and q after division by F_i. If $F_{i+1} = F_i$, then F_i is a Gröbner basis of the ideal $\mathcal{I}$.

Proof

According to Proposition 4.1.25 it suffices to show that this algorithm terminates. So we only have to find a $k \in \mathbb{N}$ such that $F_k = F_{k+1}$.

By definition, $F_i \subseteq F_{i+1}$, and thus $\langle \text{in}(F_i) \rangle \subseteq \langle \text{in}(F_{i+1}) \rangle$. Since $\mathcal{K}[x_1, \ldots, x_n]$ is a Noetherian ring, the ascending chain

$$\langle \text{in}(F_0) \rangle \subseteq \langle \text{in}(F_1) \rangle \subseteq \cdots$$

becomes stationary. Hence there exists a $k \in \mathbf{N}$ such that

$$\langle \text{in}(F_k) \rangle = \langle \text{in}(F_{k+1}) \rangle.$$

Let $r \in F_{k+1} \backslash F_k$. Then there are polynomials p and q in F_k such that r is a remainder of $S(p, q)$ after division by F_k. However, $\text{in}(r) \in \langle \text{in}(F_k) \rangle$, and thus Proposition 4.1.13 implies that $r = 0$. This contradicts the definition of F_{k+1}, and we conclude that $F_{k+1} \backslash F_k = \varnothing$. ∎

The algorithm of Theorem 4.1.26 only describes the main idea behind the construction of Gröbner bases. The algorithms implemented in most computer algebra packages contain a lot of refinements to prevent unnecessary divisions and adding of polynomials to a basis. For more details on this subject we refer to [7] and [2].

Remark 4.1.27 Application of the Gröbner basis algorithm to an ideal generated by a finite set of polynomials $F = \{f_1, \ldots, f_k\}$ in $\mathcal{R}$ does not only yield a Gröbner basis $G = \{g_1, \ldots, g_\ell\}$. Implicitly, also the relations between the polynomials in F and G are computed. All polynomials in G are $\mathcal{R}$-linear combinations of $f_1, \ldots, f_k$. Sometimes these polynomial coefficients relating the polynomials of G to F are needed in subsequent computations. By an accurate bookkeeping of the interrelations between polynomials, it is possible to obtain these coefficients during the Gröbner basis algorithm. For this extra information one has to pay a price: the necessary computer space and time are increased considerably.

It was already noted in Remark 4.1.18 that an ideal does not have a unique Gröbner basis with respect to a given term ordering. This degree of freedom can be eliminated by imposing some extra conditions on Gröbner bases.

Definition 4.1.28 Let $\mathcal{I}$ be an ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$ with term ordering $>$, and assume that $\mathcal{I}$ is nonzero. A *reduced Gröbner basis* of $\mathcal{I}$ is a Gröbner basis of $\mathcal{I}$ such that

(i) $\forall p \in G : \text{lc}(p) = 1$,

(ii) For all $p \in G$, no monomial of p lies in $\langle \text{in}(G \backslash \{p\}) \rangle$.

Proposition 4.1.29 *Let $\mathcal{I}$ be a nonzero polynomial ideal in $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$. For a given term ordering $>$ on $\mathcal{R}$, the ideal $\mathcal{I}$ has a unique reduced Gröbner basis.* ∎

For the proof of this result we refer to [14, p.91]. There a proof of the existence and uniqueness of a reduced Gröbner basis is given, and also an algorithm for its computation.

The algorithms in most computer algebra packages like Maple and Reduce are based on the definition of reduced Gröbner bases, sometimes with a slightly different convention on the leading coefficients. In the Gröbner basis algorithm of these

programs, the reduction process is carried out automatically. Note that reduced Gröbner bases w.r.t. different term orderings may be different. Also the computational expenses differ. The computation of a reduced Gröbner basis w.r.t the pure lexicographic ordering is typically far more time consuming than the computation w.r.t. the graded lexicographic (tdeg) ordering or the graded reverse lexicographic (grevlex) ordering. In the next subsection it turns out that this observation has important implications for some of the applications of the Gröbner basis method, especially for elimination.

4.1.6 Application of Gröbner bases

The Gröbner basis method can be applied to solve several questions on polynomial ideals and their varieties. For example, the result of Corollary 4.1.21 yields a method to decide on the membership problem in an algorithmic way. Also the problem whether two polynomial ideals are equal is not difficult to solve. Since a reduced Gröbner basis is unique w.r.t. the chosen term ordering, it suffices to test whether the reduced Gröbner bases of the two ideals are identical. However, to one of the main applications of the Gröbner basis method we have paid too little attention thus far: the use of Gröbner bases for elimination.

Let $f_1, \ldots, f_r$ be polynomials in $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$, and let $\bar{\mathcal{K}}$ denote the algebraic closure of $\mathcal{K}$. Suppose that we want to find a solution in $\bar{\mathcal{K}}^n$ of the system of polynomial equations $f_1(x_1, \ldots, x_n) = f_2(x_1, \ldots, x_n) = \cdots = f_r(x_1, \ldots, x_n) = 0$. So we are interested in the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I} := \langle f_1, \ldots, f_r \rangle$ generated by the polynomials $f_1, \ldots, f_r$ (see also Appendix A.2). A Gröbner basis of $\mathcal{I}$ contains a lot of information on this variety $\mathcal{V}(\mathcal{I})$.

Proposition 4.1.30 *Let $f_1, \ldots, f_r$ be polynomials in $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$, and $\mathcal{I} = \langle f_1, \ldots, f_r \rangle$. Let G be a Gröbner basis of $\mathcal{I}$ w.r.t. a given term ordering $>$. Then*

(i) $\mathcal{V}(\mathcal{I}) = \varnothing \iff G \cap \mathcal{K} \neq \varnothing$,

(ii) $\mathcal{V}(\mathcal{I})$ *is zero-dimensional, i.e.* $\mathcal{V}(\mathcal{I})$ *is non-empty and contains finitely many points, if and only if*

$$\forall i \in \{1, \ldots, n\}\ \exists p \in G\ \exists j \in \mathbf{N} : \frac{\mathrm{in}(p)}{\mathrm{lc}(p)} = x_i^j. \tag{4.11}$$

Sketch of the proof

(i) is a very straightforward corollary of the Hilbert Nullstellensatz. A proof of *(ii)* can be found in [7, p.209] or [14, p. 232]. It relies on the fact that $\mathcal{V}(\mathcal{I})$ is zero-dimensional if and only if $\mathcal{K}[x_1, \ldots, x_n]/\mathcal{I}$ is a finite (but not zero) dimensional vector space over $\mathcal{K}$. ∎

According to Proposition 4.1.30, a Gröbner basis of an ideal indicates whether the corresponding variety contains zero, a finite number, or infinitely many elements. When the lexicographic ordering is used, the Gröbner basis method can even be applied for the elimination of indeterminates.

Theorem 4.1.31 *Let $\mathcal{I}$ be a nonzero ideal in the polynomial ring $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$. Assume that the term ordering on $\mathcal{R}$ is fixed by the ranking $x_1 \succ x_2 \succ \cdots \succ x_n$, and the pure lexicographic ordering on $\mathbf{N}_0^n$ as defined in Definition 4.1.7 (i). Let G be a Gröbner basis of $\mathcal{I}$ w.r.t. this term ordering. Then for all $k \in \{1, \ldots, n\}$,*

$$G_k := G \cap \mathcal{K}[x_k, \ldots, x_n] \tag{4.12}$$

is a Gröbner basis of the ideal

$$\mathcal{I}_k := \mathcal{I} \cap \mathcal{K}[x_k, \ldots, x_n] \tag{4.13}$$

in the polynomial ring $\mathcal{K}[x_k, \ldots, x_n]$.

Proof

Let $k \in \{1, \ldots, n\}$ and $q \in \mathcal{I}_k$. Because of the lexicographic term ordering, we know that for all polynomials $p \in \mathcal{R}$, the inequality $\deg(p) \leq \deg(q)$ implies that $p \in \mathcal{K}[x_k, \ldots, x_n]$. Since $q \in \mathcal{I}$, Propositions 4.1.13 and 4.1.20 imply that q is an admissible combination of G:

$$q = \sum_{p \in G, \alpha \in \mathbf{N}_0^n} c(\alpha, p) \cdot x^\alpha \cdot p.$$

If $c(\alpha, p) \neq 0$, then $\deg(x^\alpha \cdot p) \leq \deg(q)$. Hence, $c(\alpha, p) \neq 0$ implies that $x^\alpha \cdot p \in \mathcal{K}[x_k, \ldots, x_n]$. Therefore q is an admissible combination of G_k, and it follows from Proposition 4.1.20 that G_k is a Gröbner basis of the ideal $\mathcal{I}_k$ in $\mathcal{K}[x_k, \ldots, x_n]$. ∎

We conclude that application of the Gröbner basis algorithm with respect to a lexicographic term ordering yields a set of generators of an ideal that is in *triangular form*. First a number of polynomials in indeterminates of low rank is obtained, and later on also indeterminates of higher rank occur. If the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ is zero-dimensional, and the Gröbner basis G is reduced, G_n contains exactly one polynomial, and all elements of $\mathcal{V}(\mathcal{I})$ can be computed by backward substitution. This is possible because the zero-dimensionality of $\mathcal{I}$ guarantees that $G_{k+1} \subsetneq G_k$ for all $k \in \{1, \ldots, n-1\}$. In this perspective, the Gröbner basis algorithm looks very similar to Gauß-elimination, where a triangular form for a system of linear equations is obtained to facilitate the search of common zeros.

Example 4.1.32 Consider the system of polynomial equations

$$\begin{cases} x_3^2 - 5x_1 &= -1, \\ x_1x_2 - 4x_1x_3 + 6x_3 &= -2, \\ x_1^2x_2 + 3x_2x_3 &= 2. \end{cases}$$

To find a solution, we compute a Gröbner basis of the ideal $\langle f_1, f_2, f_3 \rangle$ in $\mathbf{R}[x_1, x_2, x_3]$, where $f_1 = x_3^2 - 5x_1 + 1$, $f_2 = x_1x_2 - 4x_1x_3 + 6x_3 + 2$ and $f_3 = x_1^2x_2 + 3x_2x_3 - 2$. We use the lexicographic term ordering with ranking $x_1 \succ x_2 \succ x_3$. In this way the Gröbner basis $\{g_1, g_2, g_3\}$ is obtained, where

$$\begin{aligned} g_1 &= 5x_1 - x_3^2 - 1, \\ g_2 &= 75x_2 - 4x_3^6 + 22x_3^4 - 290x_3^3 + 26x_3^2 + 2010x_3 + 750, \\ g_3 &= 2x_3^7 - 9x_3^5 + 145x_3^4 - 24x_3^3 - 1010x_3^2 - 388x_3 - 30. \end{aligned}$$

To find a solution, we first compute all solutions to the equation $g_3 = 0$. Numerically this is relatively easy because g_3 is a univariate polynomial in the indeterminate x_3. Substitution of these solutions in the equations $q_1 = 0$ and $q_2 = 0$ yields the corresponding values of x_1 and x_2.

Remark 4.1.33 Note that the ranking of the indeterminates prescribes in what order the indeterminates are eliminated. When there are some variables of special interest, the ranking may be chosen in such a way that all other indeterminates are eliminated first.

The elimination procedure of Theorem 4.1.31 is not only useful for the computation of common zeros of polynomials; it also helps to find a Gröbner basis of the intersection of two polynomial ideals. This method is based on the following result.

Lemma 4.1.34 *Let $\mathcal{I}$ and $\mathcal{J}$ be ideals in the polynomial ring $\mathcal{K}[x_1, \ldots, x_n]$, and let t denote another indeterminate. Then*

$$\mathcal{I} \cap \mathcal{J} = (t \cdot \mathcal{I} + (1-t) \cdot \mathcal{J}) \cap \mathcal{K}[x_1, \ldots, x_n]. \tag{4.14}$$

■

For a proof of this result we refer to [14, p. 186].

Remark 4.1.35 Suppose that $\mathcal{I} = \langle p_1, \ldots, p_k \rangle$ and $\mathcal{J} = \langle q_1, \ldots, q_\ell \rangle$ are ideals in the polynomial ring $\mathcal{K}[x_1, \ldots, x_n]$. Then the ideals $t \cdot \mathcal{I}$, $(1-t) \cdot \mathcal{J}$ and $t \cdot \mathcal{I} + (1-t) \cdot \mathcal{J}$ belong to the extended polynomial ring $\mathcal{K}[x_1, \ldots, x_n, t]$. To obtain a Gröbner basis of $\mathcal{I} \cap \mathcal{J}$, we first compute a Gröbner basis G of

$$t \cdot \mathcal{I} + (1-t) \cdot \mathcal{J} = \langle t \cdot p_1, \ldots, t \cdot p_k, (1-t) \cdot q_1, \ldots, (1-t) \cdot q_\ell \rangle$$

w.r.t. the lexicographic term ordering with the rank of t higher than that of all other indeterminates. Combining Lemma 4.1.34 and Theorem 4.1.31, we conclude that $G \cap \mathcal{K}[x_1, \ldots, x_n]$ is a Gröbner basis of $\mathcal{I} \cap \mathcal{J}$ w.r.t. the lexicographic term ordering.

The elimination method using Gröbner bases w.r.t. a pure lexicographic ordering also has an important shortcoming. As already mentioned in Subsection 4.1.5, the computation of a Gröbner basis w.r.t. a pure lexicographic ordering is very time consuming, certainly when we compare it with other term orderings. Therefore this part of the elimination method is not very attractive from the computational point of view. However, for zero-dimensional ideals this problem can be circumvented. In this case it is possible to construct a univariate polynomial in the ideal under consideration with help of an arbitrary Gröbner basis.

Proposition 4.1.36 *Let $\mathcal{I}$ be a nonzero ideal in $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$, and let G be a Gröbner basis of $\mathcal{I}$ w.r.t. an arbitrary term ordering. Let $i \in \{1, \ldots, n\}$ and consider a univariate polynomial p in the indeterminate x_i of degree k:*

$$p(x_i) = \sum_{j=0}^{k} a_j x_i^j.$$

Let for each $j \in \{0, 1, \ldots, k\}$ *the monomial* x_i^j *be written as*

$$x_i^j = q_j + r_j,$$

where q_j *is an admissible combination of* G *and* r_j *is a remainder of* x_i^j *after division by* G. *Then*

$$p \in \mathcal{I} \iff \sum_{j=0}^{k} a_j r_j = 0. \tag{4.15}$$

Proof

"$\Rightarrow$" Assume that $p \in \mathcal{I}$. Then

$$\sum_{j=0}^{k} a_j x_i^j = \sum_{j=0}^{k} a_j q_j + \sum_{j=0}^{k} a_j r_j \in \mathcal{I}.$$

The admissible combination q_j of G is an element of $\mathcal{I}$ for all $j \in \{0, 1, \ldots, k\}$. Therefore $\sum_{j=0}^{k} a_j q_j \in \mathcal{I}$ and thus

$$\sum_{j=0}^{k} a_j r_j \in \mathcal{I}.$$

Since all r_j $(j = 0, 1, \ldots, k)$ are remainders after division by G, we know that for every $j = 0, 1, \ldots, k$ either $r_j = 0$ or $\mathrm{in}(r_j) \notin (\mathrm{in}(G))$. This implies that the polynomial $\sum_{j=0}^{k} a_j r_j$ does not contain a nonzero admissible combination of G, and therefore it is a remainder after division by G. Since G is a Gröbner basis of $\mathcal{I}$, we conclude that $\sum_{j=0}^{k} a_j r_j = 0$.

"$\Leftarrow$" Assume that $\sum_{j=0}^{k} a_j r_j = 0$. Since q_j $(j = 0, 1, \ldots, k)$ is an admissible combination of G, it follows that $\sum_{j=0}^{k} a_j q_j + \sum_{j=0}^{k} a_j r_j \in \mathcal{I}$. Hence $p \in \mathcal{I}$. ∎

Zero-dimensional ideals always contain univariate polynomials in each of their indeterminates. Using Proposition 4.1.36 we obtain the following result (see e.g. [4] and [7]).

Proposition 4.1.37 *Let* $\mathcal{I}$ *be a zero-dimensional ideal in* $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$, *and let* G *be a Gröbner basis of* $\mathcal{I}$ *w.r.t. an arbitrary term ordering. Let* $i \in \{1, \ldots, n\}$. *Compute the remainders* r_j *of* x_i^j $(j = 0, 1, \ldots)$ *after division by* G, *until these remainders become linearly dependent over* $\mathcal{K}$:

$$\sum_{j=0}^{k} a_j r_j = 0.$$

Then $\sum_{j=0}^{k} a_j x_i^j$ *is a univariate polynomial in the indeterminate* x_i *in* $\mathcal{I}$ *of lowest possible degree.* ∎

In the computer algebra package Maple, the method of Proposition 4.1.37 is available as an algorithm under the name `finduni`. Application of this algorithm speeds up the elimination procedure considerably.

4.1.7 Complexity issues and closing remarks

The computation of a Gröbner basis is extremely time and space consuming, certainly for somewhat larger problems. This observation is not very surprising because the problems that can be solved by this method are very complex themselves. A lot of research has been done in the investigation of this complexity issue for Gröbner bases. To get a little feeling for the problem, we mention a few results.

It is clear that the time and space consumption of an algorithm depends on the complexity of the input applied to it. For Gröbner bases, the complexity of the input is measured by the number of indeterminates of the polynomial ring under consideration, and by the degrees of the input polynomials. The number of input polynomials is often omitted because this number occurs as a linear term in the general expression. Denoting the number of indeterminates by n, and the degree by d, the complexity of the Gröbner basis computation was estimated by several authors, and under various assumptions. They obtain the following expressions

$$d^{O(n)}, \qquad d^{O(n^2)}, \qquad d^{2^{O(n)}}, \tag{4.16}$$

(see e.g. [64], [63], [70] and the references therein). These estimates illustrate that the computation of a Gröbner basis can be extremely involved. This is probably due to the fact that an enormous amount of intermediate polynomials have to be computed. These polynomials become more and more complex in each subsequent step of the algorithm. Moreover, formulae (4.16) indicate that especially the number of indeterminates in the polynomial ideal is critical for the complexity of Gröbner basis computations.

Next, we look at the outcome of all these computations: the Gröbner bases themselves. From Proposition 4.1.29 we recall that a reduced Gröbner basis of an ideal is unique w.r.t. the chosen term ordering. This implies that before the Gröbner basis computation w.r.t. a given term ordering starts, the number of elements of the corresponding reduced Gröbner basis is fixed. One of the shortcomings of the Gröbner basis method is that this number cannot be predicted beforehand. However, there exist some upper bounds for this number; an overview of existing results is given in [2, p. 513]. Unfortunately, an exact solution is not known yet, and the final reduced Gröbner basis can become very complex. Not only the number of elements of a Gröbner basis may become very large, also the degrees of these polynomials in some of the indeterminates, and their coefficients may grow very rapidly. This makes the final solution very difficult to overlook. Also possible subsequent numerical computations suffer from this complex structure of Gröbner bases: numerical computation of the zeros of a polynomial becomes more involved for polynomials of high degree and with large coefficients.

Finally, we mention once more that this section only contains the main ideas of the Gröbner basis algorithm. There is much more to say about refinements of the algorithm, and on other applications. It is impossible to give a list of representative references on this subject here. Instead we refer once more to [14] and [2] and to the bibliographies at the end of these books.

4.2 Characteristic sets

The Gröbner basis method is not the only method in constructive commutative algebra to manipulate polynomial ideals. There are several alternatives, for example the classical technique of resultants, and the characteristic sets algorithm. In this section we give a rather extensive introduction to this last method. Although this method is based on a completely different philosophy, there are also some resemblances with Gröbner bases. On the one hand, the characteristic set of a polynomial ideal characterizes the ideal in a completely different way than a Gröbner basis; in general it is not a generating set. Instead, the concept of so-called ascending chains is the key idea. On the other hand, the main ingredients of the characteristic sets method look very similar to those of Gröbner bases: there is a (partial) ordering on polynomials, and a (pseudo) division algorithm to compute polynomials of lower rank. In this perspective, the characteristic sets method can be seen as another sort of generalization of the Euclidean algorithm to polynomials in more than one indeterminate.

The notion of characteristic sets was introduced by J.F. Ritt in his work on differential algebra (see [78], [79]). Later on, E.R. Kolchin gave a more rigorous basis to the work of Ritt, at least from the algebraic point of view, in his exposition on differential algebra (see [59]). Although characteristic sets are intended to deal with differential algebra in the first place, they can also be very useful in ordinary commutative algebra. This was already pointed out by Ritt in [79], where he devoted one chapter to the non-differential case. However, in the field of constructive commutative algebra there was only little interest in Ritt's characteristic sets method. Through the work of Wu-Wen-Tsun (see [101]), it became clear that also in this field characteristic sets might be a powerful tool, and new interest was awakened in the subject. Finally, the implementation by D. Wang of the characteristic sets algorithm in the computer algebra package Maple (see [96] and [97]), made it possible to test the method on its practical merits.

One of the main difficulties in the study of characteristic sets is the fact that the definitions used by the authors mentioned above do not coincide. Ritt and Kolchin use a definition which has very strong properties and is theoretically very interesting, but with this definition characteristic sets are difficult to compute. Although Ritt suggests an algorithm, it is a partly non-constructive one. On the other hand, the definition of Wu-Wen-Tsun is based on the constructive part of Ritt's algorithm. It has the advantage that it is rather easily computable (this is in fact what Wang's implementation does), but the outcome is not as powerful as Ritt's original characteristic sets.

In this section we give an overview of both approaches simultaneously. One of the main topics is the internal structure of characteristic sets. We describe to what extent a characteristic set uniquely determines a polynomial ideal. In this way it is also possible to clear up the confusion, caused by the before mentioned disagreement on the definition issue. Furthermore we investigate some of the applications of the characteristic sets method in commutative algebra. Except for elimination, characteristic sets can also be used to solve several questions on polynomial ideals. The method is especially suitable for the representation of a radical ideal as a finite intersection of prime ideals.

In this section we follow the same lines as in [35]. This paper contains an overview of the characteristic sets method, mainly based on [79], [101] and [98]. It especially emphasizes the difference between Ritt- and Wu-characteristic sets. Moreover, some results are derived in a slightly different way. Another overview of the characteristic sets method can be found in [10].

Finally we remark that in this whole section we confine ourselves to the non-differential case.

4.2.1 Ritt-characteristic sets

The purpose of this subsection is merely to give the framework in which characteristic sets can be defined. We shall take the same approach as was originally taken by Ritt. To do so, we have to introduce a few, partly new, concepts; some of them we already encountered in the section on Gröbner bases, often in a slightly different form.

Let $\mathcal{K}$ be a fixed basic field of characteristic 0, and consider the set of indeterminates $\{x_1, x_2, \ldots, x_n\}$. In the same fashion as in Definition 4.1.3, a ranking is defined, inducing an ordering on these indeterminates. It describes which indeterminates are considered to be more important than others. To facilitate the notation in the subsequent subsections, we assume (without loss of generality) that this ranking is fixed in the following way:

$$x_1 \prec x_2 \prec \ldots \prec x_n. \tag{4.17}$$

In the characteristic sets method this ordering is used to introduce a *partial* ordering on the polynomials in $\mathcal{K}[x_1, \ldots, x_n]$. So here we encounter the first difference with Gröbner bases.

Definition 4.2.1 Let $f \in \mathcal{K}[x_1, \ldots, x_n]$. Then the *class* of f is defined as:

$$\text{class}(f) := \begin{cases} -1 & \text{if } f = 0, \\ 0 & \text{if } f \in \mathcal{K}\backslash\{0\}, \\ \max\{i \mid \deg_{x_i}(f) > 0\} & \text{if } f \in \mathcal{K}[x_1, \ldots, x_n]\backslash\mathcal{K}, \end{cases} \tag{4.18}$$

where $\deg_{x_i}(f)$ denotes the degree of f, considered as a polynomial in x_i with coefficients in $\mathcal{K}[x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$. So if f is not a constant, the class of f is the index of the largest variable (in the ordering (4.17)) that actually occurs in f.

Definition 4.2.2 Let $f, g \in \mathcal{K}[x_1, \ldots, x_n]$. Then f has a higher *rank* than g (notation $f \succ g$) if one of the following properties is satisfied:

(i) $\text{class}(f) > \text{class}(g)$,

(ii) $\text{class}(f) = \text{class}(g) = p$ for certain $p > 0$, and $\deg_{x_p}(f) > \deg_{x_p}(g)$.

If for two polynomials f and g neither $f \succ g$ nor $g \succ f$, then f and g are said to have the same rank, and we write $f \sim g$.

Note that the ranking defined in Definition 4.2.2 only imposes a *partial ordering* on the polynomial ring $\mathcal{K}[x_1, \ldots, x_n]$. For example, the polynomial $x_1^3 + x_1 x_2^2$ has lower rank than $x_1 + x_3$, but has the same rank as $x_1^2 + 2x_1^3 x_2^2 + x_1 x_2$.

Definition 4.2.3 Let f be a polynomial in $\mathcal{K}[x_1, \ldots, x_n]$, with $\text{class}(f) = p \geq 0$. A polynomial g is called *reduced* with respect to f (notation $g \lhd f$) if one of the following conditions is satisfied:

(i) $p = 0$ and $g = 0$,

(ii) $p > 0$ and $\deg_{x_p}(g) < \deg_{x_p}(f)$.

Of course it is also possible that a pair of polynomials is reduced with respect to each other. An example is the following:

$$\begin{aligned} f &= x_1^3 + 3x_1^2 + 5x_1 - 2 \\ g &= (x_1^2 + 2x_1)x_2^2 + x_1x_2 + 1. \end{aligned}$$

In a polynomial ring in one indeterminate, a generator of an ideal $\mathcal{I}$ is a nonzero polynomial that is reduced with respect to all other nonzero polynomials of the ideal. In this situation the Euclidean algorithm constructs a sequence of polynomials in which each successor is reduced with respect to its predecessor. The same ideas are applicable to polynomial rings in more than one indeterminate. When a polynomial f is not reduced with respect to a polynomial g, we can achieve this property by carrying out a same sort of division algorithm. However, when $\text{class}(g) = p$, then g is not necessarily monic, considered as a polynomial in x_p, and the normal Euclidean division algorithm cannot be applied. To overcome this problem we introduce an alternative way for computing the remainder of one polynomial with respect to another polynomial.

Definition 4.2.4 Let f be a polynomial in $\mathcal{K}[x_1, \ldots, x_n]$, and assume that $p = \text{class}(f) > 0$, and and $\deg_{x_p}(f) = d$. Then f can uniquely be written as

$$f = \sum_{i=0}^{d} \alpha_i x_p^i$$

with $\alpha_i \in \mathcal{K}[x_1, \ldots, x_{p-1}]$. The coefficient α_d, belonging to the highest power x_p^d of f is called the *initial* of f, denoted as

$$I_f = \alpha_d.$$

Note that there is a big difference between initials and initial terms as defined in (4.4); an initial term is always a monomial, while an initial may be a polynomial with more than one term. Using initials, the alternative division algorithm, called *pseudo-division*, is given by:

Definition 4.2.5 Let $f, g \in \mathcal{K}[x_1, \ldots, x_n]$, and assume that $g \neq 0$. Then there exist an integer $\nu \in \mathbb{N} \cup \{0\}$ and a polynomial q such that $I_g^\nu f - qg$ is reduced with respect to g. The *pseudo-remainder* of f with respect to g is the polynomial R in the formula

$$I_g^\nu f = qg + R,$$

where ν is chosen as small as possible, while R remains reduced with respect to g. In this way, R is uniquely defined. Notation: $R = \text{prem}(f, g)$.

Definition 4.2.5 is illustrated by the following example. Choose:

$$\begin{aligned} f &= x_1x_2^2 + x_1^2 \\ g &= x_1x_2 + 1. \end{aligned}$$

Then $x_1(x_1x_2^2 + x_1^2) = (x_1x_2 + 1)(x_1x_2 - 1) + 1 + x_1^3$. So $\text{prem}(f, g) = x_1^3 + 1$.

With the foregoing definitions, we can introduce the key idea of the characteristic sets method: the notion of ascending chains.

Definition 4.2.6 An *ascending chain* $\mathcal{A}$ is a sequence of polynomials in the ring $\mathcal{K}[x_1, \ldots, x_n]$,

$$\mathcal{A} = (f_1, \ldots, f_r),$$

which satisfies one of the following properties:

(i) $r = 1$ and $f_1 \neq 0$;

(ii) $r > 1, 0 < \text{class}(f_1) < \text{class}(f_2) < \cdots < \text{class}(f_r)$, and moreover, f_j is reduced with respect to f_i for each $j > i$ $(i = 1, \ldots, r-1)$.

Note that when $\mathcal{A}$ is an ascending chain, the set $F = \{f_1, \ldots, f_r\}$ is auto-reduced, i.e. for all pairs $f_i, f_j \in F$, $i \neq j$, f_i is reduced with respect to f_j and f_j is reduced with respect to f_i. By definition, an ascending chain has a triangular form and contains at most n polynomials. In the sequel it turns out that this makes the characteristic sets method especially suitable for elimination purposes.

The concepts of reducedness and pseudo-remainders may be extended to ascending chains in the following way:

Definition 4.2.7 Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in $\mathcal{K}[x_1, \ldots, x_n]$. A polynomial $g \in \mathcal{K}[x_1, \ldots, x_n]$ is called *reduced w.r.t.* $\mathcal{A}$ if g is reduced w.r.t. all polynomials $f_1, \ldots, f_r$.

Definition 4.2.8 Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in $\mathcal{K}[x_1, \ldots, x_n]$, and assume that for $i = 1, \ldots, r$: $p_i := \text{class}(f_i) > 0$. Denote the initial of f_i by I_i. Let g be an arbitrary polynomial and define $R_r := g$. Now it is possible to form successively the remainders $R_{r-1}, \ldots, R_0$ by pseudo dividing R_i by f_i:

$$\begin{aligned} I_r^{\nu_r} R_r &= q_r f_r + R_{r-1} \\ I_{r-1}^{\nu_{r-1}} R_{r-1} &= q_{r-1} f_{r-1} + R_{r-2} \\ &\vdots \\ I_1^{\nu_1} R_1 &= q_1 f_1 + R_0 \end{aligned}$$

Defining $R := R_0$, we can combine these equations in order to get:

$$I_1^{\nu_1} \cdots I_r^{\nu_r} g = \tilde{q}_1 f_1 + \cdots + \tilde{q}_r f_r + R, \tag{4.19}$$

and we write $R = \text{prem}(g, \mathcal{A})$. R is called the *pseudo-remainder* of g with respect to the ascending chain $\mathcal{A}$.

If $\mathcal{A}$ only consists of a nonzero constant, $\text{prem}(g, \mathcal{A}) := 0$ for all $g \in \mathcal{K}[x_1, \ldots, x_n]$.

It is obvious that a pseudo-remainder with respect to an ascending chain is reduced w.r.t. that chain.

Next, the partial ordering $\prec$ on polynomials is extended to ascending chains.

Definition 4.2.9 Let $\mathcal{A}$ and $\mathcal{B}$ be ascending chains in $K[x_1, \ldots, x_n]$,

$$\begin{aligned} \mathcal{A} &= (f_1, \ldots, f_r), \\ \mathcal{B} &= (g_1, \ldots, g_s). \end{aligned}$$

Then $\mathcal{B}$ has lower rank than $\mathcal{A}$ (notation $\mathcal{B} \prec \mathcal{A}$) if either *(i)* or *(ii)* below holds true:

(i) $\exists j \leq \min(r, s)$ such that $f_1 \sim g_1, \ldots, f_{j-1} \sim g_{j-1}$ while $g_j \prec f_j$.

(ii) $s > r$ and $f_1 \sim g_1, \ldots, f_r \sim g_r$.

With this definition it is possible for a given set of ascending chains, to introduce the notion of minimal ascending chains. The next two lemmas state this observation in a more formal way.

Lemma 4.2.10 *In every nonempty set of ascending chains in $K[x_1, \ldots, x_n]$ there is an ascending chain which is not of higher rank than any other chain in the set.*

Proof

We here give the Ritt's proof (see also [79, p. 4]).

Let V be an arbitrary set of ascending chains in $K[x_1, \ldots, x_n]$. Define V_1 as the set of all ascending chains in V, such that the first polynomial of a chain in V_1 does not have higher rank than the first polynomial of any other chain in V. If all chains in V_1 contain only one polynomial, we are ready: every chain in V_1 has the desired property. Otherwise, let V_2 be the set of all chains in V_1, such that the rank of the second polynomial of a chain in V_2 is not higher than the rank of the second polynomial of any chain in V_1. Again, if all chains in V_2 have two elements, we are done: every chain in V_2 has the desired property. Otherwise we construct V_3 in the same way as before. In this way we continue, but because an ascending chain contains at most n polynomials, the process terminates after at most n steps. ∎

The next lemma is a specialization of the previous result.

Lemma 4.2.11 *Let $(\mathcal{A}_k)_{k \in \mathbf{N}}$ be a sequence of ascending chains in $K[x_1, \ldots, x_n]$ such that for all $k \in \mathbf{N}$ we have that either $\mathcal{A}_{k+1} \prec \mathcal{A}_k$ or $\mathcal{A}_{k+1} \sim \mathcal{A}_k$. This means that the rank of the ascending chains never increases. Then there exists an index $\tilde{k}$ such that for all $k > \tilde{k}$:*

$$\mathcal{A}_k \sim \mathcal{A}_{\tilde{k}}.$$

So, for all $k \geq \tilde{k}$, $\mathcal{A}_k$ is a minimal ascending chain for the sequence. ∎

A proof of this result may be found in [101].

Now, let F be a (not necessarily finite) set of polynomials, containing at least one nonzero polynomial. An ascending chain $\mathcal{A}$ is said to belong to F if every polynomial of $\mathcal{A}$ is an element of F.

Definition 4.2.12 Let F be a finite or infinite set of polynomials, containing a nonzero polynomial. Then an ascending chain $\mathcal{A}$ is called a *(Ritt)-characteristic set* of F if the following two conditions hold:

(i) $\mathcal{A}$ belongs to F,

(ii) if $\mathcal{B}$ is an ascending chain and $\mathcal{B}$ belongs to F, then the rank of $\mathcal{B}$ is not lower than the rank of $\mathcal{A}$.

From Lemma 4.2.10 it is immediately clear that every non-empty polynomial set $F \neq \{0\}$ has a characteristic set. However, this characteristic set is not necessarily unique. Moreover, the definition is valid for both finite and infinite polynomial sets. So characteristic sets of a finite polynomial set are defined in the same way as characteristic sets of the polynomial ideal generated by this finite polynomial set, although they are, in general, completely different.

Remark 4.2.13 The definition of characteristic sets as given above was introduced by Ritt in [78]. It is different from the one Wu-Wen-Tsun gives in [101]. The differences will be pointed out later on in Subsection 4.2.3.

4.2.2 Some results on Ritt-characteristic sets

In this subsection we derive some results, mainly due to Ritt (see [79]), that show why characteristic sets are such an interesting tool in constructive commutative algebra. We are especially interested in the relationship between polynomial ideals and their characteristic sets. The main aim of this subsection is to give an overview of the properties of characteristic sets, once they have been calculated. The computation of characteristic sets is the subject of the next subsection.

The first lemma gives an other characterization of characteristic sets.

Lemma 4.2.14 *Let F be a non-empty (finite or infinite) set of polynomials in $K[x_1, \ldots, x_n]$, and assume that $F \neq \{0\}$. Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain that belongs to F. Then:*

$$\mathcal{A} \text{ is a (Ritt)-characteristic set of } F,$$

$$\Longleftrightarrow$$

$$\forall p \in F : [\, p \text{ is reduced w.r.t. } \mathcal{A} \Longrightarrow p = 0 \,].$$

Proof

"$\Rightarrow$" Suppose that $\mathcal{A}$ is a (Ritt)-characteristic set of F, and let $p \in F$, p reduced w.r.t. $\mathcal{A}$.

If $\mathcal{A}$ only consists of a nonzero constant, $p = 0$ is the only polynomial that is reduced w.r.t. $\mathcal{A}$.

Now suppose that $\text{class}(f_1) > 0$, and assume that p is nonzero. If the class of p is higher than that of f_r, we can get an ascending chain of lower order than $\mathcal{A}$ by adjoining p to $\mathcal{A}$ (recall Definition 4.2.9 *(ii)*). Since $\mathcal{A}$ is a characteristic set, this cannot happen. So the class of p is lower than or equal to that of f_r. Let $j = \min\{i \mid \text{class}(f_i) \geq \text{class}(p)\}$. Then $\text{class}(f_j) \geq \text{class}(p)$, but since p is reduced

w.r.t. $\mathcal{A}$, we must have $p \prec f_j$. Also $\text{class}(f_{j-1}) < \text{class}(p)$, so $(f_1, \ldots, f_{j-1}, p)$ is an ascending chain of lower rank than $\mathcal{A}$. This contradicts the fact that $\mathcal{A}$ is a characteristic set.

"$\Leftarrow$" Let $\mathcal{A}$ be an ascending chain such that for all $p \in F$ for which p is reduced w.r.t. $\mathcal{A}$ it follows that $p = 0$. Assume that $\mathcal{A}$ is not a characteristic set, but the ascending chain $\mathcal{B} = (g_1, \ldots, g_s)$ is. Then $\mathcal{B}$ has lower rank than $\mathcal{A}$. First assume that $f_1 \sim g_1, \ldots, f_r \sim g_r$, and $s > r$. Then g_{r+1} is reduced with respect to $\mathcal{A}$, so $g_{r+1} = 0$, which contradicts the assumption that $s > r$. So there exists an $j \leq \min(r, s)$ such that $f_1 \sim g_1, \ldots, f_{j-1} \sim g_{j-1}$, while $g_j \prec f_j$. Then g_j is reduced w.r.t. $\mathcal{A}$ and therefore $g_j = 0$. Again we derive a contradiction. ∎

With help of the previous lemma it is quite easy to prove the following:

Lemma 4.2.15 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be a (Ritt)-characteristic set of a finite or infinite polynomial set F in $K[x_1, \ldots, x_n]$, and assume that* $\text{class}(f_1) > 0$. *Let I_i denote the initial of f_i. Then*

$$\forall i = 1, \ldots, r : I_i \notin F.$$

Proof

Let $i \in \{1, \ldots, r\}$. By the definition of ascending chains, I_i is reduced w.r.t. $\mathcal{A}$. Now suppose that $I_i \in F$. Since $\mathcal{A}$ is a characteristic set of F, it follows from Lemma 4.2.14 that $I_i = 0$, which is a contradiction. ∎

Using the two foregoing lemmas it is possible to derive the relationship between characteristic sets of polynomial ideals and pseudo-remainders. Eventually this leads to a full characterization of prime polynomial ideals based on their characteristic sets. First we need the following theorem.

Theorem 4.2.16 *Let $\mathcal{I}$ be a nonzero ideal in $K[x_1, \ldots, x_n]$, and let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain that belongs to $\mathcal{I}$. Then*

$$\mathcal{A} \text{ is a (Ritt)-characteristic set of } \mathcal{I}$$

$$\Longleftrightarrow$$

$$\forall p \in \mathcal{I} : \text{prem}(p, \mathcal{A}) = 0.$$

Proof

"$\Leftarrow$" Assume that for all $p \in \mathcal{I}$ we have $\text{prem}(p, \mathcal{A}) = 0$, and suppose that $\mathcal{A}$ is not a (Ritt)-characteristic set of $\mathcal{I}$. Then, according to Lemma 4.2.14, there exists a nonzero polynomial $p \in \mathcal{I}$ that is reduced with respect to $\mathcal{A}$. Therefore $\text{prem}(p, \mathcal{A}) = p \neq 0$. This contradicts our assumption, and thus $\mathcal{A}$ has to be a (Ritt)-characteristic set of $\mathcal{I}$.

"$\Rightarrow$" To prove necessity, assume that $\mathcal{A}$ is a characteristic set of $\mathcal{I}$. Let $p \in \mathcal{I}$ and $R = \text{prem}(p, \mathcal{A})$. Then, according to formula (4.19), there exist integers $\nu_1, \ldots, \nu_r \in \mathbb{N} \cup \{0\}$ and polynomials $\alpha_1, \ldots, \alpha_r$ such that

$$I_1^{\nu_1} \cdots I_r^{\nu_r} p = \sum_{i=0}^{r} \alpha_i f_i + R.$$

Since $f_i \in \mathcal{I}$ $(i = 1, \ldots, r)$, it follows that $R \in \mathcal{I}$. Moreover, by the definition of the pseudo-remainder, R is reduced w.r.t. $\mathcal{A}$. Using Lemma 4.2.14 we immediately conclude that $R = 0$, and thus we have $\mathrm{prem}(p, \mathcal{A}) = 0$. Since $p \in \mathcal{I}$ was arbitrary, this proves our claim. ∎

If $\mathcal{I}$ is a *prime* polynomial ideal, it is even completely determined by each of its (Ritt)-characteristic sets via pseudo-remainder computations.

Theorem 4.2.17 *Let $\mathcal{P}$ be a (nonzero) prime polynomial ideal in $\mathcal{K}[x_1, \ldots, x_n]$, and suppose that $\mathcal{A} = (f_1, \ldots, f_r)$ is an ascending chain that belongs to $\mathcal{P}$. Then*

$$\mathcal{A} \text{ is a Ritt-characteristic set of } \mathcal{P}$$

$$\Longleftrightarrow$$

$$\mathcal{P} = \{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\}.$$

Proof

"$\Rightarrow$" Suppose that $\mathcal{A}$ is a characteristic set of $\mathcal{P}$. Then it follows from Theorem 4.2.16 that $\forall p \in \mathcal{P} : \mathrm{prem}(p, \mathcal{A}) = 0$. So $\mathcal{P} \subset \{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\}$.

On the other hand, let $p \in \mathcal{K}[x_1, \ldots, x_n]$ and suppose that $\mathrm{prem}(p, \mathcal{A}) = 0$. Then there exist integers $\nu_1, \ldots, \nu_r$ and polynomials $\alpha_1, \ldots, \alpha_r$ such that

$$I_1^{\nu_1} \cdots I_r^{\nu_r} p = \sum_{i=0}^{r} \alpha_i f_i.$$

Now, $f_1, \ldots, f_r \in \mathcal{P}$, so $I_1^{\nu_1} \cdots I_r^{\nu_r} p \in \mathcal{P}$. By assumption $\mathcal{P}$ is prime, and thus it follows that one of the factors of $I_1^{\nu_1} \cdots I_r^{\nu_r} p$ is an element of $\mathcal{P}$. Since $\mathcal{A}$ is a characteristic set, Lemma 4.2.15 indicates that $I_1, \ldots, I_r$ do not belong to $\mathcal{P}$. So $p \in \mathcal{P}$, and we have proven that $\{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\} \subset \mathcal{P}$. Together with the other inclusion this yields that $\mathcal{P} = \{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\}$.

"$\Leftarrow$" Suppose $\mathcal{P} = \{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\}$. Then certainly

$$\forall p \in \mathcal{P} : \mathrm{prem}(p, \mathcal{A}) = 0.$$

So, according to Theorem 4.2.16, $\mathcal{A}$ is a characteristic set of $\mathcal{P}$. ∎

The result of Theorem 4.2.17 is very important. It indicates that a prime polynomial ideal is completely determined by each of its characteristic sets. What particular characteristic set $\mathcal{A}$ is chosen, does not make any difference. The prime polynomial ideal consists precisely of the polynomials that have pseudo-remainder zero w.r.t. $\mathcal{A}$. Moreover, it is clear that an ascending chain $\mathcal{A}$ can be the characteristic set of at most one prime polynomial ideal.

In Theorem 4.2.17, the condition of primality on the ideal $\mathcal{P}$ is really necessary. This is illustrated in the next example.

Example 4.2.18 Consider the polynomial ring $\mathbf{R}[x, y]$ with ranking $x \prec y$ on the indeterminates. Let $F = \{xy\}$, and denote by $\langle F \rangle$ the ideal in $\mathbf{R}[x, y]$ generated by the polynomial xy. $\langle F \rangle$ is not a prime ideal because $xy \in \langle F \rangle$, but neither x nor y is an element of $\langle F \rangle$. It is immediately clear that $\mathcal{A} = (xy)$ is a (Ritt)-characteristic set of $\langle F \rangle$.

Next, consider the polynomial $y \notin \langle F \rangle$. Since the initial of xy is x, we have

$$x \cdot (y) = 1 \cdot (xy) + 0,$$

and thus $\text{prem}(y, \mathcal{A}) = 0$. We conclude that for a polynomial ideal $\mathcal{I}$ that is not prime, having (Ritt)-characteristic set $\mathcal{A}$, there may exist polynomials p outside $\mathcal{I}$ still satisfying $\text{prem}(p, \mathcal{A}) = 0$.

Compared to Gröbner bases, it looks rather restrictive that characteristic sets can only deal with *prime* polynomial ideals. Nevertheless, this is a very interesting class of ideals in view of the following problem.

Let $F = \{f_1, \ldots, f_m\}$ be a set of polynomials in $\mathcal{K}[x_1, \ldots, x_n]$, and suppose that we want to compute all common zeros of the polynomials in F. In the same way as in Subsection 4.1.6, this problem may be restated as follows: we are interested in determining the variety $\mathcal{V}(\langle F \rangle)$ of the ideal $\langle F \rangle$ in the algebraic closure $\bar{\mathcal{K}}$ of $\mathcal{K}$. From the Hilbert-Nullstellensatz it follows that the variety of $\langle F \rangle$ and the variety of the radical of $\langle F \rangle$, $\text{rad}(\langle F \rangle)$, are identical. Moreover, a radical ideal is the intersection of a finite number of prime ideals (see Corollary A.1.17 in Appendix A.1). Given a finite set of polynomials F, there exist prime polynomial ideals $\mathcal{P}_1, \ldots, \mathcal{P}_k$ such that

$$\text{rad}(\langle F \rangle) = \mathcal{P}_1 \cap \mathcal{P}_2 \cap \cdots \cap \mathcal{P}_k.$$

With each prime polynomial ideal we can associate a characteristic set $\mathcal{A}_i$. Given a set F, we are interested in a method to compute characteristic sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ for $\mathcal{P}_1, \ldots, \mathcal{P}_k$. Since each characteristic set has a triangular form, we obtain a relatively simple representation of the variety of $\langle F \rangle$ in this way. This also leads to a solution of the original problem: the determination of all common zeros of the polynomials in F.

4.2.3 Computation of characteristic sets

The application of characteristic sets for the solution of some of the problems mentioned in the previous subsection is only possible, if there exists a constructive method for computing characteristic sets. In this subsection, we describe an algorithm that plays a very important role in the characteristic sets framework. It can be considered as the first step in the computation of (Ritt)-characteristic sets of prime polynomial ideals.

For a finite set of polynomials it is relatively easy to find a characteristic set. This can be done, for instance, using the following algorithm.

Algorithm 4.2.19 Let $F \neq \{0\}$ be a *finite* set of polynomials. Then the following procedure computes a (Ritt)-characteristic set of F:

f_1 := a polynomial of lowest rank in $F \backslash \{0\}$;
$\mathcal{A} := (f_1)$;
if $\text{class}(f_1) \neq 0$ then
  while $(\exists p \in F \backslash \{0\} : p$ is reduced w.r.t. $\mathcal{A})$ do
    g := a polynomial of lowest rank in $F \backslash \{0\}$ that is reduced w.r.t. $\mathcal{A}$;

```
        A := (A, g);
    od;
fi;
```

The correctness of this algorithm, which originates from Ritt, can be verified using the definition of (Ritt)-characteristic sets.

The computation of characteristic sets for polynomial ideals is much more involved. However, for polynomial ideals that are *radical*, Ritt also suggests a solution (see [79, pp. 88-90]). The crucial part of this algorithm was later used by Wu-Wen-Tsun to define and compute characteristic sets in his own sense. This is the source of a lot of confusion, because Ritt-characteristic sets and Wu-characteristic sets are *not* the same notions in general. Of course, there are also a lot of relationships, and one of the main aims of this section is to bring up both the resemblances and differences.

To introduce the framework of Wu-characteristic sets, we follow the same lines as Wang in [96]. We start with some new definitions.

Definition 4.2.20 Let $F = \{f_1, \ldots, f_k\}$ be a *finite* polynomial set in $\mathcal{K}[x_1, \ldots, x_n]$, containing at least one nonzero polynomial. Then a Ritt-characteristic set of F is also called a *basic set* of F.

Definition 4.2.21 Let $F = \{f_1, \ldots, f_k\}$ be a finite polynomial set in $\mathcal{K}[x_1, \ldots, x_n]$, containing at least one nonzero polynomial. Then a *medial set* of F is an ascending chain that belongs to the polynomial ideal $\langle F \rangle$, generated by the polynomials $f_1, \ldots, f_k$ in F, and with rank not higher than that of the basic set of F.

Definition 4.2.22 Let $F = \{f_1, \ldots, f_k\}$ be a finite polynomial set in $\mathcal{K}[x_1, \ldots, x_n]$, containing at least one nonzero polynomial. A medial set $\mathcal{A}$ of F is called a *Wu-characteristic set* of F if

$$\forall f \in F : \mathrm{prem}(f, \mathcal{A}) = 0. \tag{4.20}$$

Note that a Wu-characteristic set of a finite polynomial set F does not belong to F, but to the ideal $\langle F \rangle$ generated by the polynomials in F. Moreover, it is clear that a Wu-characteristic set of F is highly related to both $\langle F \rangle$ and the Ritt-characteristic sets of $\langle F \rangle$. If an ideal is characterized by a finite set F of generators, one therefore often speaks (with some abuse of terminology) of the Wu-characteristic set of $\langle F \rangle$.

From Definition 4.2.22 and Theorem 4.2.16 it follows that a (Ritt)-characteristic set of the ideal $\langle F \rangle$, generated by a finite set of polynomials $F \neq \{0\}$, is also a Wu-characteristic set of F, but not the other way around. This is illustrated by the following example.

Example 4.2.23 Take the polynomial ring $\mathbb{R}[x, y]$ with ordering $x \prec y$ and let $f_1 = x^2$, $f_2 = xy + 1$. Define $F := \{f_1, f_2\}$. Since f_1 and f_2 have no common zeros, the Hilbert Nullstellensatz yields $\langle F \rangle = \mathbb{R}[x, y]$. So $\mathcal{A} = (1)$ is a (Ritt)-characteristic set of $\langle F \rangle$. On the other hand it is easily verified that the ascending chain $\mathcal{B} = (x^2, xy + 1)$ is a Wu-characteristic set of F. Now $\mathcal{A}$ is also a Wu-characteristic set of F, but $\mathcal{B}$ is not a (Ritt)-characteristic set of $\langle F \rangle$ because $\mathcal{A}$ is an ascending chain in $\langle F \rangle$ of lower rank than $\mathcal{B}$.

From Example 4.2.23 it is obvious that the concept of Wu-characteristic sets for finite polynomial sets F is weaker than that of (Ritt)-characteristic sets for the corresponding ideals $\langle F\rangle$. However, this has the advantage that Wu-characteristic sets can be computed more easily. Moreover, given a finite polynomial set $F \neq \{0\}$, a lot of the properties of the variety of the polynomial ideal $\langle F\rangle$ remain valid for the Wu-characteristic sets of F. Since a Wu-characteristic set is an ascending chain, its triangular structure is preserved. This makes the Wu-characteristic sets method very suitable for the computation of the solutions of a system of polynomial equations. Moreover, we may expect that the computation of a Wu-characteristic set for a finite polynomial set $F \neq \{0\}$ is the first step in the computation of a Ritt-characteristic set of $\langle F\rangle$.

To explain the main ideas, we need still another definition.

Definition 4.2.24 Let $f \in \mathcal{K}[x_1, \ldots, x_n]$ and let $\bar{\mathcal{K}}$ be the algebraic closure of $\mathcal{K}$. A point $\alpha = (\alpha_1, \ldots, \alpha_n) \in \bar{\mathcal{K}}^n$ is called an *(extended) zero* of f if

$$f(\alpha_1, \ldots, \alpha_n) = 0.$$

In the same way we define for a polynomial set F and a polynomial g in $\mathcal{K}[x_1, \ldots, x_n]$:

$$\text{Zero}(F) = \{\alpha \in \bar{\mathcal{K}}^n \mid \forall f \in \langle F\rangle : f(\alpha) = 0\}, \tag{4.21}$$

$$\text{Zero}(F/g) = \{\alpha \in \bar{\mathcal{K}}^n \mid \forall f \in \langle F\rangle : f(\alpha) = 0 \wedge g(\alpha) \neq 0\}. \tag{4.22}$$

When F contains only a nonzero constant, we call F *contradictory*. In this case $\text{Zero}(F) = \varnothing$.

Note that $\text{Zero}(F)$ is the variety of the ideal $\langle F\rangle$, generated by the polynomials in F, but $\text{Zero}(F/g)$ is not necessarily a variety.

We now give the algorithm to compute a Wu-characteristic set as stated by Wang in [96]. It is a generalization of the so-called "Ritt principle", because the original ideas stem from the work of Ritt (see [79, pp. 88-90]).

Algorithm 4.2.25 Given a finite set of polynomials $F = \{f_1, \ldots, f_k\}$ in the ring $\mathcal{K}[x_1, \ldots, x_n]$, containing at least one nonzero polynomial. Then the following algorithm is applied:

$Q := F; R := F;$
while $R \neq \varnothing$ **do**
  $C :=$ a *medial set* of Q;
  if C is *contradictory*
    then $R := \varnothing$;
    else $R := \{\text{prem}(q, C) \mid q \in Q \backslash C\} \backslash \{0\}$;
      $Q := Q \cup C \cup R$;
  fi;
od;

Proposition 4.2.26 *Let $F = \{f_1, \ldots, f_k\}$ be a finite set of polynomials in the ring $\mathcal{K}[x_1, \ldots, x_n]$, containing at least one nonzero polynomial, and apply Algorithm 4.2.25. Then:*

(i) The algorithm terminates.

(ii) After termination of the algorithm, C is a Wu-characteristic set of F.

(iii) If $C = (c_1, \ldots, c_r)$ and I_i denotes the initial of c_i $(i = 1, \ldots, r)$, then

$$\text{Zero}(C/g) \subset \text{Zero}(F) \subset \text{Zero}(C), \tag{4.23}$$

$$\text{Zero}(F) = \text{Zero}(C/g) \cup \bigcup_{i=1}^{r} \text{Zero}(F_i), \tag{4.24}$$

where $g = I_1 \cdots I_r$ and $F_i = F \cup \{I_i\}$ $(i = 1, \ldots, r)$. ∎

To prove Proposition 4.2.26, we need the following lemma from Wang (see [96]).

Lemma 4.2.27 *Let $F \neq \{0\}$ be a finite set of polynomials in $\mathcal{K}[x_1, \ldots, x_n]$ and $\mathcal{M} = (m_1, \ldots, m_s)$ a medial set of F, with class$(m_1) > 0$. Let m be a nonzero polynomial reduced with respect to $\mathcal{M}$. Define $\tilde{F} := F \cup \{m_1, \ldots, m_s\} \cup \{m\}$. Then any medial set $\tilde{\mathcal{M}}$ of $\tilde{F}$ will have lower rank than $\mathcal{M}$.*

Proof

Let $\tilde{\mathcal{B}}$ be a basic set of $\tilde{F}$, and $\hat{\mathcal{B}}$ a basic set of $\hat{F} := F \cup \{m_1, \ldots, m_s\}$. Then clearly $\hat{\mathcal{B}}$ does not have higher rank than $\mathcal{M}$.

Now assume that $\hat{\mathcal{B}}$ has the same rank as $\mathcal{M}$. Since m is reduced with respect to $\mathcal{M}$, it is also reduced with respect to $\hat{\mathcal{B}}$. But because $m \neq 0$, it follows from Lemma 4.2.14 that $\tilde{\mathcal{B}}$, which is a basic set of $\hat{F} \cup \{m\}$, must have strictly lower rank than $\hat{\mathcal{B}}$. In this way we find

$$\tilde{\mathcal{M}} \preceq \tilde{\mathcal{B}} \prec \hat{\mathcal{B}} \sim \mathcal{M}.$$

On the other hand, if $\hat{\mathcal{B}}$ has lower rank than $\mathcal{M}$, it is immediately clear that

$$\tilde{\mathcal{M}} \preceq \tilde{\mathcal{B}} \preceq \hat{\mathcal{B}} \prec \mathcal{M}.$$

In either case $\tilde{\mathcal{M}}$ has strictly lower rank than $\mathcal{M}$. ∎

Proof of Proposition 4.2.26

(i) Termination: To prove the termination of Algorithm 4.2.25 it is sufficient to show that the while-loop will be carried out only a finite number of times.

Let $\mathcal{M}_j$ denote the medial set C in the j^{th} while-loop. If a certain $\mathcal{M}_j$ is contradictory, i.e. if it contains only a nonzero constant, the algorithm terminates immediately. Otherwise it follows from Lemma 4.2.27 that $\mathcal{M}_{j+1}$ is an ascending chain of lower rank than $\mathcal{M}_j$. So $\mathcal{M}_1, \mathcal{M}_2, \ldots$ is a sequence of ascending chains with strictly decreasing ranks. According to Lemma 4.2.11 such a sequence can only contain a finite number of chains, say ℓ. This means that $\mathcal{M}_\ell$ is a medial set of the final finite polynomial set Q_ℓ. The pseudo-remainder of an arbitrary polynomial in Q_ℓ with respect to $\mathcal{M}_\ell$ must therefore be zero, i.e. $R = \varnothing$. Thus the algorithm terminates.

(ii) Correctness: To prove the correctness of the Algorithm 4.2.25, we first show that the ascending chain $\mathcal{C}$, for which the algorithm terminates, belongs to the polynomial ideal $\langle F\rangle$. Denote by Q_j and R_j the sets Q and R at the start of the j^{th} while-loop, and let $\mathcal{M}_j$ be the ascending chain $\mathcal{C}$ calculated in the j^{th} while-loop. We start to prove by induction that

$$\forall j \in \{1,\ldots,\ell\} : \langle Q_j\rangle \subset \langle F\rangle, \langle R_j\rangle \subset \langle F\rangle \text{ and } \langle \mathcal{M}_j\rangle \subset \langle F\rangle. \tag{4.25}$$

j = 1: $Q_1 = F$, $R_1 = F$ and $\mathcal{M}_1$ is a medial set of F, so clearly $\langle Q_1\rangle \subset \langle F\rangle$, $\langle R_1\rangle \subset \langle F\rangle$ and $\langle \mathcal{M}_1\rangle \subset \langle F\rangle$.

induction step: Assume that $\langle Q_j\rangle \subset \langle F\rangle$, $\langle R_j\rangle \subset \langle F\rangle$ and $\langle \mathcal{M}_j\rangle \subset \langle F\rangle$. Then $R_{j+1} = \{\text{prem}(q,\mathcal{M}_j) \mid q \in Q_j \backslash \mathcal{M}_j\}$. Let $v \in R_{j+1}$ and $\mathcal{M}_j = (m_1,\ldots,m_s)$, with initials $I_1,\ldots,I_s$ respectively. Then there exists a polynomial $p \in Q_j\backslash\mathcal{M}_j$, integers $\nu_1,\ldots,\nu_s$ and polynomials $\alpha_1,\ldots,\alpha_s$ such that

$$I_1^{\nu_1}\cdots I_s^{\nu_s} p = \sum_{i=1}^{s} \alpha_i m_i + v.$$

Hence

$$v = -\sum_{i=1}^{s} \alpha_i m_i + I_1^{\nu_1}\cdots I_s^{\nu_s} p \in \langle \mathcal{M}_j \cup Q_j\rangle.$$

Since both $\langle \mathcal{M}_j\rangle \subset \langle F\rangle$ and $\langle Q_j\rangle \subset \langle F\rangle$, it follows that $v \in \langle F\rangle$, and because $v \in R_{j+1}$ was arbitrary, we conclude that $\langle R_{j+1}\rangle \subset \langle F\rangle$. Recall that $Q_{j+1} = Q_j \cup \mathcal{M}_j \cup R_{j+1}$. Since the three sets in the decomposition of Q_{j+1} belong to $\langle F\rangle$, it is clear that $\langle Q_{j+1}\rangle \subset \langle F\rangle$. Finally, $\mathcal{M}_{j+1}$ is a medial set of Q_{j+1}. So $\mathcal{M}_{j+1}$ belongs to $\langle Q_{j+1}\rangle$. Since $\langle Q_{j+1}\rangle \subset \langle F\rangle$, we certainly have that $\langle \mathcal{M}_{j+1}\rangle \subset \langle F\rangle$.

Since Algorithm 4.2.25 terminates with $\mathcal{C} = \mathcal{M}_\ell \subset \langle F\rangle$, the ascending chain $\mathcal{C}$ belongs to $\langle F\rangle$. Moreover, $\mathcal{C}$ is a medial set of Q_ℓ. Now $Q_1 = F$ and $Q_j \subset Q_{j+1}$ for all j, so $F \subset Q_\ell$. For all polynomials $q \in Q_\ell$ we have $\text{prem}(q,\mathcal{C}) = 0$. So in particular

$$\forall f \in F : \text{prem}(f,\mathcal{C}) = 0,$$

and $\mathcal{C}$ is a Wu-characteristic set of F.

(iii) Zero-inclusions (4.23) and (4.24): Finally, to prove the zero-inclusions, we need the following result:

$$\forall j \in \{1,\ldots,\ell-1\} : \text{Zero}(Q_j) = \text{Zero}(Q_{j+1}). \tag{4.26}$$

The correctness of this claim can be seen as follows.

"$\supset$" Let $\alpha \in \text{Zero}(Q_{j+1})$. Let $f \in Q_j$. Because $Q_{j+1} = Q_j \cup \mathcal{M}_j \cup R_{j+1}$ it follows that $f \in Q_{j+1}$, so $f(\alpha) = 0$ and $\alpha \in \text{Zero}(Q_j)$.

"$\subset$" Let $\alpha \in \text{Zero}(Q_j)$. Since $\mathcal{M}_j$ is a medial set of Q_j, all elements of $\mathcal{M}_j$ belong to $\langle Q_j \rangle$, so $\alpha \in \text{Zero}(\mathcal{M}_j)$. Let $v \in R_{j+1}$. Then there exists a polynomial $p \in Q_j \backslash \mathcal{M}_j$, integers $\nu_1, \ldots, \nu_s$ and polynomials $\beta_1, \ldots, \beta_s$ such that

$$I_1^{\nu_1} \cdots I_s^{\nu_s} p = \sum_{i=1}^{s} \beta_i m_i + v.$$

Since $\alpha \in \text{Zero}(\mathcal{M}_j)$ and $\alpha \in \text{Zero}(Q_j)$, it follows that α is a zero of v. Now $v \in R_{j+1}$ was arbitrary, hence $\alpha \in \text{Zero}(R_{j+1})$. Let $f \in Q_{j+1} = Q_j \cup \mathcal{M}_j \cup R_{j+1}$. Then f is an element of at least one of the sets Q_j, $\mathcal{M}_j$ or R_{j+1}. However, in either case we know that α is a zero of f. So $\alpha \in \text{Zero}(Q_{j+1})$.

Since $Q_1 = F$, it follows that for all $j \in \{1, \ldots, \ell\}$:

$$\text{Zero}(F) = \text{Zero}(Q_1) = \text{Zero}(Q_j).$$

Next, because $\mathcal{C}$ is a medial set of Q_ℓ, all its elements belong to $\langle Q_\ell \rangle$. Let $\alpha \in \text{Zero}(F)$. Then $\alpha \in \text{Zero}(Q_\ell)$. Since $\mathcal{C} \subset \langle Q_\ell \rangle$, it is obvious that $\alpha \in \text{Zero}(\mathcal{C})$. This proves that $\text{Zero}(F) \subset \text{Zero}(\mathcal{C})$.

On the other hand, suppose $\mathcal{C} = (c_1, \ldots, c_r)$ with initials $I_1, \ldots, I_r$. Define

$$g := \prod_{i=1}^{r} I_i.$$

Let $\alpha \in \text{Zero}(\mathcal{C}/g)$ and $p \in Q_\ell$. Then $\text{prem}(p, \mathcal{C}) = 0$, so there exist integers $\nu_1, \ldots, \nu_r$ and polynomials $\beta_1, \ldots, \beta_r$ such that

$$I_1^{\nu_1} \cdots I_r^{\nu_r} p = \sum_{i=1}^{r} \beta_i c_i. \tag{4.27}$$

Since $\alpha \in \text{Zero}(\mathcal{C})$, α is a zero of the right-hand side of (4.27), so α is a zero of $I_1^{\nu_1} \cdots I_r^{\nu_r} p$ too. By assumption α is not a zero of one of the initials $I_1^{\nu_1}, \ldots, I_r^{\nu_r}$, hence α must be a zero of p. Now $p \in Q_\ell$ was arbitrary, so this proves that $\alpha \in \text{Zero}(Q_\ell) = \text{Zero}(F)$. Therefore $\text{Zero}(\mathcal{C}/g) \subset \text{Zero}(F)$.

To prove (4.24), first recall that we already have shown that $\text{Zero}(\mathcal{C}/g) \subset \text{Zero}(F)$. Since for all $i \in \{1, \ldots, r\}$ we have $F \subset F_i$, it is clear that $\text{Zero}(F_i) \subset \text{Zero}(F)$. Hence

$$\left[\text{Zero}(\mathcal{C}/g) \cup \bigcup_{i=1}^{r} \text{Zero}(F_i)\right] \subset \text{Zero}(F). \tag{4.28}$$

Let now $\alpha \in \text{Zero}(F) \subset \text{Zero}(\mathcal{C})$. Assume that $\alpha \notin \text{Zero}(\mathcal{C}/g)$. Then there has to exist an initial I_i such that α is a zero of I_i. But then α is a zero of F_i. So we also have

$$\text{Zero}(F) \subset \left[\text{Zero}(\mathcal{C}/g) \cup \bigcup_{i=1}^{r} \text{Zero}(F_i)\right]. \tag{4.29}$$

Formulae (4.28) and (4.29) together yield (4.24).

This completes the proof. ∎

Example 4.2.28 Consider the same system of polynomial equations as in Example 4.1.32:

$$\begin{cases} x_3^2 - 5x_1 &= -1, \\ x_1x_2 - 4x_1x_3 + 6x_3 &= -2, \\ x_1^2x_2 + 3x_2x_3 &= 2. \end{cases}$$

To find a solution, we consider the ideal $\langle f_1, f_2, f_3 \rangle$ in $\mathbf{R}[x_1, x_2, x_3]$, where $f_1 = x_3^2 - 5x_1 + 1$, $f_2 = x_1x_2 - 4x_1x_3 + 6x_3 + 2$ and $f_3 = x_1^2x_2 + 3x_2x_3 - 2$, and compute a Wu-characteristic set of $\{f_1, f_2, f_3\}$ with respect to the ordering $x_1 \succ x_2 \succ x_3$ of indeterminates, using Algorithm 4.2.25. In this way the Wu-characteristic set (h_1, h_2, h_3) is obtained, where

$$\begin{aligned} h_1 &= 2x_3^7 - 9x_3^5 + 145x_3^4 - 24x_3^3 - 1010x_3^2 - 388x_3 - 30, \\ h_2 &= (1 + x_3^2) \cdot x_2 - 4x_3^3 + 26x_3 + 10, \\ h_3 &= 5x_1 - x_3^2 - 1. \end{aligned}$$

Since the initial of h_3 is a nonzero constant in $\mathbf{R}$, and the initial $x_3^2 + 1$ of h_2 has no zeros in common with h_1, we conclude from formula (4.23) that the set of common zeros of the polynomials f_1, f_2 and f_3 is equal to the set of common zeros of the Wu-characteristic set (h_1, h_2, h_3). To find a solution, we first compute all solutions to the equation $h_1 = 0$. This is relatively easy because h_1 is a univariate polynomial in the indeterminate x_3. Substitution of these solutions in the equations $h_2 = 0$ and $h_3 = 0$ yields the corresponding values of x_1 and x_2.

Note that in Algorithm 4.2.25 the method to find a medial set C of Q is not specified. The user has the freedom to choose a method himself. The most obvious choice is to take a basic set of Q. Such a basic set (which is a (Ritt)-characteristic set of the finite polynomial set Q) can be computed with Algorithm 4.2.19. Therefore Algorithm 4.2.25 and 4.2.19 together constitute a fully constructive method for the computation of a Wu-characteristic set of the polynomial set F.

Of course it is also possible to calculate a medial set C of Q in another way. Several suggestions are made in the paper of Wang (see [96]). In that paper the performances of the various methods are also compared. Wang was the first one who stated this so called "generalized characteristic sets algorithm" as we did it above. It is a generalization of an algorithm mentioned by Wu-Wen-Tsun in [101], which, in turn, was inspired by the original ideas of Ritt.

An implementation of the algorithm, with various possibilities to choose the medial set C, already exists for the computer algebra package Maple (see [97]). The program is developed by Wang and is available by anonymous ftp, together with a manual.

4.2.4 Irreducible ascending chains

In the previous subsection we explained how a Wu-characteristic set of a finite set of polynomials F can be computed. These Wu-characteristic sets turned out to be

ascending chains in the ideal $\langle F \rangle$, but did not have as strong properties as Ritt-characteristic sets of $\langle F \rangle$. Nevertheless, both the definitions of Ritt- as well as Wu-characteristic sets suggest that there is a close relationship between these two concepts. This relationship can be determined completely for so-called irreducible characteristic sets. In this subsection it is explained how this can be done. Moreover, the study of irreducible ascending chains yields the main ideas for the treatment of reducible characteristic sets that are considered in the next subsection.

The proofs of the main results of this subsection are rather technical and long. To preserve the continuity of our exposition, and to illuminate the main ideas more clearly, these proofs are omitted here. However, in Appendix C they are elaborated in full detail.

In the title and the introduction of this subsection we frequently encountered the term "irreducible ascending chain". For univariate polynomials the concept of irreducibility is well known: let $\mathcal{K}$ be an arbitrary field and let $p \in \mathcal{K}[x]$ be a polynomial such that $\deg(p) > 0$; then p is called *irreducible* if it is impossible to factorize p as $p = q_1 \cdot q_2$, where $q_1, q_2 \in \mathcal{K}[x]$ are polynomials with $\deg(q_i) > 0$ $(i = 1, 2)$. This notion of irreducibility can be generalized to ascending chains. This idea was first used by Ritt (see [79, pp. 88-90]), but we prefer to give the more formal definition according to Wu-Wen-Tsun (see [101, pp. 233-234]).

Definition 4.2.29 Consider an ascending chain

$$\mathcal{A} = (f_1, \ldots, f_r),$$

in $\mathcal{K}[x_1, \ldots, x_n]$ and assume that $\mathrm{class}(f_i) = p_i$, with

$$0 < p_1 < p_2 < \cdots < p_r.$$

Define for $i = 1, \ldots, r$: $y_i := x_{p_i}$ and $m_i := \deg_{x_{p_i}}(f_i)$, and denote the other original x-indeterminates in the original order as $u_1, \ldots, u_d$. $d := n - r$ is called the *dimension* of the ascending chain $\mathcal{A}$.

Write the polynomials $(f_1, \ldots, f_r)$ in $\mathcal{A}$ as

$$\mathcal{A} \begin{cases} f_1 &= c_{10} y_1^{m_1} + c_{11} y_1^{m_1 - 1} + \cdots + c_{1m_1} \\ f_2 &= c_{20} y_2^{m_2} + c_{21} y_2^{m_2 - 1} + \cdots + c_{2m_2} \\ &\vdots \\ f_r &= c_{r0} y_r^{m_r} + c_{r1} y_r^{m_r - 1} + \cdots + c_{rm_r} \end{cases}$$

where the coefficients $c_{i0} \neq 0$ are the initials of the polynomials f_i, and each coefficient c_{ij} is itself a polynomial in $u_1, \ldots, u_d, y_1, \ldots, y_{i-1}$ with coefficients in $\mathcal{K}$. Since f_i is reduced with respect to $f_1, \ldots, f_{i-1}$, the degrees of c_{ij} in $y_1, \ldots, y_{i-1}$ are less than $m_1, \ldots, m_{i-1}$ respectively.

Let the transcendental extension field $\mathcal{K}(u_1, \ldots, u_d)$ of $\mathcal{K}$, obtained by adjoining the symbols $u_1, \ldots, u_d$ to $\mathcal{K}$, be denoted by $\mathcal{K}_0$. Then the ascending chain $\mathcal{A}$ is called *irreducible* if the following r conditions are all satisfied:

(i) Let $\tilde{f}_1$ denote the polynomial f_1, considered as a polynomial in $\mathcal{K}_0[y_1]$, so with coefficients in $\mathcal{K}_0$. Then $\tilde{f}_1$ is irreducible in $\mathcal{K}_0[y_1]$.

(ii) Let the algebraic extension field of $\mathcal{K}_0$, obtained by adjoining an (extended) zero η_1 of $\tilde{f}_1 = 0$, be denoted by $\mathcal{K}_1 := \mathcal{K}_0(\eta_1)$. Then the polynomial $\tilde{f}_2$, obtained by substituting η_1 for y_1 in f_2, is irreducible in $\mathcal{K}_1[y_2]$.

(iii) Let the algebraic extension field of $\mathcal{K}_1$, obtained by adjoining an (extended) zero η_2 of $\tilde{f}_2 = 0$, be denoted by $\mathcal{K}_2 := \mathcal{K}_1(\eta_2)$. Then the polynomial $\tilde{f}_3$, obtained by substituting η_1 for y_1 and η_2 for y_2 in f_3, is irreducible in $\mathcal{K}_2[y_3]$.

$\vdots$

(r) Let the algebraic extension field of $\mathcal{K}_{r-2}$, obtained by adjoining an (extended) zero η_{r-1} of $\tilde{f}_{r-1} = 0$, be denoted by $\mathcal{K}_{r-1} := \mathcal{K}_{r-2}(\eta_{r-1})$. Then the polynomial $\tilde{f}_r$, obtained by substituting $(\eta_1, \ldots, \eta_{r-1})$ for $(y_1, \ldots, y_{r-1})$ in f_r, is irreducible over $\mathcal{K}_{r-1}[y_r]$.

Let η_r be an (extended) zero of $\tilde{f}_r = 0$ and $\mathcal{K}_r := \mathcal{K}_{r-1}(\eta_r)$ the algebraic extension obtained by adjoining η_r to $\mathcal{K}_{r-1}$. Then the u_i and the η_j are all elements of $\tilde{\mathcal{K}} := \mathcal{K}_r$, and the point

$$\tilde{\eta} = (u_1, \ldots, u_d, \eta_1, \ldots, \eta_r) \tag{4.30}$$

can be considered as a point of the affine space $\tilde{\mathcal{K}}^{d+r} = \tilde{\mathcal{K}}^n$. $\tilde{\eta}$ is called a *generic point* of $\mathcal{A}$. $\tilde{\mathcal{K}}$ is called a *generating field* of $\mathcal{A}$.

From Definition 4.2.29 we see that the polynomials of an irreducible ascending chain are themselves irreducible in a very specific sense. Intuitively it is clear that this has to be related with the primality of the polynomial ideal corresponding to this ascending chain $\mathcal{A}$. In the next proposition this relationship between irreducible ascending chains and prime polynomial ideals is elaborated. Moreover, it gives an alternative characterization of irreducible ascending chains.

Proposition 4.2.30 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in $\mathcal{K}[x_1, \ldots, x_n]$, and rename the indeterminates in the same way as in Definition 4.2.29. Then we have:*

$$\mathcal{A} = (f_1, \ldots, f_r) \text{ is irreducible}$$

$$\Longleftrightarrow$$

$$\forall j = 1, \ldots, r : \langle f_1, \ldots, f_j \rangle \text{ is a prime ideal in } \mathcal{K}_0[y_1, \ldots, y_j].$$

(This means that $\langle f_1 \rangle$ is a prime ideal in $\mathcal{K}_0[y_1]$, $\langle f_1, f_2 \rangle$ is a prime ideal in $\mathcal{K}_0[y_1, y_2]$ and so on, until the final condition: $\langle f_1, \ldots, f_r \rangle$ is a prime ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$).

■

For a proof of this result, we refer to Appendix C.

The concept of irreducible ascending chains also enables us to give another characterization of the property of a polynomial to have pseudo-remainder zero with respect to a chain. This is very important, because when a polynomial p and an ascending chain $\mathcal{A}$ are given, the equality $\text{prem}(p, \mathcal{A}) = 0$ contains a lot of information, but this information is very hard to extract. For irreducible ascending chains however, there exists an easy translation, given in the next proposition. A proof of this result may be found in [101, p. 234, Lemma 3].

Proposition 4.2.31 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in the polynomial ring $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ (in the notation of Definition 4.2.29), and assume that $\mathcal{A}$ is irreducible. Let*

$$\tilde{\eta} = (u_1, \ldots, u_d, \eta_1, \ldots, \eta_r)$$

be a generic point, as defined in (4.30). Let $p \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$. Then

$$\mathrm{prem}(p, \mathcal{A}) = 0$$

$$\Longleftrightarrow$$

$\tilde{\eta}$ is an extended zero of p. ∎

This result of Wu-Wen-Tsun describes the relationship between a polynomial p with the property $\mathrm{prem}(p, \mathcal{A}) = 0$, and the variety $\mathcal{V}(\mathcal{A})$ of the irreducible ascending chain $\mathcal{A}$. On the other hand, we know from Proposition 4.2.30 that the polynomials of an irreducible ascending chain $\mathcal{A}$ generate a prime polynomial ideal in the ring $\mathcal{K}_0[y_1, \ldots, y_r]$, and from Theorem 4.2.17 that an ascending chain $\mathcal{A} = (f_1, \ldots, f_r)$ is a (Ritt-) characteristic set of the prime ideal $\langle f_1, \ldots, f_r \rangle$ if and only if

$$\langle f_1, \ldots, f_r \rangle = \{p \mid \mathrm{prem}(p, \mathcal{A}) = 0\}.$$

Therefore it is intuitively clear that in the ring $\mathcal{K}_0[y_1, \ldots, y_r]$ there is a strong link between irreducible ascending chains and (Ritt-) characteristic sets of prime polynomial ideals. However, the exact relationship is still unclear, and, moreover, we are merely interested in (Ritt-) characteristic sets over the ring $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$. To solve this unsatisfactory situation, we introduce two ideals that clarify the structure of the problems mentioned above.

Definition 4.2.32 Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in the polynomial ring $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ (in the notation of Definition 4.2.29). Then $\tilde{\mathcal{F}}$ is defined as

$$\tilde{\mathcal{F}} := \{\sum_{i=1}^{r} \alpha_i f_i \mid \alpha_i \in \mathcal{K}_0[y_1, \ldots, y_r]\}. \tag{4.31}$$

So $\tilde{\mathcal{F}}$ is the ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$ generated by the polynomials $f_1, \ldots, f_r$. On the other hand we define $\mathcal{F}$ as

$$\mathcal{F} := \{p \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r] \mid \exists \alpha_i \in \mathcal{K}_0[y_1, \ldots, y_r] \text{ s.t. } p = \sum_{i=1}^{r} \alpha_i f_i\}. \tag{4.32}$$

It is not difficult to see that $\mathcal{F}$ is an ideal in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$.

From the definition it follows that every polynomial $p \in \mathcal{F}$, is also an element of $\tilde{\mathcal{F}}$. So in the ring $\mathcal{K}_0[y_1, \ldots, y_r]$ we have: $\mathcal{F} \subset \tilde{\mathcal{F}}$. However, the following property is more important:

Lemma 4.2.33 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in the polynomial ring $K[u_1, \ldots, u_d, y_1, \ldots, y_r]$ (in the notation of Definition 4.2.29). Let $\mathcal{F}$ and $\tilde{\mathcal{F}}$ be defined as in (4.32) and (4.31). Then*

$$\mathcal{F} \text{ is a prime ideal in } K[u_1, \ldots, u_d, y_1, \ldots, y_r],$$
$$\Longleftrightarrow$$
$$\tilde{\mathcal{F}} \text{ is a prime ideal in } K_0[y_1, \ldots, y_r].$$

■

For the proof of Lemma 4.2.33 we again refer to Appendix C.

At this point it is possible to determine the exact relationship between irreducible ascending chains on the one hand, and Ritt-characteristic sets of prime polynomial ideals on the other hand.

Theorem 4.2.34 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in the polynomial ring $K[u_1, \ldots, u_d, y_1, \ldots, y_r]$ (in the notation of Definition 4.2.29). Define the ideals $\mathcal{F}$ and $\tilde{\mathcal{F}}$ as in (4.32) and (4.31), respectively. Then the following three statements are equivalent:*

(i) The ascending chain $\mathcal{A}$ is irreducible,

(ii) $\tilde{\mathcal{F}}$ is a prime ideal in $K_0[y_1, \ldots, y_r]$, and $\mathcal{A}$ is a (Ritt-) characteristic set of $\tilde{\mathcal{F}}$,

(iii) $\mathcal{F}$ is a prime ideal in $K[u_1, \ldots, u_d, y_1, \ldots, y_r]$, and $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$.

■

A detailed proof of this result is given in appendix C.

Theorem 4.2.34 is a very interesting result. Given an irreducible ascending chain $\mathcal{A}$, it gives a characterization of a prime ideal $\mathcal{F}$ such that $\mathcal{A}$ is a (Ritt-) characteristic set of this prime ideal $\mathcal{F}$. Moreover, from Theorem 4.2.17 we know that the prime ideal of which $\mathcal{A}$ is a (Ritt-) characteristic set is unique, at least in the polynomial ring under consideration. So, for an irreducible chain $\mathcal{A}$, Theorem 4.2.17 shows that

$$\{p \mid \text{prem}(p, \mathcal{A}) = 0\}.$$

is an alternative description of the prime ideal $\mathcal{F}$, defined in (4.32). In fact, the polynomials in $\mathcal{A}$ "generate" this prime ideal in a very special way.

Theorem 4.2.34 does not only give a clearer view on the structure of irreducible characteristic sets of prime ideals, it also enables us to describe the link between Wu- and Ritt-characteristic sets, at least in the irreducible case.

Theorem 4.2.35 *Let $P = \{p_1, \ldots, p_m\}$ be a finite polynomial set in $K[x_1, \ldots, x_n]$ containing a nonzero polynomial, and let $\mathcal{A} = (f_1, \ldots, f_r)$ a Wu-characteristic set of P. Rename the indeterminates $x_1, \ldots, x_n$ in the same way as in Definition 4.2.29. Suppose that $\mathcal{A}$ is an irreducible ascending chain in the polynomial ring*

$\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$. *Let $\mathcal{F}$ denote the prime ideal as defined in (4.32). Then $\langle P \rangle \subset \mathcal{F}$ and $\mathcal{A}$ is a Ritt-characteristic set of $\langle P \rangle$.*

Moreover, if the dimension of the irreducible ascending chain $\mathcal{A}$ is zero (so all indeterminates occur once as leading variable, i.e. $d = 0$, $r = n$, $x_i = y_i$ and $\mathcal{K}_0 = \mathcal{K}$), then we even have $\langle P \rangle = \mathcal{F}$.

Proof

Since $\mathcal{A}$ is an irreducible ascending chain, $\mathcal{A}$ is a Ritt-characteristic set of $\mathcal{F}$. So, according to Theorem 4.2.17,

$$\mathcal{F} = \{p \mid \operatorname{prem}(p, \mathcal{A}) = 0\}.$$

Let $p \in P$. Since $\mathcal{A}$ is a Wu-characteristic set of P, $\operatorname{prem}(p, \mathcal{A}) = 0$. So $p \in \mathcal{F}$. This proves that $P \subset \mathcal{F}$, and therefore also $\langle P \rangle \subset \mathcal{F}$.

Now suppose that $\mathcal{A}$ is not a Ritt-characteristic set of $\langle P \rangle$. Since $\mathcal{A}$ is an ascending chain belonging to $\langle P \rangle$, there exists an ascending chain $\mathcal{B}$ in $\langle P \rangle$ of lower rank than $\mathcal{A}$. Then $\mathcal{B}$ also belongs to $\mathcal{F}$, and it is an ascending chain of lower rank than $\mathcal{A}$. We conclude that $\mathcal{A}$ is not a Ritt-characteristic set of $\mathcal{F}$, and this contradicts the result of Theorem 4.2.34. So necessarily $\mathcal{A}$ is a Ritt-characteristic set of $\langle P \rangle$.

Finally, assume that the dimension of $\mathcal{A}$ is zero. Let $p \in \mathcal{F}$. Since $\mathcal{K}_0 = \mathcal{K}$, it follows from Definition 4.2.32 that there exist polynomials $\beta_i \in \mathcal{K}[y_1, \ldots, y_n]$ such that

$$p = \sum_{i=i}^{n} \beta_i(y_1, \ldots, y_n) f_i(y_1, \ldots, y_i).$$

Since all $f_i \in \langle P \rangle$, it is clear that $p \in \langle P \rangle$. So $\mathcal{F} \subset \langle P \rangle$. The inclusion $\langle P \rangle \subset \mathcal{F}$ always holds (also when the dimension of $\mathcal{A}$ is greater than zero), and thus the proof is complete. ∎

So, given a finite set of polynomials $P \neq \{0\}$, we can compute a Wu-characteristic set of P rather easily with Algorithm 4.2.25. When the resulting chain is irreducible, this chain is also a Ritt-characteristic set of the ideal $\langle P \rangle$, generated by the polynomials in P.

Corollary 4.2.36 *Let $P = \{p_1, \ldots, p_m\}$ be a finite polynomial set containing a nonzero polynomial, and assume that $\langle P \rangle$ is a prime ideal. Let $\mathcal{A} = (f_1, \ldots, f_r)$ be a Wu-characteristic set of P, and suppose that $\mathcal{A}$ is irreducible. Let $\mathcal{F}$ be defined as in (4.32). Then $\langle P \rangle = \mathcal{F}$.*

Proof

Since $\mathcal{A}$ is irreducible, $\mathcal{A}$ is a Ritt-characteristic set of the prime ideal $\mathcal{F}$, but also of $\langle P \rangle$. Both $\mathcal{F}$ and $\langle P \rangle$ are prime ideals in the same ring $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$, and according to Theorem 4.2.17, $\mathcal{A}$ can only be the characteristic set of one prime polynomial ideal at a time. Therefore we must have $\langle P \rangle = \mathcal{F}$. ∎

If $\mathcal{A}$ is an irreducible Wu-characteristic set of a finite polynomial set P, and the ideal $\langle P \rangle$ generated by P is not prime, and the dimension of $\mathcal{A}$ is greater than zero, then the equality $\mathcal{F} = \langle P \rangle$ does not hold in general. This is illustrated in the following example.

Example 4.2.37 Let $K = \mathbf{R}$, and consider the polynomials in the indeterminates u and y, with the ordering $u \prec y$. Define $p_1(u, y) := uy$ and $P := \{p_1\}$. Clearly, $\langle P \rangle$ is not a prime ideal. Now $\mathcal{A} = (p_1)$ is a Wu-characteristic set of P, and, moreover, it is irreducible. According to Theorem 4.2.35 $\mathcal{A}$ is a Ritt-characteristic set of $\langle P \rangle$.

On the other hand, because $\frac{1}{u}(uy) = y$, the prime ideal $\mathcal{F}$, as defined in (4.32), is

$$\mathcal{F} = \langle y \rangle = \{yg(u, y) \mid g(u, y) \in K[u, y]\}.$$

$\mathcal{A}$ is also a Ritt-characteristic set of $\mathcal{F}$, and $\langle P \rangle \subset \mathcal{F}$, but clearly $\mathcal{F} \not\subset \langle P \rangle$.

From Example 4.2.37 we see that in general an irreducible Wu-characteristic set $\mathcal{A}$ of a polynomial set P may describe a larger polynomial ideal (namely the prime ideal $\mathcal{F}$) than the ideal $\langle P \rangle$ generated by the polynomials in P. However, this in a sense superfluous part of $\mathcal{F}$ is not unnecessarily large.

Corollary 4.2.38 *Let $P = \{p_1, \ldots, p_m\}$ be a finite set of polynomials containing a nonzero polynomial, and let $\mathcal{A} = (f_1, \ldots, f_r)$ be a Wu-characteristic set of P. Suppose that $\mathcal{A}$ is irreducible, and define the prime ideal $\mathcal{F}$ as in (4.32). Let $\mathcal{G}$ be a prime ideal in $K[x_1, \ldots, x_n]$ such that*

$$\langle P \rangle \subset \mathcal{G} \subset \mathcal{F}.$$

Then $\mathcal{G} = \mathcal{F}$.

Proof

We only have to prove that $\mathcal{F} \subset \mathcal{G}$. Since $\mathcal{A}$ is an irreducible ascending chain, we know from Theorem 4.2.34 that $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$. We show that $\mathcal{A}$ is also a (Ritt-) characteristic set of $\mathcal{G}$.

First of all, because $\mathcal{A}$ is a Wu-characteristic set of P, we know that $\mathcal{A}$ belongs to $\langle P \rangle$. So $\mathcal{A}$ also belongs to $\mathcal{G}$. Now suppose that $\mathcal{A}$ is not a (Ritt-) characteristic set of $\mathcal{G}$. Then there exists an ascending chain $\mathcal{B}$, belonging to $\mathcal{G}$, of lower rank than $\mathcal{A}$. Since $\mathcal{G} \subset \mathcal{F}$, this chain $\mathcal{B}$ also belongs to $\mathcal{F}$. So $\mathcal{B}$ is an ascending chain in $\mathcal{F}$ of lower rank than $\mathcal{A}$. This contradicts the fact that $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$, and we conclude that $\mathcal{A}$ is also a (Ritt-) characteristic set of $\mathcal{G}$.

Finally, $\mathcal{F}$ and $\mathcal{G}$ are prime ideals in the same ring $K[x_1, \ldots, x_n]$, and $\mathcal{A}$ is a (Ritt-) characteristic set of both $\mathcal{F}$ and $\mathcal{G}$, so according to Theorem 4.2.17 we have $\mathcal{F} = \mathcal{G}$. ∎

It is also possible to describe the part of $\mathcal{F}$ which does not belong to $\langle P \rangle$ more explicitly. Although it is rather difficult to do this in terms of polynomial sets, the relationship can be characterized completely with help of the varieties of $\langle P \rangle$ and $\mathcal{F}$. The next proposition, which originates from the work of Wu (see [101]), states this result.

Proposition 4.2.39 *Let $P = \{p_1, \ldots, p_m\}$ be a finite polynomial set containing a nonzero polynomial, and suppose that $\mathcal{A} = (f_1, \ldots, f_r)$ is a Wu-characteristic set of P. Assume that $\mathcal{A}$ is irreducible. Let I_i denote the initial of f_i $(i = 1, \ldots, r)$, and let $\mathcal{F}$ be the prime ideal as defined in (4.32). Then*

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}) \cup \bigcup_{i=1}^{r} \mathcal{V}(P \cup \{I_i\}). \tag{4.33}$$

Proof

"$\supset$" We know already that $P \subset \mathcal{F}$, so $\mathcal{V}(\mathcal{F}) \subset \mathcal{V}(P)$. Moreover, for all $i \in \{1, \ldots, r\} : P \subset P \cup \{I_i\}$, so $\mathcal{V}(P \cup \{I_i\}) \subset \mathcal{V}(P)$. Hence

$$\left[\bigcup_{i=1}^{r} \mathcal{V}(P \cup \{I_i\})\right] \subset \mathcal{V}(P).$$

"$\subset$" Let $\alpha \in \mathcal{V}(P)$. Assume first that there exists an initial I_j such that α is a zero of I_j. Then $\alpha \in \mathcal{V}(P \cup \{I_j\})$ and so

$$\alpha \in \mathcal{V}(P \cup \{I_j\}) \subset \bigcup_{i=1}^{r} \mathcal{V}(P \cup \{I_i\}).$$

Next assume that α is not a zero of one of the initials $I_1, \ldots, I_r$. Let $p \in \mathcal{F}$. Since $\mathcal{A}$ is a Ritt-characteristic set of $\mathcal{F}$, $\mathrm{prem}(p, \mathcal{A}) = 0$, and there exist integers $\nu_1, \ldots, \nu_r$ and polynomials $\beta_1, \ldots, \beta_r$ such that

$$I_1^{\nu_1} \cdots I_r^{\nu_r} p = \sum_{i=1}^{r} \beta_i f_i. \tag{4.34}$$

Since $\mathcal{A}$ belongs to $\langle P \rangle$, α is a zero of all polynomials f_i, and therefore α is a zero of the right-hand side of (4.34). By assumption α is not a zero of $I_1, \ldots, I_r$, so it must be a zero of p. $p \in \mathcal{F}$ was arbitrary, so $\alpha \in \mathcal{V}(\mathcal{F})$.

We conclude that in both cases

$$\alpha \in \mathcal{V}(\mathcal{F}) \subset \mathcal{V}(\mathcal{F}) \cup \bigcup_{i=1}^{r} \mathcal{V}(P \cup \{I_i\}).$$

This completes the proof. ∎

The result of Proposition 4.2.39 is illustrated by our example.

Example 4.2.40 In the situation of Example 4.2.37, so with $p_1 = uy$, $I_1 = u$, $P = \{p_1\}$ and $\mathcal{F} = \langle y \rangle$, the varieties in (4.33) become:

$$\begin{aligned} \mathcal{V}(P) &= \{(u, y) \mid u = 0 \vee y = 0\} = \\ &= \{(u, 0) \mid u \in \bar{\mathcal{K}}\} \cup \{(0, y) \mid y \in \bar{\mathcal{K}}\}, \\ \mathcal{V}(\mathcal{F}) &= \{(u, 0) \mid y \in \bar{\mathcal{K}}\}, \\ \mathcal{V}(P \cup \{I_1\}) &= \{(0, y) \mid y \in \bar{\mathcal{K}}\}. \end{aligned}$$

(where $\bar{\mathcal{K}}$ denotes the algebraic closure of $\mathcal{K}$). Indeed we have

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}) \cup \mathcal{V}(P \cup \{I_1\}).$$

Proposition 4.2.39 is very important for the following problem. Let P be a finite polynomial set, and suppose we are interested in the set of all common zeros (in the algebraic closure $\bar{\mathcal{K}}$ of $\mathcal{K}$) of the polynomials in P. So we want to compute the variety of P in $\bar{\mathcal{K}}$.

Suppose that after the computation of a Wu-characteristic set, this set turns out to be an irreducible ascending chain. Then we can decompose the variety of P as described in Proposition 4.2.39. This yields a prime ideal $\mathcal{F}$ (with (Ritt-)

characteristic set $\mathcal{A}$), and finite polynomial sets $P_i := P \cup \{I_i\}$, $(i = 1, \ldots, r)$. Since the initial I_i is reduced with respect to $\mathcal{A}$, $\langle P_i \rangle$ will have a (Ritt-) characteristic set of lower rank than $\mathcal{A}$. (Recall the "$\Rightarrow$" part of the proof of Lemma 4.2.14.) For all $i \in \{1, \ldots, r\}$ we can compute Wu-characteristic sets of P_i. If they all turn out to be irreducible, we can decompose the corresponding varieties again. This process can be continued until the polynomial sets $\tilde{P}_i$ we get become contradictory: the polynomials in $\tilde{P}_i$ generate the whole ring, and the variety $\mathcal{V}(\tilde{P}_i)$ of $\tilde{P}_i$ is empty. In this way we end up with a sequence of (Ritt-) characteristic sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and a decomposition

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}_1) \cup \cdots \cup \mathcal{V}(\mathcal{F}_k), \tag{4.35}$$

where $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are prime polynomial ideals and for all $i \in \{1, \ldots, k\}$: $\mathcal{A}_i$ is a (Ritt-) characteristic set of $\mathcal{F}_i$. Since an irreducible ascending chain uniquely determines a prime polynomial ideal, the sequence of characteristic sets $(\mathcal{A}_1, \ldots, \mathcal{A}_k)$ is an unambiguous description of the variety $\mathcal{V}(P)$.

Of course the process described above has one serious drawback. In each step we have to assume that all Wu-characteristic sets we have computed in that step are irreducible. This is a quite restrictive condition, but fortunately, also when we encounter reducible ascending chains during the computation, the process can be carried out in almost the same way. The next subsection is devoted to this subject.

4.2.5 Decomposition of varieties and radical ideals

This subsection has a double purpose. On the one hand we generalize the results on the decomposition of varieties obtained in the previous subsection to the case of reducible ascending chains. On the other hand, we are interested in a reformulation of these results in terms of polynomial ideals. Such a reformulation is possible, and in this way the characteristic sets method becomes a constructive method for the computation of the Lasker-Noether decomposition of a radical ideal, representing it as a finite intersection of prime ideals (for some theoretical background we refer to Appendix A, Theorem A.1.16, Corollary A.1.17 and Theorem A.2.8). The key to the solution is the observation that the decomposition process explained in subsection 4.2.4 for irreducible ascending chains, may be carried out in a similar way in the reducible case. To explain this idea, we follow the same lines as Wu-Wen-Tsun in [101].

Consider an ascending chain $\mathcal{A} = (f_1, \ldots, f_r)$ in $\mathcal{K}[x_1, \ldots, x_n]$ and rename the indeterminates as in Definition 4.2.29. Assume that $\mathcal{A}$ is reducible. Then there exists an integer $k \in \mathbf{N}$ such that

$$\mathcal{A}_{k-1} = (f_1, f_2, \ldots, f_{k-1})$$

is irreducible with generic point $\tilde{\eta}_{k-1} = (u_1, \ldots, u_d, \eta_1, \ldots, \eta_{k-1})$ and the polynomial $\tilde{f}_k$, obtained by substitution of $(\eta_1, \ldots, \eta_{k-1})$ for $(y_1, \ldots, y_{k-1})$ in f_k, is a reducible polynomial in $\mathcal{K}_{k-1}[y_k]$, where $\mathcal{K}_{k-1}$ is the algebraic extension field $\mathcal{K}_0(\eta_1, \ldots, \eta_{k-1})$, obtained by successively adjoining $\eta_1, \ldots, \eta_{k-1}$ to $\mathcal{K}_0$. Let the irreducible factorization of the polynomial $\tilde{f}_k$ in $\mathcal{K}_{k-1}[y_k]$ be given by

$$\tilde{f}_k = \tilde{g}_1 \cdots \tilde{g}_h. \tag{4.36}$$

So $h \geq 2$, and all polynomials $\tilde{g}_i \in \mathcal{K}_{k-1}[y_k]$ $(i = 1, \ldots, h)$ are irreducible over $\mathcal{K}_{k-1}[y_k]$.

The coefficients of the polynomial $\tilde{g}_i$ $(i = 1, \ldots, h)$ in (4.36) are elements of $\mathcal{K}_{k-1}$. From the theory of algebraic field extensions (see for example [104, p. 56]), it follows that each coefficient c_j of $\tilde{g}_i$ may be written as

$$c_j = \frac{\beta_j}{\gamma_j},$$

with $\beta_j \in \mathcal{K}[u_1, \ldots, u_d, \eta_1, \ldots, \eta_{k-1}]$ and $\gamma_j \in \mathcal{K}[u_1, \ldots, u_d]$. After multiplication of (4.36) with the product of all denominator polynomials γ_j we obtain:

$$d\tilde{f}_k = \bar{g}_1 \cdots \bar{g}_h, \tag{4.37}$$

where $d \in \mathcal{K}[u_1, \ldots, u_d]$ and all $\bar{g}_i \in \mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-1}][y_k]$ $(i = 1, \ldots, h)$. Since $d \in \mathcal{K}[u_1, \ldots, u_d]$, d is trivially reduced with respect to $\mathcal{A}_k = (f_1, \ldots, f_k)$.

Let $i \in \{1, \ldots, h\}$, and consider the polynomial $\bar{g}_i$. In this polynomial we can successively substitute back $(y_{k-1}, \ldots, y_1)$ for $(\eta_{k-1}, \ldots, \eta_1)$ in the following way.

$$\bar{g}_i \in \mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-1}][y_k].$$

Regard $\bar{g}_i$ as a polynomial in $\mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-2}][\eta_{k-1}, y_k]$, (so as a polynomial in the indeterminates η_{k-1} and y_k with coefficients in $\mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-2}]$), and replace η_{k-1} by y_{k-1}. This yields a polynomial:

$$\bar{g}_{2i} \in \mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-2}][y_{k-1}, y_k].$$

The polynomial $\tilde{f}_{k-1}$, obtained by substituting $(\eta_1, \ldots, \eta_{k-2})$ for $(y_1, \ldots, y_{k-2})$ in f_{k-1}, is irreducible with extended zero η_{k-1}. So $\tilde{f}_{k-1}$ is a minimal polynomial in y_{k-1} over $\mathcal{K}_0(\eta_1, \ldots, \eta_{k-2})$. Thus from the theory of algebraic field extensions (see [104, p. 56]), it follows that

$$\deg_{y_{k-1}}(\bar{g}_{2i}) = \deg_{\eta_{k-1}}(\bar{g}_i) < \deg_{y_{k-1}}(\tilde{f}_{k-1}) \leq \deg_{y_{k-1}}(f_{k-1}),$$

and we see that $\bar{g}_{2i}$ is reduced with respect to f_{k-1}, because y_{k-1} is the indeterminate of highest rank occurring in f_{k-1}. Moreover, the indeterminate y_k is not influenced by this substitution.

In this way we continue: regard $\bar{g}_{2i}$ as a polynomial in the indeterminates η_{k-2}, y_{k-1} and y_k with coefficients in $\mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-3}]$ and replace η_{k-2} by y_{k-2}. This yields a polynomial:

$$\bar{g}_{3i} \in \mathcal{K}[u_1, \ldots, u_d][\eta_1, \ldots, \eta_{k-3}][y_{k-2}, y_{k-1}, y_k].$$

Since the polynomial $\tilde{f}_{k-2}$, obtained by substituting $(\eta_1, \ldots, \eta_{k-3})$ for $(y_1, \ldots, y_{k-3})$ in f_{k-2}, is irreducible with extended zero η_{k-2}, the results in [104] yield again that

$$\deg_{y_{k-2}}(\bar{g}_{3i}) = \deg_{\eta_{k-2}}(\bar{g}_{2i}) < \deg_{y_{k-2}}(\tilde{f}_{k-2}) \leq \deg_{y_{k-2}}(f_{k-2}).$$

Of course, the indeterminates y_k and y_{k-1} are not influenced by this successive substitution process. So finally we obtain a polynomial g_i such that

$$g_i \in \mathcal{K}[u_1, \ldots, u_d][y_1, \ldots, y_k],$$

and for all $j \in \{1, \ldots, k-1\}$:

$$\deg_{y_j}(g_i) = \deg_{\eta_j}(\bar{g}_{k+1-j,i}) < \deg_{y_j}(\tilde{f}_j) \leq \deg_{y_j}(f_j),$$

while

$$\deg_{y_k}(g_i) = \deg_{y_k}(g_i(\eta_1, \ldots, \eta_{k-1}, y_k)) = \deg_{y_k}(\bar{g}_i) < \deg_{y_k}(\tilde{f}_k) \leq \deg_{y_k}(f_k).$$

We conclude that g_i is reduced with respect to $\mathcal{A}_k$.

The same process can be carried out for each polynomial $\bar{g}_i$ $(i = 1, \ldots, h)$ separately. So we have found a constructive method that enables us to compute explicitly polynomials $g_i \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_k]$ $(i = 1, \ldots, h)$ such that all g_i are reduced with respect to $\mathcal{A}_k$ and

$$g_i(u_1, \ldots, u_d, \eta_1, \ldots, \eta_{k-1}, y_k) = \bar{g}_i. \tag{4.38}$$

Next consider the polynomial

$$g_1 \cdots g_h - df_k.$$

This is a polynomial in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_k]$, but, for the moment, we consider it as a polynomial in the indeterminate y_k:

$$g_1 \cdots g_h - df_k = \sum_{j=0}^{N} b_j y_k^j,$$

with coefficients $b_j \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_{k-1}]$. Let β_j $(j = 0, 1, \ldots, N)$ denote the element in $\mathcal{K}_{k-1}$, obtained by substituting $(\eta_1, \ldots, \eta_{k-1})$ for $(y_1, \ldots, y_{k-1})$ in b_j. Then for all $j \in \{0, 1, \ldots, N\}$ we have $b_j = 0$, because (recall (4.37))

$$\bar{g}_1 \cdots \bar{g}_h = d\tilde{f}_k.$$

Therefore $\tilde{\eta}_{k-1} = (u_1, \ldots, u_d, \eta_1, \ldots, \eta_{k-1})$ is an extended zero of all polynomials b_j $(j = 0, 1, \ldots, N)$, and application of Proposition 4.2.31 (recall that the ascending chain $\mathcal{A}_{k-1}$ is irreducible with generic point $\tilde{\eta}_{k-1}$) yields that

$$\forall j \in \{0, 1, \ldots, N\} : \operatorname{prem}(b_j, \mathcal{A}_{k-1}) = 0.$$

So for each $j \in \{0, 1, \ldots, N\}$ there exist integers $\nu_{j,1}, \ldots, \nu_{j,k-1} \in \mathbb{N}$ and polynomials $q_{ji} \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_{k-1}]$ such that

$$I_1^{\nu_{j,1}} \cdots I_{k-1}^{\nu_{j,k-1}} b_j = \sum_{i=1}^{k-1} q_{ji} f_i,$$

where I_i denotes the initial of f_i. Define for all $i = 1, \ldots, k-1$: $\nu_i := \max(\nu_{j,i} \mid j = 0, 1, \ldots, N\}$. Then there exist polynomials $\tilde{q}_i \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_k]$ such that

$$I_1^{\nu_1} \cdots I_{k-1}^{\nu_{k-1}} (g_1 \cdots g_h - df_k) = \sum_{i=1}^{k-1} \tilde{q}_i f_i.$$

So, defining $\tilde{q}_k := I_1^{\nu_1} \cdots I_{k-1}^{\nu_{k-1}} d$, we finally get

$$I_1^{\nu_1} \cdots I_{k-1}^{\nu_{k-1}} g_1 \cdots g_h = \sum_{i=1}^{k} \tilde{q}_i f_i. \tag{4.39}$$

Proposition 4.2.41 *Let $P = \{p_1, \ldots, p_m\}$ be a finite set of polynomials in the ring $\mathcal{K}[x_1, \ldots, x_n]$ containing a nonzero polynomial, and let $\mathcal{A} = (f_1, \ldots, f_r)$ be a Wu-characteristic set of P. Assume that for all $i = 1, \ldots, r$: $\mathrm{class}(f_i) > 0$ and denote the initial of f_i by I_i. Rename the indeterminates $x_1, \ldots, x_n$ as $u_1, \ldots, u_d, y_1, \ldots, y_k$ in the same way as in Definition 4.2.29. Assume that $\mathcal{A}$ is reducible. Then there exists a $k \in \mathbf{N}$ such that $\mathcal{A}_{k-1} = (f_1, \ldots, f_{k-1})$ is irreducible with generic point $\tilde{\eta}_{k-1} \in \mathcal{K}_{k-1}^{d+k-1}$, while the polynomial $\tilde{f}_k = f_k(\tilde{\eta}_{k-1}, y_k)$, obtained by substituting $\tilde{\eta}_{k-1}$ for $(u_1, \ldots, u_d, y_1, \ldots, y_{k-1})$ in f_k, is reducible over $\mathcal{K}_{k-1}$. Let $\tilde{g}_1 \cdots \tilde{g}_h$ be an irreducible factorization of $\tilde{f}_k$, and define the polynomials $\bar{g}_1, \ldots, \bar{g}_h$ and d as in (4.37). Construct polynomials $g_1, \ldots, g_h$ in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_k]$ such that all g_i are reduced with respect to $\mathcal{A}_k$ and satisfy (4.38). Then the variety of P can be decomposed as*

$$\mathcal{V}(P) = \bigcup_{i=1}^{k-1} \mathcal{V}(P \cup \{I_i\}) \cup \bigcup_{j=1}^{h} \mathcal{V}(P \cup \{g_j\}). \tag{4.40}$$

Moreover, in this situation the following two statements hold:

(i) $\forall i \in \{1, \ldots, k-1\}$*: any medial set of $P \cup \mathcal{A} \cup \{I_i\}$ has lower rank than $\mathcal{A}$.*

(ii) $\forall j \in \{1, \ldots, h\}$*: any medial set of $P \cup \mathcal{A} \cup \{g_j\}$ has lower rank than $\mathcal{A}$.*

Proof

We start with the proof of equality (4.40).

"$\supset$" Clearly for all $i = 1, \ldots, k-1$ we have $P \subset P \cup \{I_i\}$, so $\mathcal{V}(P \cup \{I_i\}) \subset \mathcal{V}(P)$, and completely analogous for all $j = 1, \ldots, h$: $P \subset P \cup \{g_j\}$, so $\mathcal{V}(P \cup \{g_j\}) \subset \mathcal{V}(P)$. This proves "$\supset$".

"$\subset$" Let $\alpha \in \mathcal{V}(P)$. Since $\mathcal{A}$ belongs to $\langle P \rangle$, α is a zero of all polynomials f_i in $\mathcal{A}$. So α is a zero of the right-hand side of (4.39). Therefore α must be a zero of one of the factors of the left-hand side of (4.39). Thus α is a zero of some I_i or some g_j and we conclude that there exists an $i \in \{1, \ldots, k-1\}$ or an $j \in \{1, \ldots, h\}$ such that $\alpha \in \mathcal{V}(P \cup \{I_i\})$ or $\alpha \in \mathcal{V}(P \cup \{g_j\})$.

To prove *(i)*, let $1 \leq i \leq k-1$. Then the initial I_i is reduced with respect to $\mathcal{A}$. So, according to Lemma 4.2.27 any medial set of $P \cup \mathcal{A} \cup \{I_i\}$ has lower rank than $\mathcal{A}$.

Finally, because for all $j \in \{1, \ldots, h\}$, g_j is reduced with respect to $\mathcal{A}_k$, g_j is also reduced with respect to $\mathcal{A}$. Hence, *(ii)* is proved with completely the same argument. ∎

Combining Proposition 4.2.41 and Proposition 4.2.39, it is possible to derive a constructive method for the decomposition of the variety of an arbitrary polynomial ideal into irreducible varieties (see Definition A.2.6). The prime polynomial ideals corresponding to these irreducible varieties (recall Proposition A.2.7) are determined by their (Ritt-) characteristic sets.

Start with a polynomial set $P = \{p_1, \ldots, p_m\}$ and compute a Wu-characteristic set of P with Algorithm 4.2.25. If it is contradictory, $\langle P \rangle$ is the whole ring and the algorithm terminates because $\mathcal{V}(P)$ is empty. Otherwise, the Wu-characteristic set $\mathcal{A} = (f_1, \ldots, f_r)$ of P is an ascending chain with for all $i \in \{1, \ldots, r\}$: $\mathrm{class}(f_i) > 0$.

First assume that $\mathcal{A}$ is an irreducible ascending chain. Then we decompose $\mathcal{V}(P)$ as in formula (4.33):

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}) \cup \bigcup_{i=1}^{r} \mathcal{V}(P \cup \{I_i\}), \tag{4.41}$$

where $\mathcal{F}$ is the prime ideal as defined in (4.32) and $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$.

If $\mathcal{A}$ is reducible, we apply Proposition 4.2.41 and write

$$\mathcal{V}(P) = \bigcup_{i=1}^{k-1} \mathcal{V}(P \cup \{I_i\}) \cup \bigcup_{j=1}^{h} \mathcal{V}(P \cup \{g_j\}). \tag{4.42}$$

In both the reducible and the irreducible case we can decompose the varieties of the polynomial sets $P \cup \{I_i\}$ and $P \cup \{g_j\}$ in completely the same way. Because $\mathcal{A}$ belongs to $\langle P \rangle$, $P \cup \{I_i\}$ and $P \cup \mathcal{A} \cup \{I_i\}$ generate the same ideal. The same holds true for $P \cup \{g_j\}$ and $P \cup \mathcal{A} \cup \{g_j\}$. Therefore we continue the process with these larger polynomial sets. From Proposition 4.2.41 it follows that $P \cup \mathcal{A} \cup \{I_i\}$ and $P \cup \mathcal{A} \cup \{g_j\}$ must have Wu-characteristic sets of lower rank than $\mathcal{A}$. In this way, the rank of the Wu-characteristic sets we are computing is strictly decreasing. So at a certain moment this process terminates, because then all the Wu-characteristic sets under consideration are contradictory. At that moment we have found a decomposition

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}_1) \cup \cdots \cup \mathcal{V}(\mathcal{F}_l), \tag{4.43}$$

where $\mathcal{F}_1, \ldots, \mathcal{F}_l$ are prime polynomial ideals and the computed ascending chains $\mathcal{A}_1, \ldots, \mathcal{A}_l$ are irreducible (Ritt-) characteristic sets of $\mathcal{F}_1, \ldots, \mathcal{F}_l$ respectively. So $\mathcal{F}_1, \ldots, \mathcal{F}_l$ are completely determined by $\mathcal{A}_1, \ldots, \mathcal{A}_l$ and we have found a constructive proof for Theorem A.2.8, the decomposition theorem for algebraic varieties stated in Appendix A.2.

Corollary 4.2.42 *Let $P = \{p_1, \ldots, p_m\}$ be a polynomial set containing a nonzero polynomial, and assume that $\langle P \rangle$ is a prime ideal. Let $\mathcal{A} = (f_1, \ldots, f_r)$ be a Ritt-characteristic set of $\langle P \rangle$. Then $\mathcal{A}$ is an irreducible ascending chain.*

Proof

Let $\mathcal{A}$ be a Ritt-characteristic set of the prime ideal $\langle P \rangle$, and assume that $\mathcal{A}$ is reducible. Then we can carry out the decomposition process described above to arrive at the decomposition (4.43):

$$\mathcal{V}(\langle P \rangle) = \mathcal{V}(\mathcal{F}_1) \cup \cdots \cup \mathcal{V}(\mathcal{F}_l),$$

where $\mathcal{F}_i (i = 1, \ldots, l)$ are prime ideals with characteristic sets $\mathcal{A}_1, \ldots, \mathcal{A}_l$ respectively. From the description of the decomposition process it follows immediately that all ascending chains $\mathcal{A}_1, \ldots, \mathcal{A}_l$ are irreducible and have strictly lower rank than $\mathcal{A}$.

Now $\langle P \rangle$ is a prime ideal, and therefore the variety $\mathcal{V}(P)$ is irreducible (see Proposition A.2.7). This implies that there exists a $k \in \{1, \ldots, l\}$ such that

$$\mathcal{V}(P) = \mathcal{V}(\mathcal{F}_k).$$

Since $\langle P \rangle$ and $\mathcal{F}_k$ are both prime ideals, it follows that $\langle P \rangle = \mathcal{F}_k$ (see for example [105, pp. 160-161]). Therefore $\mathcal{A}_k$ is an ascending chain in $\langle P \rangle$ $(= \mathcal{F}_k)$ of lower rank than $\mathcal{A}$. This contradicts the assumption that $\mathcal{A}$ is a Ritt-characteristic set of $\langle P \rangle$, and we conclude that $\mathcal{A}$ must be irreducible. ∎

Using the results of Appendix A.2, formula (4.43) can be translated back into terms of polynomial ideals. Let $\langle P \rangle$ denote the ideal generated by the finite set of polynomials P. Then, according to Corollary A.2.11, $\mathrm{Id}(\mathcal{V}(P)) = \sqrt{\langle P \rangle}$. Moreover, because $\mathcal{F}_1, \ldots, \mathcal{F}_\ell$ are prime ideals, we have $\mathrm{Id}(\mathcal{V}(\mathcal{F}_i)) = \mathcal{F}_i$ $(i = 1, \ldots, \ell)$. Successive application of (A.5) finally yields

$$\sqrt{\langle P \rangle} = \mathcal{F}_1 \cap \cdots \cap \mathcal{F}_l, \tag{4.44}$$

where the prime ideals $\mathcal{F}_1, \ldots, \mathcal{F}_l$ are completely determined by their irreducible (Ritt-) characteristic sets $\mathcal{A}_1, \ldots, \mathcal{A}_l$. In this way we have derived a constructive method to carry out the Lasker-Noether decomposition theorem for radical ideals as described in Corollary A.1.17.

Formula (4.44) may also be used to solve the membership problem for radical ideals. Suppose we have a finite polynomial set $P = \{p_1, \ldots, p_m\}$ and a polynomial $g \in \mathcal{K}[x_1, \ldots, x_n]$, and we want to know whether or not $g \in \sqrt{\langle P \rangle}$. Then we first decompose the radical of $\langle P \rangle$ as before:

$$\sqrt{\langle P \rangle} = \mathcal{F}_1 \cap \cdots \cap \mathcal{F}_l.$$

This decomposition is determined by the (Ritt-) characteristic sets $\mathcal{A}_1, \ldots, \mathcal{A}_l$ of the prime polynomial ideals $\mathcal{F}_1, \ldots, \mathcal{F}_l$ respectively. Now clearly $g \in \sqrt{\langle P \rangle}$ if and only if

$$\forall i = 1, \ldots, l : g \in \mathcal{F}_i.$$

Since $\mathcal{A}_i$ is a (Ritt-) characteristic set of the prime ideal $\mathcal{F}_i$, we know from Theorem 4.2.17 that in this case the membership problem is easy to solve: $g \in \mathcal{F}_i \iff \mathrm{prem}(g, \mathcal{A}_i) = 0$. Combining these results we obtain:

$$g \in \sqrt{\langle P \rangle} \iff \forall i \in \{1, \ldots, l\} : \mathrm{prem}(g, \mathcal{A}_i) = 0. \tag{4.45}$$

Finally we have to make a critical remark. All parts of the decomposition method described in this subsection were constructive, except one: the factorization of the polynomial $\tilde{f}_k$ over $\mathcal{K}_{k-1}$ (formula (4.36)). In general this is a quite difficult problem, although lately Wang has given a possible solution to this problem (see [99]). Nevertheless, the factorization question remains the most important bottleneck in the characteristic sets method.

4.2.6 Complexity issues

There is not much known yet on the complexity of the characteristic sets algorithm. Like Gröbner basis computation, the computation of a Wu-characteristic set can become very time and space consuming. In [28] some efficient algorithms are given together with bounds on the complexity. Often it is assumed that the computation of a characteristic set is somewhat less involved than the computation of Gröbner bases because a Gröbner basis of a polynomial ideal has stronger properties than a Ritt- or Wu-characteristic set.

4.3 A comparison of Gröbner bases and characteristic sets

After the introduction of two different methods from constructive commutative algebra in the previous sections, the question arises which method is the most favourable one to use. It is almost impossible to give an exhaustive answer to this question because it very much depends on the particular situation which method is preferable. In this section we summarize the results of this chapter and compare both methods. In this way we obtain some insight in the considerations that lead to a choice for one of the methods in some specific situations.

In Section 4.1 we have seen that a Gröbner basis is a set of *generators* of a polynomial ideal with some very useful properties. Using Gröbner bases it is possible to decide on the membership problem for arbitrary polynomial ideals. Operations on ideals, like summation and intersection, are easily carried out explicitly with help of Gröbner bases. Also for the determination of varieties of polynomial ideals, Gröbner bases may be applied. If a pure lexicographic term ordering is used, a triangular form is computed, and with backward substitution the variety of an ideal is obtained. Therefore the Gröbner basis algorithm is a constructive method that may handle both polynomial ideals and their varieties.

The applicability of characteristic sets is far more restricted. In general they are not a generating subset of a polynomial ideal. Only irreducible ascending chains form an exception; in some sense they generate a specific prime ideal, as explained in Definition 4.2.32 and Theorem 4.2.34. Only the membership problem for prime polynomial ideals can be solved directly. Moreover, the decomposition of a characteristic set in irreducible ascending chains enables us to solve the same question for radical ideals too. However, the main application of characteristic sets is the computation of the variety of a polynomial ideal. Since by definition ascending chains have a triangular form, characteristic sets are especially suitable for this purpose. A Wu-characteristic set obtained with Algorithm 4.2.25 is often enough to determine the variety of a polynomial ideal.

Although the differences between Gröbner bases and characteristic sets are often very significant, both methods are sometimes also quite related to each other. When for an ideal in the polynomial ring $K[x_1, \ldots, x_n]$ a Gröbner basis is computed with respect to the pure lexicographic term ordering and a ranking $x_1 \prec x_2 \prec \cdots \prec x_n$ of the indeterminates, the resemblances with the Ritt-Wu algorithm are often quite remarkable. Of course there remain differences, both in the ordering of polynomials (total vs. partial) and in the reduction process (remainders vs. pseudo-remainders), but the outcome of both algorithms may look very similar. This is illustrated by the next result.

Proposition 4.3.1 *Let $\mathcal{I}$ be a zero-dimensional ideal in $K[x_1, \ldots, x_n]$, and let G be a reduced Gröbner basis of $\mathcal{I}$ w.r.t. the pure lexicographic term ordering with ranking*

$$x_1 \prec x_2 \prec \cdots \prec x_n.$$

Assume that G contains exactly n polynomials. Then G is a Ritt-characteristic set of the ideal $\mathcal{I}$ w.r.t. the same ordering of indeterminates.

Proof

Let $G = \{g_1, \ldots, g_n\}$ be a reduced Gröbner basis of $\mathcal{I}$ w.r.t. the pure lexicographic term ordering with ranking $x_1 \prec x_2 \prec \cdots \prec x_n$. According to Proposition 4.1.30 *(ii)* there exists for every indeterminate x_i $(i = 1, \ldots, n)$ a polynomial $p \in \mathcal{I}$ such that $\mathrm{in}(p) = c_i \cdot x_i^{n_i}$ with $c_i \in \mathcal{K}$ and $n_i \in \mathbb{N}$. Since G is a Gröbner basis of $\mathcal{I}$ we have $\mathrm{in}(p) \in \langle \mathrm{in}(G) \rangle$, and thus there exists a $g \in G$ such that $\mathrm{in}(g)$ divides $\mathrm{in}(p)$. So G has a triangular form, and renumbering the polynomials in G we have that for every $i \in \{1, \ldots, n\}$ the polynomial $g_i \in G$ is an element of $\mathcal{K}[x_1, \ldots, x_i]$ and $\mathrm{in}(g_i) = \tilde{c}_i \cdot x_i^{m_i}$ with $\tilde{c}_i \in \mathcal{K}$ and $m_i \in \mathbb{N}$.

First we show that $\mathcal{A} := (g_1, \ldots, g_n)$ is an ascending chain in $\mathcal{K}[x_1, \ldots, x_n]$. Let $j, k \in \{1, \ldots, n\}$ with $j < k$. It is obvious that the polynomial g_j is reduced with respect to g_k. To prove that g_k is also reduced with respect to g_j we consider an arbitrary monomial q of g_k. Since G is a *reduced* Gröbner basis of $\mathcal{I}$, we know that $q \notin \langle \{\tilde{c}_i \cdot x_i^{m_i} \mid i = 1, \ldots, n\} \rangle$. So in particular $\deg_{x_j}(q) < m_j = \deg_{x_j}(g_j)$. This implies that g_k is reduced w.r.t. g_j and thus $\mathcal{A}$ is an ascending chain.

Finally, assume that there exists a nonzero polynomial $p \in \mathcal{I}$ that is reduced w.r.t. $\mathcal{A}$, i.e.

$$\forall i \in \{1, \ldots, n\} : \deg_{x_i}(p) < \deg_{x_i}(g_i) = m_i.$$

Of course, the same inequalities hold for $\mathrm{in}(p)$. Therefore $\mathrm{in}(p) \notin \langle \mathrm{in}(G) \rangle$. This contradicts the fact that G is a Gröbner basis of $\mathcal{I}$. We conclude that the only polynomial $p \in \mathcal{I}$ that is reduced with respect to $\mathcal{A}$ is the zero-polynomial. According to Lemma 4.2.14, this implies that $\mathcal{A}$ is a Ritt-characteristic set of $\mathcal{I}$. ∎

Proposition 4.3.1 indicates that for zero-dimensional ideals the differences between Gröbner bases and characteristic sets are not so apparent. This is illustrated by Examples 4.1.32 and 4.2.28; it is easily verified that the reduced Gröbner basis (g_3, g_2, g_1) of the zero-dimensional polynomial ideal in this example has the same rank as the Wu-characteristic set (h_1, h_2, h_3). Therefore they are both Ritt-characteristic sets of the polynomial ideal under consideration. The condition in Proposition 4.3.1 on the number of polynomials in the reduced Gröbner basis of a zero-dimensional ideal is always satisfied if $\mathcal{I}$ is a prime ideal (see [30, Proposition 5.9]). But also when $\mathcal{I}$ is not prime, this condition seems not very restrictive; if a zero-dimensional ideal is in so-called general position (see [30, Section 7]), the number of polynomials in its reduced Gröbner basis is often equal to n.

When an ideal is not zero-dimensional, its reduced Gröbner basis with respect to the pure lexicographic term ordering is sometimes still related to a characteristic set. In [26, Chapter 4] the relationship between these two concepts is studied in more detail. Also in [10, pp. 83-85], some results on this subject are given.

If in a particular situation there are only minor differences between Gröbner bases and characteristic sets, the complexity of the algorithm to compute one of them is an important argument for the choice of one of these methods. Often it is claimed (see e.g. [10, p. 89]) that the characteristic set algorithm is generally faster than the corresponding Gröbner basis computation. However, the Gröbner basis algorithm is more mature: it is better known and developed, and more generally available than the characteristic sets algorithm. Moreover, some questions on polynomial ideals are only solvable using Gröbner bases.

The most important argument for the choice of a specific method is the application for which it is used. For the determination of the variety of a polynomial ideal, the computation of a characteristic set is often enough. This method has also won its spurs in the field of mechanical geometry theorem proving (see [10]). The Gröbner basis method on the other hand, is especially suitable for manipulations with and operations on polynomial ideals. Because of its transparent structure, a wide variety of problems in this field can be solved (see e.g. [14, Chapter 4]). Moreover, Gröbner bases can also be used for mechanical geometry theorem proving and for the determination of varieties of polynomial ideals. Especially in this last application, Gröbner bases have the advantage that some partial information on a variety is obtainable from a Gröbner basis w.r.t. an *arbitrary* term ordering. If an ideal is zero-dimensional, univariate polynomials in the ideal can be obtained from this Gröbner basis using the method of Proposition 4.1.37. Since the determination of a Gröbner basis w.r.t. the (reverse) graded lexicographic term ordering is much faster than w.r.t. the pure lexicographic term ordering, this procedure speeds up the computations considerably.

In the rest of this thesis, Gröbner bases are used almost exclusively to solve our questions on polynomial ideals. Our motivation for this choice is twofold. First, we are interested in both polynomial ideals and their varieties, and we have just seen that for the study of polynomial ideals, the Gröbner basis method is a more appropriate tool. Second, because of the direct correspondence between Gröbner bases and polynomial ideals, this method has a great flexibility. In the next chapter it turns out that most of the problems we are interested in may be reformulated as questions on polynomial ideals in several different ways. With the characteristic sets method only one (or none) of these solution strategies can be used in general. With Gröbner bases on the contrary, every reformulation leads to a different constructive solution method for the problem under consideration. The flexibility of the Gröbner basis method enables us to test several solution strategies and to judge them on their merits and efficiency. In this way it is possible to find effective solution methods for the problems we are interested in. The general applicability and the greater flexibility of the Gröbner basis method are the main incentives for the use of Gröbner bases in the next chapter.

Chapter 5

Testing reachability and stabilizability

In Chapter 2, conditions were derived for the reachability and stabilizability (w.r.t. a given Hurwitz set $\mathcal{D}$) of a system over a commutative ring $\mathcal{R}$. Both conditions look very similar; for a system $\Sigma = (A, B, C, D)$ over an integral domain $\mathcal{R}$, they can be formulated as right-invertibility conditions on the matrix $(zI - A|B)$. In this chapter we present a method to test these conditions explicitly. Our approach is based on the introduction of a polynomial ideal that enables us to test the right-invertibility of the matrix $(zI - A|B)$ over various rings simultaneously. In this way, reachability and stabilizability can be treated at the same time.

The idea behind our approach is as follows. First we introduce some ideals in the ring $\mathcal{R}[z]$ related to the the matrix $(zI - A|B)$. Next, the conditions on the right-invertibility of the matrix $(zI - A|B)$ are restated as relatively simple conditions on these polynomial ideals. The main step is the explicit determination of these ideals using the Gröbner basis method introduced in Chapter 4. With this tool from constructive commutative algebra, we obtain explicit algorithms to test the reachability of systems over polynomial rings. For stabilizability, we mainly confine ourselves to time-delay systems with point delays.

The results of the first part of this chapter (Sections 5.1 and 5.2) were already mentioned in [40], but here we give a more detailed elaboration and also include the proofs. Section 5.3 is mainly based on [39].

5.1 Right-invertibility and polynomial ideals

Let $\mathcal{R}$ be an integral domain, and consider a system $\Sigma = (A, B, C, D)$ over the ring $\mathcal{R}$. According to Definition 2.2.2 and Theorem 2.2.3, the reachability of this system only depends on the matrix pair (A, B): the matrix $(zI - A|B)$ has to be right-invertible over $\mathcal{R}[z]$. When a stability defining Hurwitz set $\mathcal{D}$ in $\mathcal{R}[z]$ is fixed, the same observation can be made for stabilizability by dynamic state feedback and for detectability. For stabilizability by dynamic state feedback it is required that the matrix $(zI - A|B)$ is right-invertible over the ring $\mathcal{R}_{\mathcal{D}}(z)$ of all stable transfer functions (see Theorem 2.8.2). According to Theorem 2.9.3, the condition

for detectability is completely dual: the matrix $\begin{pmatrix} zI - A \\ C \end{pmatrix}$ has to be left-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. Finally we recall that also the separation principle carries over to the case of systems over rings: if a system is detectable and stabilizable by dynamic state feedback, it is also stabilizable by dynamic output feedback.

From this summary of some of the main results of Chapter 2 we conclude that all these results yield conditions on the matrix pairs (A, B) and (C, A). Moreover, the conditions on (C, A) are dual to those on (A, B). So, to verify these conditions explicitly, one only needs a method for the determination of the right-invertibility of the matrix $(zI - A|B)$ over the polynomial ring $\mathcal{R}[z]$, and the ring of stable transfer functions $\mathcal{R}_{\mathcal{D}}(z)$, respectively. Since the matrices C and D of the system $\Sigma = (A, B, C, D)$ are not involved in this, we omit them throughout this chapter. Instead we use the convention that a system over a ring $\mathcal{R}$ is given by a matrix pair $\Sigma = (A, B)$ with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$.

We start with the introduction of an ideal that plays a key role in this chapter. It enables us to give a very straightforward characterization of the reachability and stabilizability of a system.

Definition 5.1.1 Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Then the ideal $\mathcal{I}$ in $\mathcal{R}[z]$ associated with Σ, is defined as

$$\mathcal{I} = \{\varphi(z) \in \mathcal{R}[z] \mid \exists M(z) \in \mathcal{R}[z]^{(n+m)\times n} \text{ s.t. } (zI - A|B) \cdot M(z) = \varphi(z) \cdot I\}. \quad (5.1)$$

Since the set $\mathcal{I}$ is closed under addition, and also under multiplication by arbitrary elements from $\mathcal{R}[z]$, $\mathcal{I}$ is indeed an ideal in $\mathcal{R}[z]$. Recalling the proof of Theorem 2.8.2 we conclude that if $\varphi(z)$ is a monic element of $\mathcal{I}$, then there exists a dynamic state feedback compensator for the system $\Sigma = (A, B)$ such that the characteristic polynomial $\det(zI - \hat{A})$ of the closed-loop system is equal to $(\varphi(z))^n$. So, to some extent the monic elements of $\mathcal{I}$ characterize the characteristic polynomials of all closed-loop systems that are obtainable by dynamic state feedback.

The next result explains our interest in the ideal $\mathcal{I}$.

Proposition 5.1.2 *Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, and let $\mathcal{I}$ be the ideal in $\mathcal{R}[z]$ associated with Σ as defined in (5.1). Let $\mathcal{D}$ be a Hurwitz set in $\mathcal{R}[z]$. Then*

(i) $\Sigma = (A, B)$ is reachable $\iff$ $\mathcal{I} = \mathcal{R}[z]$,

(ii) $\Sigma = (A, B)$ is stabilizable w.r.t. $\mathcal{D}$ $\iff$ $\mathcal{I} \cap \mathcal{D} \neq \varnothing$.

Proof

(i) Assume that $\Sigma = (A, B)$ is reachable. Then the matrix $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$, hence $1 \in \mathcal{I}$. On the other hand, if $\mathcal{I} = \mathcal{R}[z]$ then $1 \in \mathcal{I}$, and thus $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$.

(ii) Assume that $\Sigma = (A, B)$ is stabilizable w.r.t. the Hurwitz set $\mathcal{D}$, so $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. So there exists a matrix $M(z)$ over $\mathcal{R}_{\mathcal{D}}(z)$ such that

$$(zI - A|B) \cdot M(z) = I.$$

Multiplying this equality by the least common multiple $\varphi(z)$ of all denominators of the entries of $M(z)$, we obtain

$$(zI - A|B) \cdot \hat{M}(z) = \varphi(z) \cdot I,$$

where $\varphi(z) \in \mathcal{D}$, and $\hat{M}(z)$ a matrix over the polynomial ring $\mathcal{R}[z]$. Hence $\varphi(z) \in \mathcal{I}$, and thus $\mathcal{I} \cap \mathcal{D} \neq \emptyset$.

To prove the implication in the opposite direction, let $\varphi(z) \in \mathcal{I} \cap \mathcal{D}$. Since $\varphi(z) \in \mathcal{I}$, there exists a matrix $M(z)$ over $\mathcal{R}[z]$ such that

$$(zI - A|B) \cdot M(z) = \varphi(z) \cdot I.$$

Dividing both left- and right-hand side by $\varphi(z)$, we conclude that $\frac{1}{\varphi(z)} \cdot M(z)$ is a right-inverse of $(zI - A|B)$ over $\mathcal{R}_{\mathcal{D}}(z)$. ∎

When the ideal $\mathcal{I}$ associated to the system Σ is known, conclusions on the reachability and stabilizability of Σ are more easily drawn. So, to test the reachability and stabilizability of a system Σ, we first want to determine the corresponding ideal $\mathcal{I}$ explicitly. If $\mathcal{R}$ is a polynomial ring, this can be done using the Gröbner basis techniques of Section 4.1. Unfortunately, the definition of the ideal $\mathcal{I}$ in (5.1) is not very suitable for computation. The Gröbner basis algorithm can only be applied when a set of generators of a polynomial ideal is given. However, the ideal $\mathcal{I}$ is described in a completely different way. Therefore it is necessary to make a small detour. We first introduce some ideals that are closely related to $\mathcal{I}$, or are even equal to $\mathcal{I}$. They have the advantage that they can be determined exactly using the Gröbner basis method. In this way we obtain enough information to decide on the reachability and stabilizability of the system under consideration.

In this section we confine ourselves to the definition of the ideals used throughout this chapter, and to the relations between them. This can be done for systems over arbitrary integral domains $\mathcal{R}$. In the next section we specialize to the case that $\mathcal{R}$ is a polynomial ring, and show how these ideals can be computed using the Gröbner basis algorithm.

Definition 5.1.3 Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Let e_i $(i = 1, \ldots, n)$ denote the i^{th} unit vector in $\mathcal{R}^n$, and define for $i \in \{1, \ldots, n\}$:

$$\mathcal{H}_i := \{\varphi(z) \in \mathcal{R}[z] \mid \exists \psi(z) \in \mathcal{R}[z]^{n+m} \text{ s.t. } (zI - A|B) \cdot \psi(z) = \varphi(z) \cdot e_i\}. \quad (5.2)$$

Then the ideal $\mathcal{H}$ associated with Σ is defined as

$$\mathcal{H} := \bigcap_{i=1}^{n} \mathcal{H}_i. \quad (5.3)$$

In the same way as for the ideal $\mathcal{I}$, we may prove that all $\mathcal{H}_i$ $(i = 1, \ldots, n)$ are ideals in $\mathcal{R}[z]$. They can be considered as a column-wise definition of the ideal $\mathcal{I}$. Therefore it is easily verified that

$$\mathcal{H} = \mathcal{I}. \quad (5.4)$$

In Section 2.8 we already encountered an ideal $\mathcal{J}$ in $\mathcal{R}[z]$ that was used for the same purpose as the ideal $\mathcal{I}$ of Definition 5.1.1: the reformulation of the condition for stabilizability of a system, as a condition on the intersection of the ideal $\mathcal{J}$ and the Hurwitz set $\mathcal{D}$. Here we repeat the definition of this ideal.

Definition 5.1.4 Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Define $N := \binom{n+m}{n} - 1$ and denote all $n \times n$ minors of the matrix $(zI - A|B)$ by $\alpha_0(z), \ldots, \alpha_N(z)$. Then the ideal $\mathcal{J}$ associated with Σ is the ideal in $\mathcal{R}[z]$ generated by all these $n \times n$ minors of $(zI - A|B)$:

$$\mathcal{J} := \langle \alpha_0(z), \ldots, \alpha_N(z) \rangle. \tag{5.5}$$

The next result describes the relationship between the ideals $\mathcal{I}$ and $\mathcal{J}$.

Lemma 5.1.5 *Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Define the ideals $\mathcal{I}$ and $\mathcal{J}$ associated with Σ as in (5.1) and (5.5), respectively. Then*

$$\mathcal{J} \subset \mathcal{I} \subset \sqrt{\mathcal{J}}. \tag{5.6}$$

Proof

The proof of this result is based on the same ideas as the proof of Proposition 2.8.5.

"$\mathcal{J} \subset \mathcal{I}$" Let $\alpha(z)$ be one of the $n \times n$ minors of $(zI - A|B)$. Then there exists an $n \times n$ submatrix $K(z)$ of $(zI - A|B)$ such that $\alpha(z) = \det(K(z))$, and according to Cramer's rule we have

$$K(z) \cdot \operatorname{adj}(K(z)) = \det(K(z)) \cdot I = \alpha(z) \cdot I.$$

Extending the matrix $\operatorname{adj}(K(z))$ with zero rows on the right places, we obtain an $(n+m) \times n$ matrix $\tilde{K}(z)$ over $\mathcal{R}[z]$ such that

$$(zI - A|B) \cdot \tilde{K}(z) = \alpha(z) \cdot I.$$

Hence $\alpha(z) \in \mathcal{I}$. Since $\alpha(z)$ was an arbitrary $n \times n$ minor of $(zI - A|B)$, it follows that all principal minors of $(zI - A|B)$ belong to $\mathcal{I}$, so $\mathcal{J} \subset \mathcal{I}$.

"$\mathcal{I} \subset \sqrt{\mathcal{J}}$" Let $\varphi(z) \in \mathcal{I}$. Then there exists a matrix $M(z) \in \mathcal{R}[z]^{(n+m)\times n}$ such that

$$(zI - A|B) \cdot M(z) = \varphi(z) \cdot I. \tag{5.7}$$

Let $\alpha_0(z), \ldots, \alpha_N(z)$ denote all $n \times n$ minors of the matrix $(zI - A|B)$. Taking determinants on both right- and left-hand side of (5.7), and using the Binet-Cauchy formula (see e.g. [29, p. 9] or formula (2.41)), we find polynomials $\beta_0(z), \ldots, \beta_N(z) \in \mathcal{R}[z]$ (the $n \times n$ minors of the matrix $M(z)$) such that

$$\sum_{i=0}^{N} \alpha_i(z)\beta_i(z) = (\varphi(z))^n.$$

We conclude that $(\varphi(z))^n \in \mathcal{J}$, and thus by definition $\varphi(z) \in \sqrt{\mathcal{J}}$. ∎

It is important to note that the inclusions in the opposite directions do not hold in general. This is shown by the following two counterexamples.

Example 5.1.6 Let $\mathcal{R} = \mathbb{R}$, and consider the system $\Sigma = (A, B)$ with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then $(z-1) \in \mathcal{I}$ because

$$(zI - A|B) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} z-1 & 0 & 0 \\ 0 & z-1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = (z-1) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

However, the only nonzero 2×2 minor of $(zI - A|B)$ is $(z-1)^2$. Hence $(z-1) \notin \mathcal{J}$, and we conclude that in this case $\mathcal{I} \not\subset \mathcal{J}$.

Example 5.1.7 Let $\mathcal{R} = \mathbb{R}$, and consider the system $\Sigma = (A, B)$ with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The only nonzero 2×2 minor of $(zI - A|B)$ is z^2. Hence $\sqrt{\mathcal{J}} = \langle z \rangle$. Next we compute $\mathcal{I}$. Let $\varphi(z) \in \mathcal{I}$. Then there exist polynomials $m_{ij}(z) \in \mathbb{R}[z]$ $(i = 1, 2;\ j = 1, 2, 3)$ such that

$$(zI - A|B) \begin{pmatrix} m_{11}(z) & m_{12}(z) \\ m_{21}(z) & m_{22}(z) \\ m_{31}(z) & m_{32}(z) \end{pmatrix} = \begin{pmatrix} z & -1 & 0 \\ 0 & z & 0 \end{pmatrix} \begin{pmatrix} m_{11}(z) & m_{12}(z) \\ m_{21}(z) & m_{22}(z) \\ m_{31}(z) & m_{32}(z) \end{pmatrix} = \begin{pmatrix} \varphi(z) & 0 \\ 0 & \varphi(z) \end{pmatrix}.$$

In this way we obtain the equations

$$\begin{cases} z \cdot m_{11}(z) - m_{21}(z) &= \varphi(z), \\ z \cdot m_{12}(z) - m_{22}(z) &= 0, \\ z \cdot m_{21}(z) &= 0, \\ z \cdot m_{22}(z) &= \varphi(z). \end{cases}$$

So $m_{21}(z) = 0$, and $\varphi(z) = z \cdot m_{11}(z) = z \cdot m_{22}(z) = z^2 \cdot m_{12}(z)$. We conclude that all polynomials in $\mathcal{I}$ are of the form $z^2 \cdot m_{12}(z)$, with $m_{12}(z)$ an arbitrary polynomial in $\mathbb{R}[z]$, and thus $\mathcal{I} = \langle z^2 \rangle$. So in this particular case we have $\sqrt{\mathcal{J}} = \langle z \rangle \not\subset \langle z^2 \rangle = \mathcal{I}$.

Finally we introduce another ideal related to $\mathcal{I}$. At the moment, this relationship is not very clear, but in the next sections the computation of this ideal turns out to be very similar to the computation of $\mathcal{H}$.

Definition 5.1.8 Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Let $i \in \{1, \ldots, n\}$, and introduce $n - i$ new *indeterminates* $q_{i+1}, \ldots, q_n$. Define the polynomials $p_{i_1}, \ldots, p_{i_{n+m}}$ in $\mathcal{R}[z, q_{i+1}, \ldots, q_n]$ as the components of the following $(n+m)$-dimensional row vector:

$$(p_{i_1} \cdots p_{i_{n+m}}) := (\underbrace{0 \cdots 0}_{i-1} | 1 | q_{i+1} \cdots q_n) \cdot (zI - A \mid B). \tag{5.8}$$

The ideal $\mathcal{L}_i$ in $\mathcal{R}[z]$ is defined as the intersection of $\mathcal{R}[z]$ and the ideal in the extended polynomial ring $\mathcal{R}[z, q_{i+1}, \ldots, q_n]$ generated by the polynomials $p_{i_1}, \ldots, p_{i_{n+m}}$:

$$\mathcal{L}_i := \langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle \cap \mathcal{R}[z]. \tag{5.9}$$

The intersection of all ideals $\mathcal{L}_i$ $(i = 1, \ldots, n)$ is called $\mathcal{L}$:

$$\mathcal{L} := \bigcap_{i=1}^{n} \mathcal{L}_i. \tag{5.10}$$

The connection between the ideal $\mathcal{L}$ and the ideal $\mathcal{I}$ is elaborated in the next lemma.

Lemma 5.1.9 *Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$. Define the ideals $\mathcal{I}$, $\mathcal{L}_i$ $(i = 1, \ldots, n)$ and $\mathcal{L}$ associated with Σ, as in Definition 5.1.1 and 5.1.8, respectively. Then*

$$\forall i \in \{1, \ldots, n\} : \mathcal{I} \subset \mathcal{L}_i. \tag{5.11}$$

Consequently,

$$\mathcal{I} \subset \mathcal{L}. \tag{5.12}$$

Proof

Let $i \in \{1, \ldots, n\}$ and let $\varphi(z) \in \mathcal{I}$. Then there exists a matrix $M(z) \in \mathcal{R}[z]^{(n+m)\times n}$ such that

$$(zI - A|B) \cdot M(z) = \varphi(z) \cdot I.$$

Pre-multiply this equation by the n-dimensional row vector $(\underbrace{0 \cdots 0}_{i-1}|1|q_{i+1} \cdots q_n)$ consisting of $(i-1)$ zeros, one 1, and $(n-i)$ indeterminates $q_{i+1}, \ldots, q_n$. In this way we obtain

$$(p_{i_1} \cdots p_{i_{n+m}}) \cdot M(z) = (\underbrace{0 \cdots 0}_{i-1}|\varphi(z)|q_{i+1}\varphi(z) \cdots q_n\varphi(z)), \tag{5.13}$$

where $p_{i_1}, \ldots, p_{i_{n+m}}$ are defined as in (5.8). Let $(\mu_{1,i}(z) \cdots \mu_{n+m,i}(z))^T$ denote the i^{th} column of the matrix $M(z)$. Then the i^{th} component of (5.13) indicates that

$$\varphi(z) = \sum_{j=1}^{n+m} p_{i_j} \mu_{j,i}(z)$$

is an element of the ideal $\langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle$ in the polynomial ring $\mathcal{R}[z, q_{i+1}, \ldots, q_n]$. By assumption $\varphi(z) \in \mathcal{R}[z]$ and we conclude that

$$\varphi(z) \in \langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle \cap \mathcal{R}[z] = \mathcal{L}_i.$$

Since $\varphi(z) \in \mathcal{I}$ was arbitrary, this proves the claim. ∎

Again, the inclusion in the opposite direction does not hold in general. This is illustrated in the following example.

Example 5.1.10 Let $\mathcal{R} = \mathbf{R}$, and consider the same system $\Sigma = (A, B)$ as in Example 5.1.7, i.e.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Then we know that $\mathcal{I} = \langle z^2 \rangle$. We now compute $\mathcal{L}_1$ and $\mathcal{L}_2$.

Let q_2 be an indeterminate. Then

$$(1 \mid q_2) \cdot (zI - A|B) = (1 \mid q_2) \begin{pmatrix} z & -1 & 0 \\ 0 & z & 0 \end{pmatrix} = (z \mid q_2 z - 1 \mid 0),$$

and the ideal $\langle z, q_2 z - 1 \rangle$ generated by the elements of this vector is the complete ring $\mathbf{R}[z, q_2]$, because $1 = -(q_2 z - 1) + q_2 \cdot z \in \langle z, q_2 z - 1 \rangle$. So $\mathcal{L}_1 = \langle z, q_2 z - 1 \rangle \cap \mathbf{R}[z] = \mathbf{R}[z]$.

For the ideal $\mathcal{L}_2$ we compute

$$(0 \mid 1) \cdot (zI - A|B) = (0 \mid 1) \begin{pmatrix} z & -1 & 0 \\ 0 & z & 0 \end{pmatrix} = (0 \mid z \mid 0).$$

Hence $\mathcal{L}_2 = \langle z \rangle$ and we conclude that

$$\mathcal{L} = \mathcal{L}_1 \cap \mathcal{L}_2 = \langle z \rangle.$$

Therefore $\mathcal{I} = \langle z^2 \rangle \subset \langle z \rangle = \mathcal{L}$, but $\mathcal{L} \not\subset \mathcal{I}$.

The next proposition summarizes most of the results we derived on the relationships between the ideals introduced in this section.

Proposition 5.1.11 *Let $\Sigma = (A, B)$ be a system over an integral domain $\mathcal{R}$, and let $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ and $\mathcal{L}$ be the ideals associated with Σ, introduced in Definitions 5.1.1, 5.1.3, 5.1.4 and 5.1.8, respectively. Then*

$$\mathcal{J} \subset \mathcal{I} = \mathcal{H} \subset \mathcal{L}. \tag{5.14}$$

■

If we specialize somewhat further, and assume that the ring $\mathcal{R}$ under consideration is a polynomial ring over a field $\mathcal{K}$, we can also determine the connections between the varieties of the ideals $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ and $\mathcal{L}$. In fact, all these varieties turn out to be equal.

Proposition 5.1.12 *Let $\mathcal{K}$ be an arbitrary field, and let $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$ be the ring of all polynomials in the indeterminates $s_1, \ldots, s_k$ with coefficients in $\mathcal{K}$. Let $\Sigma = (A, B)$ be a system over $\mathcal{R}$, and consider the ideals $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ and $\mathcal{L}$ associated with Σ, introduced in Definitions 5.1.1, 5.1.3, 5.1.4 and 5.1.8, respectively. Then*

$$\mathcal{V}(\mathcal{J}) = \mathcal{V}(\mathcal{I}) = \mathcal{V}(\mathcal{H}) = \mathcal{V}(\mathcal{L}). \tag{5.15}$$

Proof

From the inclusions (5.14) and formula (A.2) in Appendix A.2, it follows immediately that

$$\mathcal{V}(\mathcal{L}) \subset \mathcal{V}(\mathcal{H}) = \mathcal{V}(\mathcal{I}) \subset \mathcal{V}(\mathcal{J}).$$

Moreover, according to (5.6), we have $\mathcal{I} \subset \sqrt{\mathcal{J}}$, and thus $\mathcal{V}(\sqrt{\mathcal{J}}) \subset \mathcal{V}(\mathcal{I})$. From formula (A.7) we recall that an ideal and its radical have the same variety, and thus we conclude that $\mathcal{V}(\mathcal{J}) = \mathcal{V}(\sqrt{\mathcal{J}}) \subset \mathcal{V}(\mathcal{I})$. This proves the equality $\mathcal{V}(\mathcal{I}) = \mathcal{V}(\mathcal{J})$. So it remains to be shown that

$$\mathcal{V}(\mathcal{I}) \subset \mathcal{V}(\mathcal{L}) = \bigcup_{i=1}^{n} \mathcal{V}(\mathcal{L}_i).$$

Assume that A and B are $n \times n$ and $n \times m$ matrices over $\mathcal{R}$, respectively. Denote all minors of the matrix $(zI - A|B)$ by $\alpha_0(z), \ldots, \alpha_N(z)$. By definition, all these minors are elements of the polynomial ring $\mathcal{R}[z] = \mathcal{K}[z, s_1, \ldots, s_k]$.

Let $(\bar{z}, \bar{s}_1, \ldots, \bar{s}_k) \in \mathcal{V}(\mathcal{I})$, and denote the k-tuple $(\bar{s}_1, \ldots, \bar{s}_k)$ of elements in the algebraic closure $\bar{\mathcal{K}}$ of $\mathcal{K}$ by $\bar{s}$. Then $(\bar{z}, \bar{s}) \in \mathcal{V}(\mathcal{J})$ and thus $(\bar{z}, \bar{s})$ is a common zero of all $n \times n$ minors of the matrix $(zI - A|B)$. This implies that

$$\operatorname{rank}(\bar{z}I - A(\bar{s})|B(\bar{s})) < n,$$

where $A(\bar{s})$ and $B(\bar{s})$ denote the matrices over $\bar{\mathcal{K}}$ that are obtained after substitution of $\bar{s}_j$ for the indeterminate s_j $(j = 1, \ldots, n)$ in the matrices A and B. We conclude that there exists an $i \in \{1, \ldots, n\}$ and a (normalized) vector in $\bar{\mathcal{K}}^n$ of the form

$$(\underbrace{0 \cdots 0}_{i-1}|1|\bar{q}_{i+1} \cdots \bar{q}_n)$$

such that

$$(\underbrace{0 \cdots 0}_{i-1}|1|\bar{q}_{i+1} \cdots \bar{q}_n) \cdot (\bar{z}I - A(\bar{s})|B(\bar{s})) = 0.$$

Therefore, the point $(\bar{z}, \bar{s}, \bar{q}_{i+1}, \ldots, \bar{q}_n) \in \bar{\mathcal{K}}^{1+k+n-i}$ is a common zero of the polynomials $p_{i_1}, \ldots, p_{i_{n+m}}$ in $\mathcal{R}[z, q_{i+1}, \ldots, q_n]$ as defined in (5.8).

Let now $p \in \mathcal{L}_i$. Then $p \in \langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle \cap \mathcal{R}[z]$, and there exist polynomials $\beta_j \in \mathcal{R}[z, q_{i+1}, \ldots, q_n]$ $(j = 1, \ldots, n+m)$ such that

$$p(z, s_1, \ldots, s_k) = \sum_{j=1}^{n+m} \beta_j(z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n) \cdot p_{i_j}(z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n).$$

Substitution of the point $(\bar{z}, \bar{s}_1, \ldots, \bar{s}_k, \bar{q}_{i+1}, \ldots, \bar{q}_n)$ on both right- and left-hand side yields $p(\bar{z}, \bar{s}_1, \ldots, \bar{s}_k) = 0$, so $(\bar{z}, \bar{s})$ is a zero of the polynomial p. Since $p \in \mathcal{L}_i$ was arbitrary, $(\bar{z}, \bar{s}) \in \mathcal{V}(\mathcal{L}_i)$ and we conclude that

$$\mathcal{V}(\mathcal{I}) \subset \mathcal{V}(\mathcal{L}_i) \subset \bigcup_{i=1}^{n} \mathcal{V}(\mathcal{L}_i) = \mathcal{V}(\mathcal{L}).$$

This completes the proof. ∎

Proposition 5.1.12 indicates that for a system over a polynomial ring $\mathcal{K}[s_1, \ldots, s_k]$ it does not matter which of the ideals $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ or $\mathcal{L}$ is used to determine the variety $\mathcal{V}(\mathcal{I})$. This is important for some of the later applications that only require the computation of the variety $\mathcal{V}(\mathcal{I})$. It gives us some flexibility: we are free to choose the method that is computationally the most efficient. However, for the computation of the ideals themselves, our freedom is somewhat more restricted because in general there are some differences between these ideals. Nevertheless, the ideals $\mathcal{J}$ and $\mathcal{L}$ contain important information on the ideal $\mathcal{I}$.

5.2 Gröbner basis computations

Let $\mathcal{K}$ be an arbitrary field, and $s_1, \ldots, s_k$ be a k-tuple of indeterminates. In the rest of this chapter we only consider systems over polynomial rings $\mathcal{R}$ of the form $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$. For this type of rings it is possible to determine the ideals $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ and $\mathcal{L}$ associated with a system $\Sigma = (A, B)$ explicitly using the Gröbner basis techniques of Section 4.1. In this section it is shown how Gröbner bases of these ideals can be computed.

A method for the determination of a Gröbner basis of the ideal $\mathcal{J}$ associated with a system $\Sigma = (A, B)$ is not difficult to obtain. Since $\mathcal{J}$ is defined as the ideal generated by the $n \times n$ minors of the matrix $(zI - A|B)$, straightforward application of Buchberger's algorithm (see Theorem 4.1.26) yields the desired result.

Algorithm 5.2.1 Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and let $>$ be a term ordering on the monomials of $\mathcal{K}[z, s_1, \ldots, s_k]$.

Step 1 Determine the finite subset F of $\mathcal{K}[z, s_1, \ldots, s_k]$ consisting of all $n \times n$ minors of the matrix $(zI - A|B)$,

Step 2 Compute a Gröbner basis G of the ideal generated by the polynomials in F w.r.t the term ordering $>$, using the Gröbner basis algorithm of Theorem 4.1.26.

Proposition 5.2.2 *Consider a system $\Sigma = (A, B)$ over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, and let $\mathcal{J}$ be the ideal associated with Σ, as defined in Definition 5.1.4. Let $>$ be a term ordering on $\mathcal{K}[z, s_1, \ldots, s_k]$. Then the set G obtained in Algorithm 5.2.1 is a Gröbner basis of $\mathcal{J}$ w.r.t. the term ordering $>$.* ∎

The determination of a Gröbner basis for the ideals $\mathcal{L}_i$ $(i = 1, \ldots, n)$ and their intersection $\mathcal{L}$ is somewhat more delicate. During the computation we first have to introduce some new indeterminates that are eliminated later on. As we have seen in Theorem 4.1.31, Gröbner bases w.r.t. a pure lexicographic term ordering are very useful for this purpose.

Algorithm 5.2.3 Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and let π be a ranking of the indeterminates $z, s_1, \ldots, s_k$. Let $i \in \{1, \ldots, n\}$ be given.

Step 1 Introduce $n-i$ new indeterminates $q_{i+1}, \ldots, q_n$ and construct the n-dimensional row vector

$$(\underbrace{0 \cdots 0}_{i-1} | 1 | q_{i+1} \cdots q_n).$$

Step 2 Compute

$$(p_{i_1} \cdots p_{i_{n+m}}) := (\underbrace{0 \cdots 0}_{i-1} | 1 | q_{i+1} \cdots q_n) \cdot (zI - A|B),$$

and consider $p_{i_1}, \ldots, p_{i_{n+m}}$ as elements of the polynomial ring

$$\mathcal{K}[z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n].$$

Step 3 Fix a ranking π_e on the indeterminates $z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n$ in such a way that

(i) the new indeterminates $q_{i+1}, \ldots, q_n$ are of higher rank than the original indeterminates $z, s_1, \ldots, s_k$,

(ii) in the extended ranking π_e, restricted to the indeterminates $z, s_1, \ldots, s_k$, the ordering of the indeterminates $z, s_1, \ldots, s_k$ is the same as in the original ranking π.

Step 4 Compute a Gröbner basis G_i of the ideal $\langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle$ in the polynomial ring $\mathcal{K}[z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n]$ w.r.t. the pure lexicographic term ordering with ranking π_e, using Theorem 4.1.26.

Step 5 Determine $\tilde{G}_i := G_i \cap \mathcal{K}[z, s_1, \ldots, s_k]$.

Proposition 5.2.4 *Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and consider the ideals $\mathcal{L}_i$ $(i = 1, \ldots, n)$ associated with Σ, as defined in Definition 5.1.8. Let π be a ranking of the indeterminates $z, s_1, \ldots, s_k$. Then for every $i \in \{1, \ldots, n\}$ the set $\tilde{G}_i$, obtained in Algorithm 5.2.3, is a Gröbner basis of $\mathcal{L}_i$ w.r.t. the pure lexicographic term ordering with ranking π of indeterminates.*

Proof

Let $i \in \{1, \ldots, n\}$. According to Definition 5.1.8, the ideal $\mathcal{L}_i$ is defined as

$$\mathcal{L}_i = \langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle \cap \mathcal{K}[z, s_1, \ldots, s_k].$$

So we want to apply the Gröbner basis method to eliminate the indeterminates $q_{i+1}, \ldots, q_n$. Recalling Theorem 4.1.31, this is possible by computing a Gröbner basis G_i of the ideal $\langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle$ w.r.t. the pure lexicographic term ordering. For this purpose the ranking of the indeterminates $q_{i+1}, \ldots, q_n$ has to be higher than the ranking of the indeterminates $z, s_1, \ldots, s_k$. Since this condition is satisfied in Step 3 *(i)*, and because the mutual ordering of the indeterminates $z, s_1, \ldots, s_k$ is preserved (see condition *(ii)* of Step 3), the result of Theorem 4.1.31 implies that

$\tilde{G}_i = G_i \cap \mathcal{K}[z, s_1, \ldots, s_k]$ is a Gröbner basis of $\mathcal{L}_i$ w.r.t. the pure lexicographic term ordering with ranking π of indeterminates. ∎

To obtain a Gröbner basis of the ideal $\mathcal{L} = \bigcap_{i=1}^{n} \mathcal{L}_i$, we first compute for every $i \in \{1, \ldots, n\}$ a Gröbner basis of $\mathcal{L}_i$ w.r.t. the pure lexicographic term ordering, using for every $i \in \{1, \ldots, n\}$ the same ranking π of indeterminates. Subsequent determination of the intersection is then possible with standard Gröbner basis techniques. Successive application of Lemma 4.1.34 and Remark 4.1.35 on $\mathcal{L}_1, \ldots, \mathcal{L}_n$ yields a Gröbner basis of the ideal $\mathcal{L}$ w.r.t. the pure lexicographic term ordering with ranking π.

The derivation of a method for the computation of a Gröbner basis for the ideals $\mathcal{H}_i$ $(i = 1, \ldots, n)$ and $\mathcal{H}$, introduced in Definition 5.1.3, is much more involved. First we have to characterize the ideals $\mathcal{H}_i$ $(i = 1, \ldots, n)$ in a somewhat different way, that is more suitable for computation. For this purpose we need the following definition.

Definition 5.2.5 Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Introduce an n-dimensional row vector $(q_1 \cdots q_n)$ of new indeterminates and define

$$(p_1 \cdots p_{n+m}) = (q_1 \cdots q_n) \cdot (zI - A|B), \tag{5.16}$$

where the elements $p_1, \ldots, p_{n+m}$ are polynomials in the ring $\mathcal{R}[z, q_1, \ldots, q_n]$. The ideal $\mathcal{P}$ in $\mathcal{R}[z, q_1, \ldots, q_n]$ is defined as,

$$\mathcal{P} := \langle p_1, \ldots, p_{n+m} \rangle, \tag{5.17}$$

and for every $i \in \{1, \ldots, n\}$:

$$\mathcal{P}_i := \mathcal{P} \cap \mathcal{R}[z, q_i]. \tag{5.18}$$

Lemma 5.2.6 *Let $\Sigma = (A, B)$ be a linear system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and consider the ideals $\mathcal{H}_i$ $(i = 1, \ldots, n)$ and $\mathcal{P}_i$ $(i = 1, \ldots, n)$ associated with Σ, introduced in Definition 5.1.3 and Definition 5.2.5, respectively. Then for every $i \in \{1, \ldots, n\}$ we have*

$$\mathcal{H}_i = \{\varphi(z) \in \mathcal{R}[z] \mid q_i \cdot \varphi(z) \in \mathcal{P}_i\}, \tag{5.19}$$

In the proof of Lemma 5.2.6 we use the following result that is stated separately because it is also very useful in subsequent sections.

Lemma 5.2.7 *Let $\tilde{\mathcal{R}}$ be an integral domain, and let M be an $n \times \ell$ matrix over $\tilde{\mathcal{R}}$. Introduce a vector $(q_1 \cdots q_n)$ of indeterminates, and let $\psi(q_1, \ldots, q_n) \in \tilde{\mathcal{R}}[q_1, \ldots, q_n]^{\ell}$ and $\xi \in \tilde{\mathcal{R}}^n$ be two vectors satisfying the equation*

$$(q_1 \cdots q_n) \cdot M \cdot \psi(q_1, \ldots, q_n) = (q_1 \cdots q_n) \cdot \xi. \tag{5.20}$$

Define $\psi_0 \in \tilde{\mathcal{R}}^{\ell}$ as the ℓ-dimensional vector over $\tilde{\mathcal{R}}$ that is obtained after substitution of $q_1 = q_2 = \cdots = q_n = 0$ in $\psi(q_1, \ldots, q_n)$. Then

$$M \cdot \psi_0 = \xi. \tag{5.21}$$

Proof

Let $\psi(q_1,\ldots,q_n) \in \tilde{\mathcal{R}}[q_1,\ldots,q_n]^\ell$ and $\xi \in \tilde{\mathcal{R}}^n$ be vectors such that (5.20) is satisfied. Introduce a new indeterminate λ. Since $(q_1 \cdots q_n)$ is a row vector of indeterminates, (5.20) holds for an arbitrary choice of q_i $(i = 1,\ldots,n)$, so in particular it remains valid when we replace $(q_1 \cdots q_n)$ by $(\lambda q_1 \cdots \lambda q_n)$:

$$(\lambda q_1 \cdots \lambda q_n) \cdot M \cdot \psi(\lambda q_1,\ldots,\lambda q_n) = (\lambda q_1 \cdots \lambda q_n) \cdot \xi.$$

After subtracting $(\lambda q_1 \cdots \lambda q_n) \cdot \xi$ on both left- and right-hand side, and factoring out the common term λ, we obtain

$$\lambda \cdot (q_1 \cdots q_n) \cdot [M\psi(\lambda q_1,\ldots,\lambda q_n) - \xi] = 0. \tag{5.22}$$

Now we regard the left-hand side of (5.22) as a polynomial in the indeterminate λ with coefficients in $\tilde{\mathcal{R}}[q_1,\ldots,q_n]$. It is obvious that the constant term of this polynomial is zero. Next, consider the linear term in λ. The coefficient of this term is obtained by substitution of $\lambda = 0$ in $(q_1 \cdots q_n) \cdot [M\psi(\lambda q_1,\ldots,\lambda q_n) - \xi]$, and is therefore equal to $(q_1 \cdots q_n) \cdot [M\psi_0 - \xi]$. From formula (5.22) it follows that also this coefficient is zero:

$$(q_1 \cdots q_n) \cdot [M\psi_0 - \xi] = 0. \tag{5.23}$$

Since $M\psi_0 - \xi$ is a vector in $\tilde{\mathcal{R}}^n$ and $(q_1 \cdots q_n)$ is a vector of indeterminates, (5.23) implies that $M\psi_0 - \xi = 0$. This completes the proof. ∎

The importance of Lemma 5.2.7 is not difficult to explain. In the sequel the result is applied to the situation $\tilde{\mathcal{R}} = \mathcal{R}[z]$, where $\mathcal{R}$ denotes the polynomial ring $\mathcal{K}[s_1,\ldots,s_k]$. In this case, the lemma states that if $M(z) \in \mathcal{R}[z]^{n\times\ell}$ and $\xi(z) \in \mathcal{R}[z]^n$ are given, it is not really necessary to look for solutions x of the equation

$$(q_1 \cdots q_n) \cdot M(z) \cdot x = (q_1 \cdots q_n) \cdot \xi(z)$$

in the extended module $\mathcal{R}[z,q_1,\ldots,q_n]^\ell$. A solution $x \in \mathcal{R}[z,q_1,\ldots,q_n]^\ell$ exists if and only if there exists a solution in $\mathcal{R}[z]^\ell$. Moreover, a method is given to eliminate the superfluous indeterminates $q_1,\ldots,q_n$ from a solution in the extended module $\mathcal{R}[z,q_1,\ldots,q_n]^\ell$ containing these indeterminates. This idea is also used in the proof of Lemma 5.2.6.

Proof of Lemma 5.2.6

Let $i \in \{1,\ldots,n\}$.

"$\subset$" Let $\varphi(z) \in \mathcal{H}_i$. Then there exist polynomials $\psi_i(z) \in \mathcal{R}[z]$ $(i = 1,\ldots,n+m)$ such that

$$(zI - A|B) \cdot \begin{pmatrix} \psi_1(z) \\ \vdots \\ \psi_{n+m}(z) \end{pmatrix} = \varphi(z) \cdot e_i,$$

where e_i denotes the i^{th} unit vector in $\mathcal{R}^n$. Pre-multiplication of this equality by the row vector $(q_1 \cdots q_n)$ of indeterminates yields

$$(p_1 \cdots p_{n+m}) \cdot \begin{pmatrix} \psi_1(z) \\ \vdots \\ \psi_{n+m}(z) \end{pmatrix} = q_i \cdot \varphi(z),$$

where $p_j \in \mathcal{R}[z, q_1, \ldots, q_n]$ are defined as in (5.16). We conclude that $q_i \cdot \varphi(z) = \sum_{j=1}^{n+m} p_j \cdot \psi_j(z) \in \mathcal{P} \cap \mathcal{R}[z, q_i] = \mathcal{P}_i$.

"$\supset$" Let $\varphi(z) \in \mathcal{R}[z]$ be such that $q_i \cdot \varphi(z) \in \mathcal{P}_i$. Then there exist polynomials $\beta_j(z, q_1, \ldots, q_n) \in \mathcal{R}[z, q_1, \ldots, q_n]$ $(j = 1, \ldots, n+m)$ such that

$$q_i \cdot \varphi(z) = \sum_{j=1}^{n+m} \beta_i(z, q_1, \ldots, q_n) \cdot p_i.$$

Collecting the coefficients $\beta_j(z, q_1, \ldots, q_n)$ $(j = 1, \ldots, n+m)$ in an $n+m$-dimensional vector $(\beta(z, q_1, \ldots, q_n) = (\beta_1(z, q_1, \ldots, q_n) \cdots \beta_{n+m}(z, q_1, \ldots, q_n))^T$ over the polynomial ring $\mathcal{R}[z, q_1, \ldots, q_n]$, and using definition (5.16) of the vector $(p_1 \cdots p_{n+m})$, the previous formula can be restated as

$$(q_1 \cdots q_n) \cdot (zI - A|B) \cdot \beta(z, q_1, \ldots, q_n) = (q_1 \cdots q_n) \cdot (\varphi(z) \cdot e_i).$$

Let $\beta_0(z) \in \mathcal{R}[z]^{n+m}$ denote the vector obtained after substitution of $q_1 = q_2 = \cdots = q_n = 0$ in $\beta(z, q_1, \ldots, q_n)$. Then application of Lemma 5.2.7 with $\tilde{\mathcal{R}} = \mathcal{R}[z] = \mathcal{K}[z, s_1, \ldots, s_k]$ yields

$$(zI - A|B) \cdot \beta_0(z) = \varphi(z) \cdot e_i,$$

and thus, according to Definition 5.1.3, we have $\varphi(z) \in \mathcal{H}_i$. ∎

The characterization of the ideals $\mathcal{H}_i$ $(i = 1, \ldots, n)$ given in Lemma 5.2.6 indicates that for the determination of these ideals, elimination of indeterminates plays a crucial role. However, to compute $\mathcal{H}_i$ for a fixed value of i, the indeterminates $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n$ on the one hand, and the indeterminate q_i on the other hand, have to be treated differently. The indeterminates $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n$ can be eliminated in the same way as in Algorithm 5.2.3, using an appropriate ranking of the indeterminates and applying the Gröbner basis algorithm w.r.t. the pure lexicographic term ordering. Elimination of the indeterminate q_i is more delicate because only the coefficients of the polynomials in $\mathcal{R}[z, q_i]$ that are linear in q_i belong to $\mathcal{H}_i$. The next algorithm describes how this computation may be carried out.

Algorithm 5.2.8 Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$ with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and let π be a ranking of the indeterminates $z, s_1, \ldots, s_k$. Let $i \in \{1, \ldots, n\}$ be given.

Step 1 Introduce n new indeterminates $q_1, \ldots, q_n$ and construct the n-dimensional row vector $(q_1 \cdots q_n)$.

Step 2 Compute

$$(p_1 \cdots p_{n+m}) := (q_1 \cdots q_n) \cdot (zI - A|B),$$

and consider $p_1, \ldots, p_{n+m}$ as elements of the polynomial ring

$$\mathcal{K}[z, s_1, \ldots, s_k, q_1, \ldots, q_n].$$

Step 3 Choose a ranking π_e of the indeterminates $z, s_1, \ldots, s_k, q_1, \ldots, q_n$ satisfying the following conditions:

(i) the rank of the indeterminates $q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n$ is higher than the rank of the indeterminate q_i,

(ii) The rank of q_i is higher than the rank of the indeterminates $z, s_1, \ldots, s_k$,

(iii) The ordering of the indeterminates $z, s_1, \ldots, s_k$ in the extended ranking π_e, is the same as in the original ranking π.

Step 4 Determine a Gröbner basis G_i of the ideal $\langle p_1, \ldots, p_{n+m} \rangle$ in the polynomial ring $\mathcal{K}[z, s_1, \ldots, s_k, q_1, \ldots, q_n]$ w.r.t. the pure lexicographic term ordering with ranking π_e, using Theorem 4.1.26.

Step 5 Compute $\tilde{G}_i := G_i \cap \mathcal{K}[z, s_1, \ldots, s_k, q_i]$.

Step 6 Determine the set

$$H_i := \{\varphi(z) \in \mathcal{K}[z, s_1, \ldots, s_k] \mid q_i \cdot \varphi(z) \in \tilde{G}_i\}.$$

Proposition 5.2.9 *Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$ with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$, and let π be a ranking of the indeterminates $z, s_1, \ldots, s_k$. Then for every $i \in \{1, \ldots, n\}$, Algorithm 5.2.8 yields a finite subset H_i of $\mathcal{R}[z]$ with the property*

$$\langle H_i \rangle_{\mathcal{R}[z]} = \mathcal{H}_i, \tag{5.24}$$

i.e. the ideal in $\mathcal{R}[z]$ generated by the polynomials in H_i is equal to the ideal $\mathcal{H}_i$ associated with the system Σ, introduced in Definition 5.1.3. Moreover, H_i is a Gröbner basis of $\mathcal{H}_i$ w.r.t. the pure lexicographic term ordering with ranking π.

Proof

Let $i \in \{1, \ldots, n\}$.

"$\langle H_i \rangle \subset \mathcal{H}_i$" Let $\varphi(z) \in H_i$. Then $q_i \cdot \varphi(z)$ is an element of the Gröbner basis of the ideal $\mathcal{P} = \langle p_1, \ldots, p_{n+m} \rangle$ as defined in Definition 5.2.5. Since the polynomial $\varphi(z)$ only contains the indeterminates $z, s_1, \ldots, s_k$, we even have

$$q_i \cdot \varphi(z) \in \mathcal{P} \cap \mathcal{R}[z, q_i] = \mathcal{P}_i.$$

According to Lemma 5.2.6, this implies that $\varphi(z) \in \mathcal{H}_i$. We conclude that $H_i \subset \mathcal{H}_i$, hence also $\langle H_i \rangle \subset \mathcal{H}_i$.

"$\langle H_i \rangle \supset \mathcal{H}_i$" Let $\varphi(z) \in \mathcal{H}_i$. Then it follows from Lemma 5.2.6 that $q_i \cdot \varphi(z) \in \mathcal{P}_i \subset \mathcal{P}$. According to Definition 5.2.5, this ideal $\mathcal{P}$ in $\mathcal{R}[z, q_1, \ldots, q_n]$ is contained in $\langle q_1, \ldots, q_n \rangle$.

In Step 4 of Algorithm 5.2.8 we compute a Gröbner basis G_i of $\mathcal{P}$ w.r.t. the pure lexicographic term ordering with ranking π_e of indeterminates, satisfying conditions *(i)*, *(ii)* and *(iii)* listed in Step 3 of the algorithm. According to Theorem 4.1.31, this implies that $\tilde{G}_i = G_i \cap \mathcal{R}[z, q_i]$ is a Gröbner basis of the ideal $\mathcal{P}_i$ w.r.t. the pure lexicographic term ordering with the indeterminate q_i of higher rank than the indeterminates $z, s_1, \ldots, s_k$.

Since $q_i \cdot \varphi(z) \in \mathcal{P}_i$, and $\tilde{G}_i$ is a Gröbner basis of $\mathcal{P}_i$ w.r.t. this specified lexicographic term ordering, we know from Proposition 4.1.20 that $q_i \cdot \varphi(z)$ is an admissible combination of the polynomials in $\tilde{G}_i$:

$$q_i \cdot \varphi(z) = \sum_{g \in \tilde{G}_i, \alpha_1 \in \mathbf{N}_0, \alpha_2 \in \mathbf{N}_0^{k+1}} c(\alpha_1, \alpha_2, g) q_i^{\alpha_1} x^{\alpha_2} \cdot g, \tag{5.25}$$

where $c(\alpha_1, \alpha_2, g) \in \mathcal{K}$ and x^{α_2} denotes the monomial in $\mathcal{K}[z, s_1, \ldots, s_k]$ with multi-degree α_2. Moreover, if $c(\alpha_1, \alpha_2, g) \neq 0$, then

$$\deg(q_i^{\alpha_1} x^{\alpha_2} \cdot g) \leq \deg(q_i \cdot \varphi(z)). \tag{5.26}$$

Since $\tilde{G}_i \subset \mathcal{P}$, and every polynomial in $\mathcal{P}$ is a linear combination of $q_1, \ldots, q_n$, all elements $g \in \tilde{G}_i$ have the property

$$\deg_{q_i}(g) \geq 1.$$

So, because of (5.26), only polynomials of the form $g = q_i \cdot h$ with $h \in H_i$ occur in the admissible combination (5.25), and also $\alpha_1 = 0$. Therefore (5.25) may be rewritten as

$$q_i \cdot \varphi(z) = \sum_{h \in H_i, \alpha_2 \in \mathbf{N}_0^{k+1}} \hat{c}(\alpha_2, h) q_i x^{\alpha_2} \cdot h. \tag{5.27}$$

If $\hat{c}(\alpha_2, h) \neq 0$, the inequality $\deg(q_i x^{\alpha_2} \cdot h) \leq \deg(q_i \varphi(z))$ is still satisfied, and implies that also $\deg(x^{\alpha_2} h) \leq \deg(\varphi(z))$, because the degree in q_i of both polynomials is equal to 1. Dividing (5.27) by q_i we obtain

$$\varphi(z) = \sum_{h \in H_i, \alpha_2 \in \mathbf{N}_0^{k+1}} \hat{c}(\alpha_2, h) x^{\alpha_2} h,$$

and we conclude that $\varphi(z) \in \langle H_i \rangle_{\mathcal{R}[z]}$. Moreover, each polynomial $\varphi(z) \in \mathcal{H}_i$ is an admissible combination of elements of H_i, and therefore we know from Proposition 4.1.20 that H_i is a Gröbner basis of $\mathcal{H}_i$ w.r.t. the pure lexicographic term ordering with ranking π. ∎

After computation of Gröbner bases of the ideals $\mathcal{H}_i$ $(i = 1, \ldots, n)$, a Gröbner basis of the ideal $\mathcal{H} = \bigcap_{i=1}^{n} \mathcal{H}_i$ may be obtained by repeated application of the Gröbner basis technique for the determination of the intersection of two polynomial ideals as explained in Remark 4.1.35. Since $\mathcal{H} = \mathcal{I}$, this yields a constructive method for the computation of the ideal $\mathcal{I}$ associated with a system $\Sigma = (A, B)$.

Remark 5.2.10 The algorithms developed in this section explain our preference for the Gröbner basis method (in comparison with the characteristic sets method) to a great extent. For the computations of the ideals we are interested in, it is often necessary to introduce new indeterminates that are eliminated later on. For this purpose characteristic sets are of no use; these computations can only be carried out with Gröbner basis techniques. However, in the next sections it turns out that in some applications only the varieties of ideals have to be determined. In these cases, the characteristic sets method may be an alternative for Gröbner basis computations.

We end this section with an example that illustrates the effectiveness of the algorithms developed in this section.

Example 5.2.11 Let $\mathcal{R} = \mathbf{R}[s]$, and consider the system $\Sigma = (A, B)$ with

$$A = \begin{pmatrix} 1 & 5s+2 \\ s^2 - s & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 3s+2 \\ 2s^3 - s - 1 \end{pmatrix}.$$

The ideals $\mathcal{I}$, $\mathcal{J}$, $\mathcal{H}$ and $\mathcal{L}$ associated with Σ are ideals in the polynomial ring $\mathcal{R}[z] = \mathbf{R}[s, z]$. We apply the algorithms derived in this section to compute a Gröbner basis of the ideals $\mathcal{J}$, $\mathcal{H}$ and $\mathcal{L}$ with respect to the pure lexicographic term ordering with the ranking of indeterminates fixed by $s \succ z$. Using the computer algebra package Maple V.2 for the actual Gröbner basis computations, we obtain the following results.

The Gröbner basis of $\mathcal{J}$ is $J = \{j_1, j_2\}$, where

$$\begin{aligned} j_1 &= 1890s + 136z^6 + 236z^5 - 798z^4 - 609z^3 + 2655z^2 + 270z - 1890, \\ j_2 &= 8z^7 + 4z^6 - 42z^5 + 21z^4 + 108z^3 - 99z^2. \end{aligned}$$

The Gröbner basis of $\mathcal{L}_1$ is $L_1 = \{\ell_{11}, \ell_{12}\}$, where

$$\begin{aligned} \ell_{11} &= 90s - 8z^4 - 4z^3 + 18z^2 - 15z + 9, \\ \ell_{12} &= 8z^5 + 4z^4 - 42z^3 + 21z^2 + 108z - 99. \end{aligned}$$

The Gröbner basis of $\mathcal{L}_2$ is $L_2 = \{\ell_{21}, \ell_{22}\}$, where

$$\begin{aligned} \ell_{21} &= s - 1, \\ \ell_{22} &= z. \end{aligned}$$

The Gröbner basis of $\mathcal{H}_1$ is $H_1 = \{h_{11}, h_{12}\}$, where

$$\begin{aligned} h_{11} &= 90s + 8z^5 - 4z^4 - 46z^3 + 39z^2 + 93z - 90, \\ h_{12} &= 8z^6 + 4z^5 - 42z^4 + 21z^3 + 108z^2 - 99z. \end{aligned}$$

The Gröbner basis of $\mathcal{H}_2$ is $H_2 = \{h_{21}, h_{22}\}$, where

$$\begin{aligned} h_{21} &= 1890s + 136z^6 + 236z^5 - 798z^4 - 609z^3 + 2655z^2 + 270z - 1890, \\ h_{22} &= 8z^7 + 4z^6 - 42z^5 + 21z^4 + 108z^3 - 99z^2. \end{aligned}$$

To obtain a Gröbner basis of $\mathcal{L} = \mathcal{L}_1 \cap \mathcal{L}_2$ and of $\mathcal{H} = \mathcal{H}_1 \cap \mathcal{H}_2$, with respect to the same pure lexicographic term ordering, the procedure explained in Remark 4.1.35 is carried out. This yields the following results.

The Gröbner basis of $\mathcal{L}$ is $L = \{\ell_1, \ell_2\}$, where

$$\begin{aligned} \ell_1 &= 90s + 8z^5 - 4z^4 - 46z^3 + 39z^2 + 93z - 90, \\ \ell_2 &= 8z^6 + 4z^5 - 42z^4 + 21z^3 + 108z^2 - 99z. \end{aligned}$$

The Gröbner basis of $\mathcal{H}$ is $H = \{h_1, h_2\}$, where

$$h_1 = 1890s + 136z^6 + 236z^5 - 798z^4 - 609z^3 + 2655z^2 + 270z - 1890,$$

$$h_2 = 8z^7 + 4z^6 - 42z^5 + 21z^4 + 108z^3 - 99z^2.$$

From these results it is obvious that $\mathcal{J} = \mathcal{H}$ because the Gröbner bases J and H of these ideals are completely the same. Moreover, the elements of the Gröbner basis H are admissible combinations of the polynomials in L:

$$h_1 = 21 \cdot \ell_1 + 17 \cdot \ell_2, \tag{5.28}$$

$$h_2 = z \cdot \ell_2. \tag{5.29}$$

Completely in accordance with Proposition 5.1.11 we conclude that $\mathcal{J} = \mathcal{H} \subsetneq \mathcal{L}$.

With this example it is also possible to illustrate that the varieties of the ideals $\mathcal{J}$, $\mathcal{H}$ and $\mathcal{L}$ are the same. It is immediately clear that $\mathcal{V}(\mathcal{L}) \subset \mathcal{V}(\mathcal{H})$ because $\mathcal{H} \subset \mathcal{L}$. Let now $\alpha \in \mathcal{V}(\mathcal{H})$. Then $h_1(\alpha) = h_2(\alpha) = 0$. Since $z = 0$ is a zero of both h_2 and ℓ_2, it follows from (5.29) that h_2 and ℓ_2 have the same set of zeros. So $\ell_2(\alpha) = 0$, and substitution of α in (5.28) yields $\ell_1(\alpha) = 0$. Therefore $\alpha \in \mathcal{V}(\mathcal{L})$, and since the equality $\mathcal{V}(\mathcal{J}) = \mathcal{V}(\mathcal{H})$ is trivial, we have shown that $\mathcal{V}(\mathcal{J}) = \mathcal{V}(\mathcal{H}) = \mathcal{V}(\mathcal{L})$.

5.3 Testing reachability

After the development of algorithms for the computation of the ideals $\mathcal{J}$, $\mathcal{H}$ and $\mathcal{L}$ associated with a system $\Sigma = (A, B)$, explicit methods to test the reachability of a system are not difficult to obtain. According to Proposition 5.1.2 *(i)*, a necessary and sufficient condition for the reachability of a system over an integral domain $\mathcal{R}$ is that the ideal $\mathcal{I}$ associated with the system is equal to the ring $\mathcal{R}[z]$. If we consider a system over the polynomial ring $\mathcal{K}[s_1, \ldots, s_k]$, a Gröbner basis G of the ideal $\mathcal{H} = \mathcal{I}$, may be obtained with Algorithm 5.2.8. Recalling Definition 4.1.17, the ideal $\mathcal{I}$ equals $\mathcal{R}[z]$ if and only if its Gröbner basis G contains a nonzero element of the field $\mathcal{K}$. However, also the other algorithms of Section 5.2 can be used to test the reachability of a system over a polynomial ring because of the equality of the varieties of the ideals $\mathcal{I}$, $\mathcal{H}$, $\mathcal{J}$ and $\mathcal{L}$ that was proved in Proposition 5.1.12.

Lemma 5.3.1 *Consider a linear system $\Sigma = (A, B)$ over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, and let $\mathcal{I}$, $\mathcal{J}$ and $\mathcal{L}$ be the ideals associated with Σ as defined in Section 5.1. Then*

$$\mathcal{I} = \mathcal{R}[z] \iff \mathcal{J} = \mathcal{R}[z] \iff \mathcal{L} = \mathcal{R}[z]. \tag{5.30}$$

Proof

If $\mathcal{I} = \mathcal{R}[z]$, then $\mathcal{V}(\mathcal{I}) = \varnothing$. So, according to Proposition 5.1.12, also $\mathcal{V}(\mathcal{J}) = \varnothing$ and $\mathcal{V}(\mathcal{L}) = \varnothing$. Application of Corollary A.2.10 of the Hilbert Nullstellensatz yields that $\mathcal{J} = \mathcal{R}[z]$ and $\mathcal{L} = \mathcal{R}[z]$. The other implications are proved in a completely analogous way. ∎

Lemma 5.3.1 indicates that necessary and sufficient conditions for the reachability of a system can also be given in terms of the ideals $\mathcal{J}$ and $\mathcal{L}$. According to Definition 4.1.17, the Gröbner bases of these ideals have to contain a nonzero element of the field $\mathcal{K}$.

Remark 5.3.2 With the result of Lemma 5.3.1, it is possible to give an alternative proof of Theorem 2.2.4. This theorem states that the reachability of a system $\Sigma = (A, B)$ over the polynomial ring $\mathcal{K}[s_1, \ldots, s_k]$ is equivalent to the following pointwise rank condition:

$$\forall(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1} : \operatorname{rank}(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k) \mid B(\hat{s}_1 \ldots, \hat{s}_k)) = n, \quad (5.31)$$

where n denotes the size of the matrix A. The proof of the necessity of condition (5.31) remains the same, but using the ideals $\mathcal{I}$ and $\mathcal{J}$, the proof of sufficiency can be facilitated. If (5.31) holds, there is for every point $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ an $n \times n$ minor of the matrix $(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k))$ that is unequal to zero. So the $n \times n$ minors of this polynomial matrix do not have a common zero. Therefore $\mathcal{V}(\mathcal{J}) = \varnothing$, and according to the Hilbert Nullstellensatz this implies that $\mathcal{J} = \mathcal{R}[z]$. Subsequent application of Lemma 5.3.1 and Proposition 5.1.2 *(i)* yields that $\mathcal{I} = \mathcal{R}[z]$ and thus that $\Sigma = (A, B)$ is reachable.

If the ideal $\mathcal{J}$ is used to verify the reachability of a system $\Sigma = (A, B)$, application of Algorithm 5.2.1 suffices to draw a conclusion. The system is reachable if and only if the Gröbner basis of the ideal $\mathcal{J}$ associated with Σ contains a nonzero element of the field $\mathcal{K}$. Note that the outcome of Algorithm 5.2.1 is independent of the conclusion on the reachability of the system under consideration. In any case a Gröbner basis of the ideal $\mathcal{J}$ w.r.t. the chosen term ordering is obtained.

It is obvious that the same observation holds for the reachability tests based on the ideals $\mathcal{L}$ and $\mathcal{H}$. However, if a system is reachable, the computations of the Gröbner bases of $\mathcal{L}$ and $\mathcal{H}$ take a very special form. This additional structure can be used to adjust Algorithms 5.2.3 and 5.2.8 in such a way that the computations for testing the reachability of a system are speeded up considerably. For this increased computational efficiency we have to pay a price: if the system $\Sigma = (A, B)$ is not reachable, the output of the modified algorithms is not a Gröbner basis of the ideals $\mathcal{L}$ or $\mathcal{H}$ any more.

The next proposition indicates that for reachable systems, it is not necessary to carry out Algorithm 5.2.8 for each $i \in \{1, \ldots, n\}$ separately to obtain a Gröbner basis of the ideal $\mathcal{H}$. In this situation one Gröbner basis computation suffices to draw our conclusion.

Proposition 5.3.3 *Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, and let $\mathcal{P}$ be the ideal in $\mathcal{R}[z, q_1, \ldots, q_n]$ associated with Σ, introduced in Definition 5.2.5. Then*

$$(zI - A|B) \text{ is right-invertible over } \mathcal{R}[z],$$

$$\Longleftrightarrow$$

The reduced Gröbner basis of $\mathcal{P}$ is $G = \{q_1, \ldots, q_n\}$, independent of the chosen term ordering.

Proof

"$\Leftarrow$" Suppose that $\{q_1, \ldots, q_n\}$ is the reduced Gröbner basis of $\mathcal{P}$, and let $i \in \{1, \ldots, n\}$. Then it is obvious that $q_i \in \mathcal{P} \cap \mathcal{R}[z, q_i] = \mathcal{P}_i$, and thus, according to

Lemma 5.2.6, $1 \in \mathcal{H}_i$. So for all $i \in \{1, \ldots, n\}$ we have $\mathcal{H}_i = \mathcal{R}[z]$, and therefore also

$$\mathcal{H} = \bigcap_{i=1}^{n} \mathcal{H}_i = \mathcal{R}[z].$$

Since $\mathcal{H} = \mathcal{I}$, this implies that $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$.

"$\Rightarrow$" Suppose that $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$, and let the matrix $M(z) \in \mathcal{R}[z]^{(n+m)\times n}$ be a right-inverse. Recalling Definition 5.2.5, the ideal $\mathcal{P}$ in $\mathcal{R}[z, q_1, \ldots, q_n]$ is generated by the polynomials $p_1, \ldots, p_{n+m}$ determined by

$$(p_1 \cdots p_{n+m}) = (q_1 \cdots q_n) \cdot (zI - A|B). \tag{5.32}$$

So, in particular, each polynomial p_i $(i = 1, \ldots, n+m)$ is an $\mathcal{R}[z]$-linear combination of the polynomials q_j $(j = 1, \ldots, n)$, and we conclude that in the ring $\mathcal{R}[z, q_1, \ldots, q_n]$ we have

$$\langle p_1, \ldots, p_{n+m} \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]} \subset \langle q_1, \ldots, q_n \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]}.$$

Multiplying formula (5.32) from the right by $M(z)$, we obtain

$$(p_1 \cdots p_{n+m}) \cdot M(z) = (q_1 \cdots q_n).$$

Hence, each polynomial q_j $(j = 1, \ldots, n)$ may be written as an $\mathcal{R}[z]$-linear combination of the polynomials $p_1, \ldots, p_{n+m}$. This yields

$$\langle q_1, \ldots, q_n \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]} \subset \langle p_1, \ldots, p_{n+m} \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]}.$$

We conclude that

$$\mathcal{P} = \langle p_1, \ldots, p_{n+m} \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]} = \langle q_1, \ldots, q_n \rangle_{\mathcal{R}[z,q_1,\ldots,q_n]},$$

and thus $G = \{q_1, \ldots, q_n\}$ is a set of monomials generating the ideal $\mathcal{P}$. According to Remark 4.1.19, G is a Gröbner basis of $\mathcal{P}$, independent of the chosen term ordering. Since the polynomials in G also satisfy the conditions *(i)* and *(ii)* in Definition 4.1.28, this Gröbner basis is reduced. ∎

Using Proposition 5.3.3, we obtain the following algorithm for testing the reachability of a system over a polynomial ring.

Algorithm 5.3.4 Let $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$.

Step 1 Introduce n new indeterminates $q_1, \ldots, q_n$ and construct the n-dimensional row vector $(q_1 \cdots q_n)$.

Step 2 Compute

$$(p_1 \cdots p_{n+m}) := (q_1 \cdots q_n) \cdot (zI - A|B).$$

Step 3 Determine the reduced Gröbner basis G of the ideal $\langle p_1, \ldots, p_{n+m} \rangle$ in the polynomial ring $\mathcal{R}[z, q_1, \ldots, q_n]$ w.r.t. an arbitrary term ordering.

Step 4 If $G = \{q_1, \ldots, q_n\}$, then $\Sigma = (A, B)$ is reachable; otherwise $\Sigma = (A, B)$ is not reachable.

Proposition 5.3.3 and Algorithm 5.3.4 are very important from the computational point of view. If we want to test the reachability of the system $\Sigma = (A, B)$ using the method based on the ideal $\mathcal{H}$, Algorithm 5.3.4 shows that only *one* Gröbner basis w.r.t. an *arbitrary* term ordering has to be computed. From Proposition 5.2.9 we know that in general the computation of a Gröbner basis of the ideal $\mathcal{H}$ requires n Gröbner basis computations w.r.t. the pure lexicographic term ordering, plus the determination of the intersection of n different ideals. In Subsection 4.1.5 we already mentioned that the computation of a Gröbner basis w.r.t. the pure lexicographic term ordering is typically more time consuming than the determination of a Gröbner basis w.r.t. the graded lexicographic or graded inverse lexicographic term ordering. Therefore Algorithm 5.3.4 speeds up the computations in two different ways: the number of Gröbner bases that have to be computed is reduced from more than n to only 1, and moreover, it is allowed to determine this Gröbner basis w.r.t. an arbitrary term ordering.

The result of Proposition 5.3.3 also has an intuitive interpretation. From Theorem 2.2.4 we know that the matrix $(zI - A|B)$ is right-invertible over the polynomial ring $\mathcal{R}[z]$, if and only if it has pointwise full rank. So when we substitute an arbitrary point $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ for the indeterminates $z, s_1, \ldots, s_k$, the left kernel of the matrix

$$(\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k)|B(\hat{s}_1, \ldots, \hat{s}_k)) \tag{5.33}$$

is equal to $\{0\}$. To test this condition, we want to obtain all solutions of the equation

$$(q_1 \cdots q_n) \cdot (zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k)) = (0 \cdots 0). \tag{5.34}$$

If for every point $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ the matrix (5.33) has full rank, the only solution to equation (5.34) is $q_1 = q_2 = \cdots = q_n = 0$. This is indicated by the reduced Gröbner basis of the ideal $\mathcal{P}$: it consists of the polynomials $q_1, \ldots, q_n$ only.

On the other hand, if there exists a point $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ in which the rows of the matrix (5.33) are $\bar{\mathcal{K}}$-linearly dependent, there exists a nonzero row vector $(\hat{q}_1 \cdots \hat{q}_n) \in \bar{\mathcal{K}}^n$ such that

$$(\hat{q}_1 \cdots \hat{q}_n) \cdot (\hat{z}I - A(\hat{s}_1, \ldots, \hat{s}_k)|B(\hat{s}_1, \ldots, \hat{s}_k)) = (0 \cdots 0).$$

Since the variety of the ideal $\mathcal{P}$ is completely determined by its Gröbner basis, this indicates that in this situation $\{q_1, \ldots, q_n\}$ cannot be a Gröbner basis of $\mathcal{P}$.

If for a certain point $(\hat{z}, \hat{s}_1, \ldots, \hat{s}_k) \in \bar{\mathcal{K}}^{k+1}$ the left kernel of the matrix (5.33) is unequal to $\{0\}$, the dimension of this left kernel is larger than or equal to 1. This implies that an element $(\hat{q}_1 \cdots \hat{q}_n)$ of the left kernel remains an annihilating row vector for the matrix (5.33) after multiplication by an arbitrary constant in $\bar{\mathcal{K}}$. Therefore it is possible to normalize the row vector $(\hat{q}_1 \cdots \hat{q}_n)$ in such a way that its first nonzero component becomes equal to 1. We conclude that the rank of the matrix (5.33) is strictly smaller than n if and only if there exists an $i \in \{1, \ldots, n\}$ and a row vector in $\bar{\mathcal{K}}^n$ of the form

$$(\underbrace{0 \cdots 0}_{i-1}|1|\tilde{q}_{i+1} \cdots \tilde{q}_n) \tag{5.35}$$

such that

$$(\underbrace{0\cdots 0}_{i-1}|1|\tilde{q}_{i+1}\cdots\tilde{q}_n)\cdot(\hat{z}I-A(\hat{s}_1,\ldots,\hat{s}_k)|B(\hat{s}_1,\ldots,\hat{s}_k))=(0\cdots 0).$$

The reachability test using the ideals $\mathcal{L}_i$ $(i=1,\ldots,n)$ and $\mathcal{L}$ is based on this equivalence. For every $i\in\{1,\ldots,n\}$ an n-dimensional row vector of the form

$$(\underbrace{0\cdots 0}_{i-1}|1|q_{i+1}\cdots q_n)$$

is introduced, in which $q_{i+1},\ldots,q_n$ are considered as indeterminates. Next, the polynomials $p_{i_1},\ldots,p_{i_{n+m}}$ are computed using formula (5.8):

$$(p_{i_1}\cdots p_{i_{n+m}})=(\underbrace{0\cdots 0}_{i-1}|1|q_{i+1}\cdots q_n)\cdot(zI-A(s_1,\ldots,s_k)|B(s_1,\ldots,s_k)).$$

If there exists an $i\in\{1,\ldots,n\}$ for which the ideal $\langle p_{i_1},\ldots,p_{i_{n+m}}\rangle$ has a non-empty variety, the corresponding system $\Sigma=(A,B)$ is not reachable because there exists a point $(\hat{z},\hat{s}_1,\ldots,\hat{s}_k)\in\bar{\mathcal{K}}^{k+1}$ in which the matrix $(\hat{z}I-A(\hat{s}_1,\ldots,\hat{s}_k)|B(\hat{s}_1,\ldots,\hat{s}_k))$ is not of full row rank since its left kernel contains a nonzero element. By definition (5.9), the point $(\hat{z},\hat{s}_1,\ldots,\hat{s}_k)$ is an element of $\mathcal{V}(\mathcal{L}_i)$, and thus $\mathcal{V}(\mathcal{L})=\cup_{i=1}^n\mathcal{V}(\mathcal{L}_i)\neq\varnothing$. On the other hand, if for all $i\in\{1,\ldots,n\}$, the variety of the ideal $\langle p_{i_1},\ldots,p_{i_{n+m}}\rangle$ is empty, we know already that $\mathcal{V}(\mathcal{L})=\cup_{i=1}^n\mathcal{V}(\mathcal{L}_i)=\varnothing$, and may conclude that $\Sigma=(A,B)$ is reachable. This conclusion may also be drawn directly. There does not exist a point $(\hat{z},\hat{s}_1,\ldots,\hat{s}_k)\in\bar{\mathcal{K}}^{k+1}$ for which the left kernel of the matrix $(\hat{z}I-A(\hat{s}_1,\ldots,\hat{s}_k)|B(\hat{s}_1,\ldots,\hat{s}_k))$ contains an element of the form (5.35). Hence, for all $(\hat{z},\hat{s}_1,\ldots,\hat{s}_k)\in\bar{\mathcal{K}}^{k+1}$, the matrix $(\hat{z}I-A(\hat{s}_1,\ldots,\hat{s}_k)|B(\hat{s}_1,\ldots,\hat{s}_k))$ has full row rank.

From the Hilbert Nullstellensatz it follows that for every $i\in\{1,\ldots,n\}$: $\mathcal{V}(\mathcal{L}_i)=\varnothing$ if and only if $\mathcal{L}_i=\mathcal{R}[z]$. This condition is satisfied if the ideal generated by the polynomials $p_{i_1},\ldots,p_{i_{n+m}}$, computed in Step 2 of Algorithm 5.2.3, is equal to the polynomial ring $\mathcal{R}[z,q_{i+1},\ldots,q_n]$. Therefore it is not necessary to eliminate the indeterminates $q_{i+1},\ldots,q_n$ explicitly, using a Gröbner basis w.r.t. the appropriate pure lexicographic term ordering. Instead it suffices to determine a Gröbner basis of the ideal $\langle p_{i_1},\ldots,p_{i_{n+m}}\rangle$ w.r.t. an arbitrary term ordering, and to verify whether this Gröbner basis contains a nonzero element of the field $\mathcal{K}$. Since in general Gröbner basis computations w.r.t. the graded lexicographic or graded inverse lexicographic term ordering are considerably faster than Gröbner basis computations w.r.t. the pure lexicographic term ordering, this modification increases the computational efficiency of the reachability test based on the ideal $\mathcal{L}$.

Summarizing the considerations elaborated above, we obtain the following algorithm to test the reachability of a system over a polynomial ring.

Algorithm 5.3.5 Let $\Sigma=(A,B)$ be a system over the polynomial ring $\mathcal{R}=\mathcal{K}[s_1,\ldots,s_k]$, with $A\in\mathcal{R}^{n\times n}$ and $B\in\mathcal{R}^{n\times m}$.

Step 1 Compute for every $i\in\{1,\ldots,n\}$ the set G_i in the following way:

> **Step I** Introduce $n-i$ new indeterminates $q_{i+1},\ldots,q_n$, and construct the n-dimensional row vector $(\underbrace{0\cdots 0}_{i-1}|1|q_{i+1}\cdots q_n)$.

Step II Compute

$$(p_{i_1} \cdots p_{i_{n+m}}) = (\underbrace{0 \cdots 0}_{i-1}|1|q_{i+1} \cdots q_n) \cdot (zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k)),$$

and consider p_{i_j} $(j = 1, \ldots, n+m)$ as elements of the polynomial ring $\mathcal{K}[z, s_1, \ldots, s_k, q_{i+1}, \ldots, q_n]$.

Step III Determine a Gröbner basis G_i of the ideal $\langle p_{i_1}, \ldots, p_{i_{n+m}} \rangle$ w.r.t. an arbitrary term ordering.

Step 2 If for every $i \in \{1, \ldots, n\}$ the Gröbner basis G_i contains a nonzero element of $\mathcal{K}$, then $\Sigma = (A, B)$ is reachable. Otherwise, Σ is not reachable.

As we have seen before, Algorithm 5.3.5 proceeds from Algorithm 5.3.4 by a normalization of the elements in the left kernel of the matrix $(zI - A|B)$. Instead of searching for all elements in the left kernel of a matrix, we are only interested in those elements that are normalized in a very specific way. Of course this normalization can be carried out in several different ways. For example, instead of normalizing the first nonzero component of the vector, it is also possible to normalize the last nonzero component, and to consider vectors of the form

$$(q_1 \cdots q_{n-i}|1|\underbrace{0 \cdots 0}_{i-1}). \tag{5.36}$$

In fact, a different ordering of the components of the n-dimensional vector yields another normalization, and Algorithm 5.3.5 may be modified accordingly. However, these modifications do not change the algorithm essentially. Permutation of the rows of the matrix $(zI - A|B)$ has exactly the same outcome. Therefore it is obvious that the right-invertibility of the matrix $(zI - A|B)$ over the polynomial ring $\mathcal{R}[z]$ is not influenced by these modifications.

Comparing Algorithms 5.3.4 and 5.3.5, we see that in Algorithm 5.3.4 only one Gröbner basis has to be computed, while in Algorithm 5.3.5 n Gröbner basis computations are required. So at first sight, Algorithm 5.3.5 does not seem very attractive from the computational point of view. However, Algorithm 5.3.5 also has an important advantage. From Subsection 4.1.7. we recall that the complexity of the Gröbner basis algorithm is highly dependent on the number of indeterminates in the polynomial ring under consideration. In Algorithm 5.3.4 a Gröbner basis of an ideal in a polynomial ring with $n + k + 1$ indeterminates is computed. In Algorithm 5.3.5 on the other hand, n Gröbner bases are calculated, with $n + k + 1 - j$ indeterminates $(j = 1, \ldots, n)$, respectively. Since in this case the number of indeterminates is smaller, this method may be faster, despite the fact that more Gröbner bases have to be calculated. Note that in each step these computations become less involved because the number of indeterminates is strictly decreasing. Moreover, if the system $\Sigma = (A, B)$ is not reachable, it is very likely that this is detected in the first step, after the computation of only one Gröbner basis. If G_1 does not contain a nonzero element of $\mathcal{K}$ it is already impossible that $\Sigma = (A, B)$ is reachable. From the proof of Proposition 2.2.5 on the genericity of reachability for systems over rings,

given in [67], it follows that generically the fact that a system is not reachable is detected in the first step of Algorithm 5.3.5. Therefore we expect that in this situation Algorithm 5.3.5 is the most favourable option.

We end this section with a comparison of the performances of the algorithms we developed to test the reachability of a system. For this purpose we applied these algorithms to two different examples. These experiments were made in the computer algebra package Maple V, running on a Sun/Sparc workstation with a 25MHz processor. To compute Gröbner bases, we used the function *gbasis* from the *grobner* package, with the graded lexicographic term ordering and an automatic ranking of the indeterminates (for any details, see [9, pp. 469-478]). All timings are given in CPU seconds without excluding the time for garbage collection. This garbage collection took place every 1Mb. Statistics on the use of memory are not given because the required amount of memory was not critical in these examples.

Example 5.3.6 Consider the matrices A, B and B_1 over the polynomial ring $\mathbf{R}[s]$, given by

$$A = \begin{pmatrix} 2s-3 & -s^2+3s-8 & -2s+6 \\ s+3 & s^2+4 & 2s^2-5s+4 \\ -7s+9 & -8s+7 & s+4 \end{pmatrix},$$

$$B = \begin{pmatrix} s^2+3s-2 & 2 \\ 0 & 5s+1 \\ 4s+7 & -s+2 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} s^2+3s-2 \\ 0 \\ 4s+7 \end{pmatrix}.$$

So B_1 consists of the first column of B. Based on the genericity conditions of Proposition 2.2.5, we expect $\Sigma = (A, B)$ to be reachable, but $\Sigma_1 = (A, B_1)$ not to be reachable.

The reachability of both $\Sigma = (A, B)$ and $\Sigma_1 = (A, B_1)$ is now tested with four different methods:

Method 1 Computation of a Gröbner basis of the ideal $\mathcal{J}$ associated to the system.

Method 2 Algorithm 5.3.4.

Method 3 Algorithm 5.3.5.

Method 4 A modification of Algorithm 5.3.5 in which the normalization is carried out in the opposite direction, using row vectors of the form (5.36).

The results of the application of these methods on this particular example are given in Table 5.1. First the conclusion (reachable/not reachable) is given, then the computing time (in CPU seconds) needed to arrive at the result. The computer time needed in Method 1 to verify the reachability of $\Sigma = (A, B)$ was highly variable. The indicated value is the mean of four samples.

Before drawing any conclusions, we first show that the proposed methods can easily handle systems over polynomial rings in more than one indeterminate.

Table 5.1	$\Sigma = (A, B)$	$\Sigma_1 = (A, B_1)$
Method 1	reachable $56.1 \cdot 10^2$	not reachable 3.3
Method 2	reachable 17.8	not reachable 98.4
Method 3	reachable 8.8	not reachable 27.1
Method 4	reachable 9.0	not reachable 19.9

Example 5.3.7 Consider the matrices A, B and B_1 over the polynomial ring $\mathbf{R}[s_1, s_2]$ given by

$$A = \begin{pmatrix} s_1+1 & s_2+s_1 & s_2-s_1+3 \\ s_2-1 & s_2+1 & s_1-5 \\ s_1^2+1 & 1 & s_2+1 \end{pmatrix},$$

$$B = \begin{pmatrix} s_1+s_2 & s_1-1 & 0 \\ 1 & s_2^2+1 & s_1-3 \\ 0 & s_1-s_2 & s_2+2 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} s_1+s_2 & s_1-1 \\ 1 & s_2^2+1 \\ 0 & s_1-s_2 \end{pmatrix}.$$

Now B_1 consists of the first two columns of B. After application of the same methods as mentioned in Example 5.3.6, Table 5.2 is obtained. The results confirm our expectations based on the genericity conditions of Proposition 2.2.5: $\Sigma = (A, B)$ is reachable, $\Sigma_1 = (A, B_1)$ is not.

Table 5.2	$\Sigma = (A, B)$	$\Sigma_1 = (A, B_1)$
Method 1	reachable $158.6 \cdot 10^3$	not reachable 21.3
Method 2	reachable 55.7	not reachable 362.9
Method 3	reachable 20.3	not reachable 43.7
Method 4	reachable 19.7	not reachable 145.3

When we study Table 5.1 and Table 5.2, they look very similar, and yield almost the same conclusions on the performances of the different methods. The most striking result is the extreme behaviour of Method 1. Although this method is the fastest option in the non-reachable case, it is very slow in detecting the reachability of a system. Therefore it was applied only once in Example 5.3.7. In the reachable case, Method 2 is already much better, but this method is relatively slow for systems that are not reachable. Method 3 and 4, based on Algorithm 5.3.5, behave very well in both cases. They are the fastest option in the reachable case, and if a system is not reachable, this is also detected within a reasonable amount of time. Finally we see that there is a small difference between Method 3 and Method 4 due

to the modification of the normalization procedure. Since the same modification can be carried out by a suitable permutation of the rows of $(zI - A|B)$, this difference probably depends on the particular structure of the system under consideration.

The behaviour of the different algorithms, listed in Table 5.1 and 5.2, is typical for their performances. With several other test-examples the same conclusions were obtained. The results on the complexity of the Gröbner basis computations, given in Subsection 4.1.7, explain these conclusions only to a limited extent. As expected, Algorithm 5.3.5 is somewhat faster than Algorithm 5.3.4, certainly for non-reachable systems, because all Gröbner basis computations in Algorithm 5.3.5 involve a strictly smaller number of indeterminates than the Gröbner basis computation in Algorithm 5.3.4. However, this reasoning does not hold for the ideal $\mathcal{J}$. Although the determination of a Gröbner basis for this ideal requires the smallest number of indeterminates, the computation can be very time consuming. Probably this is caused by the other factor that determines the complexity of Gröbner basis computations: the degree of the polynomials the algorithm starts with. For the computation of a Gröbner basis of $\mathcal{J}$, all principal minors of the matrix $(zI - A|B)$ are required. In this step, the degrees of the polynomials that are involved in the subsequent Gröbner basis computation grow very rapidly. In the Algorithms 5.3.4 and 5.3.5 on the other hand, all polynomials under consideration remain linear both in z and in all indeterminates that are introduced additionally. Although this observation may explain the difference in performance between Method 1 on the one side, and Methods 2, 3 and 4 on the other side, one problem remains unsolved: why is there in the reachable and non-reachable case such a huge difference in the performance of Method 1, based on the ideal $\mathcal{J}$? Unfortunately, the results of Section 4.1.7 do not give a clear answer to this question.

Remark 5.3.8 Note that in the reachability tests based on the ideals $\mathcal{J}$ and $\mathcal{L}$, we are only interested in the question whether the varieties of these ideals are empty. For this purpose it is also possible to use the characteristic sets algorithm because this method is very suitable for the determination of the variety of a polynomial ideal.

Remark 5.3.9 The methods for testing the right-invertibility of the matrix $(zI - A|B)$ over the polynomial ring $\mathcal{K}[z, s_1, \ldots, s_k]$ developed in this chapter, do not depend on the specific structure of this matrix. Therefore all methods are also applicable to verify the right-invertibility of arbitrary matrices over polynomial rings. However, the conclusions on the performances of the different algorithms do not hold in this more general situation, because they may be influenced by the particular structure of the matrix $(zI - A|B)$. In comparison with the other indeterminates, the indeterminate z plays a very special role, because it only occurs in the term $z \cdot I$. When this feature is lost, the performances of the algorithms may change considerably.

5.4 Computation of a right-inverse of $(zI - A|B)$

The reachability of a system $\Sigma = (A, B)$ over a polynomial ring $\mathcal{R}$ is not only interesting for its own sake. In the design of feedback compensators, the reachability of a system is often assumed, in order to design a controller achieving some

requirements that are specified beforehand. In the construction of these compensators, right-inverses of the matrices $(zI-A|B)$ and $(B|AB|\cdots|A^{n-1}B)$ often occur explicitly. In Section 2.8 for example, we have seen how a stabilizing feedback compensator can be obtained from a right-inverse of the matrix $(zI-A|B)$ over the ring $\mathcal{R}_{\mathcal{D}}(z)$ of all stable transfer functions. If a system is reachable, a right-inverse over $\mathcal{R}[z]$ exists, and independent of the choice of a specific Hurwitz set $\mathcal{D}$, this right-inverse can be used in the construction of a stabilizing compensator, given in the proof of Theorem 2.8.2. Another example is the input-output decoupling problem. In [18] it is shown how this problem is solvable for systems over unique factorization domains, under the condition of reachability. The construction of an input-output decoupling compensator relies on the availability of a right-inverse of the matrix $(B|AB|\cdots|A^{n-1}B)$. We conclude that the information that a system $\Sigma = (A,B)$ is reachable is often not enough for the design of feedback compensators. We are also interested in the computation of a right-inverse of the matrices $(zI-A|B)$ and $(B|AB|\cdots|A^{n-1}B)$. In this section it is shown that with a small modification of Algorithm 5.3.4 it is possible to compute these right-inverses.

Let $\Sigma = (A,B)$ be a system over a polynomial ring $\mathcal{R} = \mathcal{K}[s_1,\ldots,s_k]$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$, and assume that Σ is reachable. In Algorithm 5.3.4 this property is verified with help of an n-dimensional row vector $(q_1\cdots q_n)$ of indeterminates. Defining

$$(p_1\cdots p_{n+m}) = (q_1\cdots q_n)\cdot(zI-A|B),$$

the matrix $(zI-A|B)$ is right-invertible over $\mathcal{R}[z]$ if and only if the reduced Gröbner basis G of the ideal $\mathcal{P} = \langle p_1,\ldots,p_{n+m}\rangle$ is $\{q_1,\ldots,q_n\}$, independent of the chosen term ordering.

According to Remark 4.1.27, application of the Gröbner basis algorithm to the set of polynomials $\{p_1,\ldots,p_{n+m}\}$ does not only yield a reduced Gröbner basis $\{q_1,\ldots,q_n\}$, but also a set of coefficients in $\mathcal{R}[z,q_1,\ldots,q_n]$ describing the relationship between the polynomials $p_1,\ldots,p_{n+m}$ on the one side, and the polynomials $q_1,\ldots,q_n$ on the other side. So, for every $i \in \{1,\ldots,n\}$ we may obtain polynomials $m_{ji} \in \mathcal{R}[z,q_1,\ldots,q_n]$ $(j = 1,\ldots,n+m)$ such that

$$q_i = \sum_{j=1}^{n+m} m_{ji}\cdot p_j. \tag{5.37}$$

Let $M(z,q_1,\ldots,q_n)$ denote the $(n+m)\times n$ matrix over $\mathcal{R}[z,q_1,\ldots,q_n]$ with m_{ji} as $(j,i)^{\text{th}}$ entry. Then it follows from (5.37) that

$$(q_1\cdots q_n) = (p_1\cdots p_{n+m})\cdot M(z,q_1,\ldots,q_n).$$

Substitution of the definition of $(p_1\cdots p_{n+m})$ in this formula yields

$$(q_1\cdots q_n)\cdot(zI-A|B)\cdot M(z,q_1,\ldots,q_n) = (q_1\cdots q_n)\cdot I.$$

Next we apply Lemma 5.2.7 (with $\tilde{\mathcal{R}} = \mathcal{R}[z]$) to all columns of $M(z,q_1,\ldots,q_n)$, and conclude that the matrix $M_0(z)$ over $\mathcal{R}[z]$, obtained after substitution of $q_1 = q_2 = \cdots = q_n = 0$ in the matrix $M(z,q_1,\ldots,q_n)$ is a right-inverse of $(zI-A|B)$:

$$(zI-A|B)\cdot M_0(z) = I.$$

In general, a right-inverse of the matrix $(zI - A|B)$ is not unique. However, if one particular right-inverse is obtained, it is possible to give a characterization of all right-inverses.

Proposition 5.4.1 *Let P be a $p \times q$ matrix over an integral domain $\tilde{\mathcal{R}}$, and assume that $M \in \tilde{\mathcal{R}}^{q \times p}$ is a right-inverse of the matrix P. Then*

$$N \in \tilde{\mathcal{R}}^{q \times p} \text{ is a right-inverse of } P,$$

$$\Longleftrightarrow$$

$$\exists G \in \tilde{\mathcal{R}}^{q \times p} \text{ such that } N = M + (I - MP)G.$$

Proof

"$\Leftarrow$" Assume that N is of the form $N = M + (I - MP)G$, with G a $q \times p$ matrix over $\tilde{\mathcal{R}}$. Then

$$P \cdot N = PM + P(I - MP)G = I + PG - PMPG = I + PG - PG = I.$$

"$\Rightarrow$" Let N be a right-inverse of P over $\tilde{\mathcal{R}}$, and define $G := N - M$. Using the fact that $PN = I$, we then have

$$\begin{aligned} M + (I - MP)G &= M + (I - MP)(N - M) = \\ &= M + (N - M) - MP(N - M) = \\ &= N - MPN + MPM = N - M + M = N. \end{aligned}$$ ∎

Let now $\Sigma = (A, B)$ be a system over the polynomial ring $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Assume that we have obtained a right-inverse $M_0(z)$ over $\mathcal{R}[z]$ of the matrix $(zI - A|B)$, using the Gröbner basis method described at the beginning of this section. When we define

$$C(z) := I - M_0(z) \cdot (zI - A|B),$$

we know from Proposition 5.4.1 (with $\tilde{\mathcal{R}} = \mathcal{R}[z]$) that every right-inverse $N(z)$ of $(zI - A|B)$ over $\mathcal{R}[z]$ has the following form:

$$N(z) = M_0(z) + C(z) \cdot G(z), \tag{5.38}$$

where $G(z)$ is an arbitrary $(n+m) \times n$ matrix over $\mathcal{R}[z]$. This degree of freedom may be used to find a right-inverse that meets the requirements of our control purposes.

From Section 2.8 we recall that the degree in the indeterminate z of the right-inverse $M_0(z)$ of the matrix $(zI - A|B)$ is important for the construction of a feedback compensator. When we decompose the matrix $M_0(z)$ into two blocks, $Q(z)$ and $P(z)$, containing the first n, and the last m rows of $M_0(z)$, respectively,

$$M_0(z) = \begin{pmatrix} Q(z) \\ P(z) \end{pmatrix},$$

we are interested in a solution with $\deg_z(P(z)) \leq n - 1$. To satisfy this condition, Lemma 2.8.1 is used. First we write $P(z)$ as

$$P(z) = \chi_A(z) \cdot P_1(z) + P_2(z),$$

in such a way that $\deg_z(P_2(z)) < \deg(\chi_A(z)) = n$. Then it follows from Lemma 2.8.1 that the matrix $\hat{M}(z) = \begin{pmatrix} \hat{Q}(z) \\ \hat{P}(z) \end{pmatrix}$, with

$$\begin{aligned} \hat{Q}(z) &= Q(z) + \mathrm{adj}(zI - A) \cdot B \cdot P_1(z), \\ \hat{P}(z) &= P_2(z), \end{aligned}$$

is a right-inverse of $(zI - A|B)$, satisfying the condition $\deg_z(\hat{P}(z)) \leq n - 1$. Moreover, since

$$(zI - A) \cdot \hat{Q}(z) + B \cdot \hat{P}(z) = I,$$

this implies that $\deg_z(\hat{Q}(z)) \leq n - 2$, and thus we have obtained a right-inverse $\hat{M}(z)$ such that $\deg_z(\hat{M}(z)) \leq n - 1$.

Remark 5.4.2 It is also possible to compute the right inverse $\begin{pmatrix} \hat{Q}(z) \\ \hat{P}(z) \end{pmatrix}$ using formula (5.38). When $M_0(z) = \begin{pmatrix} Q(z) \\ P(z) \end{pmatrix}$, and we choose

$$G(z) = \begin{pmatrix} \mathrm{adj}(zI - A)BP_1(z) \\ -\chi_A(z)P_1(z) \end{pmatrix},$$

it is easily verified that the matrix $N(z)$ in (5.38) is equal to $\begin{pmatrix} \hat{Q}(z) \\ \hat{P}(z) \end{pmatrix}$.

In some situations, for example for the solution of the input-output decoupling problem given in [18], we are not interested in a right-inverse of the matrix $(zI - A|B)$ only, but also want to obtain a right-inverse of the matrix $(B|AB|\cdots|A^{n-1}B)$. In the proof of Theorem 2.2.3 it was shown how this can be done. For systems $\Sigma = (A, B)$ over an integral domain $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$, there is a one-to-one correspondence between right-inverses of $(B|AB|\cdots|A^{n-1}B)$ over $\mathcal{R}$ on the one hand, and right-inverses of $(zI - A|B)$ over $\mathcal{R}[z]$ with degree in the indeterminate z smaller than or equal to $n - 1$ on the other hand. Moreover, a constructive method was given to obtain one type of right-inverse from the other and vice versa.

Note that a right-inverse of $(B|AB|\cdots|A^{n-1}B)$ may also be computed directly, using the Gröbner basis method described at the beginning of this section for the matrix $(zI - A|B)$. It is obvious that this method also works for the determination of a right-inverse of an arbitrary polynomial matrix. The specific structure of the matrix $(zI - A|B)$ is not used explicitly. It is not clear beforehand which method to determine a right-inverse of $(B|AB|\cdots|A^{n-1}B)$ is preferable. Although the direct method involves a Gröbner basis computation in a polynomial ring that does not contain the indeterminate z, the degrees of the entries of the matrix $A^{n-1}B$ may grow very rapidly with the size n of the matrix A. Since both the degree and the number of indeterminates of the polynomials under consideration influence the complexity of the Gröbner basis algorithm, a trade-off is difficult to make.

Sometimes it is not necessary to proceed that far. In a lot of cases there exists an $i < n$ such that the matrix $(B|AB|\cdots|A^{i-1}B)$ is already right-invertible over $\mathcal{R}$.

Let ℓ denote the minimal value of i for which the matrix $(B|AB|\cdots|A^{i-1}B)$ is right-invertible. In the discrete-time interpretation of a system over a ring, as given in Example 2.1.2, ℓ represents the maximal number of steps that is required to go from one arbitrary point in the state space to another point, via the shortest route. The value of ℓ is easily obtained by subsequent application of one of the algorithms of the previous section on the matrices $(B|AB|\cdots|A^{i-1}B)$ $(i = 1, \ldots, n)$. Then ℓ is simply the smallest value of i for which the corresponding matrix is right-invertible over $\mathcal{R}$. With help of a right-inverse of $(B|AB|\cdots|A^{\ell-1}B)$ over $\mathcal{R}$, a right-inverse $M(z)$ over $\mathcal{R}[z]$ of $(zI - A|B)$ can be computed, satisfying the property $\deg_z(M(z)) = \ell - 1$. Moreover, it follows from the proof of Theorem 2.2.3 that

$$\ell - 1 = \min\{\deg_z(M(z)) | M(z) \in \mathcal{R}[z]^{(n+m)\times n} \text{ s.t. } (zI - A|B)\cdot M(z) = I\}. \quad (5.39)$$

The next example illustrates how the methods introduced in this section work out in a concrete situation.

Example 5.4.3 Let $\mathcal{R} = \mathbf{R}[s_1, s_2]$, and consider the system $\Sigma = (A, B)$ over $\mathcal{R}$ with

$$A = \begin{pmatrix} 2 & s_1 + 3 \\ s_2 - 5 & -4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & s_1 + s_2 & -4 \\ 2 & s_2 - 4 & s_1 \end{pmatrix}.$$

Using one of the methods of the previous section it is easily verified that $\Sigma = (A, B)$ is reachable. To obtain a right-inverse of $(zI - A|B)$ over $\mathcal{R}[z]$, we apply the method introduced at the beginning of this section. We compute a reduced Gröbner basis of the ideal $\mathcal{P}$ associated with Σ, w.r.t. the graded reverse lexicographic term ordering, and determine simultaneously the relationship between the original polynomials, and the polynomials $\{q_1, q_2\}$ in the Gröbner basis. Collecting these coefficients in a matrix, and substituting $q_1 = q_2 = 0$, we obtain the matrix

$$M(z) := \frac{1}{15} \cdot \begin{pmatrix} 2 & -1 \\ 4 & -2 \\ 2s_2 - 2z + 6s_1 - 17 & -s_2 + z - 3s_1 + 16 \\ -2 & 1 \\ -12 & 6 \end{pmatrix},$$

which is indeed a right-inverse of $(zI - A|B)$ over $\mathcal{R}[z]$.

Since $\deg_z(M(z)) = 1$ is smaller than the size of the matrix A, a right-inverse of $(B|AB)$ can be computed directly, with the method explained in the proof of Theorem 2.2.3. Using the same terminology as in the proof of this theorem, we have

$$N_0 = \frac{1}{15} \cdot \begin{pmatrix} -2 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad N_1 = \frac{1}{15} \cdot \begin{pmatrix} 2s_2 + 6s_1 - 17 & -s_2 - 3s_1 + 16 \\ -2 & 1 \\ -12 & 6 \end{pmatrix},$$

and it is easily verified that the matrix $\begin{pmatrix} N_1 \\ N_0 \end{pmatrix}$ is a right-inverse of $(B|AB)$ over $\mathcal{R}$.

Note that in this situation $\ell = 2$, because the matrix B is not right-invertible over $\mathcal{R}$. This implies that the degree in z of every right-inverse of $(zI - A|B)$ is at least $\ell - 1 = 1$. So the matrix $M(z)$ is a right-inverse of minimal degree in z.

Although a complete characterization of all right-inverses of the matrix $(zI - A|B)$ has been given, it is difficult in general to obtain a right-inverse that is relatively simple. In this respect, Example 5.4.3 is a little misleading. In most cases, the entries of the right-inverses that are obtained using Gröbner basis techniques, are complicated polynomials of relatively high degree. With the method based on formula (5.39) for the index ℓ, it is possible to minimize the degree of the right-inverse in the indeterminate z. This degree yields some information on the complexity of the dynamics of the compensators that may be constructed from this right-inverse. However, it is questionable whether this is the only goal we want to achieve. Therefore, simplification of right-inverses is a very difficult subject. The problems that arise in this field do not originate from the computational complexity of the problem only. The objectives that we want to realize by means of a simplification method are not very clear either, and therefore we lack an appropriate foundation for a good and effective algorithm.

Remark 5.4.4 A right-inverse of the matrix $(zI - A|B)$ may also be computed using the method based on the ideal $\mathcal{J}$. Let $\alpha_0(z), \ldots, \alpha_N(z)$ denote all principal minors of the matrix $(zI - A|B)$. Then $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$ if and only if the reduced Gröbner basis of the ideal $\langle \alpha_0(z), \ldots, \alpha_N(z) \rangle$ is $\{1\}$. From Remark 4.1.27 it follows that the Gröbner basis computation required to obtain this result, implicitly determines coefficients $\beta_0(z), \ldots, \beta_N(z)$ such that

$$\sum_{i=0}^{N} \alpha_i(z) \cdot \beta_i(z) = 1.$$

With these coefficients a right-inverse of $(zI - A|B)$ may be obtained using the same ideas as in the proof of Proposition 2.8.5.

First note that each minor $\alpha_i(z)$ $(i = 0, 1, \ldots, N)$ is the determinant of an $n \times n$ submatrix $K_i(z)$ of $(zI - A|B)$, where n denotes the size of the matrix A. Extending $\mathrm{adj}(K_i(z))$ with zero rows on the right places we obtain for each $i \in \{0, 1, \ldots, N\}$ an $(n + m) \times n$ matrix $\tilde{K}_i(z)$, satisfying Cramer's rule in the following way

$$(zI - A|B) \cdot \tilde{K}_i(z) = \det(K_i(z)) \cdot I = \alpha_i(z) \cdot I.$$

Defining $M(z) := \sum_{i=0}^{N} \beta_i(z) \cdot \tilde{K}_i(z)$ we have

$$(zI - A|B)M(z) = \sum_{i=0}^{N} \beta_i(z)(zI - A|B)\tilde{K}_i(z) = (\sum_{i=0}^{N} \beta_i(z)\alpha_i(z)) \cdot I = I,$$

and thus $M(z)$ is a right-inverse of $(zI - A|B)$ over $\mathcal{R}[z]$.

This alternative approach, based on a Gröbner basis computation for the ideal $\mathcal{J}$, is not a serious alternative for the method based on the ideal $\mathcal{H}$, explained at the beginning of this section. In the previous section we have seen that in the reachable case the reachability test based on the ideal $\mathcal{J}$ is computationally not very efficient. From the computational point of view, Algorithm 5.3.4 is more appropriate for this purpose because it is much faster. Therefore it is obvious that Algorithm 5.3.4 is also preferable for the determination of a right-inverse of $(zI - A|B)$.

5.5 Testing stabilizability of time-delay systems

Despite the fact that we have obtained constructive methods for the computation of the ideal $\mathcal{I}$ associated with a system Σ over a polynomial ring, the stabilizability question for this kind of systems remains very difficult to answer in general. According to Proposition 5.1.2 *(ii)*, a system $\Sigma = (A, B)$ over an integral domain $\mathcal{R}$ is stabilizable w.r.t. a given Hurwitz set $\mathcal{D}$ if and only if the ideal $\mathcal{I}$ associated with Σ contains a stable polynomial, i.e. a polynomial in $\mathcal{D}$. The ideal $\mathcal{I}$ may be manipulated using Gröbner basis techniques; the difficulty of the stabilizability problem arises from the Hurwitz set $\mathcal{D}$. Our knowledge of this set is restricted to the conditions *(i)* to *(iv)* of Definition 2.5.2. Basically, a Hurwitz set is a multiplicative and saturated subset of monic polynomials in $\mathcal{R}[z]$. In particular, a Hurwitz set $\mathcal{D}$ is not an ideal, and in general we cannot find a finite number of polynomials that in one way or the other characterizes the set $\mathcal{D}$ as a whole. So we simply lack a suitable representation for a Hurwitz set $\mathcal{D}$.

Especially the problem of *constructing* a polynomial in the intersection $\mathcal{I} \cap \mathcal{D}$ is very hard. We are not able to solve this problem for systems over (polynomial rings) in full generality, and therefore confine ourselves to systems with time-delays. In this case, additional information is available, because every indeterminate in the polynomial ring corresponds to a delay operator with some time-delay τ. Using this time-delay character of the system, it is possible to proceed further. In this section we only consider the existence question, and develop some algorithms to test the stabilizability of delay systems. The Gröbner basis methods of Section 5.2 are very useful for this purpose. In the next chapter we return to the problem of constructing a $\mathcal{D}$-stable polynomial in the ideal $\mathcal{I}$, and show how a solution to this problem leads to a stabilizing feedback compensator.

Let $\mathcal{R} = \mathcal{K}[s_1, \ldots, s_k]$ be a polynomial ring, and consider the examples of Hurwitz sets in $\mathcal{R}[z]$ given in Section 2.5. Then we see that often Hurwitz sets $\mathcal{D}$ may be characterized alternatively by a subset $W \subset \bar{\mathcal{K}}^{k+1}$, in which the polynomials of $\mathcal{D}$ are not allowed to have zeros. So in this situation $\mathcal{D}$ takes the form

$$\mathcal{D} = \{p \in \mathcal{K}[z, s_1, \ldots, s_k] \mid p \text{ is monic in } z, \text{ and } \forall \alpha \in W : p(\alpha) \neq 0\}. \qquad (5.40)$$

Also the Hurwitz sets describing the stability of time-delay systems can be written in this particular form.

Let $\Sigma = (A, B)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ as in Example 2.1.3. So the indeterminate s_i corresponds to the time-delay operator with time-delay τ_i $(i = 1, \ldots, k)$. Let $\mathbb{C}_g$ be a stability domain satisfying conditions *(i)* to *(iv)* of Definition 3.1.2, and define $\overline{\mathbb{C}_b} := \mathbb{C} \backslash \mathbb{C}_g$. According to (3.5), the Hurwitz set $\mathcal{D}_g$, describing $\mathbb{C}_g$-stability for the time-delay system Σ, is given by

$$\mathcal{D}_g := \{p(z, s_1, \ldots, s_k) \in \mathbb{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } \forall z \in \mathbb{C} : p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) = 0 \Rightarrow z \in \mathbb{C}_g\}.$$

Define the subset $W_g \subset \mathbb{C}^{k+1}$ as

$$W_g = \{(\lambda, e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda}) \in \mathbb{C}^{k+1} \mid \lambda \in \overline{\mathbb{C}_b}\}. \qquad (5.41)$$

Then $\mathcal{D}_g$ can be redefined as

$$\mathcal{D}_g = \{p \in \mathbf{R}[z, s_1, \ldots, s_k] \mid p \text{ is monic in } z, \text{ and } \forall \alpha \in W_g : p(\alpha) \neq 0\}, \quad (5.42)$$

and therefore it is characterized by the subset W_g, in completely the same way as in (5.40).

Also the concept of pointwise stability for systems with time-delays fits into this special framework. According to the results of Section 3.4, this type of stability is useful for the study of stabilizability independent of delay. In this case, the Hurwitz set $\mathcal{D}_p$ is given by (3.23):

$$\mathcal{D}_p := \{p(z, s_1, \ldots, s_k) \in \mathbf{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } \forall z \in \overline{\mathbb{C}^+} \, \forall (s_1, \ldots, s_k) \in \overline{\mathcal{U}}^k : \; p(z, s_1, \ldots, s_k) \neq 0\},$$

where $\mathcal{U}$ denotes the open unit disc $\{s \in \mathbb{C} \mid |s| < 1\}$. This Hurwitz set is already in the form of (5.40): when we define

$$W_p = \{(z, s_1, \ldots, s_k) \in \mathbb{C}^{k+1} \mid z \in \overline{\mathbb{C}^+}, \; s_i \in \overline{\mathcal{U}} \; (i = 1, \ldots, k)\}, \quad (5.43)$$

then $\mathcal{D}_p$ may be written as

$$\mathcal{D}_p = \{p \in \mathbf{R}[z, s_1, \ldots, s_k] \mid p \text{ is monic in } z, \text{ and } \forall \alpha \in W_p : p(\alpha) \neq 0\}. \quad (5.44)$$

Since Hurwitz sets of the form (5.40) are characterized by the points in which they are not allowed to have zeros, it is apparent that in this situation the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ associated with the system Σ, contains important information on the connection between the ideal $\mathcal{I}$ and the Hurwitz set $\mathcal{D}$. For systems with time-delays, the stabilizability question w.r.t. the Hurwitz sets $\mathcal{D}_g$ and $\mathcal{D}_p$ is even immediately solvable when $\mathcal{V}(\mathcal{I})$ is known.

Theorem 5.5.1 *Consider a system $\Sigma = (A, B)$ over the polynomial ring $\mathcal{R} = \mathbf{R}[s_1, \ldots, s_k]$, and let $\mathcal{I}$ be the ideal associated with Σ, introduced in Definition 5.1.1. Let $(\tau_1, \ldots, \tau_k)$ be a k-tuple of time-delays, and let $\mathbb{C}_g$ be a stability domain. Define the sets $W_g \subset \mathbb{C}^{k+1}$ and $W_p \subset \mathbb{C}^{k+1}$ as in (5.41) and (5.43), respectively, and consider the corresponding Hurwitz sets $\mathcal{D}_g$ and $\mathcal{D}_p$. Then*

(i) $\Sigma = (A, B)$ is stabilizable w.r.t. $\mathcal{D}_g$ if and only if $\mathcal{V}(\mathcal{I}) \cap W_g = \varnothing$,

(ii) $\Sigma = (A, B)$ is stabilizable w.r.t. $\mathcal{D}_p$ if and only if $\mathcal{V}(\mathcal{I}) \cap W_p = \varnothing$.

Proof

(i) Suppose that $\Sigma = (A, B)$ is stabilizable w.r.t. $\mathcal{D}_g$. Then, according to Proposition 5.1.2 *(ii)*, there exists a polynomial $p \in \mathcal{I} \cap \mathcal{D}_g$. Let $\alpha \in \mathcal{V}(\mathcal{I})$. Then $p(\alpha) = 0$ because $p \in \mathcal{I}$. However, since $p \in \mathcal{D}_g$, $p(\alpha) = 0$ implies that $\alpha \notin W_g$, and we conclude that $\mathcal{V}(\mathcal{I}) \cap W_g = \varnothing$.

Next, assume that $\mathcal{V}(\mathcal{I}) \cap W_g = \varnothing$. Let $\mathcal{J}$ be the ideal associated with Σ, generated by all principal minors of the matrix $(zI - A|B)$. According to Proposition 5.1.12, we have $\mathcal{V}(\mathcal{I}) = \mathcal{V}(\mathcal{J})$, and therefore also $\mathcal{V}(\mathcal{J}) \cap W_g = \varnothing$. Denoting the size of the matrix A by n, this implies that the $n \times n$ minors of the matrix $(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k))$ do not have a common zero in W_g. Hence, for every

point $\alpha \in W_g$, there exists an $n \times n$ minor that is nonzero in α. Using definition (5.41) of W_g, we conclude that

$$\forall \lambda \in \mathbf{C} \backslash \mathbf{C}_g : \operatorname{rank}(\lambda I - A(e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda}) \mid B(e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda})) = n.$$

Successive application of Theorem 3.2.8 and Theorem 2.8.2 yields that $\Sigma = (A, B)$ is stabilizable w.r.t. $\mathcal{D}_g$.

(ii) The proof of *(ii)* proceeds completely analogously and is therefore omitted. We only remark that in the last step Proposition 3.2.14 is used instead of Theorem 3.2.8.

■

The proof of Theorem 5.5.1 is based on the equivalence

$$\mathcal{I} \cap \mathcal{D} \neq \varnothing \iff \mathcal{V}(\mathcal{I}) \cap W = \varnothing.$$

It is important to note that this equivalence does not hold for all Hurwitz sets of the form (5.40). The proof of Theorem 5.5.1 relies on the results of Chapter 3, where the stabilizability condition for time-delay systems is restated as a pointwise rank condition on the matrix

$$(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \mid B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})).$$

In general, such a reformulation is not possible, and in these situations the intersection $\mathcal{V}(\mathcal{I}) \cap W$ does not yield enough information to decide on the stabilizability of the corresponding system.

Theorem 5.5.1 is very important for testing the stabilizability of time-delay systems. First the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ associated with the system $\Sigma = (A, B)$ is computed, using one of the methods developed in Section 5.2. Next, it is verified whether $\mathcal{V}(\mathcal{I})$ contains any point that belongs to the set W. If $\mathcal{V}(\mathcal{I}) = \varnothing$, this is a trivial task. In this case, the system $\Sigma = (A, B)$ is reachable, which may be verified with one of the methods of Section 5.3. If $\mathcal{V}(\mathcal{I})$ is zero-dimensional and contains only a finite number of points, the intersection problem is also conceptually easy to solve, but in this case there are some numerical pitfalls. The problem gets more complicated if the variety $\mathcal{V}(\mathcal{I})$ has positive dimension and contains infinitely many elements. Below these last two cases are elaborated in more detail.

$\mathcal{V}(\mathcal{I})$ is zero-dimensional

If $\mathcal{V}(\mathcal{I})$ is zero-dimensional, in principle all points in $\mathcal{V}(\mathcal{I})$ can be computed with one of the Gröbner basis methods of Section 5.2. Algorithm 5.2.1, based on the ideal $\mathcal{J}$, is probably the best choice because we have seen in Section 5.3 that this method has a good performance for non-reachable systems. Since we are interested in the variety of an ideal, we choose the pure lexicographic term ordering, because this yields a generating set of polynomials that is in triangular form. Then the set of common zeros is determined with a backward substitution process. Summarizing we obtain the following algorithm:

Algorithm 5.5.2 Let $\Sigma = (A, B)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbf{R}[s_1, \ldots, s_k]$, where the indeterminate s_i corresponds to the time-delay operator with time-delay

τ_i $(i = 1, \ldots, k)$. Assume that the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ associated with Σ is zero-dimensional. Let $\mathbf{C}_g$ be a stability domain satisfying condition *(i)* to *(iv)* of Definition 3.1.2, and define W_g and $\mathcal{D}_g$ accordingly, cf. (5.41) and (5.42), respectively.

Step 1 Determine with Algorithm 5.2.1 a Gröbner basis of the ideal $\mathcal{J}$ associated with Σ, w.r.t. a pure lexicographic term ordering.

Step 2 Determine (possibly numerically) all points in the variety of $\mathcal{J}$ by backward substitution.

Step 3 If no point of $\mathcal{V}(\mathcal{J})$ belongs to W_g, then Σ is stabilizable; otherwise Σ is not stabilizable.

Algorithm 5.5.2 has a very useful property. The first two steps are always the same, independent of the choice of the stability domain $\mathbf{C}_g$. So, when Step 2 is accomplished, and the variety $\mathcal{V}(\mathcal{J})$ is obtained, it is possible to verify with respect to what stability domains the system is stabilizable. Also changes in the lengths of the time-delays $\tau_1, \ldots, \tau_k$ only influence the conclusions of Step 3. This enables us to study the sensitivity to uncertainties in the lengths of the time-delays occurring in the system. It is even possible to test whether a time-delay system is pointwise stabilizable. In this case the variety $\mathcal{V}(\mathcal{J})$ determined in Step 2 is not allowed to contain an element of the set W_p defined in (5.43). So, given a time-delay system $\Sigma = (A, B)$, the variety $\mathcal{V}(\mathcal{J})$ implicitly describes all sorts of stability that are attainable from Σ using dynamic state feedback.

Unfortunately, Algorithm 5.5.2 has some serious drawbacks. First of all, in Step 1 a Gröbner basis computation w.r.t. a pure lexicographic term ordering is required, and this computation may be rather time consuming. However, the numerical difficulties that occur in Step 2 are more important. In somewhat larger examples, the Gröbner basis of the ideal $\mathcal{J}$ typically contains polynomials of high degree and with very large coefficients. So in most cases, the backward substitution process has to be carried out numerically. This can be very tricky because the zeros of one polynomial, that are computed numerically with a certain level of accuracy, are substituted in the subsequent polynomial. In this way $k + 1$ steps are taken, and therefore it is obvious that if no precautions are taken, huge fault propagations may occur. To overcome this problem, at least two different solutions are possible. First of all, one may try to estimate the errors that are made in the numerical computation of the zeros of the subsequent polynomials. By a careful substitution of these zeros in the next polynomial (using for example Horner's Algorithm (see e.g. [87, p.44])), it is possible to diminish the fault propagation, and to estimate the accuracy of our computations. This enables us to obtain a more reliable answer to the stabilizability question.

In the next algorithm we present an alternative way to test the stabilizability of a time-delay system. It is based on completely different ideas, and depends heavily on the structure of the problem under consideration. Although this algorithm lacks the flexibility of Algorithm 5.5.2 to treat several types of stability simultaneously, this method is very fast, and avoids the numerical difficulties mentioned above.

Algorithm 5.5.3 Let $\Sigma = (A, B)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$, where the indeterminate s_i corresponds to the time-delay operator with time-delay τ_i $(i = 1, \ldots, k)$. Assume that the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ associated with the system Σ is zero-dimensional. Let $\mathbf{C}_g$ be a stability domain satisfying condition *(i)* to *(iv)* of Definition 3.1.2, and define W_g and $\mathcal{D}_g$ accordingly, cf. (5.41) and (5.42), respectively.

Step 1 Determine with Algorithm 5.2.1 a Gröbner basis of the ideal $\mathcal{J}$ associated with Σ, w.r.t. an arbitrary term ordering.

Step 2 Apply the construction method of Proposition 4.1.37 to obtain a univariate polynomial p in $\mathcal{J}$ that only contains the indeterminate z.

Step 3 Compute (numerically) all zeros of the polynomial p, and determine the set $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ consisting of all zeros of p that belong to $\mathbf{C} \backslash \mathbf{C}_g$.

Step 4 Determine for every $i \in \{1, \ldots, N\}$ the rank of the matrix

$$(\lambda_i I - A(e^{-\tau_1 \lambda_i}, \ldots, e^{-\tau_k \lambda_i}) \mid B(e^{-\tau_1 \lambda_i}, \ldots, e^{-\tau_k \lambda_i})), \tag{5.45}$$

using the singular value decomposition.

Step 5 If for every $i \in \{1, \ldots, N\}$ the matrix (5.45) has full row rank, the time-delay system is stabilizable w.r.t. $\mathcal{D}_g$; otherwise Σ is not stabilizable w.r.t. $\mathcal{D}_g$.

The major part of Algorithm 5.5.3 speaks for itself. The algorithm mainly relies on the special meaning of the indeterminate z. By assumption $\mathcal{V}(\mathcal{J})$ is zero-dimensional, and thus it follows from Proposition 4.1.30 that the ideal $\mathcal{J}$ contains a univariate polynomial p in the indeterminate z. Denoting by $\tilde{\Lambda}$ the finite set of all zeros of the polynomial p, it is easily verified that the only points in $\mathbf{C}^{k+1}$ that can be contained in $\mathcal{V}(\mathcal{J}) \cap W_g$ are given by

$$\{(\lambda, e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda}) \in \mathbf{C}^{k+1} \mid \lambda \in \tilde{\Lambda} \backslash \mathbf{C}_g\}. \tag{5.46}$$

By definition, all these candidate points belong to W_g. To verify whether one of these points also belongs to the variety $\mathcal{V}(\mathcal{J})$, it suffices to check whether the matrix $(zI - A(s_1, \ldots, s_k) | B(s_1, \ldots, s_k))$ is of full row rank after substitution of $(\lambda, e^{-\tau_1 \lambda}, \ldots, e^{-\tau_k \lambda})$ for the indeterminates $(z, s_1, \ldots, s_k)$. In this way a matrix over the complex numbers is obtained, and the rank condition may be tested with the singular value decomposition: the matrix (5.45) has full row rank if and only if all its singular values are nonzero. However, if one of the singular values is very small, the matrix (5.45) almost loses rank, and we have to be very careful with our conclusion on stabilizability. This is one of the main advantages of the singular value decomposition; except for an answer (yes/no) to our question, it also gives an indication of the reliability and sensitivity of this answer.

The only numerical difficulty of Algorithm 5.5.3 occurs in Step 3. Here all zeros of a univariate polynomial are computed, and in general this has to be done numerically. Although these zeros can only be determined with a certain level

of accuracy, the danger of fault propagation is relatively small. First, each point $\lambda \in \mathbb{C}\backslash\mathbb{C}_g$ that is a solution to the equation $p = 0$, is completed to an element of the set (5.46), by addition of the components $e^{-\tau_i\lambda}$ $(i = 1, \ldots, k)$. Since the entries of the matrix $(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k))$ are polynomials of relatively low degree and with relatively small coefficients, substitution of $(\lambda, e^{-\tau_1\lambda}, \ldots, e^{-\tau_k\lambda})$ for the indeterminates $(z, s_1, \ldots, s_k)$ does not cause difficult numerical problems in general. Finally, there exist reliable numerical methods for the determination of the singular value decomposition required in Step 4 of the algorithm. Although the final conclusion on the stabilizability of the system is mainly based on these singular values, also the accuracy of the computations in the previous steps has to be taken into account.

Note that in Step 3 of Algorithm 5.5.3, the stability domain $\mathbb{C}_g$ is used explicitly. Therefore this method can only test the stabilizability of a delay system with respect to one stability domain at a time. However, compared with Algorithm 5.5.2, Algorithm 5.5.3 is much faster. First of all it is not necessary to compute a Gröbner basis of $\mathcal{J}$ w.r.t. a pure lexicographic term ordering. Instead, the Gröbner basis computations may be carried out w.r.t. the graded (reverse) lexicographic term ordering, which is computationally much more efficient. The determination of a univariate polynomial using the method of Proposition 4.1.37 is relatively easy, and also the other (numerical) steps are computationally not very demanding. It is not necessary to take as many numerical precautions as are required in Algorithm 5.5.2. Therefore, if $\mathcal{V}(\mathcal{I})$ is zero-dimensional, Algorithm 5.5.3 seems the most appropriate method to test the stabilizability of the corresponding time-delay system.

$\mathcal{V}(\mathcal{I})$ is positive dimensional

If the variety of the ideal $\mathcal{I}$ contains infinitely many elements, it is much more difficult to test stabilizability. In this case the ideal $\mathcal{I}$ contains no univariate polynomials in general, and the variety $\mathcal{V}(\mathcal{I})$ cannot be obtained using Gröbner basis computations only. Therefore, the Algorithms 5.5.2 and 5.5.3 cannot be applied because they are mainly based on the manipulation of the polynomials in the ideal $\mathcal{I}$.

However, if we also consider so-called *exponential polynomials* (in Section 6.1 this kind of functions is studied in more detail), a slight modification of Algorithm 5.5.3 yields a solution method. For a system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system $\Sigma = (A, B)$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$, the characteristic function is given by the exponential polynomial

$$\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})). \tag{5.47}$$

In the case of commensurable time-delays, and when the stability domain is $\mathbb{C}^-$, we know from Lemma 3.3.33 that this function only has a finite number of zeros in $\mathbb{C}\backslash\mathbb{C}^-$. In fact, it is not difficult to prove that also for systems with incommensurable time-delays, and for arbitrary stability domains $\mathbb{C}_g$, satisfying the conditions of Definition 3.1.2, the function (5.47) only has a finite number of zeros in $\mathbb{C}\backslash\mathbb{C}_g$. In Section 6.1 this claim is elaborated in more detail.

Let $\mathbb{C}_g$ be a stability domain, and define W_g as in (5.41). Given a time-delay system Σ, we denote by Λ the set of all zeros of the corresponding characteristic

function (5.47) that are contained in $\mathbb{C}\backslash\mathbb{C}_g$. Every $\lambda \in \Lambda$ corresponds to a point

$$(\lambda, e^{-\tau_1\lambda}, \ldots, e^{-\tau_k\lambda}) \in W_g,$$

and in the same way as in Algorithm 5.5.3, the finite set

$$\{(\lambda, e^{-\tau_1\lambda}, \ldots, e^{-\tau_k\lambda}) \in \mathbb{C}^{k+1} \mid \lambda \in \Lambda\}$$

contains all elements of $\mathbb{C}^{k+1}$ that may belong to $\mathcal{V}(\mathcal{I}) \cap W_g$. Therefore, subsequent application of Step 4 and Step 5 of Algorithm 5.5.3 yields a conclusion on the stabilizability of the delay system under consideration.

The main difficulty with the approach described above is that the computation of the zeros of $\det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ in $\mathbb{C}\backslash\mathbb{C}_g$ has to be carried out numerically. In Algorithm 5.5.3 this problem was circumvented using Gröbner basis techniques. However, if $\mathcal{V}(\mathcal{I})$ is infinite-dimensional, this detour does not work in general. On the other hand, the numerical approach is also applicable if $\mathcal{V}(\mathcal{I})$ is zero-dimensional.

From the observations elaborated above, it is obvious that for a given time-delay system Σ, it is important to know what the dimension of the associated polynomial ideal $\mathcal{I}$ is. With Gröbner basis techniques it can be verified whether this ideal is zero-dimensional or not. In the first case, the same Gröbner basis can be used subsequently in Algorithm 5.5.3. However, if it turns out that the variety $\mathcal{V}(\mathcal{I})$ contains infinitely many elements, the Gröbner basis computation is of no use, and we may proceed with the more involved numerical approach given above.

For this reason it would be very useful to know what the dimension of the variety $\mathcal{V}(\mathcal{I})$ is, before a particular algorithm is chosen and the actual computations are started. However, this is impossible because a Gröbner basis is really required to make sure that a given ideal is zero-dimensional. Nevertheless, a good indication for the dimension of the variety $\mathcal{V}(\mathcal{I})$ may be obtained with the following conjecture.

Conjecture 5.5.4 *Let $\mathcal{K}$ be a field of characteristic zero, and let $\mathcal{R}$ denote the polynomial ring $\mathcal{K}[s_1, \ldots, s_k]$. Consider all systems $\Sigma = (A, B)$ over $\mathcal{R}$, with $A \in \mathcal{R}^{n\times n}$ and $B \in \mathcal{R}^{n\times m}$, and let $\mathcal{I}$ be the ideal associated with Σ, defined in Definition 5.1.1. Considering the concept of genericity in the Zariski topology we have:*

(i) If $k < m$, then generically $\mathcal{V}(\mathcal{I}) = \varnothing$.

(ii) If $k \geq m$, then generically the dimension of the variety $\mathcal{V}(\mathcal{I})$ is equal to $k - m$.

According to personal communication with M.S. Ravi, a proof of this result may be expected in the near future. The conjecture can be seen as an extension of Proposition 2.2.5 on the genericity of reachability for systems over polynomial rings. A careful study of the proof of this proposition in [67], leads to the conclusion that almost the same ideas can be used to prove this more general result.

Conjecture 5.5.4 gives an indication how difficult the investigation of the stabilizability of a time-delay system is. Let $\Sigma = (A, B)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. If $k < m$, so if the number of incommensurable time-delays is strictly smaller than the number of inputs, $\mathcal{V}(\mathcal{I})$ is generically empty. Therefore the system $\Sigma = (A, B)$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$ is generically reachable.

This was already proved in [67]. If the number of incommensurable time-delays k is equal to the number of inputs m, the variety $\mathcal{V}(\mathcal{I})$ is generically zero-dimensional, and in this case the Algorithms 5.5.2 and 5.5.3 are applicable. Note that for systems with commensurable time-delays, we have $k = 1 \leq m$, and thus the testing of reachability is generically possible without computation of the zeros of an exponential polynomial. However, if $k > m$, so if the number of incommensurable time-delays is strictly larger than the number of inputs, the variety $\mathcal{V}(\mathcal{I})$ generically contains infinitely many points. In this situation the stabilizability test is much more involved, and we have to take recourse to rather difficult numerical computations to obtain a solution.

To illustrate the methods developed in this section, we end with two examples that illustrate the effectiveness of Algorithms 5.5.2 and 5.5.3.

Example 5.5.5 Let $\Sigma = (A, B)$ be a system with a commensurable time-delay τ, modeled as a system over the ring $\mathbb{R}[s]$. The indeterminate s corresponds to a time-delay operator with time-delay τ. The matrices A and B are given by

$$A = \begin{pmatrix} 1 & 4s^3 + 7s^2 + 8s + 2 \\ s^2 - s & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 3s + 2 \\ 2s^3 - s - 1 \end{pmatrix}.$$

To test the stabilizability of this system, we compute a Gröbner basis of the ideal $\mathcal{J}$ associated with Σ, w.r.t. the pure lexicographic term ordering, with the ranking of the indeterminates fixed by $z \prec s$. Using the computer algebra package Maple V.2 for the actual computations, we obtain the Gröbner basis $J = \{j_1, j_2\}$ of $\mathcal{J}$, where

$$\begin{aligned} j_1 &= 2120685s - 2441024z^8 - 636640z^7 + 6049648z^6 - 339936z^5 + \\ &\quad -19183784z^4 - 2756516z^3 + 21327952z^2 + 100985z - 2120685, \\ j_2 &= z^2(z-1)(32z^6 + 80z^5 + 16z^4 - 72z^3 + 172z^2 + 508z + 309). \end{aligned}$$

The Gröbner basis J shows that already in relatively small and simple examples, the polynomials in the Gröbner basis of $\mathcal{J}$ may have high degree. Also the coefficients may grow very rapidly. However, in this particular example this last effect can be eliminated by a normalization of the coefficients, because they are all in the same order of magnitude.

Although the Gröbner basis J looks rather complicated, a conclusion on the stabilizability of the time-delay system Σ is not difficult to obtain. It is easily verified that the points $(\hat{z}, \hat{s}) = (0, 1)$ and $(\bar{z}, \bar{s}) = (1, 0)$ are elements of $\mathcal{V}(\mathcal{J})$. Since $0 \notin \mathbb{C}^-$, and $e^{-\tau 0} = 1$, the point $(0, 1)$ is always an element of the set W_g corresponding to the stability domain $\mathbb{C}_g$, independent of the actual value of τ and the choice of stability domain $\mathbb{C}_g$. So for all values τ of the length of the time-delay, the system Σ is <u>not</u> stabilizable.

Note that if the point $(\bar{z}, \bar{s}) = (1, 0)$ would have been the only element of $\mathcal{V}(\mathcal{J})$, then the corresponding system would have been stabilizable independent of the value of τ because for all $\tau > 0$: $e^{-\tau} \neq 0$. Therefore $(1, 0) \notin W_g$ for any arbitrary stability domain $\mathbb{C}_g$. However, in this situation the system is not pointwise stabilizable because $(1, 0)$ is an element of the set W_p defined in (5.43).

Example 5.5.6 Consider a system $\Sigma = (A, B)$ with two incommensurable time-delays $\tau_1 = 1$ and $\tau_2 = \pi$. The system is modeled as a system over the ring $\mathbf{R}[s_1, s_2]$, where the indeterminates s_1 and s_2 correspond to time-delay operators with time-delays τ_1 and τ_2, respectively. The matrices A and B are given by

$$A = \begin{pmatrix} s_1+1 & s_2+s_1 & s_2-s_1+3 \\ s_2-1 & s_2+1 & s_1-5 \\ s_1^2+1 & 1 & s_2+1 \end{pmatrix}, \qquad B = \begin{pmatrix} s_2+s_1 & 0 \\ 1 & s_1-3 \\ 0 & s_2+2 \end{pmatrix}.$$

We start with the computation of a Gröbner basis of the ideal $\mathcal{J}$ associated with Σ, w.r.t. the graded lexicographic term ordering. In this way we obtain a Gröbner basis that looks rather difficult, but satisfies condition *(ii)* of Proposition 4.1.30. Therefore the variety $\mathcal{V}(\mathcal{J})$ is zero-dimensional, and we may apply Algorithm 5.5.3 to test the stabilizability of the system Σ with respect to the stability domain $\mathbb{C}_g = \mathbb{C}^-$.

The Gröbner basis of the ideal $\mathcal{J}$ can be used in the construction method of Proposition 4.1.37 to compute a univariate polynomial $p \in \mathcal{J}$ in the indeterminate z, that is of minimal degree. In the computer algebra package Maple V.2 this method is implemented in the function *finduni* of the *grobner*-package. After application of this function we obtain the following polynomial p:

$$\begin{aligned} p \;=\; & (2z-5)(2z^{12}+21z^{11}-235z^{10}-1136z^9-73z^8+9966z^7+20199z^6+ \\ & -11053z^5-87368z^4-89565z^3+51062z^2+156327z+83361). \end{aligned}$$

All 13 zeros of the polynomial p are simple, and five of them are elements of $\overline{\mathbb{C}^+}$. Let Λ denote this set of closed right half plane zeros of p. For each $\lambda \in \Lambda$ we substitute $(\lambda, e^{-\lambda}, e^{-\pi\lambda})$ for the indeterminates (z, s_1, s_2) in the matrix $(zI - A(s_1, s_2)|B(s_1, s_2))$, and compute the three singular values of the matrix $(\lambda I - A(e^{-\lambda}, e^{-\pi\lambda}) \mid B(e^{-\lambda}, e^{-\pi\lambda}))$. The results are listed in the table given below. From

RHP-zero	singular values		
1.3720	0.69586	7.5180	39.662
2.2186	0.69920	11.223	42.780
2.3808	0.85132	12.043	43.438
2.5000	0.98710	12.666	43.948
8.9196	36.191	80.184	124.78

these computations we see that for every $\lambda \in \Lambda$, the point $(\lambda, e^{-\lambda}, e^{-\pi\lambda}) \in \mathcal{W}_g$ is not an element of $\mathcal{V}(\mathcal{J})$ because the corresponding singular values stay away from zero. So $\mathcal{V}(\mathcal{J}) \cap \mathcal{W}_g = \varnothing$, and the time-delay system Σ is stabilizable w.r.t. $\mathbb{C}^-$.

Remark 5.5.7 The methods to test stabilizability developed in this section are mainly based on the computation of the varieties of polynomial ideals; the ideals themselves are only of secondary interest. Therefore the characteristic sets algorithm might be a good alternative for the required Gröbner basis computations, especially in Algorithm 5.5.2.

Chapter 6

Stabilization of time-delay systems

This chapter is devoted to the constructive part of the stabilization problem. In Chapter 3 and 5 we already considered several aspects of the stabilizability question, but only addressed the existence issue. We concluded that a time-delay system is stabilizable if and only if

$$\mathcal{I} \cap \mathcal{D}_g \neq \varnothing,$$

where $\mathcal{I}$ is the ideal associated with the system Σ introduced in Definition 5.1.1, and $\mathcal{D}_g$ is the Hurwitz set corresponding to the stability domain $\mathbb{C}_g$. Moreover, we showed that this condition may be verified without computing an element of $\mathcal{I} \cap \mathcal{D}_g$ explicitly. However, this is only sufficient to answer the existence question; for the construction of a stabilizing feedback compensator, a polynomial in $\mathcal{I} \cap \mathcal{D}_g$ is required. In this chapter we study this more practical part of the stabilization problem, and give an overview of some of the methods known in literature for the construction of stabilizing compensators. It is not our goal to treat these methods in full detail. We confine ourselves to the main ideas behind the different approaches, and discuss their shortcomings and advantages. In this way we get some insight in the considerations that are important for the design of stabilizing feedback compensators.

Before addressing the stabilization problem, it seems natural to consider another question first: how is it verified that a time-delay system is internally stable w.r.t. a Hurwitz set $\mathcal{D}_g$? In Section 6.1 we study this problem and develop a numerical method to test the stability of the characteristic polynomial of a system. The same method can be applied to test whether a given polynomial p belongs to $\mathcal{I} \cap \mathcal{D}_g$. Using the Gröbner basis algorithm, the membership problem for the ideal $\mathcal{I}$ is not difficult to solve, and the stability may be verified with the stability test for characteristic polynomials. The algorithm is also important in subsequent sections, where it is used in the design of stabilizing feedback compensators.

Next we return to the stabilization problem itself. In Section 6.2 we discuss a universally applicable method for stabilizing time-delay systems with commensurable time-delays. It is based on the construction of a BIBO-stabilizing compensator that is approximated afterwards by finite-dimensional controllers. In Section 6.3 some alternative methods are proposed. In one of them, the infinite-dimensional systems approach is used to model time-delay systems, but nevertheless it yields

finite-dimensional stabilizing compensators. We also present another method that fits better in the algebraic framework of this thesis. It is based on the results on pole placement in Section 2.6, but in this method, the pole-placement techniques are used in a different way, which makes them applicable in a more general context. We also discuss the advantages and drawbacks of the different methods and indicate what is the most appropriate choice in some particular cases.

Although most results of this chapter may be generalized to arbitrary stability domains $\mathbb{C}_g$ satisfying conditions *(i)* to *(iv)* of Definition 3.1.2, we confine ourselves in the major part of this chapter to the classical notion of stability, i.e. to the stability domain $\mathbb{C}^-$.

6.1 A stability test for exponential polynomials

Consider a system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system $\Sigma = (A, B, C, D)$ over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. According to Definition 2.5.3 *(i)*, this system is internally stable w.r.t. the stability domain $\mathbb{C}_g$ if and only if

$$\det(zI - A) \in \mathcal{D}_g,$$

where $\mathcal{D}_g$ denotes the Hurwitz set corresponding to the stability domain $\mathbb{C}_g$. Recalling formula (3.5), $\mathcal{D}_g$ is given by

$$\mathcal{D}_g := \{p(z, s_1, \ldots, s_k) \in \mathbb{R}[z, s_1, \ldots, s_k] \mid p(z, s_1, \ldots, s_k) \text{ is monic in } z \text{ and } \forall z \in \mathbb{C}\backslash\mathbb{C}_g : p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \neq 0\}.$$

Note that $p(z, s_1, \ldots, s_k) := \det(zI - A)$ is a polynomial in $k+1$ indeterminates that is monic in the indeterminate z. So $p(z, s_1, \ldots, s_k)$ is of the form

$$p(z, s_1, \ldots, s_k) = z^n + \sum_{i=0}^{n-1} p_i(s_1, \ldots, s_k) \cdot z^i, \tag{6.1}$$

where $p_i(s_1, \ldots, s_k) \in \mathbb{R}[s_1, \ldots, s_k]$ $(i = 0, 1, \ldots, n-1)$. When we substitute $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$ for the indeterminates $s_1, \ldots, s_k$ we obtain

$$f(z) := p(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) = z^n + \sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \cdot z^i.$$

Definition 6.1.1 An *exponential polynomial* is an analytic function of the form

$$f : \mathbb{C} \to \mathbb{C} : \qquad f(z) = z^n + \sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \cdot z^i, \tag{6.2}$$

where for all $i \in \{0, 1, \ldots, n-1\}$ the function $p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$ is a polynomial in the variables $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$ with real coefficients, and all $\tau_j > 0$ $(j = 1, \ldots, k)$. The degree n of the monic leading term of $f(z)$ is called the *degree* of the exponential polynomial f.

The name exponential polynomial is mainly motivated by the strong relationship with polynomials in more than one indeterminate as explained above. However, another reason might be that in the right half plane exponential polynomials behave similar to ordinary polynomials. Therefore exponential polynomials are often called *quasi polynomials.*

Definition 6.1.2 An exponential polynomial f is called *stable* w.r.t. the stability domain $\mathbf{C}_g$ if it has no zeros outside $\mathbf{C}_g$:

$$\forall z \in \mathbf{C} \backslash \mathbf{C}_g : \ f(z) \neq 0.$$

Clearly, this definition of stability for exponential polynomials coincides with the notion of stability for time-delay systems. A time-delay system is internally stable w.r.t. the stability domain $\mathbf{C}_g$ if and only if the exponential polynomial corresponding to the characteristic polynomial of the system is stable w.r.t. $\mathbf{C}_g$. In other words, testing the stability of a delay system comes down to testing the stability of an exponential polynomial.

In this section, we derive a method for carrying out this stability test. For this purpose, we confine ourselves to the stability domain $\mathbf{C}^-$. In this particular situation, exact analytic conditions to test the stability of an exponential polynomial are known in literature (see [3, Chapter 13] or the original paper [77] by Pontryagin for the commensurable delay case, and [92] for the incommensurable delay case). However, in somewhat more complicated examples these conditions become very difficult and are impossible to check. Therefore, mostly a graphical test based on the well-known circle criterion is used instead. Note that from sheer necessity, a numerical algorithm for carrying out this test is based on the computation of only a finite number of points on the curve to which the circle criterion is applied. Therefore it cannot be guaranteed that always a correct result is obtained. In this section, we present an algorithm that solves these problems to a great extent, and yields a reliable answer in a reasonable amount of computer time. However, we start with a short introduction to the circle criterion itself, emphasizing the special form this criterion takes for exponential polynomials.

6.1.1 The circle criterion for exponential polynomials

The circle criterion is a well-known result from complex analysis, used for determining the number of zeros of an analytic function in an area enclosed by a Jordan curve. In this subsection, this criterion is specialized to the case of exponential polynomials. Given an exponential polynomial f, it is shown how the number of right half plane zeros of f is determined using the image of the imaginary axis under the function f.

The first lemma may be considered as a generalization of Lemma 3.3.33 to systems with incommensurable time-delays. However, the result is not formulated in terms of the characteristic polynomial of a time-delay system, but stated directly as a property of exponential polynomials.

Lemma 6.1.3 *An exponential polynomial has only a finite number of zeros in the closed right half plane* $\overline{\mathbf{C}^+}$.

Proof

Consider an arbitrary exponential polynomial f given by

$$f(z) = z^n + \sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \cdot z^i.$$

Since $p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$ is a polynomial in $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$ for $i = 0, 1, \ldots, n-1$, and since these exponential functions are bounded in $\overline{\mathbb{C}^+}$, we know that also $|p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})|$ is bounded in $\overline{\mathbb{C}^+}$. So, if $\mathrm{Re}\,(z) \geq 0$ and $|z|$ becomes large, then the term z^n is the dominant term of f. Therefore there exists an $R \in \mathbb{R}$ such that

$$\forall z \in \overline{\mathbb{C}^+}, |z| > R: \ |f(z)| \geq |z|^n - \sum_{i=0}^{n-1} |p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})| \cdot |z|^i > 0.$$

Hence, all right half plane zeros of f lie in the half disk $D = \{z \in \mathbb{C} \mid \mathrm{Re}\,(z) \geq 0, |z| \leq R\}$. Since f is an analytic function and D is a compact set, f has only a finite number of zeros inside D. This proves the claim. ∎

Note that exactly the same arguments can be used to prove that an exponential polynomial only has a finite number of zeros in any arbitrary closed right half plane $\mathbb{C}_\alpha^+ = \{z \in \mathbb{C} \mid \mathrm{Re}\,(z) \geq \alpha\}$. The main point is that in such a half plane the exponential functions $e^{-\tau_j z}$ $(j = 1, \ldots, k)$ are always bounded from above.

The exact number of right half plane zeros of an exponential polynomial f can be determined with the following well-known result from complex analysis (see for example [83, p. 225]).

Proposition 6.1.4 *Let J be a Jordan curve in the complex plane. Consider a function f that is analytic inside and on the curve J. Assume that f has no zeros on J. Then the number of zeros of f inside J (counting multiplicities) is given by*

$$\frac{1}{2\pi i} \oint_J \frac{f'(z)}{f(z)} dz. \tag{6.3}$$

∎

We shall apply Proposition 6.1.4 to exponential polynomials, with a Jordan curve J of the following form.

Definition 6.1.5 Let $R \in \mathbb{R}^+$. Then the half circle C_R is defined as

$$C_R := \{z \in \mathbb{C} \mid |z| = R \wedge \mathrm{Re}\,(z) \geq 0\},$$

the part I_R of the imaginary axis as

$$I_R := \{z \in \mathbb{C} \mid \mathrm{Re}\,(z) = 0 \wedge |z| < R\},$$

and the Jordan curve J_R as

$$J_R := C_R \cup I_R.$$

This Jordan curve is traversed in counter clockwise direction as depicted in Figure 6.1.

Consider an exponential polynomial f, and assume that f has no zeros on the imaginary axis. Let N_f denote the number of zeros of f in $\mathbb{C}^+$, counting multiplicities. When R becomes large enough, J_R will enclose all N_f zeros of f. Hence

$$N_f = \lim_{R \to \infty} \frac{1}{2\pi i} \oint_{J_R} \frac{f'(z)}{f(z)} dz. \tag{6.4}$$

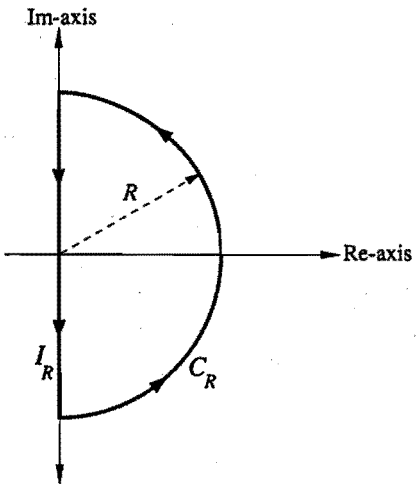


Figure 6.1: The Jordan curve J_R

Of course it is possible to give an upper bound for the value of R that guarantees that the corresponding Jordan curve J_R contains all right half plane zeros of f. For this purpose we may use similar techniques as in Section 3.3, where we elaborated on the relationship between the norm of square polynomial matrices $A(s)$ and the location of the zeros of the corresponding characteristic function $\det(zI - A(e^{-\tau z}))$. However, if we use this approach, we are compelled to compute the integral (6.3) explicitly, and this may lead to rather involved numerical computations.

Instead we choose a different approach. First we split the integral (6.4) into two parts: the integral over the half circle C_R, and over the imaginary axis I_R. When R tends to infinity, the integral over the half circle may be computed analytically. In combination with a method to determine the integral over the imaginary axis, this yields a circle criterion for exponential polynomials.

We start with the computation of the integral over the half circle C_R. For this we need some preliminary lemmas.

Lemma 6.1.6 *Let f be an exponential polynomial of the form*

$$f(z) = z^n + \sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) z^i.$$

Define

$$A(z) := \frac{d}{dz}(p_{n-1}(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})). \tag{6.5}$$

Then for large values of $|z|$, such that $\mathrm{Re}\,(z) \geq 0$ we have

$$\frac{f'(z)}{f(z)} = \frac{n}{z} + \frac{A(z)}{z} + O(\frac{1}{z^2}). \tag{6.6}$$

Proof

Since all functions $p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$ are polynomials in $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$, they are bounded in $\overline{\mathbb{C}^+}$, and the same is true for their derivatives. Hence $A(z)$ is also bounded in $\overline{\mathbb{C}^+}$. If $z \in \overline{\mathbb{C}^+}$ and $|z|$ becomes large, this implies that

$$f'(z) = nz^{n-1} + A(z) \cdot z^{n-1} + O(z^{n-2}).$$

So

$$\begin{aligned} \frac{f'(z)}{f(z)} - \frac{n}{z} - \frac{A(z)}{z} &= \frac{zf'(z) - nf(z) - A(z)f(z)}{zf(z)} = \\ &= \frac{nz^n + A(z)z^n - nz^n - A(z)z^n + O(z^{n-1})}{z^{n+1} + O(z^n)} = O(\frac{1}{z^2}). \end{aligned}$$

■

Lemma 6.1.7 *Let $\alpha \in \mathbf{R}$, $\alpha > 0$. Then*

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{C_R} \frac{e^{-\alpha z}}{z} dz = 0. \tag{6.7}$$

Proof

$$\left| \frac{1}{2\pi i} \int_{C_R} \frac{e^{-\alpha z}}{z} dz \right| = \left| \frac{1}{2\pi} \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} e^{-\alpha R(\cos(\omega) + i \sin(\omega))} d\omega \right| \leq \frac{1}{\pi} \int_0^{\frac{1}{2}\pi} e^{-\alpha R \cos(\omega)} d\omega.$$

Since $\alpha > 0$, the inequality $-\alpha R \cos(\omega) \leq \alpha R(\frac{2}{\pi}\omega - 1)$ holds on the whole interval $[0, \frac{\pi}{2}]$. Hence:

$$\frac{1}{\pi} \int_0^{\frac{1}{2}\pi} e^{-\alpha R \cos(\omega)} d\omega \leq \frac{1}{\pi} \int_0^{\frac{1}{2}\pi} e^{\alpha R(\frac{2}{\pi}\omega - 1)} d\omega = \frac{1}{2\alpha R}(1 - e^{-\alpha R}) \longrightarrow 0,$$

when R tends to infinity. ■

Proposition 6.1.8 *Let f be an exponential polynomial of degree n as defined in (6.2). Then*

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{C_R} \frac{f'(z)}{f(z)} dz = \frac{n}{2}. \tag{6.8}$$

Proof

Since we are interested in the asymptotic behaviour of (6.8) for $R \to \infty$, while the integration variable z remains in $\overline{\mathbb{C}^+}$, we may use (6.6) to prove (6.8). First of all,

$$\frac{1}{2\pi i} \int_{C_R} \frac{n}{z} dz = \frac{n}{2\pi} \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} d\omega = \frac{n}{2}. \tag{6.9}$$

$A(z)$ was defined in (6.5) as the derivative of $p_{n-1}(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$, where p_{n-1} is considered as a polynomial in $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$. Therefore $A(z)$ is a linear combination of functions of the form $e^{-\alpha z}$, with $\alpha > 0$. Applying Lemma 6.1.7, we obtain

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{C_R} \frac{A(z)}{z} dz = 0. \tag{6.10}$$

Finally,

$$\lim_{R\to\infty}\frac{1}{2\pi\imath}\int_{C_R}\frac{1}{|z|^2}dz=\lim_{R\to\infty}\frac{1}{2\pi R}\int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi}e^{\imath\omega}d\omega=\lim_{R\to\infty}\frac{1}{\pi R}=0. \tag{6.11}$$

Combination of (6.6), (6.9), (6.10) and (6.11) yields the required result. ∎

Using the previous proposition it is possible to rewrite equality (6.4). In this way we obtain the following expression for the number of zeros in $\overline{\mathbb{C}^+}$ of an exponential polynomial f of degree n:

$$N_f=\frac{n}{2}+\lim_{R\to\infty}\frac{1}{2\pi\imath}\int_{I_R}\frac{f'(z)}{f(z)}dz=\frac{n}{2}-\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{f'(\imath\omega)}{f(\imath\omega)}d\omega. \tag{6.12}$$

To compute N_f we need a method to determine the second term in (6.12). The next result indicates that this can be done by inspection of the image of the imaginary axis under the function f.

Theorem 6.1.9 *Let f be an exponential polynomial of degree n as defined in (6.2). Assume that $f(0)>0$ and that f has no zeros on the imaginary axis. Let N_f denote the number of zeros of f in $\mathbb{C}^+$, counting multiplicities. Then*

$$N_f=\frac{n}{2}-\frac{1}{\pi}\cdot\operatorname{totarg}(f(\imath\infty)), \tag{6.13}$$

where $\operatorname{totarg}(f(\imath\infty))$ *is the net increase of the argument of $f(z)$ when z traverses the imaginary axis from $z=0$ to $z=\imath\infty$.*

Proof

Since the function f is a real function on the real axis, the second term of (6.12) may be written as

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{f'(\imath\omega)}{f(\imath\omega)}d\omega=\frac{1}{\pi}\int_0^{\infty}\operatorname{Re}\left(\frac{f'(\imath\omega)}{f(\imath\omega)}\right)d\omega.$$

Define $u(\omega):=\operatorname{Re}(f(\imath\omega))$ and $v(\omega):=\operatorname{Im}(f(\imath\omega))$. Then

$$\frac{1}{\pi}\int_0^{\infty}\operatorname{Re}\left(\frac{f'(\imath\omega)}{f(\imath\omega)}\right)d\omega=\frac{1}{\pi}\int_0^{\infty}\frac{v'(\omega)u(\omega)-u'(\omega)v(\omega)}{u^2(\omega)+v^2(\omega)}d\omega=$$

$$=\frac{1}{\pi}\int_0^{\infty}\frac{d}{d\omega}\arctan\left(\frac{v(\omega)}{u(\omega)}\right)=\frac{1}{\pi}\int_0^{\infty}\frac{d}{d\omega}\arctan(\tan(\arg(f(\imath\omega))))=$$

$$=\frac{1}{\pi}\cdot(\operatorname{totarg}(f(\imath\infty))-\arg(f(0)))=\frac{1}{\pi}\operatorname{totarg}(f(\imath\infty)).$$

Substitution of this formula in (6.12) completes the proof. ∎

Theorem 6.1.9 gives rise to a graphical method for computing the number of right half plane zeros N_f of an exponential polynomial f. The image of the positive imaginary axis under the function f is required to determine the final result. The rest of this section is devoted to the question how this search along the positive imaginary axis and its image can be carried out in a numerically reliable and efficient way.

6.1.2 Bounds on the search along the imaginary axis

Let f be an exponential polynomial of degree n as defined in (6.2). Recalling the proof of Lemma 6.1.3, we know that the term z^n is the dominant term of this function f when the modulus of $z \in \overline{\mathbf{C}^+}$ is large. This implies that $\arg(f(\imath\omega))$ is convergent for ω tending to infinity because with growing ω the term $(\imath\omega)^n$ becomes more and more dominant. It is easily verified that

$$\lim_{\omega\to\infty} \arg(f(\imath\omega)) = \frac{n \bmod 4}{2} \cdot \pi. \tag{6.14}$$

Note however that the function $f(\imath\omega)$ $(\omega \geq 0)$ does not converge itself.

From the asymptotic behaviour of the argument of $f(\imath\omega)$, it follows that the function $f(\imath\omega)$ $(\omega \geq 0)$ has a *half plane of convergence* in the following sense: there exists a $K \in \mathbf{R}$ such that for all $\omega > K$, the value of $f(\imath\omega)$ remains in the half plane determined by the degree of the exponential polynomial f, as described in (6.14). Choose this K as small as possible:

$$K := \min\{\beta \in \mathbf{R} \mid \forall \omega > \beta : |\mathrm{totarg}(f(\imath\omega)) - \mathrm{totarg}(f(\imath\infty))| \leq \frac{\pi}{2}\}. \tag{6.15}$$

Then it follows that at $\omega = K$, i.e. at the moment that $f(\imath\omega)$ enters the half plane of convergence for the last time to remain there forever, the value of $\mathrm{totarg}(f(\imath\infty))$ is completely known: the difference with $\mathrm{totarg}(f(\imath K))$ is exactly $\frac{\pi}{2}$.

We conclude that for the computation of $\mathrm{totarg}(f(\imath\infty))$ only the behaviour of $f(\imath\omega)$ for $\omega \in [0, K]$ is important. Thus for an efficient test, we should look for a sharp upper bound $K_{\max}$ for K. In this way we obtain a search interval $[0, K_{\max}]$ that is as small as possible, but still contains all information that is required for the determination of the number of right half plane zeros of f.

Recall that the exponential polynomial f is of the form (6.2), and let K_1 be a positive real number such that

$$\forall \omega > K_1 : |\imath\omega|^n > \left|\sum_{i=0}^{n-1} p_i(e^{-\imath\tau_1\omega}, \ldots, e^{-\imath\tau_k\omega})(\imath\omega)^i\right|. \tag{6.16}$$

On the imaginary axis, z^n is the dominant term of f, and therefore such a $K_1 \in \mathbf{R}$ exists. For all $\omega > K_1$, the modulus of the first term $(\imath\omega)^n$ of $f(\imath\omega)$ is so large that this term determines the half plane in which $f(\imath\omega)$ is located. By definition, this is the half plane of convergence. So $K_1 \geq K$, and thus formula (6.16) may be used to obtain an upper bound $K_{\max}$ for K.

Proposition 6.1.10 *Let f be an exponential polynomial of degree n as defined in (6.2) and suppose that $n \geq 2$. Let $\alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in \mathbf{R}$ be such that*

$$\forall i \in \{0, 1, \ldots, n-1\}\ \forall \omega \in \mathbf{R} : |p_i(e^{-\imath\tau_1\omega}, \ldots, e^{-\imath\tau_k\omega})| \leq \alpha_i \tag{6.17}$$

(Since all p_i are polynomials in $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$, and for $z = \imath\omega$ we have $|e^{-\imath\tau_j\omega}| = 1$, an upper bound for α_i can be obtained by summation of all absolute values of the coefficients of p_i). Define

$$\alpha_{\max,2} := \max\{\alpha_i \mid i = 0, 1, \ldots, n-2\}, \quad K_{\max,2} := \sqrt{\alpha_{\max,2}} + \max(1, \alpha_{n-1}).$$

Then $K_{\max,2}$ is an upper bound for K.

Proof

It suffices to show that $K_{\max,2}$ satisfies (6.16). Let $\omega > K_{\max,2}$ and $z = w\omega$. Then $|z| - 1 > \sqrt{\alpha_{\max,2}}$ and we have

$$\left|\sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) z^i\right| \leq \sum_{i=0}^{n-1} \left|p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\right| |z|^i \leq$$

$$\alpha_{n-1}|z|^{n-1} + \sum_{i=0}^{n-2} \alpha_i |z|^i \leq \alpha_{n-1}|z|^{n-1} + \alpha_{\max,2} \sum_{i=0}^{n-2} |z|^i =$$

$$\alpha_{n-1}|z|^{n-1} + \alpha_{\max,2} \frac{|z|^{n-1} - 1}{|z| - 1} \leq \alpha_{n-1}|z|^{n-1} + \sqrt{\alpha_{\max,2}}(|z|^{n-1} - 1) \leq$$

$$\leq (\alpha_{n-1} + \sqrt{\alpha_{\max,2}})|z|^{n-1} \leq K_{\max,2}|z|^{n-1} < |z|^n. \qquad \blacksquare$$

For exponential polynomials of high degree and with large coefficients it is possible to derive sharper bounds for K. The next proposition states two of these alternative results.

Proposition 6.1.11 *Let f be an exponential polynomial of degree n as defined in (6.2). Choose $\alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$ such that*

$$\forall i \in \{0, 1, \ldots, n-1\}\ \forall \omega \in \mathbb{R} : \left|p_i(e^{-i\tau_1\omega}, \ldots, e^{-i\tau_k\omega})\right| \leq \alpha_i.$$

(i) If $n \geq 3$, and $\alpha_{\max,3}$ is defined as $\alpha_{\max,3} := \max\{\alpha_i \mid i = 0, 1, \ldots, n-3\}$, then $K_{\max,3} := \sqrt[3]{\alpha_{\max}} + \max(1, \alpha_{n-1} + \sqrt{\alpha_{n-2}})$ is an upper bound for K.

(ii) If $n \geq 4$, define $\alpha_{\max,4} := \max\{\alpha_i \mid i = 0, 1, \ldots, n-4\}$, and let ξ be the largest positive real solution of the polynomial equation

$$x^3 - \alpha_{n-1}x^2 - \alpha_{n-2}x - (\alpha_{n-3} + (\alpha_{\max,4})^{\frac{3}{4}}) = 0. \tag{6.18}$$

Then $K_{\max,4} := \max(1 + \sqrt[4]{\alpha_{\max,4}}, \xi)$ is an upper bound for K.

Proof

We only give a proof of *(ii)*; the proof of *(i)* is based on exactly the same considerations and is therefore omitted.

(ii) In the same way as in the proof of Proposition 6.1.10 it suffices to show that $K_{\max,4}$ satisfies (6.16). Let $\omega > K_{\max,4}$ and $z = w\omega$. Then $|z| - 1 > \sqrt[4]{\alpha_{\max,4}}$ and we have

$$\left|\sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) z^i\right| \leq$$

$$\leq \alpha_{n-1}|z|^{n-1} + \alpha_{n-2}|z|^{n-2} + \alpha_{n-3}|z|^{n-3} + \alpha_{\max,4} \cdot \sum_{i=0}^{n-4} |z|^i =$$

$$= \alpha_{n-1}|z|^{n-1} + \alpha_{n-2}|z|^{n-2} + \alpha_{n-3}|z|^{n-3} + \alpha_{\max,4} \frac{|z|^{n-3} - 1}{|z| - 1} \leq$$

$$\leq \left(\alpha_{n-1}|z|^2 + \alpha_{n-2}|z| + (\alpha_{n-3} + (\alpha_{\max,4})^{\frac{3}{4}})\right)|z|^{n-3}.$$

Since $|z| > \xi$, and ξ is the largest positive real solution of (6.18), it follows that

$$\left(\alpha_{n-1}|z|^2 + \alpha_{n-2}|z| + (\alpha_{n-3} + (\alpha_{\max,4})^{\frac{3}{4}})\right) < |z|^3.$$

Hence $\left|\sum_{i=0}^{n-1} p_i(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})z^i\right| \leq |z|^n$, and thus (6.16) is satisfied. ∎

In the same way we may continue. Based on formula (6.16), smaller upper bounds for K may be derived by taking more terms α_i separately into account. It depends on the exponential polynomial f under consideration which estimation method is the best. Clearly, the minimum of all computed upper bounds is chosen as the final value of $K_{\max}$.

6.1.3 Determination of the number of RHP-zeros

In the previous subsection it has been shown that the number of right half plane zeros of an exponential polynomial f is determined by the behaviour of the function $f(\imath\omega)$ on the finite interval $[0, K]$. Moreover, an upper bound $K_{\max}$ for K was derived. This subsection is devoted to the question how totarg$(f(\imath\infty))$ can be computed with help of the image $f(\omega)$ of f on the interval $\omega \in [0, K_{\max}]$. For this purpose we assume that f is an exponential polynomial of degree n, without zeros on the imaginary axis, and satisfying $f(0) > 0$.

Recalling formula (6.15), it is apparent that if totarg$(f(\imath K_{\max}))$ is known, the value of totarg$(f(\imath\infty))$ is easily derived: totarg$(f(\imath\infty))$ is the sum of arg$(f(\imath K_{\max}))$ and 2π times the number of complete encirclements of the curve $\Gamma = \{f(\imath\omega) \mid \omega \in [0, K_{\max}]\}$ of the origin of the complex plane. In order to count these encirclements, we have to track the curve Γ parametrized by ω, from $\omega = 0$ to $\omega = K_{\max}$. However, for the determination of the number of encirclements of the origin, it is only interesting to know where the curve Γ enters and leaves the half plane of convergence and in what direction. Since the boundary of this half plane is the real axis (when the degree n of the exponential polynomial is odd) or the imaginary axis (when n is even), only intersections with these axes have to be considered. Then one may verify that an intersection with one of the axes contributes to the number of encirclements as depicted in Figure 6.2.

The only problem left is to track the curve Γ of $f(\imath\omega)$ for $\omega \in [0, K_{\max}]$ accurately enough to ensure that all intersections with the real axis (when n is odd) or the imaginary axis (when n is even) are detected. However, in a numerical algorithm we have to confine ourselves to the computation of only a finite number of points on the curve Γ. Therefore the tracking problem may be reformulated as the question of finding a method for the selection of a finite number of points on the curve Γ in such a way that all intersections of Γ with the real and imaginary axis can be detected from this finite set of points. For this purpose, linear search in the parameter space with constant step length ℓ is most commonly used up to now. Unfortunately, this method is not always reliable as is illustrated by an example in the next section. Therefore we propose another, more reliable method to overcome this problem. The

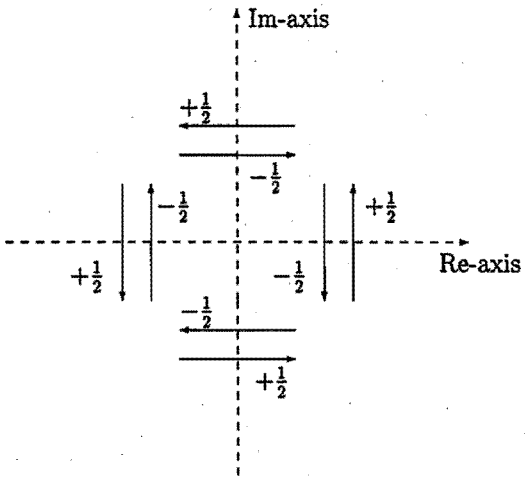


Figure 6.2: Counting the number of encirclements

main idea is to make the step length variable depending on the curvature of the curve Γ.

For a curve $\Gamma = \{(u(\omega), v(\omega)) \mid \omega \in [0, K]\}$ in a two-dimensional plane, that is parametrized by the variable ω, the *curvature* in a point $\bar{x}(\omega_0) = (u(\omega_0), v(\omega_0))$ at $\omega = \omega_0$, is given by

$$\frac{\dot{u}(\omega_0)\ddot{v}(\omega_0) - \dot{v}(\omega_0)\ddot{u}(\omega_0)}{(\dot{u}^2(\omega_0) + \dot{v}^2(\omega_0))^{3/2}}. \tag{6.19}$$

(See for example [88, pp. 13-15]; the formula can be found literally in [6, pp. 589-590]). The curvature in a point $\bar{x}(\omega_0)$ on Γ is a measure for the rate of change of the tangent in $\bar{x}(\omega_0)$, when proceeding along the curve. For example, in every point of a circle with radius R the curvature is equal to $\frac{1}{R}$. For arbitrary curves in $\mathbf{R}^2$ it is easily seen that if the curvature in a point $\bar{x}$ on the curve is equal to $k_{\bar{x}}$, then the curve Γ behaves in a small neighbourhood of $\bar{x}$ like a circle with radius $|\frac{1}{k_{\bar{x}}}|$. So, in order to track the curve accurately, we have to take small steps along the curve when the absolute value of the curvature is large, and we can take somewhat larger steps when the absolute value of the curvature is small. In this way we obtain the following rule:

$$|\text{curvature}| \times \text{step length along the curve} = \text{constant} \tag{6.20}$$

Note that the curve $\{(u(\omega), v(\omega)) \mid \omega \in [0, K]\}$ is parametrized by ω and not by its arc length. The length of the curve between ω_0 and $\omega_0 + \Delta\omega$ is given by

$$\int_{\omega_0}^{\omega_0+\Delta\omega} \sqrt{\dot{u}^2(\omega) + \dot{v}^2(\omega)}\, d\omega.$$

For small values of $\Delta\omega$, this integral is approximately $\sqrt{\dot{u}^2(\omega_0) + \dot{v}^2(\omega_0)}\Delta\omega$. Substitution of this formula and (6.19) in (6.20) yields

$$\frac{|\dot{u}(\omega_0)\ddot{v}(\omega_0) - \dot{v}(\omega_0)\ddot{u}(\omega_0)|}{\dot{u}^2(\omega_0) + \dot{v}^2(\omega_0)} \cdot \Delta\omega = C.$$

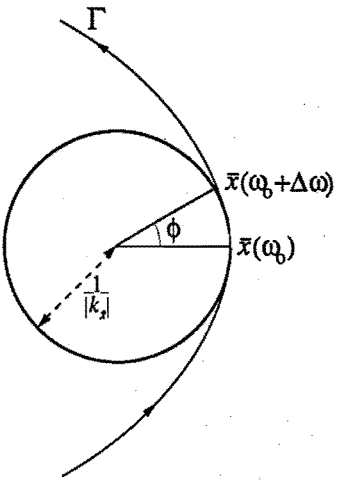


Figure 6.3: The relation between C and ϕ

In this way we obtain the following formula for the step length at $\omega = \omega_0$ in terms of the parameter ω:

$$\Delta\omega = C \cdot \frac{\dot{u}^2(\omega_0) + \dot{v}^2(\omega_0)}{|\dot{u}(\omega_0)\ddot{v}(\omega_0) - \dot{v}(\omega_0)\ddot{u}(\omega_0)|}. \tag{6.21}$$

However, in this formula there is still one unknown: the parameter C, which has the following interpretation.

Consider in Figure 6.3 the point $\bar{x}$ on the curve Γ. Assume that the curvature of Γ in $\bar{x}$ is $k_{\bar{x}}$. Then the curve Γ behaves in a neighbourhood of $\bar{x}$ like a circle with radius $\frac{1}{|k_{\bar{x}}|}$, as depicted in the figure. For small values of the angle ϕ, the step length along the curve Γ and along the circle are almost the same. So the step length along the curve Γ is approximately $\phi/|k_{\bar{x}}|$. Therefore $C = |k_{\bar{x}}| \cdot \frac{1}{|k_{\bar{x}}|} \cdot \phi = \phi$. Hence, C may be interpreted as the angle of rotation along the curve between two successively calculated points. A smaller choice of C leads to a more accurate tracking of the curve, but also to more computations. In practice one has to find a trade-off between accuracy and computational expenses. In our case, the accuracy has to be large enough to detect all intersections of the curve $\{f(\imath\omega) \mid \omega \in [0, K_{\max}]\}$ with the real and imaginary axis. The choice of the actual value of C is based on this condition, and on the interpretation of the parameter C as given above.

Remark 6.1.12 It is possible that in some points on the curve Γ the absolute value of the curvature is very small. In this case, the step length obtained with formula (6.21) is probably too large. This unwanted situation can be solved by defining an upper bound for the maximal step length beforehand. Moreover, in a practical implementation it is recommendable to restrict the growth of the step length by an (exponential) growth bound.

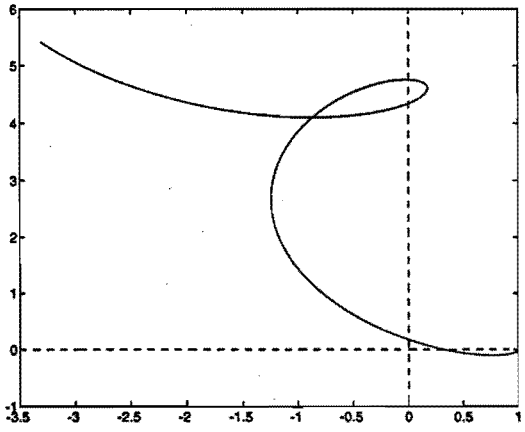


Figure 6.4: The curve Γ

6.1.4 Some examples

To illustrate the advantage of the variable step length method proposed in the previous subsection compared to linear search, we have implemented both algorithms in MATLAB. This can be done in a rather straightforward way. Given an exponential polynomial f, points on the curve $\Gamma = \{f(\imath\omega) \mid \omega \in [0, K_{\max}]\}$ are easily calculated, and formula (6.21) is used for the variable step length. Intersections with the real and imaginary axis are detected by a change of the signs of the imaginary and real parts of $f(\imath\omega)$ in two successive points. The rest of this section describes an experiment to compare the performances of both methods.

Example 6.1.13 Consider the following exponential polynomial:

$$f(z) = z^2 + (2 - e^{-z} - e^{-2z} - e^{-3z} - e^{-4z}) \cdot z + (2 - e^{-z}). \tag{6.22}$$

This exponential polynomial has degree $n = 2$, so the left half plane is the half plane of convergence. First we apply Proposition 6.1.10 to find an upper bound $K_{\max}$ for the search interval $[0, K]$. In the notation of Proposition 6.1.10, we have $\alpha_1 = 6$, $\alpha_{\max} = \alpha_0 = 3$. So $K_{\max} = K_{\max,2} = 6 + \sqrt{3}$. The interesting part of the curve $\Gamma = \{f(\imath\omega) \mid \omega \in [0, 6 + \sqrt{3}]\}$, for $\omega \in [0, 2]$, is depicted in Figure 6.4. The other part of the curve Γ, for $\omega \in [2, 6 + \sqrt{3}]$ is not very important; on this interval Γ remains in the left half plane. This indicates that the upper bound $K_{\max}$ for K is not very sharp. From Figure 6.4 it is immediately clear that $\text{totarg}(f(\imath\infty)) = \pi$ because the curve Γ crosses the *positive* imaginary axis on its way to the half plane of convergence. So, according to Theorem 6.1.9, the number N_f of right half plane zeros of f is given by:

$$N_f = \frac{n}{2} - \frac{1}{\pi} \cdot \text{totarg}(f(\imath\infty)) = \frac{2}{2} - \frac{1}{\pi} \cdot \pi = 0.$$

Hence f is a stable exponential polynomial.

The same conclusion on the number of right half plane zeros can be obtained by a linear search with step length 0.1 along the interval $[0, 6+\sqrt{3}]$. An intersection of the curve Γ with the imaginary axis is detected by a change of the signs of the real parts of two successive points. In this way only a half encirclement around the origin is counted, and in this way we obtain exactly the same result. Also the search method using variable step length, based on the curvature can be applied. In this simple case, application of this advanced method is not really necessary, but it yields the same result.

The advantage of the variable step length method becomes apparent when we want to verify the stability of exponential polynomials of high degree and with large coefficients.

Example 6.1.14 Consider the exponential polynomial

$$g(z) = (f(z))^6, \tag{6.23}$$

where f is defined by (6.22). This exponential polynomial has degree $n = 12$, so the right half plane is the half plane of convergence. Moreover, since f is a stable polynomial, g is stable too. However, in this case application of Theorem 6.1.9 is not so easy because it is very complicated to obtain the curve $\Gamma = \{g(i\omega) \mid \omega \in [0, K_{\max}]\}$ explicitly.

First we compute $K_{\max}$ using Proposition 6.1.11. With *(i)* the value $K_{\max,3} = 88.4$ is obtained. This upper bound is not very sharp; it can be improved a lot by *(iv)*. This method yields $K_{\max,4} = 45.3$, and thus we take $K_{\max} = 45.3$ as the upper bound for K in the subsequent computations.

To carry out a linear search along the interval $[0, 45.3]$, first the step length ℓ has to be determined. This is a rather difficult problem because a step length that is too large may lead to wrong conclusions on the number of right half plane zeros. Errors occur if some of the intersections with the real or imaginary axis are not detected. Therefore the choice of the step length ℓ is very critical. One of the main problems is that the possible errors made during the computation cannot be detected from the data afterwards. Moreover, the choice of an appropriate step length is dependent on the problem under consideration and almost impossible to predict beforehand. In the case of the exponential polynomial g, it turns out (using trial and error) that the step length must be smaller than or equal to $\ell = 0.008$. This implies that at least $\frac{45.3}{0.008} = 5663$ points on the curve Γ have to be computed to derive a correct conclusion on the number of right half plane zeros of g.

Application of the method with variable step length is much safer. Only the value of C has to be chosen. However, this parameter C has a clear interpretation as explained in Subsection 6.1.3. Therefore the choice it not so problem dependent. In most cases a choice of $C = 0.25 \approx \frac{\pi}{12}$ is appropriate. Also Remark 6.1.12 has to be taken into account. In this example we defined an upper bound of 0.25 for the step length, and imposed an exponential growth bound of 2. This means that the step length in a new step is at most twice as large as in the previous step. This method was started with an initial step length of 0.001, because especially at low frequencies the step length has to be very small. In this way, the number of right half plane zeros of g was correctly determined, computing only 631 points on the

curve Γ. However, the computational expenses for each step are much higher than in the linear search method. Therefore the variable step length method is not much faster than linear search. Its main advantage is the improved reliability: it takes small steps where this is required and somewhat larger steps where this is allowed.

6.1.5 Closing remarks

In this section we developed an algorithmic method for testing the stability of exponential polynomials. Although the method is based on the well-known circle criterion, the new contribution is the strategy that is used for the search along the imaginary axis. Replacing linear search by a curvature based variable step length method, the reliability of the stability test is increased considerably. This is very important for the verification of the stability of exponential polynomials of high degree. For low order exponential polynomials, this method is unnecessarily advanced; in this case the linear search method is probably good enough. However, in the next sections, high order exponential polynomials often occur in the construction of stabilizing feedback compensators for time-delay systems. Therefore, the variable step length method is a very important tool throughout the whole chapter.

6.2 A constructive approach to stabilization

In this section we present a constructive method for the stabilization of time-delay systems. According to the results in Sections 3.2 and 5.5, the computation of a polynomial in the intersection $\mathcal{I} \cap \mathcal{D}$ is required for this. Note that in this intersection the ideal $\mathcal{I}$ is completely characterized by the system $\Sigma = (A, B)$, whereas the the Hurwitz set $\mathcal{D}$ describes the notion of stability. In the proof of Theorem 3.2.8 we have seen how the construction of a polynomial in $\mathcal{I} \cap \mathcal{D}$ may be carried out. The proof of this result is completely constructive, except two parts:

- Theorem 3.2.3: the Nullstellensatz in the Banach algebra $\mathcal{A}(\Omega)$,
- Theorem 3.2.4: Mergelyan's Theorem on uniform approximation.

For the design of stabilizing compensators we need a constructive method to carry out these steps explicitly.

The approach to stabilization in this section contains both ingredients mentioned above. However, we shall not use the reformulation of the stabilization problem to the level of polynomials. Instead we look for a stable right-inverse of the matrix $(zI - A|B)$ in a more direct way using so-called Bezout factorizations. Considering the problem in a somewhat more general context, it is possible to construct a right-inverse of $(zI - A|B)$ that is not a matrix over $\mathcal{R}_{\mathcal{D}}(z)$, but still retains a specific stability property. Using an appropriate approximation of this right-inverse, we find a solution to our original problem: an internally stabilizing feedback compensator. Note that this method is based on the same ideas as the proof of Theorem 3.2.8. First the problem is solved in a more general framework, and afterwards this solution is approximated to obtain a solution in the original setting.

This section is organized as follows. In the first subsection the main concepts of the stabilization method described above are elaborated in more detail. The

next two subsections are devoted to the two constructive parts: computation of a Bezout factorization and approximation by rational functions. Finally the method is illustrated with an example.

The main ideas behind the approach to stabilization of time-delay systems as presented in this section originate from several papers by Kamen, Khargonekar and Tannenbaum. For further reading we refer to [54], [56] and [55].

Throughout this section we only consider stability in the classical sense, so the stability domain is always $\mathbb{C}^-$.

6.2.1 BIBO-stability and Bezout factorizations

The main idea of the stabilization method of this section is based on the following observation. Consider a time-delay system with commensurable time-delay τ, modeled as a system $\Sigma = (A, B)$ over the polynomial ring $\mathbb{R}[s]$, given by $A = 0$ and $B = 1 - s$. Since $z = \det(zI - A) \notin \mathcal{D}$, it is obvious that Σ is neither internally nor externally stable. Next consider the transfer function

$$T(z) = \frac{1-s}{z}$$

of the system $\Sigma = (A, B)$ over the ring $\mathbb{R}[s]$, and substitute $s = e^{-\tau z}$. Then we obtain the transfer function of the delay system in the classical sense:

$$T(z) = \frac{1-e^{-\tau z}}{z}.$$

This transfer function has no pole in $z = 0$ because there is a pole-zero cancellation between numerator and denominator. In fact, the transfer function is analytic in $\mathbb{C}^+$ (it is even analytic in $\mathbb{C}$), continuous on $\overline{\mathbb{C}^+}$, and

$$\lim_{|z|\to\infty,\, z\in\overline{\mathbb{C}^+}} T(z) = 0.$$

Therefore we conclude that the corresponding delay system is *BIBO-stable*, i.e. if the system is started with zero initial conditions, and we apply a bounded input to the system, the output of the system is also bounded. (BIBO stands for Bounded Input Bounded Output). Note however that there exists no realization of this transfer function that is internally or externally stable. Unlike systems without delays, external stability of a time-delay system (i.e external stability in the systems over rings sense cf. Definition 2.5.3 *(ii)*), and BIBO-stability of a time-delay system are not equivalent notions any more.

Using Definition 3.2.6 of the commutative Banach algebra $\mathcal{A}_0(\mathbb{C}^+)$, it is possible to characterize all BIBO-stable time-delay systems. For convenience we first repeat the definition of this Banach algebra.

Definition 6.2.1 The algebra $\mathcal{A}_0(\mathbb{C}^+)$ is defined as the set of all functions f that are analytic in $\mathbb{C}^+$, continuous on $\overline{\mathbb{C}^+}$, and satisfy

$$\exists L \in \mathbb{C} : \lim_{|z|\to\infty,\, z\in\overline{\mathbb{C}^+}} |f(z) - L| = 0,$$

i.e. f can be extended continuously to infinity.

Definition 6.2.2 Let $\Sigma = (A, B, C, D)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. Then the *transfer matrix* of the *delay system* Σ (in the classical sense) is given by

$$T(z) := C(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))^{-1} B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) + \\ + D(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}) \tag{6.24}$$

Clearly, the transfer matrix of the delay system Σ in the classical sense is obtained from the transfer matrix of the system $\Sigma = (A, B, C, D)$ in the algebraic sense (cf. Definition 2.4.2) by substitution of the exponential functions $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$ for the corresponding indeterminates $s_1, \ldots, s_k$.

Proposition 6.2.3 *Let $\Sigma = (A, B, C, D)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$, modeled as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. Then the delay system is BIBO-stable if and only if all entries of the transfer matrix $T(z)$ defined in (6.24) are elements of the Banach algebra $\mathcal{A}_0(\mathbb{C}^+)$.*

Sketch of the proof

According to Lemma 6.1.3, the characteristic function of the delay system Σ given by $p(z) = \det(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$, has only a finite number of zeros in any arbitrary right half plane. If $T(z)$ is a matrix over $\mathcal{A}_0(\mathbb{C}^+)$, all zeros of $p(z)$ in $\overline{\mathbb{C}^+}$ are cancelled out. So there exists a $\delta > 0$ such that all poles of $T(z)$ have real part smaller than $-\delta$. This guarantees that the impulse response of the delay system falls off with an exponential decay rate of at least $\frac{1}{2}\delta$. If the system is started with zero initial conditions, this implies that a bounded input results in a bounded output.

On the other hand, if the system is BIBO-stable, the transfer matrix $T(z)$ cannot have poles in $\overline{\mathbb{C}^+}$. By definition $T(z)$ is proper, and thus it is a matrix over $\mathcal{A}_0(\mathbb{C}^+)$.

■

Corollary 6.2.4 *A time-delay system that is externally stable is also BIBO-stable*

Proof

Let Σ be an externally stable time-delay system with transfer matrix $T(z)$. Then the denominator of each entry of $T(z)$ is a stable exponential polynomial. Moreover, $T(z)$ is proper, i.e. $\lim_{|z| \to \infty,\, z \in \overline{\mathbb{C}^+}} T(z)$ exists. Hence $T(z)$ is a matrix over $\mathcal{A}_0(\mathbb{C}^+)$.

■

From Corollary 6.2.4 we conclude that after substitution of the exponential functions $e^{-\tau_1 z}, \ldots, e^{-\tau_k z}$ for the indeterminates $s_1, \ldots, s_k$, a proper element $q \in \mathcal{R}_{\mathcal{D}}(z)$ becomes an element of $\mathcal{A}_0(\mathbb{C}^+)$. In this respect, the Banach algebra $\mathcal{A}_0(\mathbb{C}^+)$ describes a more general class of functions. So we may expect that right-invertibility problems are easier to solve over $\mathcal{A}_0(\mathbb{C}^+)$ than over $\mathcal{R}_{\mathcal{D}}(z)$. This is the key-idea that is used to solve the stabilization problem.

Considering systems in the frequency domain by using their transfer matrices, the stabilizability of a system is usually investigated with help of so-called Bezout factorizations. An extensive treatise on this subject is [95].

Definition 6.2.5 Let T be a $p \times m$ transfer matrix of a time-delay system as defined in (6.24). A pair (D, N) of matrices over $\mathcal{A}_0(\mathbb{C}^+)$ is called a *(left) Bezout factorization of T over $\mathcal{A}_0(\mathbb{C}^+)$* if

(i) N is a $p \times m$ matrix and D is a $p \times p$ matrix such that $\det(D) \not\equiv 0$,

(ii) there exist matrices Q and P over $\mathcal{A}_0(\mathbb{C}^+)$ such that

$$DQ + NP = I, \tag{6.25}$$

(iii) $T = D^{-1}N$.

Formula (6.25) in condition *(ii)* is called the *Bezout identity*. It is equivalent to the condition that the matrix $(D|N)$ is right-invertible over $\mathcal{A}_0(\mathbb{C}^+)$.

Using the notion of Bezout factorizations over $\mathcal{A}_0(\mathbb{C}^+)$, it is possible to give another characterization of internal stabilizability for time-delay systems. Moreover, the proof of this result provides interesting information on the design of stabilizing feedback compensators.

Theorem 6.2.6 *Let $\Sigma = (A, B)$ be a time-delay system with k incommensurable time-delays $\tau_1, \ldots, \tau_k$ modeled as a system of rank n over the ring $\mathcal{R} = \mathbb{R}[s_1, \ldots, s_k]$. Let*

$$T(z) := (zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))^{-1} B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})$$

denote the transfer matrix of Σ, and define

$$(D(z), N(z)) := \left(\frac{1}{z+1} \cdot (zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})), \frac{1}{z+1} \cdot B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\right). \tag{6.26}$$

Then

$$\Sigma \text{ is internally stabilizable by dynamic state feedback,}$$

$$\Longleftrightarrow$$

$$(D(z), N(z)) \text{ is a left Bezout factorization of } T(z) \text{ over } \mathcal{A}_0(\mathbb{C}^+).$$

Moreover, if Σ is internally stabilizable, then there exists a finite-dimensional compensator (i.e. a compensator without time-delays) that stabilizes the system.

Proof

"$\Rightarrow$" Suppose that Σ is internally stabilizable. So, in the algebraic setup of Section 2.8, the matrix $(zI - A(s_1, \ldots, s_k)|B(s_1, \ldots, s_k))$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$. According to the proof of Theorem 2.8.2, this implies that there exists a polynomial $\varphi(z, s_1, \ldots, s_k) \in \mathcal{D}$ with $\deg_z(\varphi(z, s_1, \ldots, s_k)) \geq n + 1$, and polynomial matrices Q and P over $\mathcal{R}[z]$ such that

$$(zI - A(s_1, \ldots, s_k))Q(z, s_1, \ldots, s_k) + B(s_1, \ldots, s_k)P(z, s_1, \ldots, s_k) =$$

$$= \varphi(z, s_1, \ldots, s_k) \cdot I.$$

Recalling Lemma 2.8.1, it follows that the polynomial matrix P can be chosen in such a way that $\deg_z(P(z,s_1,\ldots,s_k)) \leq n-1$. Define

$$\hat{Q}(z) := \frac{(z+1)}{\varphi(z,e^{-\tau_1 z},\ldots,e^{-\tau_k z})} \cdot Q(z,e^{-\tau_1 z},\ldots,e^{-\tau_k z}),$$

$$\hat{P}(z) := \frac{(z+1)}{\varphi(z,e^{-\tau_1 z},\ldots,e^{-\tau_k z})} \cdot P(z,e^{-\tau_1 z},\ldots,e^{-\tau_k z}).$$

Then it is obvious that $\hat{Q}(z)$ and $\hat{P}(z)$ are analytic in $\mathbb{C}^+$, continuous on $\overline{\mathbb{C}^+}$ and satisfy

$$D(z)\hat{Q}(z) + N(z)\hat{P}(z) = I.$$

Moreover, $\lim_{|z|\to\infty,\, z\in\overline{\mathbb{C}^+}} \hat{P}(z) = 0$, and $\lim_{|z|\to\infty,\, z\in\overline{\mathbb{C}^+}} D(z) = I$, so the previous formula implies that $\lim_{|z|\to\infty,\, z\in\overline{\mathbb{C}^+}} \hat{Q}(z) = I$. Hence $\hat{Q}(z)$ and $\hat{P}(z)$ are matrices over $\mathcal{A}_0(\mathbb{C}^+)$ and condition *(ii)* of Definition 6.2.5 is satisfied. Since it is clear that also conditions *(i)* and *(iii)* hold, we conclude that $(D(z),N(z))$ is a left Bezout factorization of $T(z)$.

"$\Leftarrow$" Suppose that $(D(z),N(z))$ is a left Bezout factorization of $T(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$. Then there exist matrices $Q(z)$ and $P(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$ such that

$$D(z)Q(z) + N(z)P(z) = I. \tag{6.27}$$

Without loss of generality we assume that the entries of the matrices $Q(z)$ and $P(z)$ only take real values on the real axis. Using Mergelyan's Theorem in the same way as in Proposition 3.2.7, we approximate $Q(z)$ and $P(z)$ uniformly by proper stable real rational matrices. Define $M_1 := \sup\{\|D(z)\| \mid z \in \overline{\mathbb{C}^+}\}$ and $M_2 := \sup\{\|N(z)\| \mid z \in \overline{\mathbb{C}^+}\}$. Since the entries of both matrices belong to $\mathcal{A}_0(\mathbb{C}^+)$, M_1 and M_2 are well defined. Fix $\varepsilon < \frac{1}{3\max(M_1,M_2)}$. Then there exist matrices $\hat{Q}(z)$ and $\hat{P}(z)$ over $\mathbb{R}(z) \cap \mathcal{R}_{\mathcal{D}}(z)$ such that

$$\forall z \in \overline{\mathbb{C}^+} : \|Q(z) - \hat{Q}(z)\| < \varepsilon \text{ and } \|P(z) - \hat{P}(z)\| < \varepsilon.$$

Next define

$$\Phi(z,s_1,\ldots,s_k) := \frac{1}{z+1}(zI - A(s_1,\ldots,s_k))\hat{Q}(z) + \frac{1}{z+1}B(s_1,\ldots,s_k)\hat{P}(z). \tag{6.28}$$

Since the right hand side of (6.28) only consists of proper matrices over $\mathcal{R}_{\mathcal{D}}(z)$, all entries of the matrix $\Phi(z,s_1,\ldots,s_k)$ are proper elements of $\mathcal{R}_{\mathcal{D}}(z)$. Moreover, for every $z \in \overline{\mathbb{C}^+}$ we have:

$$\begin{aligned}
\|\Phi(z,e^{-\tau_1 z},\ldots,e^{-\tau_k z}) - I\| &= \|D(z)(\hat{Q}(z) - Q(z)) + N(z)(\hat{P}(z) - P(z))\| \leq \\
&\leq \|D(z)\| \cdot \|\hat{Q}(z) - Q(z)\| + \|N(z)\| \cdot \|\hat{P}(z) - P(z)\| < \\
&< M_1 \frac{1}{3\max(M_1,M_2)} + M_2 \frac{1}{3\max(M_1,M_2)} \leq \frac{2}{3},
\end{aligned}$$

and thus $\det(\Phi(z, e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))$ has no zeros in $\overline{\mathbb{C}^+}$. It follows that its inverse $\Phi(z, s_1, \ldots, s_k)^{-1}$ is a matrix over $\mathcal{R}_{\mathcal{D}}(z)$. Define the the following two matrices over $\mathcal{R}_{\mathcal{D}}(z)$:

$$\tilde{Q}(z, s_1, \ldots, s_k) := \frac{1}{z+1}\hat{Q}(z) \cdot \Phi(z, s_1, \ldots, s_k)^{-1},$$

$$\tilde{P}(z, s_1, \ldots, s_k) := \frac{1}{z+1}\hat{P}(z) \cdot \Phi(z, s_1, \ldots, s_k)^{-1}.$$

Then multiplication of (6.28) by $\Phi(z, s_1, \ldots, s_k)^{-1}$ yields:

$$(zI - A(s_1, \ldots, s_k))\tilde{Q}(z, s_1, \ldots, s_k) + B(s_1, \ldots, s_k)\tilde{P}(z, s_1, \ldots, s_k) = I,$$

and thus $(zI - A|B)$ is right-invertible over $\mathcal{R}_{\mathcal{D}}(z)$.

Finally, using exactly the same arguments as in the proof of Theorem 2.8.2, it may be verified that an appropriate realization of the transfer matrix

$$\tilde{P}(z, s_1, \ldots, s_k)(\tilde{Q}(z, s_1, \ldots, s_k))^{-1} = \hat{P}(z)(\hat{Q}(z))^{-1},$$

yields an internally stabilizing controller for the delay system Σ. Since $\hat{P}(z)(\hat{Q}(z))^{-1}$ is the transfer matrix of a finite-dimensional system, this proves the last assertion of the theorem. ∎

Theorem 6.2.6 and especially its proof are important because they describe a method for the stabilization of time-delay systems. All steps in the proof are constructive, except the following two parts:

(i) computation of the matrices $Q(z)$ and $P(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$ that satisfy the Bezout identity (6.27),

(ii) approximation of the matrices $Q(z)$ and $P(z)$ by matrices over the ring of proper stable real rational functions.

Note that requirement *(i)* above is weaker than the original stabilizability condition of Theorem 2.8.2 because *(i)* allows us to search for a right-inverse of the matrix

$$\left(\frac{1}{z+1}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})) \mid \frac{1}{z+1}B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\right)$$

over a more general class of functions. In the next two subsections we show how the problems *(i)* and *(ii)* can be solved in a constructive way. Together with the proof of Theorem 6.2.6, this yields a constructive solution to the stabilization problem.

6.2.2 Construction of Bezout factorizations over $\mathcal{A}_0(\mathbb{C}^+)$

This subsection is devoted to the construction of matrices $Q(z)$ and $P(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$, satisfying the Bezout identity (6.27) corresponding to Bezout factorization (6.26) of the transfer matrix of a time-delay system $\Sigma = (A, B)$. For this purpose it is not necessary to consider the whole Banach algebra $\mathcal{A}_0(\mathbb{C}^+)$; a specific

subclass of $\mathcal{A}_0(\mathbb{C}^+)$ suffices to solve the problem. The method is mainly based on functions in $\mathcal{A}_0(\mathbb{C}^+)$ of the form

$$\vartheta_\alpha(z) = \frac{1 - e^{\tau(\alpha - z)}}{z - \alpha},$$

where $\alpha \in \mathbb{C}$. These functions facilitate the computations considerably because they may be applied to cancel a factor $(z - \alpha)$, and replace it by a factor containing only exponential terms.

Unfortunately, the explicit construction method is very involved and requires a lot of technicalities. Moreover, the method is only applicable for systems with commensurable time-delays, satisfying some rather mild regularity conditions. We confine ourselves to the statement of the main result and omit the proof. For a detailed discussion of the construction, we refer to [56]. The example in Subsection 6.2.4 illustrates the most important ideas behind the construction method.

We start with the introduction of some new terminology that is required for the statement of the main result.

Definition 6.2.7 Let ϑ_α denote the analytic function defined by

$$\vartheta_\alpha : \mathbb{C} \to \mathbb{C} : \qquad \vartheta_\alpha(z) := \frac{1 - e^{\tau(\alpha - z)}}{z - \alpha}, \qquad (\alpha \in \mathbb{C}). \tag{6.29}$$

Let $\vartheta_\alpha^{(i)}$ denote the i^{th} derivative of ϑ_α. Then Θ is defined as the ring of all analytic functions that are real on the real axis, generated by

$$\{c\vartheta_\alpha^{(i)} \mid \alpha \in \mathbb{C}, c \in \mathbb{R}, i \in \mathbb{N}_0\}.$$

From the previous definition it follows that an analytic function f belongs to Θ if it is generated by functions of the form (6.29) and their derivatives, and satisfies $\overline{f(\bar{z})} = f(z)$ for all $z \in \mathbb{C}$.

Lemma 6.2.8 *Let A be an $n \times n$ matrix over $\mathbb{R}$ and let $\tau > 0$. Then all entries of the matrix*

$$(zI - A)^{-1}(I - e^{-\tau z} \cdot e^{\tau A}) \tag{6.30}$$

belong to the ring Θ. ∎

The matrix (6.30) can be seen as a generalization of functions of the form (6.29) to the higher dimensional case. For a proof of the correctness of this result we refer to [56, p. 844].

Definition 6.2.9 $\mathbb{R}_U(s)$ is defined as the subring of $\mathbb{R}(s)$ consisting of all rational functions for which the denominator polynomial has no zeros inside the closed unit disc $\overline{U}$:

$$\mathbb{R}_U(s) := \left\{ \frac{a(s)}{b(s)} \mid a(s), b(s) \in \mathbb{R}[s] \text{ and } \forall \lambda \in \overline{U} : b(\lambda) \neq 0 \right\}.$$

Definition 6.2.10 Let $\tau > 0$. Then Λ_τ is defined as the ring obtained from $\mathbb{R}_{\mathcal{U}}(s)$ after substitution of the exponential function $e^{-\tau z}$ for the indeterminate s:

$$\Lambda_\tau := \mathbb{R}_{\mathcal{U}}(e^{-\tau z}).$$

The definition of Λ_τ implies that every element of Λ_τ may be considered as a function in the Banach algebra $\mathcal{A}_0(\mathbb{C}^+)$. Since for all $z \in \overline{\mathbb{C}^+}$ we have $|e^{-\tau z}| \leq 1$, a function in Λ_τ has no poles in the the closed right half plane.

Definition 6.2.11 For every $\tau > 0$, the ring $\Theta\Lambda_\tau$ is defined as the product of the rings Θ and Λ_τ.

Note that every element of $\Theta\Lambda_\tau$ is a function that is analytic in $\mathbb{C}^+$, continuous in $\overline{\mathbb{C}^+}$, and that can be extended continuously to infinity. Thus the functions in $\Theta\Lambda_\tau$ form a subclass of $\mathcal{A}_0(\mathbb{C}^+)$. Moreover, every element $f \in \Theta\Lambda_\tau$ is real on the real axis: $\forall z \in \overline{\mathbb{C}^+} : \overline{f(\overline{z})} = f(z)$.

After all these preparations we now state the main result of this subsection:

Theorem 6.2.12 *Let $\Sigma = (A, B)$ be a time-delay system with a commensurable time-delay τ, modeled as a system over the ring $\mathbb{R}[s]$. Assume that the pair (B^T, A^T) of matrices over $\mathbb{R}[s]$ is observable in the sense of Definition 2.3.2, and that the time-delay system Σ is internally stabilizable. Then there exists a constructive method for the computation of matrices $Q(z)$ and $P(z)$ of the form*

$$Q(z) = \sum_{i=0}^{\ell} Q_i(z) \cdot z^i, \qquad P(z) = \sum_{i=0}^{\ell} P_i(z) \cdot z^i,$$

with $Q_i(z), P_i(z) \in \Lambda_\tau$ $(i = 1, \ldots, \ell)$ and $Q_0(z), P_0(z) \in \Lambda_\tau\Theta$, such that

$$(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))Q(z) + B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})P(z) = I. \tag{6.31}$$

■

For a proof of Theorem 6.2.12 and a detailed description of the construction we refer to [56]. The method involves a lot of technicalities, and therefore a thorough elaboration of the complete method is omitted. Some of the main ideas behind the construction in the scalar case will be illustrated in Example 6.2.21. Lemma 6.2.8 indicates that in the higher dimensional case, the method can be applied in an analogous way.

Formula (6.31) may be considered as a polynomial Bezout identity; it is not exactly the Bezout identity we are looking for. To a certain extent, the matrices $Q(z)$ and $P(z)$ resemble polynomial matrices in the variable z. However, in general the coefficient matrices $Q_i(z)$ and $P_i(z)$ $(i = 0, 1, \ldots, \ell)$ are not constant, and may contain the variable z explicitly. Note that all $Q_i(z)$ and $P_i(z)$ $(i = 0, 1, \ldots, \ell)$ are matrices over $\mathcal{A}_0(\mathbb{C}^+)$, so in the right half plane they are uniformly bounded. This indicates that in the right half plane the matrices $Q(z)$ and $P(z)$ exhibit a polynomial character. The situation is very similar to that of exponential polynomials, with the only difference that the entries of $Q(z)$ and $P(z)$ are not necessarily monic. However,

the notion of degree is straightforwardly extended to this class of matrices: it is the degree of the highest power of z occurring in one of the entries of the matrix.

To obtain a Bezout factorization over $\mathcal{A}_0(\mathbf{C}^+)$, formula (6.31) is modified using the same arguments as in the proof of Theorem 2.8.2. Let n be the size of the matrix A, and let $\varphi(z)$ be a stable exponential polynomial of degree $n+1$. After multiplication of (6.31) by $\varphi(z)$ and a rearrangement of terms based on the result of Lemma 2.8.1, we obtain

$$(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))\tilde{Q}(z) + B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\tilde{P}(z) = \varphi(z) \cdot I,$$

where $\tilde{Q}(z)$ and $\tilde{P}(z)$ are matrices of polynomials in z with coefficients in $\Lambda_\tau\Theta \subset \mathcal{A}_0(\mathbf{C}^+)$ (so the coefficients also contain the variable z), such that $\deg_z(\tilde{P}(z)) \leq n-1$. Dividing the previous equation by $\varphi(z)$ we get

$$\left(\frac{1}{z+1}(zI - A(e^{-\tau_1 z}, \ldots, e^{-\tau_k z}))\right)\hat{Q}(z) + \left(\frac{1}{z+1}B(e^{-\tau_1 z}, \ldots, e^{-\tau_k z})\right)\hat{P}(z) = I, \tag{6.32}$$

where

$$\hat{Q}(z) := \frac{z+1}{\varphi(z)}\tilde{Q}(z) \quad \text{and} \quad \hat{P}(z) := \frac{z+1}{\varphi(z)}\tilde{P}(z). \tag{6.33}$$

By construction, the matrices $\hat{Q}(z)$ and $\hat{P}(z)$ are analytic in $\mathbf{C}^+$ and continuous on $\overline{\mathbf{C}^+}$, and moreover $\lim_{|z|\to\infty,\, z\in\overline{\mathbf{C}^+}} \hat{P}(z) = 0$. Using the Bezout identity (6.32), this implies that $\lim_{|z|\to\infty,\, z\in\overline{\mathbf{C}^+}} \hat{P}(z) = I$. So $(\hat{Q}(z), \hat{P}(z))$ is a pair of matrices over $\mathcal{A}_0(\mathbf{C}^+)$ satisfying the Bezout identity (6.27).

6.2.3 Uniform approximation in $\mathcal{A}_0(\mathbf{C}^+)$

In this subsection we describe a constructive method for the computation of a sequence of proper stable rational functions in $\mathbf{R}(z) \cap \mathcal{R}_\mathcal{D}(z)$, that approximates a given function $f \in \mathcal{A}_0(\mathbf{C}^+)$ uniformly over the closed right half plane. For this purpose, the same ideas as in Subsection 3.2.1 are used. First the problem is transformed to the unit disc, next Mergelyan's Theorem is applied, and finally the obtained approximation is transformed back to the closed right half plane. In this subsection all these steps are carried out explicitly, yielding a constructive solution to the approximation problem.

Let $f \in \mathcal{A}_0(\mathbf{C}^+)$, and assume that for all $z \in \overline{\mathbf{C}^+}$: $\overline{f(\bar{z})} = f(z)$. Let $\mathcal{U}$ denote the open unit disc $\{s \in \mathbf{C} \mid |s| < 1\}$, and consider the Möbius transformation

$$T : \overline{\mathbf{C}^+} \to \overline{\mathcal{U}}\backslash\{-1\} : \qquad T(z) := \frac{1-z}{1+z}, \tag{6.34}$$

with inverse

$$T^{-1} : \overline{\mathcal{U}}\backslash\{-1\} \to \overline{\mathbf{C}^+} : \qquad T^{-1}(s) := \frac{1-s}{1+s}. \tag{6.35}$$

Define the function $\tilde{f} : \overline{\mathcal{U}} \to \overline{\mathbb{C}^+}$ by

$$\tilde{f} : \overline{\mathcal{U}} \to \overline{\mathbb{C}^+} : \qquad \tilde{f}(s) := \begin{cases} f(T^{-1}(s)) & \text{if} \quad s \neq -1, \\ \lim_{|z| \to \infty,\, z \in \overline{\mathbb{C}^+}} f(z) & \text{if} \quad s = -1. \end{cases} \tag{6.36}$$

In the same way as in the proof of Theorem 3.2.3, it may be verified that $\tilde{f}$ is an element of the Banach algebra $\mathcal{A}(\mathcal{U})$, consisting of all functions that are analytic in $\mathcal{U}$ and continuous on $\overline{\mathcal{U}}$. Moreover,

$$\forall s \in \overline{\mathcal{U}} : \overline{\tilde{f}(\bar{s})} = \tilde{f}(s),$$

so $\tilde{f}$ is a real function on the real axis. Our goal is to approximate the function $\tilde{f}$ uniformly over $\overline{\mathcal{U}}$ by a sequence of polynomials with real coefficients.

Proposition 6.2.13 (Maximum modulus principle) *Let $g \in \mathcal{A}(\mathcal{U})$, and define*

$$M := \max\{|g(s)| \mid |s| = 1\}.$$

Then

$$\forall s \in \overline{\mathcal{U}} : |g(s)| \leq M. \qquad \blacksquare$$

For a proof of this well-known result, we refer to e.g. [83, p. 253]. The maximum modulus principle implies that for the uniform approximation of the function $\tilde{f} \in \mathcal{A}(\mathcal{U})$ it is sufficient to find a real polynomial p that approximates $\tilde{f}$ uniformly on the unit circle.

Note that on the unit circle, the function $\tilde{f}$ is periodic with period 2π. Therefore, it is natural to use the *Fourier series expansion* of $\tilde{f}(e^{\imath\omega})$ for the approximation of $\tilde{f}$. This Fourier series is given by

$$\tilde{f}(e^{\imath\omega}) \sim \sum_{n=-\infty}^{\infty} c_n e^{\imath n\omega},$$

where the *Fourier coefficients* c_n ($n \in \mathbb{Z}$) are defined by

$$c_n := \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{f}(e^{\imath\omega}) \cdot e^{-\imath n\omega} d\omega. \tag{6.37}$$

Lemma 6.2.14 *Let $g \in \mathcal{A}(\mathcal{U})$. Then the Fourier coefficients c_n ($n \in \mathbb{Z}$) of the Fourier series of $g(e^{\imath\omega})$ are given by*

$$c_n = 0 \qquad \text{for} \quad n < 0,$$

$$c_n = \frac{g^{(n)}(0)}{n!} \qquad \text{for} \quad n \geq 0.$$

Proof

Let $n \in \mathbb{Z}$, and define for every $r \in \mathbb{R}^+$ the circle Δ_r with centre 0 and radius r by $\Delta_r := \{s \in \mathbb{C} \mid |s| = r\}$. Then

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(e^{\imath\omega}) \cdot e^{-\imath n\omega} d\omega = \frac{1}{2\pi\imath} \oint_{\Delta_1} \frac{g(s)}{s^{n+1}} ds.$$

By assumption g is continuous on the compact set $\overline{\mathcal{U}}$, so in particular it is uniformly continuous. So for every $n \in \mathbb{Z}$, and for every $\varepsilon > 0$, there exists a $0 < \delta_n < 1$ such that

$$\left| \frac{1}{2\pi i} \oint_{\Delta_1} \frac{g(s)}{s^{n+1}} ds - \frac{1}{2\pi i} \oint_{\Delta_{\delta_n}} \frac{g(s)}{s^{n+1}} ds \right| < \varepsilon. \tag{6.38}$$

Now every circle Δ_{δ_n} is contained in the region $\mathcal{U}$ on which the function g is analytic. Therefore we may apply a generalized version of Cauchy's integral formula (see for example [11, p. 113]) to obtain the following results:

$$\begin{aligned} \frac{1}{2\pi i} \oint_{\Delta_{\delta_n}} \frac{g(s)}{s^{n+1}} ds &= 0 \qquad \text{for} \quad n < 0, \\ \frac{1}{2\pi i} \oint_{\Delta_{\delta_n}} \frac{g(s)}{s^{n+1}} ds &= \frac{g^{(n)}(0)}{n!} \quad \text{for} \quad n \geq 0 \end{aligned}$$

Because of (6.38) the same results hold if the contour Δ_{δ_n} is replaced by Δ_1. This completes the proof. ∎

Remark 6.2.15 From a numerical point of view the best way to compute the Fourier coefficients c_n ($n \in \mathbb{N} \cup \{0\}$) is probably the fast Fourier transform (FFT). However, in a computer algebra environment, the derivatives of a function are easily obtained, and in this situation Lemma 6.2.14 describes an attractive alternative for the numerical approach.

Although the Fourier series of a continuous function satisfying the Dirichlet conditions converges pointwise to this function, uniform convergence cannot be guaranteed. For this some extra regularity conditions (see for example Proposition 6.2.19) have to be satisfied. Since $\tilde{f} \in \mathcal{A}(\mathcal{U})$, we only know that $\tilde{f}(e^{i\omega})$ is continuous, and this is not sufficient to ensure uniform convergence. However, this problem may be solved by taking the arithmetic means of the partial sums of the Fourier series expansion.

Definition 6.2.16 Consider a Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{in\omega}$, and define its k^{th} partial sum S_k by

$$S_k := \sum_{n=-k}^{k} c_n e^{in\omega}. \tag{6.39}$$

Then the *Cesaro means* $(\bar{S}_N)_{N \in \mathbb{N}}$ of the Fourier series are the arithmetic means

$$\bar{S}_N := \frac{1}{N} \sum_{k=0}^{N-1} S_k \qquad (N \in \mathbb{N}). \tag{6.40}$$

Recall that for the function $\tilde{f} \in \mathcal{A}(\mathcal{U})$ all Fourier coefficients c_n for $n < 0$ are zero. Therefore the N-th Cesaro mean of the Fourier series of $\tilde{f}(e^{i\omega})$ may be written as

$$\bar{S}_N = \sum_{n=0}^{N-1} \frac{N-n}{N} c_n e^{in\omega} \qquad (N \in \mathbb{N}). \tag{6.41}$$

Using Cesaro means instead of Fourier series, it is possible to find a uniform approximation of the function $\tilde{f}$ over the unit circle.

Proposition 6.2.17 *Let $g : [-\pi, \pi] \to \mathbf{C}$ be a continuous function satisfying $g(\pi) = g(-\pi)$. Then the Cesaro means of the Fourier series of g converge uniformly to g on the interval $[-\pi, \pi]$.* ∎

A proof of this result that originates from Fejér may be found in e.g. [46, pp. 17-19].

Combining all previous results, we find the following constructive method for the uniform approximation of functions in $\mathcal{A}(\mathcal{U})$ by polynomials.

Corollary 6.2.18 *Let $g \in \mathcal{A}(\mathcal{U})$, and assume that g is real on the real axis. Define for every $n \in \mathbf{N}$ the polynomial g_n of degree n by:*

$$g_n : \overline{\mathcal{U}} \to \mathbf{C}: \qquad g_n(s) := \sum_{j=0}^{n} \frac{n+1-j}{n+1} \cdot \frac{g^{(j)}(0)}{j!} \cdot s^j. \tag{6.42}$$

Then $(g_n)_{n\in\mathbf{N}}$ is a sequence of real polynomials converging uniformly to g:

$$\forall \varepsilon > 0 \; \exists M \in \mathbf{N} \; \forall n > M \; \forall s \in \overline{\mathcal{U}} : \; |g(s) - g_n(s)| < \varepsilon.$$

Proof

From Proposition 6.2.17 we know that $g(e^{\imath\omega})$ is uniformly approximated by the Cesaro means of its Fourier series:

$$\bar{S}_{n+1} = \sum_{j=0}^{n} \frac{n+1-j}{n+1} c_j e^{\imath j\omega}.$$

According to Lemma 6.2.14, $c_j = \frac{g^{(j)}(0)}{j!}$, so if we define $g_n := \bar{S}_{n+1}$ and substitute s for $e^{\imath\omega}$, we obtain formula (6.42). Now the maximum modulus principle implies that g_n converges uniformly to g on $\overline{\mathcal{U}}$ for $n \to \infty$. ∎

Finally, the result of Corollary 6.2.18 has to be transformed to the closed right half plane. Let $(\tilde{f}_n)_{n\in\mathbf{N}}$ denote a sequence of polynomials with real coefficients that converge uniformly to $\tilde{f}$. Define for all $n \in \mathbf{N}$ the function $r_n \in \mathcal{A}_0(\mathbf{C}^+)$ as the composition of $\tilde{f}_n$ and the Möbius transformation T of formula (6.34): $r_n := \tilde{f}_n \circ T$. Then $(r_n)_{n\in\mathbf{N}}$ is a sequence of proper stable real rational functions in $\mathcal{A}_0(\mathbf{C}^+)$ that converge uniformly to the function $f \in \mathcal{A}_0(\mathbf{C}^+)$ we started with.

Note that the approximation method described in this subsection is applicable for arbitrary functions in $\mathcal{A}_0(\mathbf{C}^+)$. However, for the stabilization problem, this approximation method is only applied to the matrices $\hat{Q}(z)$ and $\hat{P}(z)$ defined in (6.33), satisfying the Bezout identity (6.32). These matrices are constructed in a very specific way, and therefore the entries of $\hat{Q}(z)$ and $\hat{P}(z)$ possess an important additional property: they are also analytic on the imaginary axis.

Recalling the construction of Subsection 6.2.2, we know that each entry $f(z)$ of the matrices $\hat{Q}(z)$ and $\hat{P}(z)$ may be written as $f(z) = \frac{n(z)}{d(z)}$, where $d(z)$ is a stable exponential polynomial and $n(z)$ has the form

$$n(z) = \sum_{i=0}^{\ell} a_i(z) \cdot z^i,$$

with $a_i(z) \in \Theta\Lambda_\tau$. Now Θ is a ring of functions that are analytic over $\mathbf{C}$, and the definition of Λ_τ implies that for every element $l \in \Lambda_\tau$ there exists a $\delta_l > 0$ such that l is analytic in the right half plane $\{z \in \mathbf{C} \mid \operatorname{Re} z > -\delta_l\}$. Since the coefficients $a_i(z)$ $(i = 0, 1, \ldots, \ell)$ are constructed as finite sums of products of elements of Θ and Λ_τ, all coefficients $a_i(z)$, and therefore also the numerator $n(z)$, are analytic in an open right half plane that contains $\overline{\mathbf{C}^+}$. Moreover, $d(z)$ is a stable exponential polynomial, so Lemma 6.1.3 implies that there exists a $\delta_d > 0$ such that $d(z)$ has no zeros in $\{z \in \mathbf{C} \mid \operatorname{Re} z > -\delta_d\}$. Therefore we conclude that also the function $f(z) = \frac{n(z)}{d(z)}$ is analytic in an open right half plane containing $\overline{\mathbf{C}^+}$. In particular $f(z)$ is analytic on the imaginary axis.

The importance of this observation becomes evident in the following well-known result on the Fourier series of a continuously differentiable function. For a proof we refer to [91, p. 81].

Proposition 6.2.19 *Let $g : [-\pi, \pi] \to \mathbf{C}$ be a continuously differentiable function, and assume that $g(-\pi) = g(\pi)$. Then the partial sums of the Fourier series of g converge uniformly to g.* ∎

Let $f(z)$ be an entry of one of the matrices $\hat{Q}(z)$ or $\hat{P}(z)$ in (6.33), and define $\tilde{f}$ as in (6.36). Then $\tilde{f}$ is continuously differentiable on the unit circle, and according to Proposition 6.2.19 this implies that the Fourier series of $\tilde{f}$ converges uniformly to $\tilde{f}$. Therefore the partial sums $\tilde{f}_n$ $(n \in \mathbf{N})$ of the Taylor expansion of $\tilde{f}$,

$$\tilde{f}_n(s) = \sum_{j=0}^{n} \frac{\tilde{f}^{(j)}(0)}{j!} s^j,$$

converge uniformly to $\tilde{f}$. So in our specific situation it is not necessary to take Cesaro means of the Fourier series. Since in general the convergence of the partial sums of the Fourier series is much faster than that of the corresponding Cesaro means, this modification may increase the computational efficiency considerably.

Remark 6.2.20 In the stabilization problem, the uniform convergence of a sequence of approximations of the matrices $\hat{Q}(z)$ and $\hat{P}(z)$ defined in (6.33) is sufficient to find a stabilizing controller, but this condition is not necessary. Therefore we may also apply alternative approximation methods that do not guarantee uniform convergence, e.g. Padé approximation. In each step the stability test for exponential polynomials described in Section 6.1 is used to test whether the given approximation gives rise to a stabilizing controller. If not, we have to try a higher order approximation. Uniform convergence of a sequence of approximations ensures that this process ends after a finite number of steps. Although termination is not guaranteed for approximation methods that do not converge uniformly, these alternative methods have their own merits: sometimes they lead to the construction of stabilizing compensators of low order. The next section contains an example of this unexpected phenomenon.

6.2.4 An application

Probably the best way to explain the working of the stabilization method developed in this section is by means of an example. In this subsection we apply the method on a simple scalar system. First we give a detailed discussion how a Bezout factorization over $\mathcal{A}_0(\mathbb{C}^+)$ can be computed. Next two different approximation methods are used to obtain a stabilizing feedback compensator.

Example 6.2.21 ([54]) Consider a system $\Sigma = (A, B)$ with a commensurable time-delay $\tau = 1$, modeled as a system over the ring $\mathbb{R}[s]$. The matrices A and B are the scalars:

$$A = 2, \qquad B = s,$$

and thus the transfer function of the delay system Σ is given by

$$T(z) = \frac{e^{-z}}{z-2}.$$

The system itself is neither internally nor BIBO-stable, but it is internally stabilizable by dynamic state feedback because the rank condition of Theorem 3.2.8 is satisfied:

$$\forall z \in \overline{\mathbb{C}^+} : \ \operatorname{rank}(z-2 \mid e^{-z}) = 1.$$

To obtain a Bezout factorization of $T(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$ and a solution to the polynomial Bezout identity (6.31), we first apply the construction method of Theorem 6.2.12. Note that the polynomial $(zI - A) = (z-2)$ has precisely one zero in $z = 2$. Multiplying $(z-2)$ by $\vartheta_2(z)$, this polynomial expression may be changed into an exponential expression:

$$(z-2) \cdot \vartheta_2(z) = (z-2) \cdot \frac{1 - e^{2-z}}{z-2} = 1 - e^{2-z}.$$

In this way we have obtained an expression of the same type as $B(e^{-z}) = e^{-z}$, and it is easily seen that

$$(zI - A) \cdot \vartheta_2(z) + B(e^{-z}) \cdot e^2 = (z-2) \cdot \vartheta_2(z) + e^{-z} \cdot e^2 = 1. \tag{6.43}$$

We conclude that after definition of $Q(z) := \vartheta_2(z)$ and $P(z) := e^2$, Bezout identity (6.31) is satisfied.

In this particular example the size of the matrix A is 1, so to find a Bezout factorization of $T(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$, we have to multiply (6.43) by a stable exponential polynomial of degree 2. Define $\varphi(z) := (z+1)^2$, and multiply equation (6.43) by $\varphi(z)$:

$$(z-2) \cdot (\vartheta_2(z)(z+1)^2) + e^{-z} \cdot (e^2(z+1)^2) = (z+1)^2. \tag{6.44}$$

Next we apply Lemma 2.8.1 on the polynomials $(z+1)^2$ and $(z-2)$. In the scalar case this lemma reduces to the division algorithm with remainder, and we get

$$(z+1)^2 = (z-2)(z+4) + 9.$$

Therefore we may rewrite (6.44) as

$$(z-2)\cdot\left(\vartheta_2(z)(z+1)^2+(z+4)e^{2-z}\right)+e^{-z}\cdot(9e^2)=(z+1)^2,$$

and after division by $(z+1)^2$ we obtain

$$\left(\frac{z-2}{z+1}\right)\cdot\left(\frac{\vartheta_2(z)(z+1)^2+(z+4)e^{2-z}}{z+1}\right)+\left(\frac{e^{-z}}{z+1}\right)\cdot\left(\frac{9e^2}{z+1}\right)=1. \qquad (6.45)$$

Next we show that $(D(z),N(z)):=\left(\frac{z-2}{z+1},\frac{e^{-z}}{z+1}\right)$ is a Bezout factorization of $T(z)$ over $\mathcal{A}_0(\mathbb{C}^+)$. For this purpose we define

$$\hat{Q}(z) := \frac{\vartheta_2(z)(z+1)^2+(z+4)e^{2-z}}{z+1},$$

$$\hat{P}(z) := \frac{9e^2}{z+1}.$$

It is obvious that $(D(z),N(z))$ satisfies conditions *(i)* and *(iii)* of Definition 6.2.5, and with $\hat{Q}(z)$ and $\hat{P}(z)$ also Bezout identity (6.25) holds. So we only have to prove that $\hat{Q}(z)$ and $\hat{P}(z)$ are elements of $\mathcal{A}_0(\mathbb{C}^+)$. For $\hat{P}(z)$ this is trivial, and thus we may confine ourselves to $\hat{Q}(z)$. Note that

$$(z+4)\cdot e^{2-z}=(z+4)\left((e^{2-z}-1)+1\right)=(z+4)-(z+4)(z-2)\vartheta_2(z).$$

Therefore $\hat{Q}(z)$ may be rewritten as

$$\hat{Q}(z)=\frac{\vartheta_2(z)\cdot((z+1)^2-(z+4)(z-2))+(z+4)}{z+1}=\frac{z+4+9\vartheta_2(z)}{z+1},$$

and since $\vartheta_2(z)$ is an analytic function over $\mathbf{C}$ that is uniformly bounded in $\overline{\mathbb{C}^+}$, it follows that $\hat{Q}(z)\in\mathcal{A}_0(\mathbb{C}^+)$.

At this moment a BIBO-stabilizing compensator for the system may be obtained by taking

$$C(z):=\hat{P}(z)(\hat{Q}(z))^{-1}=\frac{9e^2}{z+4+9\vartheta_2(z)}.$$

If we multiply both numerator and denominator of $C(z)$ by $(z-2)$, we may realize this transfer function by a (not internally stable) delay system with point delays. Then the transfer function of the closed-loop system becomes

$$T_{cl}(z)=\frac{e^{-z}\cdot((z+1)^2-9e^{2-z})}{(z-2)(z+1)^2},$$

and we see that the closed-loop system is BIBO-stable but neither externally nor internally stable. To solve this problem, we first have to approximate the functions $\hat{Q}(z)$ and $\hat{P}(z)$ by stable rational functions, and then construct a finite-dimensional stabilizing compensator.

Since $\hat{P}(z)$ is already a stable rational function, we only have to approximate $\hat{Q}(z)$. This function may be written as

$$\hat{Q}(z)=\frac{z+4}{z+1}+9\frac{\vartheta_2(z)}{z+1},$$

and thus only an approximation of the second term is required. For this purpose the results of Subsection 6.2.3 are used.

Define

$$\tilde{f} : \overline{U} \to \mathbf{C} : \qquad \tilde{f}(s) := \begin{cases} \frac{\vartheta_2(\frac{1-s}{1+s})}{(\frac{1-s}{1+s})+1} & \text{if} \quad s \neq -1, \\ 0 & \text{if} \quad s = -1, \end{cases}$$

and let

$$c_j := \frac{\tilde{f}^{(j)}(0)}{j!} \qquad (j \in \mathbf{N} \cup \{0\}).$$

Since $\tilde{f}$ is continuously differentiable on the unit circle, the Taylor series expansion of $\tilde{f}$ converges uniformly to $\tilde{f}$ on $\overline{U}$, and therefore the partial sums

$$S_n(z) := \sum_{j=0}^{n} c_j \cdot \left(\frac{1-z}{1+z}\right)^j \qquad (n \in \mathbf{N}),$$

converge uniformly to $\frac{\vartheta_2(z)}{z+1}$ on the closed right half plane.

For every $n \in \mathbf{N}$, the n^{th} approximation of $\hat{Q}_n(z)$ is given by

$$\hat{Q}_n(z) = \frac{z+4}{z+1} + 9 \cdot S_n(z),$$

and this approximation $\hat{Q}_n(z)$ leads to a compensator with transfer function $C_n(z)$:

$$C_n(z) \;:=\; \hat{P}(z)(\hat{Q}_n(z))^{-1} = \frac{\frac{9e^2}{z+1}}{\frac{z+4}{z+1} + 9S_n(z)} =$$

$$= \frac{9e^2}{z+4+9(z+1)S_n(z)}.$$

Note that $C_n(z)$ $(n \in \mathbf{N})$ may be realized by an n^{th} order finite-dimensional system Γ_n (so Γ_n is a minimal realization of a linear system without time-delays). This may be verified by multiplying both the numerator and the denominator of the previous expression for $C_n(z)$ by $(z+1)^{n-1}$. Moreover, since $\hat{Q}_n(z)$ converges uniformly to $\hat{Q}(z)$ $(n \to \infty)$, it is guaranteed that this n^{th} order compensator Γ_n becomes internally stabilizing for suitably large n.

Let $n \in \mathbf{N}$ and rewrite $C_n(z)$ as

$$C_n(z) = \frac{a_n(z)}{b_n(z)},$$

where $a_n(z) = 9e^2(z+1)^{n-1}$ is a polynomial in $\mathbf{R}[z]$ of degree $n-1$ and $b_n(z) = (z+1)^{n-1}(z+4) + 9(z+1)^n S_n(z)$ a polynomial of degree n. Then the transfer function of the closed-loop system of Σ and Γ_n is given by

$$T_{cl}(z) = \frac{\frac{e^{-z}}{z-2}}{1 + \frac{e^{-z}}{z-2} \cdot \frac{a_n(z)}{b_n(z)}} = \frac{e^{-z}b_n(z)}{(z-2)b_n(z) + e^{-z}a_n(z)}.$$

Since there are no pole-zero cancellations between $T(z)$ and $C_n(z)$, it is easily verified that the closed-loop system is internally stable if and only if $(z-2)b_n(z)+e^{-z}a_n(z)$ is a stable exponential polynomial. Using the stability test for exponential polynomials of Section 6.1 we find that for $n = 17$ the exponential polynomial $(z-2)b_n(z) + e^{-z}a_n(z)$ has no zeros in $\overline{\mathbf{C}^+}$. Hence Γ_{17} is a finite-dimensional internally stabilizing feedback compensator with transfer function $C_{17}(z)$. However, in comparison with the order of the delay system Σ (i.e the order in the algebraic sense), the order of this compensator is extremely high.

To find a solution that is more attractive from a practical point of view, we try another approximation method. Consider the transfer function $C(z)$ of the original compensator,

$$C(z) = \frac{9e^2}{z+4+9\vartheta_2(z)},$$

and apply the Padé approximation method around the point $z = 0$. In this way we obtain for every $n \in \mathbf{N}$ a real rational function of the form

$$C_{P,n} = \frac{\alpha_0 + \alpha_1 z + \cdots + \alpha_n z^n}{\beta_0 + \beta_1 z + \cdots + \beta_n z^n},$$

that is a good approximation of the transfer function $C(z)$ around $z = 0$: the coefficients α_i and β_i $(i = 0, 1, \ldots, n)$ are chosen in such a way that in $z = 0$ the transfer functions $C(z)$ and $C_{P,n}(z)$ and their derivatives up to a certain order take the same values. However, uniform convergence on the closed right half plane of $C_{P,n}(z)$ to $C(z)$ for $n \to \infty$ is not guaranteed. Let $\Gamma_{P,3}$ be a minimal realization of the transfer function $C_{P,3}(z)$, obtained with the Padé approximation method. Then it turns out that already this third order compensator $\Gamma_{P,3}$ is internally stabilizing the system Σ. So although Padé approximation seems not very attractive theoretically, we find a compensator that is very useful from the practical point of view.

To give a somewhat intuitive explanation of this surprising behaviour of the Padé approximation method, we have a look at the Nyquist plot of the feedback interconnection of the transfer functions $T(z)$ and $C(z)$ (i.e the transfer functions of the system Σ and the original compensator). This is the curve $\Upsilon = \{T(\imath\omega)C(\imath\omega) \mid \omega \in \mathbf{R}\}$ as depicted in Figure 6.5; it gives rise to a straightforward graphical test for the stability of the closed-loop system in the following way. Let P_T and P_C denote the number of right half plane poles of $T(z)$ and $C(z)$, respectively, and let N_Υ denote the number of clockwise encirclements of the point -1 by the Nyquist curve Υ when $\imath\omega$ traverses the imaginary axis from $-\imath\infty$ to $+\imath\infty$. Then the Nyquist criterion (a consequence of the circle criterion) states that the number of right half plane poles P_{cl} of the closed-loop system of $T(z)$ and $C(z)$ is

$$P_{cl} = N_\Gamma + P_T + P_C.$$

It is obvious that $T(z)$ has one pole in the right half plane: $z = 2$. Using the stability test for exponential polynomials we find that $C(z)$ has two poles in the right half plane. Finally, zooming in on the point -1, we see in Figure 6.6 that the Nyquist curve Υ encircles the point -1 three times in counter clockwise direction, and thus

$$P_{cl} = -3 + 1 + 2 = 0.$$

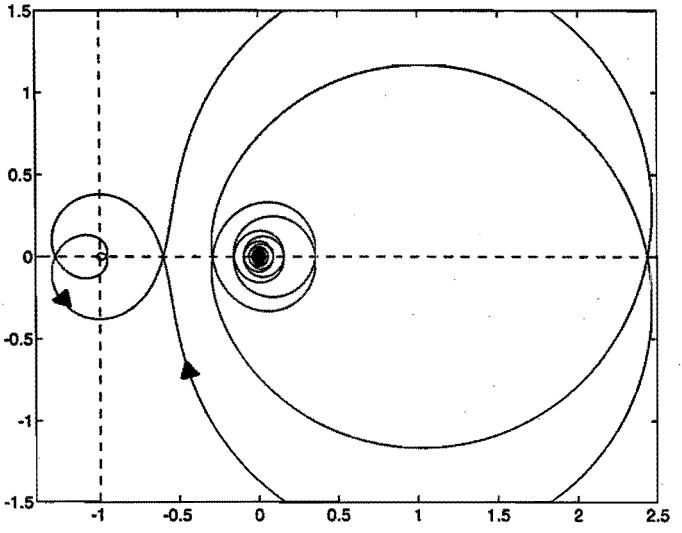


Figure 6.5: Nyquist plot of $T(z)C(z)$

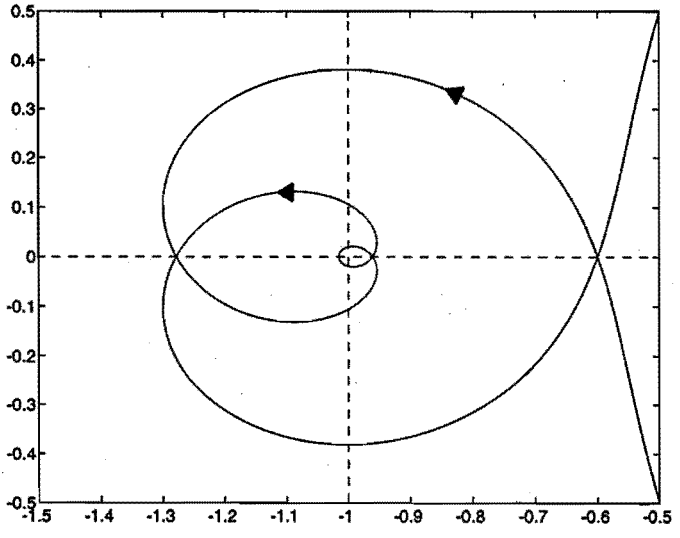


Figure 6.6: Nyquist plot of $T(z)C(z)$ around the point -1

So indeed the closed-loop system is BIBO-stable. However, the Nyquist plot also indicates that this result is not very robust. The Nyquist plot intersects the real axis in -1.0153 (at frequency $\omega = 0$) and in -0.96 (at frequency $\omega = \pm 1.3$), and therefore small perturbations of the compensator may affect the closed-loop stability. This explains why in the original approximation method such a high order compensator was required to ensure closed-loop stability: the transfer function $C(z)$ has to be approximated very accurately.

Next, consider the third order compensator $\Gamma_{P,3}$ with transfer function $C_{P,3}(z)$, computed with Padé approximation of $C(z)$ around $z = 0$. The Nyquist plot for the closed-loop interconnection of $T(z)$ and $C_{P,3}(z)$ is given in Figure 6.7. The number of clockwise encirclements of the point -1 when $w\omega$ traverses the imaginary axis from $-\imath\infty$ to $+\imath\infty$ is easily determined with help of the more detailed Figure 6.8. We conclude that $N_\Upsilon = -3$. Since the transfer function $C_{P,3}(z)$ is a real rational function, the poles of $C_{P,3}(z)$ may be determined numerically, and it turns out that two of them are contained in $\overline{\mathbb{C}^+}$. So again

$$P_{cl} = -3 + 1 + 2 = 0,$$

and the closed-loop is stable.

In this particular case, an explanation for the surprisingly good performance of the Padé approximation is not difficult to give. The Nyquist plot of Figure 6.5 indicates that especially at low frequencies the approximation of the transfer function $C(z)$ of the original compensator has to be very accurate because the encirclement of the point -1 at low frequencies is very critical. By definition, Padé approximation around $z = 0$ is a very good approximation of $C(z)$ when the modulus of z is small. This is also illustrated by Figure 6.6 and 6.8: around the point -1 both Nyquist plots are almost identical. We conclude that the Padé approximation has exactly the property that is required in this particular example.

The Nyquist plot of $T(z)$ with the transfer function $C_{17}(z)$ obtained with the original approximation method, resembles Figure 6.5 and 6.7 to a great extent and is therefore omitted. It is important to note that also this compensator has an extremely small stability margin. Also in this respect the high order compensator Γ_{17} is not better than the controller $\Gamma_{P,3}$ computed with the Padé approximation method.

Remark 6.2.22 Example 6.2.21 is taken from [54]: in this article the same system is also studied in detail. We repeated this example because the results obtained in [54] are different from ours, and therefore it seemed worthwhile to discuss the same example. In the article it is claimed that an 11^{th} order controller based on the Cesaro means approximation method described in Subsection 6.2.3 is already stabilizing. We were not able to reproduce this result; according to our computations this compensator is not stabilizing. Instead we used the Fourier series approximation because it is well known that the convergence of this method is faster. Nevertheless very high order approximations are still necessary.

The results of Example 6.2.21 lead to some new questions. Is Padé approximation always a good method for the approximation of the matrices $\hat{Q}(z)$ and $\hat{P}(z)$ in (6.33) that satisfy Bezout identity (6.32)? The example suggests that the choice

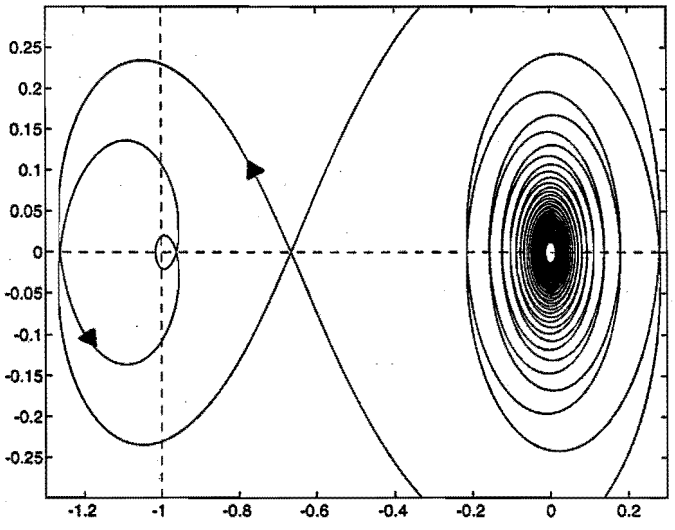


Figure 6.7: Nyquist plot of $T(z)C_{P,3}(z)$

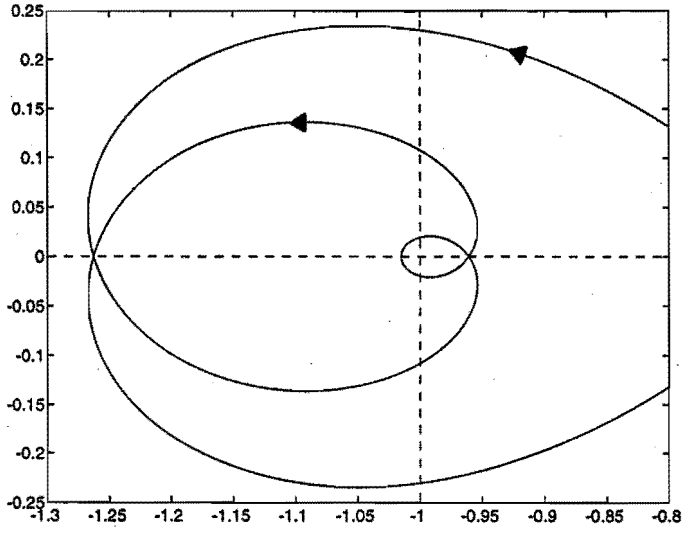


Figure 6.8: Nyquist plot of $T(z)C_{P,3}(z)$ around the point -1

of a good approximation method is problem dependent. However, because of the substantial reduction of the order of the stabilizing compensator in this particular example, further investigation of the applicability of Padé approximation seems worthwhile. The second question addresses the problem of stability robustness. We have seen that all compensators computed in Example 6.2.21 have a very small stability margin. Using some results on robust stabilization it is probably possible to find out whether this property is problem dependent, or inherent to this stabilization method. The results in [16] indicate that in this particular example only a very small stability margin can be realized. This suggests that the stabilization problem for the system in Example 6.2.21 is really very difficult. In any case, this example illustrates that a constructive solution to the stabilization problem is difficult to give, and that the stabilization method described in this section is not the final solution to this problem.

6.3 Alternative stabilization methods

The constructive approach to the stabilization problem presented in the previous section is not the only method known in literature to solve this problem. However, the method of Section 6.2 fits very well in the algebraic approach to time-delay systems and therefore we treated it in more detail. In this section we give an overview of some alternative stabilization methods. Two of them are mainly based on the infinite-dimensional systems approach mentioned in Section 1.3. The third method fits in the algebraic framework; it can be seen as a generalization of the pole-placement result of Section 2.6. This selection of alternative methods is certainly not complete. Moreover, we shall not discuss these methods in full detail, but confine ourselves to the explanation of the most important ideas. The purpose is to show that besides the more algebraic method of Section 6.2, there exist various other interesting approaches to the stabilization problem.

6.3.1 Approximation of delay systems

The first method we present is a well-known approach to the stabilization problem that can be applied to a rather general class of infinite-dimensional systems, including time-delay systems. It is based on an inversion of the ideas behind the method of Section 6.2. Instead of approximating an infinite-dimensional stabilizing controller by a finite-dimensional one, we start approximating the original time-delay system by a finite-dimensional system (i.e. a linear system without delays). Based on this approximation a finite-dimensional compensator is designed in the expectation that this compensator also stabilizes the delay system. In the sequel we only discuss the key-ideas of this approach, and for simplicity we only consider single input single output (SISO) systems. For a detailed elaboration, also in the multi input multi output (MIMO) case, we refer to the literature, for example [16], [31], [75], [33] and [34] and the references therein.

Let Σ be a time-delay system with transfer function $T(z)$. According to Lemma 6.1.3, this system has only a finite number of poles in $\overline{\mathbb{C}^+}$. Therefore we may

decompose $T(z)$ as

$$T(z) = T_{st}(z) + T_u(z),$$

where $T_u(z)$ is the transfer function of a finite-dimensional anti-stable system (i.e. $T_u(z)$ is a rational function with all its poles in $\overline{\mathbb{C}^+}$), and $T_{st}(z)$ is the transfer function of a BIBO-stable infinite-dimensional system. In fact, $T_{st}(z)$ is the transfer function of a BIBO-stable time-delay system. So in particular, $T_{st}(z)$ has only a finite number of poles in any right half plane.

The main idea is now to approximate the infinite-dimensional stable part $T_{st}(z)$ of Σ by a stable finite-dimensional system with transfer function $\hat{T}_{st}(z)$. Next we consider the robust stabilization problem for the finite-dimensional system $\hat{\Sigma}$ with transfer function $T_u(z) + \hat{T}_{st}(z)$, and regard $T_{st}(z) - \hat{T}_{st}(z)$ as an additive perturbation. If $\hat{T}_{st}(z)$ is close enough to $T_{st}(z)$, the compensator Γ with transfer function $K(z)$, designed to stabilize $T_u(z) + \hat{T}_{st}(z)$ with a prescribed additive robustness margin, will also stabilize the original system Σ.

This idea was elaborated in [16]. Recall from Definition 3.2.1 that the norm of a function $f \in \mathcal{A}_0(\mathbb{C}^+)$ is given by

$$\|f\|_\infty := \sup\{|f(\imath\omega)| \mid \omega \in \mathbb{R}\}.$$

In [16], Curtain and Glover prove that there exist a number $\varepsilon > 0$, completely determined by the anti-stable part $T_u(z)$ of the transfer function $T(z)$, and a controller Γ with transfer function $K(z)$, only depending on $T_u(z)$ and the finite-dimensional approximant $\hat{T}_{st}(z)$ of $T_{st}(z)$, with the following property. If

$$\|T_{st} - \hat{T}_{st}\|_\infty < \varepsilon, \tag{6.46}$$

then the controller Γ stabilizes the original system Σ. In fact, Γ is designed as a stabilizing compensator for the system $\hat{\Sigma}$ (with transfer function $T_u(z) + \hat{T}_{st}(z)$), that is additively robust against stable perturbations bounded in norm by ε. In particular, this implies that if $\delta > 0$ is such that

$$\|T_{st} - \hat{T}_{st}\|_\infty \leq \delta < \varepsilon,$$

and if $T(z)$ is perturbed with an additive stable perturbation $\Delta(z)$ bounded in norm by $\|\Delta\|_\infty < \varepsilon - \delta$, then Γ also stabilizes the system with transfer function $T(z) + \Delta(z)$.

Note that it is possible to split up the method in several consecutive steps that may be carried out one by one. First the number ε is determined; for this only the anti-stable part $T_u(z)$ of $T(z)$ is required. Next we have to find an approximation $\hat{T}_{st}(z)$ of the stable part $T_{st}(z)$ that satisfies (6.46). For this purpose various techniques are proposed in the literature. In [16], Hankel norm approximation is discussed, but the convergence of this method can only be guaranteed under some additional constraints (see also [75]). In [33] an alternative approximation method is presented that resembles the approximation techniques of Subsection 6.2.3. First the transfer function $T_{st}(z)$ is transformed into a function on the closed unit disc, and next Fourier series or Cesaro means are used to find an appropriate approximation. When the approximation step is completed successfully, we may construct a stabilizing compensator.

We start to solve the robust stabilization problem for the finite-dimensional anti-stable system with transfer function $T_u(z)$. Using the techniques described in [31], we construct a compensator with transfer function $K_1(z)$, that satisfies $\lim_{|z|\to\infty} 1+\hat{T}_{st}(z)K_1(z) \neq 0$, and that is robust against stable additive perturbations $\Delta(z)$ of $T_u(z)$, bounded in norm by

$$\|\Delta\|_\infty < \varepsilon.$$

Then Theorem 4.1 in [16] states that

$$K(z) = \frac{K_1(z)}{1+\hat{T}_{st}(z)K_1(z)}$$

is the transfer function of a (BIBO)-stabilizing controller for the system Σ.

It is possible to modify the method a little and to incorporate also approximations of the anti-stable part $T_u(z)$ of $T(z)$. In this case, the stabilization method is slightly changed (see [16] for the details). This modification is useful because the anti-stable part of $T(z)$ is sometimes difficult to extract from the system. An approximation method for $T_u(z)$ is developed in [34]. The techniques that are used for this are similar to the ones of Subsection 6.2.3.

The advantages of the stabilization method based on the approximation of time-delay systems are clear. Since the value of ε can be determined first, it is known beforehand how accurate the approximation of the stable part of a system has to be. Once such an approximation is obtained, it is relatively easy to compute a stabilizing compensator. Recall that in the method of the previous section we have to compute a controller in every step of the algorithm, and test whether it is already stabilizing. This condition is replaced by a condition on the approximation: if the required level of accuracy has been reached, the method ensures that the corresponding compensator is stabilizing. Moreover, the same approach is applicable to a far more general class than the time-delay systems considered in this thesis. However, this is also a disadvantage. The time-delay character of the system is never used in the construction of the compensator. In fact, the delay system is stabilized by regarding it as a finite-dimensional system of very high order. This is no problem if in the modeling of a real world system a time-delay is used to model some unknown dynamics. On the other hand, if the time-delay occurring in the system has a physical interpretation, it is preferable to maintain the time-delay character of the system. However, in the approach to the stabilization problem described in this subsection, the additional information on the algebraic structure of a time-delay system is not used, although it might be important for several control purposes.

6.3.2 A direct approach to stabilization

The next method we discuss is taken from [84] and is based on the infinite-dimensional systems approach to time-delay systems as mentioned in Section 1.3. Originally, the main purpose of this article was to prove that infinite-dimensional systems may be stabilized by *finite-dimensional* controllers. Since the proof of this result is completely constructive, the method is also applicable in the more algebraic framework of this thesis. Unlike the method of Subsection 6.3.1, the approach in [84]

is not based on the approximation of a time-delay system by a finite-dimensional system; instead it works in a more direct way. Although also in this method an approximation step is required, the main ideas and techniques that are used to derive the result are completely different from the ones we have seen up to now. In [84] the method is described for a rather general class of infinite-dimensional systems in state-space form. Since we did not introduce this terminology in this thesis, we only explain the main ideas of the approach in a finite-dimensional setting. The generalization to infinite-dimensional systems involves a lot of technicalities and is omitted. For a detailed elaboration we refer to [84].

Let $\Sigma = (A, B, C, 0)$ be an n-dimensional system over $\mathbf{R}$ with m inputs and p outputs. So Σ is a finite-dimensional system without time-delays. Assume that (A, B) is stabilizable and (C, A) is detectable. Then there exist matrices $F \in \mathbf{R}^{m\times n}$ and $G \in \mathbf{R}^{n\times p}$ such that both $A+BF$ and $A+GC$ are stable:

$$\sigma(A+BF) \subset \mathbf{C}^- \quad \text{and} \quad \sigma(A+GC) \subset \mathbf{C}^-.$$

Next consider the n^{th} order dynamic compensator Γ defined by

$$\Gamma = (A+BF+GC, -G, F, 0). \tag{6.47}$$

According to formula (2.27), the closed-loop system of Σ and Γ, depicted in Figure 2.2, is given by

$$\Sigma_{cl} = \left(\begin{pmatrix} A & -BF \\ GC & A+BF+GC \end{pmatrix}, \begin{pmatrix} B \\ 0 \end{pmatrix}, (C \mid 0), 0 \right),$$

and it is easily verified (see for example [62, Section 5.2]) that

$$\sigma \begin{pmatrix} A & -BF \\ GC & A+BF+GC \end{pmatrix} = \sigma(A+BF) \cup \sigma(A+GC) \subset \mathbf{C}^-.$$

Hence the closed-loop system is internally stable.

Note that in this construction of an internally stabilizing compensator it is not guaranteed that the system Γ is reachable and observable, and therefore it is possible that Γ is not a *minimal* realization of its transfer function

$$T_\Gamma(z) = -F(zI - (A+BF+GC))^{-1}G.$$

If not, there are pole-zero cancellations in $T_\Gamma(z)$, and there exists a system of lower order than Γ, with the same transfer function $T_\Gamma(z)$, that also stabilizes the system Σ. It is our objective to find such a low order stabilizing compensator.

Assume that $(A+BF+GC, G)$ is not reachable, and let $\mathcal{V}$ be the linear subspace of $\mathbf{R}^n$ consisting of all reachable points. Then $\mathcal{V}$ is also the reachable subspace of $(A+BF, G)$, and therefore it can be characterized as the smallest subspace $\mathcal{V}$ of $\mathbf{R}^n$ with the properties:

(i) $(A+BF)\mathcal{V} \subset \mathcal{V}$,

(ii) $\operatorname{im}(G) \subset \mathcal{V}$.

The next proposition states that in this situation there also exists a stabilizing controller for Σ of order $\dim(\mathcal{V})$.

Proposition 6.3.1 *Let $\Sigma = (A, B, C, 0)$ be a system over $\mathbf{R}$ of order n that is both stabilizable and detectable. Let F and G be matrices such that $A+BF$ and $A+GC$ are stable matrices, i.e. all their eigenvalues are contained in $\mathbb{C}^-$. Suppose that there exists a k-dimensional subspace $\mathcal{V}$ in $\mathbf{R}^n$ such that*

(i) $(A+BF)\mathcal{V} \subset \mathcal{V}$,

(ii) $\mathrm{im}(G) \subset \mathcal{V}$.

Then there exists a stabilizing compensator for Σ of order k. ∎

For a proof of this result we refer to [84, Lemma 4.2]; the idea is that the subspace $\mathcal{V}$ suffices as the state-space of a stabilizing compensator.

Next, consider the system $\Sigma = (A, B, C, 0)$, and apply a state-space transformation to decompose the system in a stable and anti-stable part. In this way obtain the following block representation of the matrices A, B and C:

$$A = \begin{pmatrix} A_u & 0 \\ 0 & A_s \end{pmatrix}, \quad \begin{pmatrix} B_u \\ B_s \end{pmatrix}, \quad C = (C_u \mid C_s), \tag{6.48}$$

with $\sigma(A_u) \subset \overline{\mathbb{C}^+}$ and $\sigma(A_s) \subset \mathbb{C}^-$. Since we assume that Σ is stabilizable and detectable, it follows that (A_u, B_u) is reachable and that (C_u, A_u) is observable. So there exist matrices F_u and G_u such that both $A_u + B_u F_u$ and $A_u + G_u C_u$ are stable matrices. Defining $F := (F_u \mid 0)$ and $G := \begin{pmatrix} G_u \\ 0 \end{pmatrix}$, we know that the corresponding compensator Γ as defined in (6.47) is internally stabilizing the system Σ. After determination of the reachable subspace $\mathcal{V}$, we may apply Proposition 6.3.1 to compute a reduced order compensator.

It is even possible to go one step further and to use the freedom in the choice of the matrices F and G to reduce the dimension of the reachable subspace $\mathcal{V}$. Suppose that F is fixed and consider a perturbation $\tilde{G}$ of the matrix G. Since the eigenvalues of $A+GC$ depend continuously on G, and since $A+GC$ is stable, we know that if $\|G-\tilde{G}\|$ is small enough, the matrix $A+\tilde{G}C$ remains stable. Now we choose $\tilde{G}$ close enough to G to ensure that $A+\tilde{G}C$ is stable, and in such a way that the dimension k of the reachable subspace of $(A+BF, \tilde{G})$ is as small as possible. Finally we apply Proposition 6.3.1 to find a controller of order k that still stabilizes Σ.

The main observation in [84] is that under certain regularity conditions (that are not very restrictive for time-delay systems), the same approach can be applied to infinite-dimensional systems. Let Σ be a time-delay system, and recall from Lemma 6.1.3 that the unstable part of such a system is always finite-dimensional. Therefore the system may be represented in the same way as in (6.48), with A_u an anti-stable finite-dimensional matrix, and A_s an infinite-dimensional operator that is stable. Now recall that in the construction of a stabilizing controller in the finite-dimensional case only the matrices A_u, B_u and C_u are involved. Although in the present situation the system as a whole is infinite-dimensional, the matrices A_u,

B_u and C_u are still finite-dimensional, and therefore exactly the same techniques as before can be applied. The crucial result of [84] is that the reduction of the compensator can be carried out in such a way that a *finite-dimensional* compensator is obtained. In fact, it is shown that there exists a finite-dimensional subspace $\mathcal{V}$ of the state-space, and operators F and $\tilde{G}$ such that both $A+BF$ and $A+\tilde{G}C$ are stable and

(i) $\forall x \in \mathcal{V}: \ (A+BF)x \in \mathcal{V}$,

(ii) $\mathrm{im}(\tilde{G}) \subset \mathcal{V}$.

Since a generalized version of Proposition 6.3.1 also holds in the case of time-delay systems, this implies that it is possible to construct a finite-dimensional controller with state-space isomorphic to $\mathcal{V}$, that stabilizes the system Σ.

In [84] the stabilization method is applied to the time-delay system

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = \begin{pmatrix} -\frac{\pi}{2}\sigma & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t),$$

$$y(t) = x_1(t),$$

where $\sigma x_1(t) = x_1(t-1)$. The example illustrates that the method can be carried out in a completely constructive way. However, even in this simple example the computations are already rather involved. This does not imply that the present approach is not interesting because also the two previous methods are computationally very demanding. Therefore it would be interesting to make a comparison of the stabilizing compensators obtained with the various methods, and of the computational costs that are involved. Such a comparison is left for future research.

Remark 6.3.2 We want to stress once more that the construction method of this section does not fit in the algebraic framework; it is completely based on the infinite-dimensional systems approach to time-delay systems. However, since the method yields a *finite-dimensional* controller, the final result is also applicable in the algebraic setting. Therefore we decided to include also this method. It illustrates that the two different approaches to time-delay systems are not completely separated. Results that are obtained within one framework, are often very useful in the other framework.

6.3.3 Generalized pole placement

In contrast to the previous approaches, the last stabilization method we discuss is based on the algebraic framework of Chapter 2. Unfortunately, this method is not generally applicable; first of all we have to confine ourselves to system with commensurable time-delays, and even then the method works only for a rather restricted subclass. However, the ideas behind this approach differ so much from what we have seen up to now, that it is certainly worthwhile to consider this method. One of the main differences with the previous methods is that we do not apply dynamic feedback as defined in Section 2.7 to stabilize a system, but use a generalized pole placement technique. This idea originates from Sontag who mentioned it in [85]; a more detailed elaboration is given in [37, Section 3.4].

Let $\Sigma = (A, B)$ be a system with a commensurable time-delay τ, modeled as a system over the ring $\mathbf{R}[s]$. So the indeterminate s corresponds to a delay operator σ with time-delay τ. Since $\mathbf{R}[s]$ is a principal ideal domain (PID), reachability of the system Σ is equivalent to pole assignability (see Proposition 2.6.5). This implies that if Σ is reachable, it is not necessary to apply dynamic state feedback to stabilize the system; instead a static state feedback (possibly containing the delay operator σ) suffices to assign the poles of the closed-loop system to some arbitrary values $p_1, \ldots, p_n \in \mathbf{R}^-$. However, the reachability condition is rather restrictive. Recalling the genericity result for reachability (see Proposition 2.2.5), we know that in this particular situation of a system over a polynomial ring in one indeterminate, reachability is a generic property if and only if the number of inputs to the system is at least two. So for time-delay systems with commensurable time-delays that have only one input, the pole placement technique described above cannot be applied generically. The method proposed in this subsection is mainly intended to weaken the rather restrictive reachability condition.

For this purpose we first recall Definition 6.2.9 of the ring $\mathbf{R}_{\mathcal{U}}(s)$:

$$\mathbf{R}_{\mathcal{U}}(s) = \left\{ \frac{a(s)}{b(s)} \mid a(s), b(s) \in \mathbb{R}[s] \text{ and } \forall \lambda \in \overline{\mathcal{U}} : b(\lambda) \neq 0 \right\},$$

where $\mathcal{U}$ denotes the open unit disc. It is evident that

$$\mathbf{R}[s] \subset \mathbf{R}_{\mathcal{U}}(s) \subset \mathbf{R}(s),$$

and therefore the system $\Sigma = (A, B)$ may also be considered as a system over the ring $\mathbf{R}_{\mathcal{U}}(s)$. The following two results clarify the use of this change of point of view.

Lemma 6.3.3 *$\mathbf{R}_{\mathcal{U}}(s)$ is a principal ideal domain.* ∎

For a proof of this result we refer to [37, pp. 33-35]. Recalling Proposition 2.6.5 we conclude that the result of Lemma 6.3.3 implies that for systems over the ring $\mathbf{R}_{\mathcal{U}}(s)$, the properties of reachability and pole assignability are equivalent. The next result can be seen as an adaptation of Theorem 2.2.4 to systems over the ring $\mathbf{R}_{\mathcal{U}}(s)$.

Theorem 6.3.4 *Let $\Sigma = (A, B)$ be a system over the ring $\mathcal{R} = \mathbf{R}_{\mathcal{U}}(s)$, with $A \in \mathcal{R}^{n \times n}$ and $B \in \mathcal{R}^{n \times m}$. Then $\Sigma = (A, B)$ is reachable if and only if*

$$\forall \hat{z} \in \mathbf{C}\ \forall \hat{s} \in \overline{\mathcal{U}} : \operatorname{rank}(\hat{z}I - A(\hat{s}) | B(\hat{s})) = n. \tag{6.49}$$

∎

The proof of Theorem 6.3.4 follows the same lines as the proof of Theorem 2.2.4. The necessity of (6.49) is trivial, and for the sufficiency part one first shows that the set $\{\mathcal{I}_\alpha \mid \alpha \in \overline{\mathcal{U}}\}$, where $\mathcal{I}_\alpha$ is defined as the ideal $\mathcal{I}_\alpha = \{p \in \mathbf{R}_{\mathcal{U}}(s) \mid p(\alpha) = 0\}$, is the set of all maximal ideals in $\mathbf{R}_{\mathcal{U}}(s)$. Subsequent application of the local-global theorem (see Appendix A.3, Theorem A.3.4) yields the desired result. The details of the proof can be found in [37].

At this point, the crucial idea behind the stabilization method of this section becomes visible. Let $\Sigma = (A, B)$ be a time-delay system, and do not consider it as a system over the ring $\mathbf{R}[s]$, but as a system over the ring $\mathbf{R}_{\mathcal{U}}(s)$. In this way, the

condition on the reachability of the system has become much weaker. However, since $\mathbf{R}_{\mathcal{U}}(s)$ is a principal ideal domain, reachability of a system over $\mathbf{R}_{\mathcal{U}}(s)$ is equivalent to pole assignability, and in this situation a generalized pole placement technique may be applied to stabilize the time-delay system.

Assume that $\Sigma = (A, B)$ (where A is an $n \times n$ and B is an $n \times m$ matrix) is reachable over $\mathbf{R}_{\mathcal{U}}(s)$, and choose $p_1, \ldots, p_n \in \mathbf{R}^-$. Then there exists an $m \times n$ matrix K over $\mathbf{R}_{\mathcal{U}}(s)$ such that

$$\det(zI - (A + BK)) = (z - p_1)(z - p_2) \cdots (z - p_n).$$

Note that the matrix K is not necessarily an element of $\mathbf{R}[s]^{m \times n}$; its entries may contain rational functions in the indeterminate s. In this case, K cannot be implemented directly as a static state feedback, and we have to look for another way to realize K.

Denote by $\hat{K}$ the proper $m \times n$ matrix over $\mathbf{R}(s)$ defined by

$$\hat{K}(s) := K(\frac{1}{s}).$$

Since K is a matrix over $\mathbf{R}_{\mathcal{U}}(s)$, all entries of $\hat{K}$ are proper rational functions. Moreover, $\hat{K}$ has no poles outside $\mathcal{U}$ and therefore it can be seen as the transfer matrix of a *stable* discrete-time system. Let $\Gamma = (F, G, H, J)$ be a realization of $\hat{K}$:

$$\Gamma \left\{ \begin{array}{rcl} v(t+\tau) & = & Fv(t) + Gw(t), \\ y(t) & = & Hv(t) + Jw(t), \end{array} \right. \tag{6.50}$$

but consider the equations (6.50) as the dynamical equations of a system evolving in *continuous time*. We apply this continuous-time system Γ as a feedback compensator for Σ. In this way Σ becomes a so-called *neutral system*. This is a system governed by a differential-difference equation in which also delays of the derivative of the evolution variable are involved. From the derivation in [37] it follows that the poles of this neutral system are located in the points $p_1, \ldots, p_n \in \mathbf{R}^-$ and on a finite number of vertical lines in $\mathbf{C}^-$ (on these lines poles of the system lie at distances of 2π from each other). Therefore there exists a $\delta > 0$ such that all poles of the neutral system are contained in the half plane $\{z \in \mathbf{C} \mid \operatorname{Re} z \leq -\delta\}$, and according to [41, Section 1.7 and Chapter 12] this implies that this system is internally stable. Now the closed-loop system consists of a stable neutral component and of the stable component Γ, and therefore also the closed-loop system is internally stable. We conclude that Γ is a stabilizing compensator for Σ.

The advantage of the present approach is obvious. The method is completely algebraic, and independent of the actual length of the commensurable time-delay τ. Moreover, all steps of the construction can be carried out explicitly. The only difficulty arises in the computation of the matrix K over $\mathbf{R}_{\mathcal{U}}(s)$ that assigns the poles of $A + BK$ to $p_1, \ldots, p_n$. However, since $\mathbf{R}_{\mathcal{U}}(s)$ is a principal ideal domain, the algorithms developed in [20] and [21] may be used for this purpose.

On the other hand, the method has some important shortcomings. First of all it is only applicable to systems with commensurable time-delays, and even then condition (6.49) on the reachability over the ring $\mathbf{R}_{\mathcal{U}}(s)$ is far more restrictive than

the stabilizability condition of Corollary 3.2.9. Moreover, the controller Γ is not a dynamical compensator as described in Section 2.7, but a continuous-time system governed by discrete-time equations. The implementation of this kind of systems is a nontrivial task. Finally, after application of the compensator Γ, the system Σ becomes a neutral system, and therefore the closed-loop system does not belong to the class of time-delay systems considered in this thesis any more. Especially this last aspect is a very unattractive feature of this generalized pole-placement technique.

6.3.4 Closing remarks

Although we proposed several constructive methods for the solution of the stabilization problem for time-delay systems, an attractive generally applicable method has not been found yet. Most of the methods we described are based on the infinite-dimensional systems approach to time-delay systems, and do not use the algebraic structure of a time-delay system with point delays. Moreover, all these methods yield finite-dimensional stabilizing compensators, and therefore the freedom of incorporating time-delays in the compensator, that is very natural from the algebraic point of view, is not exploited. Sometimes the employment of this additional degree of freedom may lead to a very simple solution to the stabilization problem, as is illustrated in the next example.

Example 6.3.5 Consider a time-delay system $\Sigma = (A, B)$ with commensurable time-delay $\tau = 1$, modeled as a system over the ring $\mathbb{R}[s]$. So the indeterminate s corresponds to the delay operator σ with time-delay 1. Suppose that A and B are given by

$$A = 1, \qquad B = s - 1.$$

Then the system Σ is stabilizable by dynamic state feedback because the rank condition of Theorem 3.2.8 is satisfied:

$$\forall z \in \overline{\mathbb{C}^+} : \ \operatorname{rank}(z - 1 \mid e^{-z} - 1) = 1.$$

Let $\Gamma = (F, G, H, J)$ be the dynamical compensator defined by

$$F = s^4 + s^3 + s^2 + s - 3, \quad G = -s^3 - 2s^2 - 3s - 5, \quad H = 1, \quad J = 0,$$

and consider the closed-loop system $\Sigma_{cl} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ of Σ and Γ as described in (2.28) - (2.31). Then

$$\hat{A} = \begin{pmatrix} 1 & -s+1 \\ -s^3 - 2s^2 - 3s - 5 & s^4 + s^3 + s^2 + s - 3 \end{pmatrix},$$

and with this formula, the characteristic polynomial of the closed-loop system is easily computed:

$$\det(zI - \hat{A}) = z^2 + (2 - s - s^2 - s^3 - s^4)z + (2 - s).$$

Now substitute e^{-z} for the indeterminate s, and recall Example 6.1.13. Then we see that the characteristic function of the closed-loop system is a stable exponential polynomial, and we conclude that Γ is an internally stabilizing compensator of Σ.

Example 6.3.5 indicates that there exist stabilizing feedback compensators that fit very well in the algebraic framework, but that cannot be obtained with one of the stabilization methods of this chapter because they contain time-delays. The method of Subsection 6.3.3 is not even applicable in Example 6.3.5 because the system Σ is not reachable over $\mathbf{R}_{\mathcal{U}}(s)$. Nevertheless there exists a solution to the stabilization problem that is very attractive from the algebraic point of view, and therefore the question arises how this kind of controllers can be obtained in general.

Unfortunately we are not able to answer this question. As we saw before, in the algebraic setup the stabilization problem comes down to the construction of a polynomial in the set

$$\mathcal{I} \cap \mathcal{D}.$$

This problem contains both an algebraic part (the ideal $\mathcal{I}$ associated with the system Σ), and a more analytic part (the Hurwitz set $\mathcal{D}$ describing stability). It is unclear how these two aspects of the stabilization problem should be combined in order to find a constructive solution to the stabilization problem. In Section 5.5 we have seen that for the stabilizability question the reformulation of the Hurwitz set $\mathcal{D}$ as the set of all monic polynomials that have no zeros in a given set W, helps to solve the problem. Together with the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$, this reformulation enables us to use both the algebraic and the time-delay character of the system. We expect that also for the stabilization problem, a reformulation in terms of the set W and the variety $\mathcal{V}(\mathcal{I})$ might help to find a constructive solution.

Chapter 7

Summary and conclusions

In this thesis, we have investigated the applicability of methods from constructive commutative algebra to time-delay systems with point delays. For this purpose we used the so-called algebraic approach to time-delay systems. After the introduction of a number of delay operators $\sigma_1, \ldots, \sigma_k$ corresponding to the incommensurable time-delays $\tau_1, \ldots, \tau_k$ occurring in a system Σ, the system equations of Σ may be written as

$$\begin{cases} \dot{x}(t) &= A(\sigma_1, \ldots, \sigma_k)x(t) + B(\sigma_1, \ldots, \sigma_k)u(t), \\ y(t) &= C(\sigma_1, \ldots, \sigma_k)x(t) + D(\sigma_1, \ldots, \sigma_k)u(t). \end{cases}$$

Replacing the delay operators $\sigma_1, \ldots, \sigma_k$ by indeterminates $s_1, \ldots, s_k$, the quadruple of matrices $(A(s_1, \ldots, s_k), B(s_1, \ldots, s_k), C(s_1, \ldots, s_k), D(s_1, \ldots, s_k))$ can be considered as a system over the polynomial ring $\mathbb{R}[s_1, \ldots, s_k]$. The philosophy behind the algebraic approach is to consider control problems first within the framework of systems over rings, and to apply the design methods obtained in this algebraic setting to the particular case of time-delay systems. The main advantage of this detour is that in the algebraic setup powerful techniques from constructive commutative algebra are available for the necessary computations.

It is important to note that the class of systems over polynomial rings is more general than the class of time-delay systems with point delays. In the rewriting process we replaced the delay operators $\sigma_1, \ldots, \sigma_k$ by indeterminates $s_1, \ldots, s_k$, and in this way we removed the delay character from the system. Although this implies that the theory of systems over rings is more generally applicable, for example to systems with unknown parameters, this extension also has a disadvantage: we have lost some valuable information of the systems we are interested in. To solve this problem, we have refined our strategy. If a problem is not solvable in the algebraic framework of systems over rings it is allowed to use the time-delay character of the system again. With this additional information it is often possible to obtain stronger results. This omitting and recovering of information played an important role throughout the thesis. Originally we tried to proceed as long as possible within the algebraic framework, and applied the additional knowledge only when it was necessary. In this way we expected to obtain a maximal profit from the methods from constructive commutative algebra that are available in the algebraic setup. However, for some problems more flexibility in the switching between the purely algebraic and the delay character of the system was required. In our treatment we

separated these aspects as far as possible. Chapters 2, 4 and 5 (except Section 5.5) were devoted to systems over (polynomial) rings in general. In Chapter 3, Section 5.5 and Chapter 6, the specialization to systems with time-delays was discussed.

In Chapter 2, an introduction was given to the main ideas behind the theory of systems over rings. It was shown how important system theoretic concepts like reachability and observability may be generalized to a purely algebraic setting. For stability, the situation was somewhat more complicated. First, the notion of stability was translated to the algebraic framework by introducing so-called Hurwitz sets: multiplicative and saturated sets of monic polynomials. Using this concept, the stabilizability problem for systems over integral domains was solved theoretically. For this solution, only the algebraic properties of a Hurwitz set mentioned above were required. However, in the application to time-delay systems, the Hurwitz set describing stability possesses a much richer structure: it is defined as the set of all polynomials that have no zeros in a specified subset of $\mathbb{C}^{k+1}$, where k denotes the number of incommensurable time-delays occurring in the systems. One of the main questions in this thesis was how the additional information on the zero structure of the polynomials in the Hurwitz set can be used for control purposes.

In Chapter 3 a partial answer to this question was given. The additional information on the time-delay character of a system was used to specialize the results on stabilizability of Chapter 2 to the delay case. In this way a generalized Hautus test for stabilizability was obtained that is applicable for a general class of stability domains. The second part of Chapter 3 was devoted to the genericity of stabilizability. The proof of this result illustrated how powerful the algebraic approach to time-delay systems can be when also the delay character of the system is used. First the algebraic structure of time-delay systems was exploited to define a natural topology on this class of systems. Within this topological framework it was shown that the set of stabilizable delay systems contains a subset that is both open and dense in the parameter-space describing all time-delay systems. This indicates that for time-delay systems the property of stabilizability is very weak; it is satisfied for almost all time-delay systems. In the proof of this result the delay character of the system was used extensively. So a flexible way of switching between algebraic and delay aspects of a system led to a proof of the genericity of stabilizability.

The last three chapters were devoted to computational aspects of systems with time-delays. In Chapter 4 an overview was given of two methods in constructive commutative algebra: Gröbner bases and characteristic sets. Both methods can be applied to manipulate polynomial ideals and to determine (theoretically) the variety of an ideal. It turned out that Gröbner bases are more suitable to deal with polynomial ideals. With this method it is possible to carry out operations on polynomial ideals explicitly. The characteristic sets method is mainly aimed at the computation of the variety of polynomial ideals. For our purposes, Gröbner bases seemed the most appropriate tool, and therefore this method was mainly used in the sequel.

In Chapter 5, the Gröbner basis method was applied to the reachability and stabilizability problem for systems over polynomial rings. Given a system $\Sigma = (A, B, C, D)$, the right-invertibility conditions on the matrix $(zI - A|B)$ derived in

Chapter 2 were transformed to conditions on a polynomial ideal associated to the system Σ. This ideal can be seen as a description of all characteristic polynomials that can be obtained from Σ after application of a dynamic state feedback. Several methods were developed to compute a Gröbner basis of this ideal and to determine its variety. Specialization of these results to the reachability problem gave rise to various algorithms to test the reachability of a system over a polynomial ring. The performances of these methods were compared, and two of these algorithms turned out to be very effective from a computational point of view. As a byproduct of one of these reachability tests, a constructive method for the computation of a polynomial right-inverse of a nonsquare polynomial matrix was obtained. This result has interesting applications because right-inverses of this type are often required in the construction of feedback compensators. An example is the input-output decoupling problem over unique factorization domains (see [18]). However, in most cases, the right-inverse obtained in the algorithm is very complex. Since also a characterization of the set of all right-inverses of a nonsquare polynomial matrix was given, it seems that only a simplification algorithm is required to obtain a right-inverse that is of practical interest. Unfortunately the situation is not that easy because the simplification issue is not only a computational problem. It is not even clear what kind of simplifications are interesting for our control purposes. Therefore we should first specify what goals we want to achieve with a simplification algorithm.

Next the stabilizability problem for time-delay systems was considered. To solve this problem, algebraic manipulations were not sufficient any more; the delay character of the system was actually required. In an algorithm for testing the stabilizability of a time-delay system, this additional information can be used at two different stages, and this led to two different verification methods. Which method is preferable depends on the system under consideration.

Finally we investigated the stabilization problem for time-delay systems. In the algebraic setup it is very difficult to find a constructive solution to this problem. This is due to the fact that in the framework of systems over rings only the algebraic properties of a the stability defining Hurwitz set can be used. Since a Hurwitz set is only a multiplicative set and not a polynomial ideal, the constructive methods from Chapter 4 are not applicable to Hurwitz sets. Therefore we were not able to solve this problem using Gröbner basis techniques. Probably the zero structure of the polynomials in the Hurwitz set has to be used explicitly to obtain a constructive solution. Maybe some progress can be made by combining this additional information on the delay character of the system, with the variety of the ideal associated to a system over a ring that was studied in Chapter 5.

Instead of an algebraically motivated solution, an overview of some existing methods in the literature to solve the stabilization problem was given. For this purpose a numerical test for the verification of the internal stability of a time-delay system was developed. This method solves some of the numerical problems that arose in the classical approach when the stability of high order time-delay systems was tested. This was very useful in the sequel because most stabilization methods involve an approximation step. This often leads to a closed-loop time-delay system of very high order.

Most of the constructive stabilization methods that were investigated were based on the notion of bounded-input bounded-output stability. In this setup it is easier

to use the delay character of the system, and therefore constructive solutions of the stabilization problem can be obtained. Next an approximation step was carried out to find internally stabilizing controllers. Moreover, some alternative methods, mostly based on the infinite-dimensional systems approach to time-delay systems, were suggested. In these algorithms approximation of infinite-dimensional systems and infinite-dimensional controllers by finite-dimensional systems and controllers played an important role.

The last observations indicate that for the study of time-delay systems both the algebraic and the infinite-dimensional systems approach are useful. These approaches stress different aspects of the same class of systems. Although it seems difficult to combine these approaches into one framework, they are both necessary for a good understanding of time-delay systems. Hopefully this thesis showed that apart from the mainstream approach based on the theory of infinite-dimensional systems, the algebraic framework is very useful for the study of time-delay systems from a computational point of view.

This thesis contains a contribution to the application of the systems over rings approach to time-delay systems, but both in the theory of systems over rings and in the field of time-delay systems, a lot of problems remain unsolved. Therefore we conclude with a small list of interesting topics and suggestions for possible future research. We already encountered some of these problems in the thesis.

(i) In Section 3.2, the condition for stabilizability developed in the setting of systems over rings was specialized to time-delay systems. In this way a right-invertibility condition over the ring $\mathcal{R}_{\mathcal{D}}(z)$ of stable transfer functions was replaced by a pointwise rank condition. In the proof of this generalization of the Hautus test to time-delay systems, the delay character of the system was used explicitly. For other applications it is unclear under what conditions on the Hurwitz set $\mathcal{D}$ such a reformulation is possible. For systems over the ring $\mathcal{K}[s_1, \ldots, s_k]$ it is likely that Hurwitz sets of the form (5.40) play an important role in answering this question. It would be interesting to know whether the results of Section 3.2 really depend on the delay character of the systems, or whether they can be extended to a more general class of Hurwitz sets.

(ii) In this thesis, the topologies for polynomial matrices and time-delay systems introduced in Section 3.3 were primarily used to prove our genericity result on stabilizability. It can be expected that the same topologies have other interesting applications. E.g. in [72], a same sort of topology was used to study the continuity of AR-representations of dynamical systems. Moreover, in this thesis, the topology on delay systems was only used in a qualitative way. May be it is possible to find a modification of the topological setup that enables us to find also some quantitative results.

(iii) In Section 5.4 we developed a constructive method for the computation of a polynomial right-inverse of a nonsquare polynomial matrix. In most cases, such a right-inverse is a very complicated object. Since we were able to characterize the class of all polynomial right-inverses, the question arises how we

can find a right-inverse that is suitable for subsequent control purposes. Before we can find an algorithm that carries out the simplifications that are required for this, we first need a characterization of the properties of the polynomial right-inverse we are interested in. These properties have to reflect the control objectives we finally want to achieve with a compensator based on this right-inverse matrix. The relationship between right-inverses and the corresponding controllers is not very well understood, and therefore this subject seems a very promising research topic.

(iv) In this thesis, a combination of computer algebra and numerical methods turned out to be a very fruitful approach for the solution of the stabilizability problem. These mixed algebraic/numerical algorithms are applicable to a much larger class of problems. One of the important questions is how the numerical sensitivity of an algorithm that consists of an exact and a numerical part can be investigated. In Section 5.5 we have seen that for example Gröbner basis computations may lead to numerically complicated problems. Therefore another question is how a mixed algebraic/numerical algorithm should be organized in order to obtain numerically reliable results.

(v) In Chapters 5 and 6, we argued that the stabilization problem for time-delay systems consists of an algebraic and a more analytic part. The ideal $\mathcal{I}$ associated with a system Σ is a completely algebraic object, whereas the Hurwitz set $\mathcal{D}$ describing stability has a more analytic character. So in order to find a stabilizing feedback compensator, it seems logical to use an algorithm that consists of both an algebraic and a numerical part. The development of constructive methods for the design of stabilizing feedback compensators, exploiting the algebraic structure of the problem, would be a very useful contribution to the theory. This problem seems to be very difficult. From the observations in Section 5.5, we expect that the variety $\mathcal{V}(\mathcal{I})$ of the ideal $\mathcal{I}$ associated with a system plays an important role in this problem. However, the question how this information can be used in the design of stabilizing controllers remains unanswered.

Appendix A

Some results from commutative algebra

In this appendix we give a short overview of some results from commutative algebra that are used throughout this thesis. It is divided into three parts. The first part mainly consists of definitions of some of the basic concepts in commutative algebra and of a few classical results in this field. In the second part we only consider polynomial rings and study the relationship between ideals and their varieties. In the third section we state and prove the local-global theorem and show how it can be applied to polynomial rings. For further reading we refer to one of the classical textbooks on commutative algebra, for example Atiyah and MacDonald ([1]), the two volumes of van der Waerden ([93] and [94]), or the work of Zariski and Samuel ([104] and [105]).

A.1 Basic definitions and results

We start with the formal definition of a (commutative) ring, the most important concept from commutative algebra used in this thesis.

Definition A.1.1 A *ring* $\mathcal{R}$ is a set with two binary operations, $+$ (addition) and $\cdot$ (multiplication) such that

(i) $\mathcal{R}$ is an Abelian group with respect to addition (i.e. $\mathcal{R}$ has a zero element denoted by 0, and every $a \in \mathcal{R}$ has an additive inverse $-a$),

(ii) multiplication is *associative*: if $a, b, c \in \mathcal{R}$ then $(ab)c = a(bc)$,

(iii) multiplication is *distributive* over addition: if $a, b, c \in \mathcal{R}$, then $a(b+c) = ab+ac$ and $(b+c)a = ba+ca$.

Moreover, if

(iv) for all $a, b \in \mathcal{R}$: $ab = ba$,

then $\mathcal{R}$ is said to be a *commutative* ring.

Note that a ring element always has an additive inverse, but not necessarily a multiplicative inverse, so division of two nonzero ring elements is not always allowed. It is even possible that there exist nonzero elements $a, b \in \mathcal{R}$ such that $a \cdot b = 0$. Such elements are called *zero divisors.*

A nonzero element of a ring $\mathcal{R}$ that is an identity with respect to multiplication is called the *identity* of the ring. If it exists, it is uniquely determined and denoted by 1.

Definition A.1.2 An *integral domain* is a commutative ring with identity and without zero divisors.

Definition A.1.3 A *field* $\mathcal{K}$ is an integral domain in which every nonzero element has a multiplicative inverse.

Unlike rings, in a field division by a nonzero element is always possible.

Definition A.1.4 A mapping T of a commutative ring $\mathcal{R}$ into a commutative ring $\mathcal{S}$ is called a *ring homomorphism*, if for any pair of elements $a, b \in \mathcal{R}$ the following two conditions are satisfied:

(i) $T(a+b) = Ta + Tb$,

(ii) $T(a \cdot b) = (Ta) \cdot (Tb)$.

Moreover, if this mapping T is *bijective* it is called a *ring isomorphism*, and the rings $\mathcal{R}$ and $\mathcal{S}$ are said to be *isomorphic.*

In a ring homomorphism, the ring operations addition and multiplication are preserved. This implies that isomorphic rings are essentially the same. There is a one-to-one correspondence between the elements of the rings, and also addition and multiplication are carried out in a completely analogous way.

Next we introduce another important concept related to rings, namely the notion of ideals.

Definition A.1.5 Let $\mathcal{R}$ be a commutative ring. An *ideal* of $\mathcal{R}$ is a non-empty subset $\mathcal{I}$ of $\mathcal{R}$ such that

(i) if $a_1, a_2 \in \mathcal{I}$ then $a_1 - a_2 \in \mathcal{I}$,

(ii) if $a \in \mathcal{I}$ and $b \in \mathcal{R}$, then $a \cdot b \in \mathcal{I}$.

Moreover, if $\mathcal{I} \neq \mathcal{R}$, then $\mathcal{I}$ is called a *proper* ideal of $\mathcal{R}$.

From the definition it is self-evident that the intersection $\mathcal{I} \cap \mathcal{J}$ of two ideals $\mathcal{I}$ and $\mathcal{J}$ of $\mathcal{R}$ is again an ideal. Also the sum

$$\mathcal{I} + \mathcal{J} := \{a + b \in \mathcal{R} \mid a \in \mathcal{I},\ b \in \mathcal{J}\} \tag{A.1}$$

satisfies conditions *(i)* and *(ii)* of Definition A.1.5, and is an ideal of $\mathcal{R}$.

Example A.1.6 Let $\mathcal{R}$ be a commutative ring and $P \subset \mathcal{R}$. Then

$$\langle P \rangle := \{ \sum_{p \in P} \alpha_p \cdot p \mid \alpha_p \in \mathcal{R} \}$$

is an ideal in $\mathcal{R}$. $\langle P \rangle$ is called the ideal *generated* by P; it is the smallest ideal in $\mathcal{R}$ that contains P.

Often, an ideal can be characterized by a finite number of elements generating the ideal.

Definition A.1.7 Let $\mathcal{I}$ be an ideal of the commutative ring $\mathcal{R}$. Then

(i) $\mathcal{I}$ is called *finitely generated* if there exists a finite set $P \subset \mathcal{R}$ such that $\langle P \rangle = \mathcal{I}$,

(ii) $\mathcal{I}$ is called a *principal ideal* if there exists an element $p \in \mathcal{R}$ such that $\langle p \rangle = \mathcal{I}$.

Definition A.1.8 An integral domain $\mathcal{R}$ in which every ideal $\mathcal{I}$ of $\mathcal{R}$ is principal, is called a *principal ideal domain*, mostly abbreviated to PID.

Examples of PID's are the ring $\mathbb{Z}$ of all integers, and the ring of all polynomials in one indeterminate with coefficients in a field.

Let $\mathcal{R}$ be a commutative ring and $\mathcal{I}$ an ideal of $\mathcal{R}$, and define $\mathcal{R}/\mathcal{I}$ as the set of all equivalence classes under the equivalence relation $\sim$, given by

$$a \sim b \Longleftrightarrow a - b \in \mathcal{I}.$$

Defining addition and multiplication by

$$\begin{aligned} \bar{a} + \bar{b} &:= \overline{a+b}, \\ \bar{a} \cdot \bar{b} &:= \overline{a \cdot b}, \end{aligned}$$

where $\bar{a}$ denotes the equivalence class of a, $\mathcal{R}/\mathcal{I}$ is itself turned into a commutative ring.

Next we consider some ideals with special properties.

Definition A.1.9 An ideal $\mathcal{I}$ of a commutative ring $\mathcal{R}$ is called a *maximal* ideal if one of the following (equivalent) conditions is satisfied:

(i) $\mathcal{I}$ is a proper ideal of $\mathcal{R}$ and there does not exist an ideal $\mathcal{J}$ of $\mathcal{R}$ such that $\mathcal{I} \subsetneq \mathcal{J} \subsetneq \mathcal{R}$,

(ii) $\mathcal{R}/\mathcal{I}$ is a field.

Proposition A.1.10 *Every proper ideal of a commutative ring $\mathcal{R}$ with identity is contained in a maximal ideal of $\mathcal{R}$.* ∎

The proof of this result is based on Zorn's Lemma and may be found in e.g [104, p. 151].

Definition A.1.11 An ideal $\mathcal{I}$ of a commutative ring $\mathcal{R}$ is called a *prime* ideal if one of the following (equivalent) conditions is satisfied:

(i) $\mathcal{I}$ is a proper ideal of $\mathcal{R}$ and for all $a, b \in \mathcal{R}$ we have $a \cdot b \in \mathcal{I} \Longrightarrow a \in \mathcal{I}$ or $b \in \mathcal{I}$,

(ii) $\mathcal{R}/\mathcal{I}$ is an integral domain.

In some sense a prime ideal of a ring $\mathcal{R}$ may be considered as a generalization of the concept of prime numbers. In the same fashion we can introduce *primary* ideals which correspond to powers of prime numbers.

Definition A.1.12 An ideal $\mathcal{I}$ of a commutative ring $\mathcal{R}$ is called a *primary* ideal if it is a proper ideal of $\mathcal{R}$ and for all $a, b \in \mathcal{R}$:

$$a \cdot b \in \mathcal{I} \Longrightarrow \text{ either } a \in \mathcal{I} \text{ or } b^n \in \mathcal{I} \text{ for some } n > 0.$$

Definition A.1.13 Let $\mathcal{I}$ be an ideal of a commutative $\mathcal{R}$. Then the *radical* of $\mathcal{I}$, denoted by $\sqrt{\mathcal{I}}$, is the ideal of $\mathcal{R}$ defined by

$$\sqrt{\mathcal{I}} := \{a \in \mathcal{R} \mid \exists n \in \mathbf{N} : a^n \in \mathcal{I}\}.$$

If $\mathcal{I} = \sqrt{\mathcal{I}}$, then $\mathcal{I}$ is called a *radical* ideal.

It is obvious that the properties on ideals defined above are highly related. These interdependences are depicted in the following implication scheme.

$$\begin{array}{ccccc} & & \text{maximal ideal} & & \\ & & \Downarrow & & \\ \text{primary ideal} & \Leftarrow & \text{prime ideal} & \Rightarrow & \text{radical ideal} \end{array}$$

The implications in the other directions do not hold in general.

In the major part of this thesis we are concerned with a special type of commutative rings, namely *Noetherian* rings.

Definition A.1.14 A *Noetherian ring* is a commutative ring $\mathcal{R}$ satisfying one of the following (equivalent) conditions:

(i) Every ideal $\mathcal{I}$ of $\mathcal{R}$ is finitely generated,

(ii) Every strictly ascending chain $\mathcal{I}_1 \subsetneqq \mathcal{I}_2 \subsetneqq \mathcal{I}_3 \subsetneqq \cdots$ of ideals of $\mathcal{R}$ is finite. Alternatively stated: Given an ascending chain $\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathcal{I}_3 \subset \cdots$ of ideals of $\mathcal{R}$, there exists an $n \in \mathbf{N}$ such that for all $j > n$: $\mathcal{I}_j = \mathcal{I}_n$.

Condition *(ii)* is called the *ascending chain condition.*

It is obvious that fields and principal ideal domains are special kinds of Noetherian rings; in fact, the class of Noetherian rings is quite extensive. This is due to the important feature that the property of being a Noetherian ring carries over from a Noetherian ring $\mathcal{R}$ to all polynomial rings over $\mathcal{R}$.

Theorem A.1.15 (Hilbert basis theorem) *If $\mathcal{R}$ is a Noetherian ring, then any polynomial ring in a finite number of indeterminates and with coefficients in $\mathcal{R}$ is a Noetherian ring.* ∎

For a proof of the Hilbert basis theorem we refer to e.g. [104, p. 201] or [1, p. 81].

When $\mathcal{R}$ is a commutative ring, the ring of all polynomials over $\mathcal{R}$ in the indeterminates $x_1, \ldots, x_n$ is denoted by $\mathcal{R}[x_1, \ldots, x_n]$. In this thesis we mostly encounter polynomial rings in which the coefficient ring $\mathcal{R}$ is a field $\mathcal{K}$. According to the Hilbert basis theorem the polynomial ring $\mathcal{K}[x_1, \ldots, x_n]$ is Noetherian, and thus any ideal in this polynomial ring is finitely generated.

In Noetherian rings primary ideals play an important role.

Theorem A.1.16 (Lasker-Noether decomposition theorem) *In a Noetherian ring every ideal admits a representation as a finite intersection of primary ideals.* ∎

If an ideal $\mathcal{I}$ of a Noetherian ring is radical, the primary ideals of this decomposition are even prime.

Corollary A.1.17 *If $\mathcal{R}$ is a Noetherian ring and $\mathcal{I}$ a radical ideal of $\mathcal{R}$, then $\mathcal{I}$ admits a representation*

$$\mathcal{I} = \bigcap_{i=1}^{k} \mathcal{P}_i,$$

where all $\mathcal{P}_i$ $(i = 1, \ldots, k)$ are prime ideals. ∎

The proofs of both Theorem A.1.16 and Corollary A.1.17 may be found in [104, pp. 208-210].

Finally we introduce the concept of modules that may be seen as a generalization of vector spaces to the ring case.

Definition A.1.18 Let $\mathcal{R}$ be a commutative ring. A set $\mathcal{M}$ is called a *module* over $\mathcal{R}$ (or an $\mathcal{R}$-module) if the following conditions hold:

(i) $\mathcal{M}$ is a commutative group (the group operation will be written as addition).

(ii) Every ordered pair (a, x) in which $a \in \mathcal{R}$ and $x \in \mathcal{M}$ is associated with a unique element of $\mathcal{M}$, denoted by ax, in such a way that the following relations hold:

$$\begin{aligned} a(x+y) &= ax + ay, \\ (a+b)x &= ax + bx, \\ (ab)x &= a(bx), \end{aligned}$$

where a and b are arbitrary elements of $\mathcal{R}$, and x, y are arbitrary elements of $\mathcal{M}$. The element ax is called the *product* of a and x.

Comparing the definition of an ideal with the definition of a module above, it turns out that an ideal $\mathcal{I}$ of a commutative ring $\mathcal{R}$ is also an $\mathcal{R}$-module. However, the concept of modules is far more general. For example, if $\mathcal{R}$ is a commutative ring, the set $\mathcal{R}^n$ consisting of all n-tuples of elements of $\mathcal{R}$ is an $\mathcal{R}$-module. Addition and multiplication are simply defined coordinate-wise.

The notion of homomorphic and isomorphic mappings, that was introduced for rings in Definition A.1.4, is easily extended to modules.

Definition A.1.19 Let $\mathcal{R}$ be a commutative ring, and let $\mathcal{M}$ and $\mathcal{N}$ be modules over $\mathcal{R}$. A mapping T of $\mathcal{M}$ into $\mathcal{N}$ is called an *$\mathcal{R}$-homomorphism* if the following conditions are satisfied

(i) $\forall x, y \in \mathcal{M} : T(x+y) = Tx + Ty$,

(ii) $\forall x \in \mathcal{M} \forall a \in \mathcal{R} : T(ax) = aT(x)$.

Moreover, if the mapping T is also *bijective*, then T is called an *$\mathcal{R}$-isomorphism.*

In the same way as for ideals, modules that can be characterized by a finite number of generating elements, are of special interest.

Definition A.1.20 Let $\mathcal{R}$ be a commutative ring, and let $\mathcal{M}$ be a module over $\mathcal{R}$. Then

(i) $\mathcal{M}$ is called a *finitely generated module* if there exists a finite number of elements $m_1, \ldots, m_k$ in $\mathcal{M}$ such that for every element $x \in \mathcal{M}$ there exists a k-tuple $(a_1, \ldots, a_k) \in \mathcal{R}^k$ such that

$$x = \sum_{i=1}^{k} a_i m_i.$$

(ii) A set B is called a *basis* of $\mathcal{M}$ if for every element $x \in \mathcal{M}$ there exist unique coefficients a_b $(b \in B)$ such that

$$x = \sum_{b \in B} a_b \cdot b.$$

(iii) $\mathcal{M}$ is called *free* if $\mathcal{M}$ admits a basis.

From Definition A.1.20 it follows that a finitely generated free $\mathcal{R}$-module $\mathcal{M}$ is isomorphic to $\mathcal{R}^n$ for some $n \in \mathbf{N}$.

A.2 Polynomial ideals and varieties

In this section we consider polynomial rings over a field $\mathcal{K}$. We are interested in the sets of all common zeros of the polynomial ideals of these rings. Such sets are called varieties and there is a strong link between polynomial ideals and their varieties. This relationship is elaborated in more detail.

Let $\mathcal{K}$ be an arbitrary field, and $\mathcal{R} := \mathcal{K}[x_1, \ldots, x_n]$ the ring of all polynomials in the indeterminates $x_1, \ldots, x_n$ with coefficients in $\mathcal{K}$. Before we can speak of the zeros of a polynomial in $\mathcal{K}[x_1, \ldots, x_n]$, we first have to specify in what kind of set we are looking for zeros.

Definition A.2.1 A field $\mathcal{K}$ is called *algebraically closed* if one of the following two equivalent conditions is satisfied:

(i) every nonconstant polynomial of $\mathcal{K}[x]$ has at least one root in $\mathcal{K}$ and thus a linear factor in $\mathcal{K}$,

(ii) every polynomial in $\mathcal{K}[x]$ splits into linear factors.

Note that in the definition of algebraically closedness only univariate polynomials are involved.

Definition A.2.2 Let $\mathcal{K}$ be a field and $\mathcal{L}$ be an extension field of $\mathcal{K}$ (i.e. $\mathcal{K} \subset \mathcal{L}$). Then

(i) An element $\alpha \in \mathcal{L}$ is called *algebraic over* $\mathcal{K}$ if there exists a nonzero polynomial $p(x) \in \mathcal{K}[x]$ such that $p(\alpha) = 0$.

(ii) The extension field $\mathcal{L}$ of $\mathcal{K}$ is called *algebraic over* $\mathcal{K}$ if every element of $\mathcal{L}$ is algebraic over $\mathcal{K}$.

Theorem A.2.3 *For every field $\mathcal{K}$ there exists an algebraically closed, algebraic extension Ω for $\mathcal{K}$. This extension field Ω is unique up to equivalent extensions.* ∎

For a proof of this result we refer to e.g. [93, Section 10.1]

According to Theorem A.2.3, algebraically closed algebraic extensions of a field $\mathcal{K}$ are essentially unique (up to certain isomorphisms), and therefore we may speak of *the algebraic closure* of $\mathcal{K}$, denoted by $\bar{\mathcal{K}}$. This is the set in which we want to find common zeros of polynomial ideals.

Definition A.2.4 Let $\mathcal{I}$ be an ideal in $\mathcal{K}[x_1, \ldots, x_n]$, and let $\bar{\mathcal{K}}$ denote the algebraic closure of $\mathcal{K}$. The (algebraic) *variety* of $\mathcal{I}$ in the affine space $\bar{\mathcal{K}}^n$ is the set

$$V := \{(\alpha_1, \ldots, \alpha_n) \in \bar{\mathcal{K}}^n \mid \forall p \in \mathcal{I} : p(\alpha_1, \ldots, \alpha_n) = 0\}.$$

The variety V of an ideal $\mathcal{I}$ is denoted by $\mathcal{V}(\mathcal{I})$.

Note that the set of all common zeros in $\bar{\mathcal{K}}^n$ of a finite number of polynomials $p_1, \ldots, p_m$ in $\mathcal{K}[x_1, \ldots, x_n]$ is simply the variety of the ideal $\langle p_1, \ldots, p_m \rangle$ generated by these polynomials.

It is also possible to go in the opposite direction and to associate an ideal with a given subset of $\bar{\mathcal{K}}^n$.

Definition A.2.5 Let V be a subset of $\bar{\mathcal{K}}^n$. Then $\mathrm{Id}(V)$ is defined as the set of all polynomials in $\mathcal{K}[x_1, \ldots, x_n]$ that vanish in every point of V:

$$\mathrm{Id}(V) := \{p \in \mathcal{K}[x_1, \ldots, x_n] \mid \forall (\alpha_1, \ldots, \alpha_n) \in V : p(\alpha_1, \ldots, \alpha_n) = 0\}.$$

$\mathrm{Id}(V)$ is an ideal of $\mathcal{K}[x_1, \ldots, x_n]$.

From these definitions it is obvious that there is a strong link between ideals and their varieties. This is illustrated by the following relations. Let $\mathcal{I}$ and $\mathcal{J}$ be ideals in $\mathcal{K}[x_1, \ldots, x_n]$, and let V and W be subsets of $\bar{\mathcal{K}}^n$. Then we have

$$\mathcal{I} \subset \mathcal{J} \implies \mathcal{V}(\mathcal{J}) \subset \mathcal{V}(\mathcal{I}), \tag{A.2}$$

$$V \subset W \implies \mathrm{Id}(W) \subset \mathrm{Id}(V), \tag{A.3}$$

$$\mathcal{V}(\mathcal{I} + \mathcal{J}) = \mathcal{V}(\mathcal{I}) \cap \mathcal{V}(\mathcal{J}), \tag{A.4}$$

$$\mathrm{Id}(V \cup W) = \mathrm{Id}(V) \cap \mathrm{Id}(W), \tag{A.5}$$

and moreover

$$\mathcal{V}(\mathcal{I} \cap \mathcal{J}) = \mathcal{V}(\mathcal{I}) \cup \mathcal{V}(\mathcal{J}), \tag{A.6}$$

$$\mathcal{V}(\mathcal{I}) = \mathcal{V}(\sqrt{\mathcal{I}}). \tag{A.7}$$

Most of these relations are self-evident; they follow directly from the definition. For a proof of (A.6) we refer to [105, p. 161].

Definition A.2.6 A variety V (defined over a field $\mathcal{K}$) is called *reducible* (over $\mathcal{K}$) if it is decomposable into two varieties V_1 and V_2 that are defined over $\mathcal{K}$ and are proper subsets of V. If such a decomposition does not exist, V is called *irreducible* (over $\mathcal{K}$).

Proposition A.2.7 *A variety V is irreducible if and only if* $\mathrm{Id}(V)$ *is a prime ideal.* ∎

A proof of Proposition A.2.7 is given in [105, p.162].

The next theorem can be seen as a restatement of the Lasker-Noether decomposition theorem for radical ideals in the terminology of algebraic varieties.

Theorem A.2.8 *Every variety V can be represented as a finite sum of irreducible varieties $V_1, \ldots, V_h$:*

$$V = \bigcup_{i=1}^{h} V_i. \tag{A.8}$$

This decomposition is unique (up to the order in which $V_1, \ldots, V_h$ are written) if it is irredundant, i.e. if $V_i \not\subset V_j$ for $i \neq j$. ∎

We refer to for example [105, pp. 162-163] or [14, pp. 204-205] for a proof of this result.

Let $\mathcal{K}$ be a field and $V \in \bar{\mathcal{K}}^n$, and consider the set $\mathcal{V}(\mathrm{Id}(V))$. From Definitions A.2.4 and A.2.5 it is clear that $V \subset \mathcal{V}(\mathrm{Id}(V))$. The inclusion in the other direction does not always hold. In fact we have

$$\mathcal{V}(\mathrm{Id}(V)) = V \iff V \text{ is a variety,} \tag{A.9}$$

so there has to exists an ideal $\mathcal{I}$ of $\mathcal{K}[x_1,\ldots,x_n]$ such that $\mathcal{V}(\mathcal{I}) = V$.

For polynomial ideals, the same question of successive determination of varieties and ideals may be considered. Let $\mathcal{I}$ be an ideal of $\mathcal{K}[x_1,\ldots,x_n]$. Using Definitions A.2.4 and A.2.5 it is not difficult to prove that $\mathcal{I} \subset \mathrm{Id}(\mathcal{V}(\mathcal{I}))$. A necessary and sufficient condition for the inclusion in the other direction to hold, is more difficult to obtain. For this purpose we need

Theorem A.2.9 (Hilbert Nullstellensatz) *Let $\mathcal{K}$ be a field and $\bar{\mathcal{K}}$ the algebraic closure of $\mathcal{K}$. Let $p, p_1,\ldots,p_q$ be polynomials in the ring $\mathcal{K}[x_1,\ldots,x_n]$. Assume that p vanishes at every common zero of $p_1,\ldots,p_q$ in $\bar{\mathcal{K}}^n$. Then there exists an exponent $r \in \mathbb{N}$ and polynomials $a_1,\ldots,a_q$ in $\mathcal{K}[x_1,\ldots,x_n]$ such that*

$$p^r = a_1p_1 + a_2p_2 + \cdots + a_qp_q.$$ ∎

A proof of the Hilbert Nullstellensatz may be found in e.g. [105, pp. 164-167] or [94, Section 16.5].

Corollary A.2.10 *Let $\mathcal{I}$ be an ideal of $\mathcal{K}[x_1,\ldots,x_n]$. Then*

$$\mathcal{V}(\mathcal{I}) = \varnothing \iff \mathcal{I} = \mathcal{K}[x_1,\ldots,x_n]. \tag{A.10}$$

∎

The Hilbert Nullstellensatz yields the answer to our question on $\mathrm{Id}(\mathcal{V}(\mathcal{I}))$.

Corollary A.2.11 *Let $\mathcal{I}$ be an ideal of $\mathcal{K}[x_1,\ldots,x_n]$. Then*

(i) $\mathrm{Id}(\mathcal{V}(\mathcal{I})) = \sqrt{\mathcal{I}}$,

(ii) $\mathrm{Id}(\mathcal{V}(\mathcal{I})) = \mathcal{I} \iff \mathcal{I}$ *is a radical ideal.* ∎

A proof of Corollary A.2.11 *(i)* is given in [14, pp. 175-176], and *(ii)* follows immediately from *(i)*.

With the Hilbert Nullstellensatz also the equivalence of the decomposition theorem for algebraic varieties and the Lasker-Noether decomposition theorem for radical polynomial ideals is established. According to Theorem A.2.8, every algebraic variety V is decomposable into a finite number of irreducible algebraic varieties $V_1,\ldots,V_h$. According to the Hilbert Nullstellensatz, $\mathrm{Id}(V)$ is a radical ideal. Moreover, Proposition A.2.7 yields that the ideals $\mathrm{Id}(V_i)$ $(i = 1,\ldots,h)$ are prime, and thus after successive application of (A.5), we obtain

$$\mathrm{Id}(V) = \bigcap_{i=1}^{h} \mathrm{Id}(V_i).$$

This is exactly the result of Corollary A.1.17 in the special case of polynomial rings.

A.3 The local-global theorem and its application

In this section, an important theorem from commutative algebra is stated and proved: the local-global theorem. In Chapter 2 it is used, together with the Hilbert Nullstellensatz, to restate a right-invertibility condition as a pointwise rank condition. This was applied to the matrix $(zI - A|B)$ over the ring $\mathcal{R}[z]$, where $\mathcal{R} = \mathbb{R}[s_1, \ldots, s_k]$. Here the question of the surjectivity of a map from one module to another is considered in a more general context. It turns out that the investigation can be facilitated a lot using the local-global theorem. The derivation of this result given below, is based on [42] and [8, pp. 76-78].

Let $\mathcal{R}$ denote a commutative ring with identity, and let $\mathcal{M}$ and $\mathcal{N}$ be $\mathcal{R}$-modules. Consider an $\mathcal{R}$-homomorphism $T : \mathcal{M} \to \mathcal{N}$. In this section the main question is: when is this map T surjective?

Definition A.3.1 Let $\mathcal{I}$ be an ideal in $\mathcal{R}$, and $\mathcal{M}$ a module over $\mathcal{R}$. Let $\mathcal{IM}$ denote the set

$$\mathcal{IM} := \{\sum_j \alpha_j m_j \mid \alpha_j \in \mathcal{I}, m_j \in \mathcal{M}\}. \tag{A.11}$$

Then the *factor module* $\mathcal{M}_\mathcal{I}$ of $\mathcal{M}$ with respect to $\mathcal{I}$ is defined as

$$\mathcal{M}_\mathcal{I} := \mathcal{M}/\mathcal{IM}. \tag{A.12}$$

So $\mathcal{M}_\mathcal{I}$ is the set of all equivalence classes under the equivalence relation $\sim$ given by

$$x \sim y \Longleftrightarrow x - y \in \mathcal{IM}.$$

Note that $\mathcal{M}_\mathcal{I}$ can be made into an $\mathcal{R}$-module by defining

$$\begin{aligned} \overline{x} + \overline{y} &:= \overline{x+y}, \\ r\overline{x} &:= \overline{rx}, \end{aligned}$$

where $\overline{x}$ denotes the equivalence class of x. It is clear that both operations above are well defined. The map $x \mapsto \overline{x}$ from $\mathcal{M}$ to $\mathcal{M}_\mathcal{I}$ is called the *canonical projection.*

Lemma A.3.2 *Let $\mathcal{M}$ be a finitely generated $\mathcal{R}$-module and $\mathcal{I}$ an ideal in $\mathcal{R}$. Then*

$$\mathcal{M}_\mathcal{I} = 0 \text{ (or equivalently } \mathcal{M} = \mathcal{IM}), \tag{A.13}$$

$\Longleftrightarrow$

$$\exists r \in \mathcal{R} : r \equiv 1 (\mathrm{mod}\ \mathcal{I}) \text{ and } r\mathcal{M} = 0. \tag{A.14}$$

Proof

"$\Leftarrow$" Let $r \in \mathcal{R}$ be such that $r\mathcal{M} = 0$ and $r \equiv 1 (\mathrm{mod}\ \mathcal{I})$. Then $r - 1 \in \mathcal{I}$. Let $x \in \mathcal{M}$. Then $rx = 0$, so $x = x - rx = (1 - r)x$. But $(1 - r) = -(r - 1) \in \mathcal{I}$ and therefore $x \in \mathcal{IM}$. So $\mathcal{M} \subset \mathcal{IM}$. Since trivially $\mathcal{IM} \subset \mathcal{M}$, we conclude that $\mathcal{M} = \mathcal{IM}$.

"$\Rightarrow$" Let $x_1, \ldots, x_n$ generate the module $\mathcal{M}$, and let $\mathcal{N}$ be a free module of dimension n, i.e. $\mathcal{N} = \mathcal{R}^n$. Take a basis $e_1, \ldots, e_n$ of $\mathcal{N}$ and define the epimorphism $X : \mathcal{N} \to \mathcal{M}$ by

$$Xe_i = x_i \qquad (i = 1, \ldots, n).$$

Since $\mathcal{M} = \mathcal{I}\mathcal{M}$ there exist elements $a_{ji} \in \mathcal{I}$ such that

$$x_i = \sum_j a_{ji} x_j \qquad (i = 1, \ldots, n).$$

Let $A : \mathcal{N} \to \mathcal{N}$ denote the homomorphism with matrix (a_{ij}) with respect to the basis $e_1, \ldots, e_n$. Then for all $i = 1, \ldots, n$ we have $Xe_i = XAe_i$ and hence

$$X = XA.$$

Next, let $p(z)$ denote the characteristic polynomial of A, i.e. $p(z) = \det(zI - A)$. Let $p(z) = \sum_{i=0}^{n} \alpha_i z^i$. Because of the Cayley-Hamilton theorem we know that $p(A) = 0$. So

$$0 = Xp(A) = X \sum_{i=0}^{n} \alpha_i A^i = \sum_{i=0}^{n} \alpha_i X = (\sum_{i=0}^{n} \alpha_i (1)^i) \cdot X = p(1) \cdot X.$$

Therefore $p(1)\mathcal{M} = 0$.

Finally, since all entries of the matrix A are elements of $\mathcal{I}$, and using the definition of the determinant, we conclude that $p(1) = \det(I - A) \equiv 1 (\mathrm{mod}\ \mathcal{I})$. So $p(1)$ satisfies both conditions in (A.14). This completes the proof. ∎

Corollary A.3.3 *Let $\mathcal{M}$ be a finitely generated $\mathcal{R}$-module and assume that for all maximal ideals $\mathcal{A}$ of $\mathcal{R}$ we have $\mathcal{M}_{\mathcal{A}} = 0$. Then $\mathcal{M} = 0$.*

Proof

Let $x \in \mathcal{M}$, and assume that $x \neq 0$. Then the ideal $\mathcal{I}$ defined by $\mathcal{I} := \{r \in \mathcal{R} \mid r \cdot x = 0\}$ is a proper ideal of $\mathcal{R}$, i.e. $\mathcal{I} \neq \mathcal{R}$. According to Proposition A.1.10, $\mathcal{I}$ is contained in a maximal ideal $\mathcal{A}$. Since $\mathcal{M}_{\mathcal{A}} = 0$, and using Lemma A.3.2, we conclude that there exists an element $r \in \mathcal{R}$ such that $r \equiv 1 (\mathrm{mod}\ \mathcal{A})$, and $r\mathcal{M} = 0$. So in particular $r \cdot x = 0$. But then $r \in \mathcal{I} \subseteq \mathcal{A}$, and also $(1 - r) \in \mathcal{A}$. Hence $1 = r + (1 - r) \in \mathcal{A}$. This contradicts the fact that $\mathcal{A}$ is a proper ideal of $\mathcal{R}$. Therefore we conclude that $x = 0$. ∎

Now, consider an $\mathcal{R}$-homomorphism $T : \mathcal{M} \to \mathcal{N}$. Let $\mathcal{I}$ be an ideal in $\mathcal{R}$, and define the map $T_{\mathcal{I}} : \mathcal{M}_{\mathcal{I}} \to \mathcal{N}_{\mathcal{I}}$ by taking quotients

$$T_{\mathcal{I}} : \mathcal{M}_{\mathcal{I}} \to \mathcal{N}_{\mathcal{I}} : \qquad \overline{m} \mapsto \overline{Tm}.$$

Then it is clear that

$$\mathrm{im}(T_{\mathcal{I}}) = (\mathrm{im}(T))_{\mathcal{I}}.$$

Hence, if T is surjective, then for any ideal $\mathcal{I}$ in $\mathcal{R}$, $T_{\mathcal{I}}$ is also surjective. The local-global theorem is a sort of converse of this result.

Theorem A.3.4 (Local-global theorem) *Let $\mathcal{M}$ and $\mathcal{N}$ be $\mathcal{R}$-modules and assume that $\mathcal{N}$ is finitely generated. Let $T : \mathcal{M} \to \mathcal{N}$ be an $\mathcal{R}$-homomorphism. If for all maximal ideals $\mathcal{A}$ in $\mathcal{R}$ the $\mathcal{R}$-homomorphism $T_{\mathcal{A}} : \mathcal{M}_{\mathcal{A}} \to \mathcal{N}_{\mathcal{A}}$ is surjective, then T is surjective.*

Proof

Since $T_{\mathcal{A}}$ is surjective, it follows that $\mathcal{N}_{\mathcal{A}}/(\mathrm{im}(T))_{\mathcal{A}} = 0$ for all maximal ideals $\mathcal{A}$ in $\mathcal{R}$. But $\mathcal{N}_{\mathcal{A}}/(\mathrm{im}(T))_{\mathcal{A}} \cong (\mathcal{N}/\mathrm{im}(T))_{\mathcal{A}}$ via the isomorphism

$$((x)_{\mathrm{im}(T)})_{\mathcal{A}} \mapsto (x_{\mathcal{A}})_{(\mathrm{im}(T))_{\mathcal{A}}}.$$

So

$$(\mathcal{N}/\mathrm{im}(T))_{\mathcal{A}} = 0$$

for all maximal ideals $\mathcal{A}$ in $\mathcal{R}$. Moreover, because $\mathcal{N}$ is finitely generated, $\mathcal{N}/\mathrm{im}(T)$ is finitely generated. After application of Corollary A.3.3 we obtain that $\mathcal{N}/\mathrm{im}(T) = 0$. Hence $\mathrm{im}(T) = \mathcal{N}$. ∎

The local-global theorem makes it possible to investigate the surjectivity of an $\mathcal{R}$-homomorphism $T : \mathcal{M} \to \mathcal{N}$ by studying the homomorphisms $T_{\mathcal{A}}$ defined on the factor modules $T_{\mathcal{A}} : \mathcal{M}_{\mathcal{A}} \to \mathcal{N}_{\mathcal{A}}$, where $\mathcal{A}$ is a maximal ideal. In general these are much easier to investigate because the quotient $\mathcal{R}/\mathcal{A}$ is a field.

The main problem in applying the local-global theorem is that we have to guarantee that for *all* maximal ideals $\mathcal{A}$ in $\mathcal{R}$, $T_{\mathcal{A}}$ is surjective. To do so, we need a complete knowledge of the maximal ideals in $\mathcal{R}$. This is often quite a difficult problem, but for the polynomial rings we are mainly interested in, an answer can be given. In this result, the Hilbert Nullstellensatz (Theorem A.2.9), stated in the previous section, plays an important role.

Proposition A.3.5 *Let $\mathcal{K}$ be a field and $\bar{\mathcal{K}}$ the algebraic closure of $\mathcal{K}$. Consider the polynomial ring $\mathcal{R} := \mathcal{K}[x_1, \ldots, x_n]$ and let $\alpha \in \bar{\mathcal{K}}^n$. Define the ideal $\mathcal{I}_\alpha$ as*

$$\mathcal{I}_\alpha := \{p(x_1, \ldots, x_n) \in \mathcal{R} | p(\alpha) = 0\}.$$

Then

$$\{\mathcal{I}_\alpha | \alpha \in \bar{\mathcal{K}}^n\}$$

is the set of all maximal ideals in $\mathcal{R}$.

Proof

First we show that for any element $\alpha \in \bar{\mathcal{K}}^n$, the ideal $\mathcal{I}_\alpha$ is maximal. Let $\alpha = (\alpha_1, \ldots, \alpha_n) \in \bar{\mathcal{K}}^n$. Since $\bar{\mathcal{K}}$ is an algebraic extension of $\mathcal{K}$, all elements $\alpha_1, \ldots, \alpha_n$ are algebraic over $\mathcal{K}$. Let $\tilde{\mathcal{K}} := \mathcal{K}(\alpha_1, \ldots, \alpha_n)$ denote the finite algebraic extension field of $\mathcal{K}$ obtained by adjunction of the elements $\alpha_1, \ldots, \alpha_n$ to $\mathcal{K}$. We prove that

$$\tilde{\mathcal{K}} \cong \mathcal{R}/\mathcal{I}_\alpha.$$

Let $\beta \in \tilde{\mathcal{K}}$. Since $\tilde{\mathcal{K}}$ is a finite algebraic extension field of $\mathcal{K}$, there exists a polynomial $p \in \mathcal{K}[x_1, \ldots, x_n]$ such that $p(\alpha_1, \ldots, \alpha_n) = \beta$. Now we consider the map $T : \tilde{\mathcal{K}} \to \mathcal{R}/\mathcal{I}_\alpha$ defined by

$$T : \tilde{\mathcal{K}} \to \mathcal{R}/\mathcal{I}_\alpha : \qquad \beta \mapsto \{p \in \mathcal{R} \mid p(\alpha) = \beta\}.$$

Since $p(\alpha) - q(\alpha) = 0$ is equivalent to $p - q \in \mathcal{I}_\alpha$, this map T is well defined, and the correspondence between elements of $\bar{\mathcal{K}}$ and elements of $\mathcal{R}/\mathcal{I}_\alpha$ is one-to-one. Moreover, it is easily verified that T maps sums and products of elements in $\bar{\mathcal{K}}$ to sums and products of the corresponding elements in $\mathcal{R}/\mathcal{I}_\alpha$. Hence T is an isomorphism from $\bar{\mathcal{K}}$ to $\mathcal{R}/\mathcal{I}_\alpha$. We conclude that $\mathcal{R}/\mathcal{I}_\alpha$ is a field, and thus by Definition A.1.9 *(ii)*, $\mathcal{I}_\alpha$ is a maximal ideal of $\mathcal{R}$.

Next, let $\mathcal{A}$ be a maximal ideal in $\mathcal{R}$. We have to prove that there exists an $\alpha \in \bar{\mathcal{K}}^n$ such that $\mathcal{A} = \mathcal{I}_\alpha$. Suppose that the polynomials in $\mathcal{A}$ do not have a common zero. Since $\mathcal{R}$ is a polynomial ring over a field, every ideal in $\mathcal{R}$ is finitely generated, and thus there exist polynomials $p_1, \ldots, p_q$ in $\mathcal{A}$ such that $p_1, \ldots, p_q$ do not have a common zero. So, according to the Hilbert Nullstellensatz there exist polynomials $a_1, \ldots, a_q$ in $\mathcal{R}$ such that

$$a_1 p_1 + a_2 p_2 + \cdots + a_q p_q = 1.$$

We conclude that $\mathcal{A} = \mathcal{R}$, and this contradicts the fact that $\mathcal{A}$ is a maximal ideal and therefore proper.

So all elements of $\mathcal{A}$ have at least one common zero, say $\alpha \in \bar{\mathcal{K}}^n$. Then $\mathcal{A} \subseteq \mathcal{I}_\alpha$. Since both $\mathcal{A}$ and $\mathcal{I}_\alpha$ are maximal ideals, we must have $\mathcal{A} = \mathcal{I}_\alpha$.

This completes the proof. ∎

If $\mathcal{K}$ is a field and $\mathcal{R} = \mathcal{K}[x_1, \ldots, x_n]$, the surjectivity of an $\mathcal{R}$-homomorphism between two $\mathcal{R}$-modules $\mathcal{M}$ and $\mathcal{N}$ is not difficult to test now. We simply have to combine Theorem A.3.4 and Proposition A.3.5. According to Theorem A.3.4 we only have to check the surjectivity modulo each maximal ideal, and Proposition A.3.5 gives a description of all maximal ideals in $\mathcal{R}$. Computation modulo the ideal $\mathcal{I}_\alpha$ boils down to substitution of the point $\alpha \in \bar{\mathcal{K}}^n$ for the indeterminates $x_1, \ldots, x_n$.

Appendix B

A theorem on realization

In this appendix we state and prove a result on the realization of a linear system over an integral domain. It is used in Section 2 for the proof of Theorem 2.8.2. The theorem in this appendix is based on a very similar result for systems over fields described in [47, pp. 403-409]. The generalization to the case of systems over rings is very straightforward.

Theorem B.1 *Let $\mathcal{R}$ be an integral domain and $P(z)$ and $Q(z)$ matrices over $\mathcal{R}[z]$ of size $m \times n$ and $n \times n$ respectively. Assume that $Q(z)$ is monic and $\deg_z(Q(z)) = k > \deg_z(P(z))$. Define the strictly proper matrix $T(z)$ over $\mathcal{R}(z)$ as $T(z) := P(z)Q(z)^{-1}$. Then there exist matrices A, B and C over $\mathcal{R}$, such that the system $\Sigma = (A, B, C, 0)$ is a realization of $T(z)$ (i.e. $T(z) = C(zI - A)^{-1}B$), with the properties:*

(i) $\Sigma = (A, B, C, 0)$ *is reachable,*

(ii) $\det(Q(z)) = \det(zI - A)$.

Proof

Since $Q(z)$ is monic and of degree k, $Q(z)$ can be written as

$$Q(z) = S(z) + Q_{lo}\Psi(z), \tag{B.1}$$

where

$$S(z) = z^k \cdot I,$$

$$\Psi(z) = \text{blockdiag}\left(\begin{pmatrix} z^{k-1} \\ z^{k-2} \\ \vdots \\ z \\ 1 \end{pmatrix} | i = 1, \ldots, n \right),$$

and Q_{lo} is an $n \times (n \cdot k)$ matrix over $\mathcal{R}$. In fact $Q(z)$ is written as $Q(z) = z^k \cdot I +$ "lower order terms". In the same way there exists a matrix $P_{lo} \in \mathcal{R}^{m \times (n \cdot k)}$ such that

$$P(z) = P_{lo} \cdot \Psi(z), \tag{B.2}$$

because $\deg_z(P(z)) < k$.

Next, define the $(n \cdot k) \times (n \cdot k)$ matrix A_c^o as

$$A_c^o := \text{blockdiag}\left(\begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & & & \vdots \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}_{k \times k} | i = 1, \ldots, n \right), \tag{B.3}$$

and the $(n \cdot k) \times n$ matrix B_c^o by

$$(B_c^o)^T := \text{blockdiag}\left(\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}_{1 \times k} | i = 1, \cdots, n \right). \tag{B.4}$$

Then

$$\begin{aligned} B_c^o \cdot S(z) &= \text{blockdiag}\left(\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{k \times 1} | i = 1, \ldots, n \right) \cdot (z^k \cdot I_{n \times n}) = \\ &= \text{blockdiag}\left(\begin{pmatrix} z^k \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{k \times 1} | i = 1, \ldots, n \right), \end{aligned}$$

and

$$\begin{aligned} (zI - A_c^o)\Psi(z) &= \text{blockdiag}\left(\begin{pmatrix} z & & & \\ -1 & z & & \\ & -1 & \ddots & \\ & & \ddots & \ddots \\ & & & -1 \quad z \end{pmatrix} \begin{pmatrix} z^{k-1} \\ z^{k-2} \\ \vdots \\ z \\ 1 \end{pmatrix} | i = 1, \ldots, n \right) = \\ &= \text{blockdiag}\left(\begin{pmatrix} z^k \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{k \times 1} | i = 1, \ldots, n \right). \end{aligned}$$

So

$$B_c^o S(z) = (zI - A_c^o)\Psi(z). \tag{B.5}$$

Define

$$A := A_c^o - B_c^o Q_{lo}, \tag{B.6}$$

$$B := B_c^o, \tag{B.7}$$

$$C := P_{lo}. \tag{B.8}$$

Clearly A, B and C are matrices over $\mathcal{R}$, and combining (B.5) and (B.1) we obtain:

$$\begin{aligned}(zI - A)\Psi(z) &= (zI - A_c^o + B_c^o Q_{lo})\Psi(z) = (zI - A_c^o)\Psi(z) + B_c^o Q_{lo}\Psi(z) = \\ &= B_c^o S(z) + B_c^o Q_{lo}\Psi(z) = B_c^o \cdot (S(z) + Q_{lo}\Psi(z)) = B_c^o Q(z).\end{aligned}$$

So, by the definition of B, we have $(zI - A)\Psi(z) = BQ(z)$, and thus

$$\Psi(z)Q(z)^{-1} = (zI - A)^{-1}B. \tag{B.9}$$

Pre-multiplying by $P_{lo} = C$ and using (B.2) gives

$$C(zI - A)^{-1}B = P_{lo}\Psi(z)Q(z)^{-1} = P(z)Q(z)^{-1} = T(z).$$

So $\Sigma = (A, B, C, 0)$ is a realization of $T(z)$.

Next we prove that $\Sigma = (A, B, C, 0)$ is reachable. Since $AB = (A_c^o - B_c^o Q_{lo})B_c^o = A_c^o B_c^o - B_c^o Q_{lo} B_c^o$, it is obvious that for all $h \in \mathbb{N}$, the columns of $(B|AB|\cdots|A^{h-1}B)$ generate the same $\mathcal{R}$-module as the columns of the matrix $(B_c^o|A_c^o B_c^o|\cdots|(A_c^o)^{h-1}B_c^o)$. So Σ is reachable if and only if the pair (A_c^o, B_c^o) is reachable. Now

$$B_c^o = \text{blockdiag}\left(\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}_{k\times 1} \Big| i = 1, \ldots, n \right),$$

$$\begin{aligned}A_c^o B_c^o &= \text{blockdiag}\left(\begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{pmatrix} \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \Big| i = 1, \ldots, n \right) = \\ &= \text{blockdiag}\left(\begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}_{k\times 1} \Big| i = 1, \ldots, n \right),\end{aligned}$$

and in the same way for $j \leq n - 1$:

$$(A_c^o)^j B_c^o = \text{blockdiag}(e_{j+1} | i = 1, \ldots, n),$$

where e_{j+1} denotes the $(j+1)^{\text{th}}$ unit vector in $\mathcal{R}^k$. From this observation it is immediately clear that the columns of $(B_c^o|A_c^o B_c^o|\cdots|(A_c^o)^{n-1}B_c^o)$ generate $\mathcal{R}^{nk}$. But then $\mathcal{R}^{nk}$ is certainly generated by the columns of the matrix $(B_c^o|A_c^o B_c^o|\cdots|(A_c^o)^{nk-1}B_c^o)$. Hence (A_c^o, B_c^o) is reachable, and we conclude that the pair (A, B) is reachable.

Finally, to prove that $\det(zI - A) = \det(Q(z))$, note first that A is an $(n\cdot k)\times(n\cdot k)$ matrix, so $\det(zI - A)$ is a monic polynomial of degree $n \cdot k$. $Q(z)$ is an $n \times n$ matrix over $\mathcal{R}[z]$ and is monic and of degree k, so $\deg_z(\det(Q(z))) = n \cdot k$.

Since the pair (A, B) is reachable, the matrix $(zI - A|B)$ is right-invertible over $\mathcal{R}[z]$ and there exist polynomial matrices $N(z)$ and $M(z)$ such that

$$(zI - A)N(z) + BM(z) = I.$$

From (B.9) we have

$$-(zI - A)\Psi(z) + BQ(z) = 0.$$

Combining both equations above, we obtain

$$\begin{pmatrix} zI - A & B \\ 0 & I \end{pmatrix} \cdot \begin{pmatrix} N(z) & -\Psi(z) \\ M(z) & Q(z) \end{pmatrix} = \begin{pmatrix} I & 0 \\ M(z) & Q(z) \end{pmatrix}. \tag{B.10}$$

All matrices on the diagonals in (B.10) are square, so taking the determinant on the right- and left hand side we get

$$\det(zI - A) \cdot \det \begin{pmatrix} N(z) & -\Psi(z) \\ M(z) & Q(z) \end{pmatrix} = \det(Q(z)).$$

Define $p(z) := \det \begin{pmatrix} N(z) & -\Psi(z) \\ M(z) & Q(z) \end{pmatrix}$. Since $N(z)$, $M(z)$, $\Psi(z)$ and $Q(z)$ are all polynomial matrices, $p(z)$ is also an element of $\mathcal{R}[z]$. Recall that both $\det(zI - A)$ and $\det(Q(z))$ are monic polynomials of degree $n \cdot k$. Then it follows immediately from the last formula that we must have that $p(z) = 1$. Hence $\det(zI - A) = \det(Q(z))$.

This completes the proof. ∎

Appendix C

Proofs of Subsection 4.2.4

This appendix is devoted to the proofs of three of the main results on irreducible ascending chains mentioned in Subsection 4.2.4.

Proposition 4.2.30 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in $K[x_1, \ldots, x_n]$, and rename the indeterminates in the same way as in Definition 4.2.29. Then we have:*

$$\mathcal{A} = (f_1, \ldots, f_r) \text{ is irreducible}$$

$$\Longleftrightarrow$$

$$\forall j = 1, \ldots, r : \langle f_1, \ldots, f_j \rangle \text{ is a prime ideal in } K_0[y_1, \ldots, y_j].$$

(This means that $\langle f_1 \rangle$ is a prime ideal in $K_0[y_1]$, $\langle f_1, f_2 \rangle$ is a prime ideal in $K_0[y_1, y_2]$ and so on, until the final condition: $\langle f_1, \ldots, f_r \rangle$ is a prime ideal in $K_0[y_1, \ldots, y_r]$).

Proof (by induction)

i = 1: (f_1) is irreducible $\Longleftrightarrow$ $\langle f_1 \rangle$ is a prime ideal in $K_0[y_1]$.

"$\Leftarrow$" Assume that $\langle f_1 \rangle$ is a prime ideal, and let p and q be polynomials in $K_0[y_1]$ such that $f_1 = p \cdot q$. Then p or q belongs to $\langle f_1 \rangle$. Without loss of generality we assume that $p \in \langle f_1 \rangle$. Then there exists an $\alpha \in K_0[y_1]$ such that $p = \alpha \cdot f_1$. Hence $f_1 = \alpha \cdot q \cdot f_1$ and thus $\alpha \cdot q = 1$. This implies that $\deg_{y_1}(q) = \deg_{y_1}(\alpha) = 0$, and we conclude that f_1 is irreducible.

"$\Rightarrow$" Assume that (f_1) is irreducible. Let $p \in \langle f_1 \rangle$. Then there exists an $r \in K_0[y_1]$ such that $p = r f_1$. Suppose there are polynomials $g, h \in K_0[y_1]$ such that $p = gh$. Then $gh = r f_1$. Since f_1 is irreducible, we must have $g \in \langle f_1 \rangle$ or $h \in \langle f_1 \rangle$. So $\langle f_1 \rangle$ is a prime ideal.

Induction step: Assume that $\mathcal{A}_j = (f_1, \ldots, f_j)$ is an irreducible ascending chain generating a prime polynomial ideal in $K_0[y_1, \ldots, y_j]$. Let $K_i = K_0(\eta_1, \ldots, \eta_i)$ $(i = 1, \ldots, j)$ denote the field extension obtained by adjoining $\eta_1, \ldots, \eta_i$ to K_0, where η_k $(k = 1, \ldots, i)$ is an extended zero of $f_k(\eta_1, \ldots, \eta_{k-1}, y_k)$ in $K_{k-1}[y_k]$. Then we have to prove the following claim:

$$\langle \mathcal{A}_{j+1} \rangle = \langle f_1, \ldots, f_{j+1} \rangle \text{ is a prime ideal in } K_0[y_1, \ldots, y_{j+1}]$$

$$\Longleftrightarrow$$

$$f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}) \text{ is irreducible in } K_j[y_{j+1}].$$

"$\Rightarrow$" Assume that $\langle \mathcal{A}_{j+1} \rangle$ is a prime ideal. Let p and q be polynomials in $\mathcal{K}_j[y_{j+1}]$ such that

$$f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}) = p(\eta_1, \ldots, \eta_j, y_{j+1}) q(\eta_1, \ldots, \eta_j, y_{j+1}). \tag{C.1}$$

To prove that $f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1})$ is irreducible in $\mathcal{K}_j[y_{j+1}]$, we show that p or q is an element of $\mathcal{K}_j$.

First substitute $(y_1, \ldots, y_j)$ for $(\eta_1, \ldots, \eta_j)$. Then we have:

$$p(y_1, \ldots, y_{j+1}) q(y_1, \ldots, y_{j+1}) = f_{j+1}(y_1, \ldots, y_{j+1}) + r(y_1, \ldots, y_{j+1}), \tag{C.2}$$

and because of (C.1), it is clear that

$$r(\eta_1, \ldots, \eta_j, y_{j+1}) = 0. \tag{C.3}$$

Note that (C.3) holds for every common zero of $f_1, \ldots, f_j$. This can be seen as follows. Let $(\xi_1, \ldots, \xi_j)$ be another common zero of $f_1, \ldots, f_j$. This is only possible if for all $i = 1, \ldots, j$, ξ_i is a conjugate of η_i. So for all $i = 1, \ldots, j$, the field extensions $\mathcal{K}_{i-1}(\eta_i)$ and $\mathcal{K}_{i-1}(\xi_i)$ are isomorphic, say through an isomorphism T_i. Thus we have

$$T_1 \cdots T_j(p(\eta_1, \ldots, \eta_j, y_{j+1}) q(\eta_1, \ldots, \eta_j, y_{j+1})) = T_1 \cdots T_j(f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1})).$$

Therefore for all $i = 1, \ldots, j$:

$$T_1 \cdots T_i(p(\eta_1, \ldots, \eta_i, \xi_{i+1}, \ldots, \xi_j, y_{j+1}) q(\eta_1, \ldots, \eta_i, \xi_{i+1}, \ldots, \xi_j, y_{j+1})) =$$

$$= T_1 \cdots T_i(f_{j+1}(\eta_1, \ldots, \eta_i, \xi_{i+1}, \ldots, \xi_j, y_{j+1})).$$

So finally:

$$p(\xi_1, \ldots, \xi_j, y_{j+1}) q(\xi_1, \ldots, \xi_j, y_{j+1}) = f_{j+1}(\xi_1, \ldots, \xi_j, y_{j+1}).$$

In combination with (C.2) this yields

$$r(\xi_1, \ldots, \xi_j, y_{j+1}) = 0.$$

Let $\bar{\mathcal{K}}_0$ denote the algebraic closure of $\mathcal{K}_0$. We have proven that

$$\forall (\xi_1, \ldots, \xi_j) \in \bar{\mathcal{K}}_0^j : \left.\begin{array}{c} f_1(\xi_1) = 0 \\ \vdots \\ f_j(\xi_1, \ldots, \xi_j) = 0 \end{array}\right\} \Longrightarrow r(\xi_1, \ldots, \xi_j, y_{j+1}) = 0. \tag{C.4}$$

Next regard $f_1, \ldots, f_j$ as polynomials in the ring $\mathcal{K}_0[y_1, \ldots, y_{j+1}]$. From (C.4) and the Hilbert Nullstellensatz it follows that there exist a $\rho \in \mathbb{N}$ and polynomials $\alpha_i(y_1, \ldots, y_{j+1}) \in \mathcal{K}_0[y_1, \ldots, y_{j+1}]$ such that

$$r^\rho(y_1, \ldots, y_{j+1}) = \sum_{i=1}^{j} \alpha_i(y_1, \ldots, y_{j+1}) f_i(y_1, \ldots, y_i).$$

So, considering $\langle f_1, \ldots, f_j \rangle$ as an ideal in $\mathcal{K}_0[y_1, \ldots, y_{j+1}]$, we conclude that r^ρ is an element of this ideal. Recall that $\langle f_1, \ldots, f_j \rangle$, considered as an ideal in $\mathcal{K}_0[y_1, \ldots, y_j]$, is a prime ideal. Then the ideal $\langle f_1, \ldots, f_j \rangle$, considered as an ideal in $\mathcal{K}_0[y_1, \ldots, y_{j+1}]$,

is also prime, and we conclude that r itself is an element of this ideal and may be written as

$$r(y_1,\ldots,y_{j+1}) = \sum_{i=1}^{j} \beta_i(y_1,\ldots,y_{j+1}) f_i(y_1,\ldots,y_i),$$

where $\beta_i(y_1,\ldots,y_{j+1})$ are polynomials in $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$. With (C.2) we conclude that

$$p(y_1,\ldots,y_{j+1})q(y_1,\ldots,y_{j+1}) \in \langle f_1,\ldots,f_{j+1}\rangle,$$

where $\langle f_1,\ldots,f_{j+1}\rangle$ is considered as an ideal in $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$. By assumption, this ideal is a prime ideal. So either p or q (or both) is an element of this ideal.

Assume without loss of generality that

$$p(y_1,\ldots,y_{j+1}) \in \langle f_1,\ldots,f_{j+1}\rangle.$$

Then there exist polynomials $\gamma_i(y_1,\ldots,y_{j+1}) \in \mathcal{K}_0[y_1,\ldots,y_{j+1}]$ such that

$$p(y_1,\ldots,y_{j+1}) = \sum_{i=1}^{j+1} \gamma_i(y_1,\ldots,y_{j+1}) f_i(y_1,\ldots,y_i).$$

Substitution of $(\eta_1,\ldots,\eta_j)$ for $(y_1,\ldots,y_j)$ yields:

$$p(\eta_1,\ldots,\eta_j,y_{j+1}) = \gamma_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}) f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}).$$

So

$$\deg_{y_{j+1}}(p(\eta_1,\ldots,\eta_j,y_{j+1})) \geq \deg_{y_{j+1}}(f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1})),$$

and this implies that $\deg_{y_{j+1}}(q(\eta_1,\ldots,\eta_j,y_{j+1})) = 0$. Hence $f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1})$ is irreducible.

"$\Leftarrow$" Assume that $f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1})$ is irreducible over $K_j[y_{j+1}]$. We have to prove that $\langle f_1,\ldots,f_{j+1}\rangle$ is a prime ideal.

Let $g \in \langle f_1,\ldots,f_{j+1}\rangle$. Then there exist polynomials $\alpha_i \in \mathcal{K}_0[y_1,\ldots,y_{j+1}]$ such that

$$g(y_1,\ldots,y_{j+1}) = \sum_{i=1}^{j+1} \alpha_i(y_1,\ldots,y_{j+1}) f_i(y_1,\ldots,y_i).$$

Assume that there exist polynomials p and q in $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$ such that

$$g(y_1,\ldots,y_{j+1}) = p(y_1,\ldots,y_{j+1})q(y_1,\ldots,y_{j+1}). \tag{C.5}$$

Substitution of $(\eta_1,\ldots,\eta_j)$ for $(y_1,\ldots,y_j)$ in (C.5) yields:

$$p(\eta_1,\ldots,\eta_j,y_{j+1})q(\eta_1,\ldots,\eta_j,y_{j+1}) = \alpha_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}) f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}).$$

Since $f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1})$ is irreducible in $\mathcal{K}_j[y_{j+1}]$, it is clear that there either exists a polynomial $\gamma(\eta_1,\ldots,\eta_j,y_{j+1}) \in \mathcal{K}_j[y_{j+1}]$ such that

$$p(\eta_1,\ldots,\eta_j,y_{j+1}) = \gamma(\eta_1,\ldots,\eta_j,y_{j+1}) f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}),$$

or a polynomial $\delta(\eta_1,\ldots,\eta_j,y_{j+1}) \in \mathcal{K}_j[y_{j+1}]$ such that

$$q(\eta_1,\ldots,\eta_j,y_{j+1}) = \delta(\eta_1,\ldots,\eta_j,y_{j+1})f_{j+1}(\eta_1,\ldots,\eta_j,y_{j+1}),$$

or both. Without loss of generality we assume the first.

Substitute $(y_1,\ldots,y_j)$ back for $(\eta_1,\ldots,\eta_j)$:

$$p(y_1,\ldots,y_{j+1}) = \gamma(y_1,\ldots,y_{j+1})f_{j+1}(y_1,\ldots,y_{j+1}) + r(y_1,\ldots,y_{j+1}).$$

Then again $r(\eta_1,\ldots,\eta_j,y_{j+1}) = 0$, and with the same arguments as in the necessity part we can prove that for every common zero $(\xi_1,\ldots,\xi_j)$ of $f_1,\ldots,f_j$ we have $r(\xi_1,\ldots,\xi_j,y_{j+1}) = 0$. Let $\bar{\mathcal{K}}_0$ denote the algebraic closure of $\mathcal{K}_0$. Then:

$$\left.\begin{array}{rc} & f_1(\xi_1) = 0 \\ \forall(\xi_1,\ldots,\xi_j) \in \bar{\mathcal{K}}_0^j: & \vdots \\ & f_j(\xi_1,\ldots,\xi_j) = 0 \end{array}\right\} \Longrightarrow r(\xi_1,\ldots,\xi_j,y_{j+1}) = 0. \qquad \text{(C.6)}$$

Regard $f_1,\ldots,f_j$ as polynomials in the ring $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$. From (C.6) and the Hilbert Nullstellensatz it follows that there exists an integer $\rho \in \mathbb{N}$ and polynomials $\beta_i(y_1,\ldots,y_{j+1}) \in \mathcal{K}_0[y_1,\ldots,y_{j+1}]$ such that

$$r^\rho(y_1,\ldots,y_{j+1}) = \sum_{i=1}^{j} \beta_i(y_1,\ldots,y_{j+1})f_i(y_1,\ldots,y_i).$$

Since $\langle f_1,\ldots,f_j\rangle$, considered as an ideal in $\mathcal{K}_0[y_1,\ldots,y_j]$, is a prime ideal, the polynomials $f_1,\ldots,f_j$ generate also a prime ideal in the ring $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$. Now r^ρ is an element of this ideal, so r itself is an element of this ideal too, and thus there exist polynomials $\delta_i(y_1,\ldots,y_{j+1}) \in \mathcal{K}_0[y_1,\ldots,y_{j+1}]$ such that r can be written as

$$r(y_1,\ldots,y_{j+1}) = \sum_{i=1}^{j} \delta_i(y_1,\ldots,y_{j+1})f_i(y_1,\ldots,y_i).$$

Hence

$$p(y_1,\ldots,y_{j+1}) = \gamma(y_1,\ldots,y_{j+1})f_{j+1}(y_1,\ldots,y_{j+1}) + r(y_1,\ldots,y_{j+1})$$

is an element of the ideal $\langle f_1,\ldots,f_{j+1}\rangle$, considered as an ideal in $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$, and we conclude that $\langle f_1,\ldots,f_{j+1}\rangle$ is a prime ideal in $\mathcal{K}_0[y_1,\ldots,y_{j+1}]$.

This completes the proof. ∎

Lemma 4.2.33 *Let $\mathcal{A} = (f_1,\ldots,f_r)$ be an ascending chain in $\mathcal{K}[u_1,\ldots,u_d,y_1,\ldots,y_r]$ (in the notation of Definition 4.2.29). Let $\mathcal{F}$ and $\tilde{\mathcal{F}}$ be defined as in (4.32) and (4.31). Then*

$$\mathcal{F} \text{ is a prime ideal in } \mathcal{K}[u_1,\ldots,u_d,y_1,\ldots,y_r],$$

$$\Longleftrightarrow$$

$$\tilde{\mathcal{F}} \text{ is a prime ideal in } \mathcal{K}_0[y_1,\ldots,y_r].$$

Proof

"$\Leftarrow$" Suppose $\tilde{\mathcal{F}}$ is a prime ideal. Let $g \in \mathcal{F}$, and assume that there exist polynomials p and q in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ such that $g = p \cdot q$. Since $\mathcal{F}$ is contained in the prime ideal $\tilde{\mathcal{F}}$, we know that $g \in \tilde{\mathcal{F}}$, and thus $p \in \tilde{\mathcal{F}}$ or $q \in \tilde{\mathcal{F}}$. Without loss of generality we assume that $p \in \tilde{\mathcal{F}}$. Then p can be written as

$$p = \sum_{i=1}^{r} \alpha_i f_i,$$

with $\alpha_i \in \mathcal{K}_0[y_1, \ldots, y_r]$. However, p itself is an element of $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$, so $p \in \mathcal{F}$ and thus $\mathcal{F}$ is a prime ideal.

"$\Rightarrow$" Next, suppose that $\mathcal{F}$ is prime. Let $g \in \tilde{\mathcal{F}}$, and assume that $g = pq$ with $p, q \in \mathcal{K}_0[y_1, \ldots, y_r]$.

Since $g \in \tilde{\mathcal{F}}$, g can be written as

$$g = \sum_{i=1}^{r} \frac{\beta_i}{\gamma_i} f_i,$$

with $\beta_i \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ and $\gamma_i \in \mathcal{K}[u_1, \ldots, u_d]$. Define $\gamma := \prod_{i=1}^{r} \gamma_i$. Then $\gamma \in \mathcal{K}[u_1, \ldots, u_d]$ and

$$\gamma g = \sum_{i=1}^{r} (\beta_i \frac{\bar{\gamma}}{\gamma_i}) f_i \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r].$$

Thus $\gamma g \in \mathcal{F}$.

In almost the same way we can prove that there exist polynomials ψ and μ in $\mathcal{K}[u_1, \ldots, u_d]$ such that ψp and μq are elements of $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$.

Multiplying the equation $g = p \cdot q$ on both left- and right-hand side by $\gamma \cdot \psi \cdot \mu$, we obtain

$$\psi\mu(\gamma g) = (\gamma\psi p)(\mu q). \tag{C.7}$$

Since $\gamma g \in \mathcal{F}$, the left-hand side of (C.7) is an element of $\mathcal{F}$. Because $\mathcal{F}$ is a prime ideal in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ and both $(\gamma\psi p)$ and (μq) are polynomials in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$, we conclude that $(\gamma\psi p) \in \mathcal{F}$ or $(\mu q) \in \mathcal{F}$ (or both).

Now $\gamma\psi$ and μ are polynomials in $\mathcal{K}[u_1, \ldots, u_d]$. If $\mu q \in \mathcal{F}$, then clearly $q \in \tilde{\mathcal{F}}$, and if $\gamma\psi p \in \mathcal{F}$, then $p \in \tilde{\mathcal{F}}$. So either p or q (or both) is an element of $\tilde{\mathcal{F}}$, and we conclude that $\tilde{\mathcal{F}}$ is a prime ideal. ∎

Theorem 4.2.34 *Let $\mathcal{A} = (f_1, \ldots, f_r)$ be an ascending chain in the polynomial ring $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$ (in the notation of Definition 4.2.29). Define the ideals $\mathcal{F}$ and $\tilde{\mathcal{F}}$ as in (4.32) and (4.31). Then the following three statements are equivalent:*

(i) The ascending chain $\mathcal{A}$ is irreducible,

(ii) $\tilde{\mathcal{F}}$ is a prime ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$, and $\mathcal{A}$ is a (Ritt-) characteristic set of $\tilde{\mathcal{F}}$,

(iii) $\mathcal{F}$ is a prime ideal in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$, and $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$.

Proof

We prove the following implication scheme: *(i)* $\Rightarrow$ *(iii)* $\Rightarrow$ *(ii)* $\Rightarrow$ *(i)*.

(i) $\Rightarrow$ (iii) Suppose that $\mathcal{A} = (f_1, \ldots, f_r)$ is an irreducible ascending chain. According to Proposition 4.2.30, this implies that the ideal $\tilde{\mathcal{F}} = \langle f_1, \ldots, f_r \rangle$ in the ring $\mathcal{K}_0[y_1, \ldots, y_r]$, generated by $f_1, \ldots, f_r$, is a prime ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$. Application of Lemma 4.2.33 yields that $\mathcal{F}$ is a prime ideal in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$.

To show that $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$, it suffices (according to Theorem 4.2.16) to prove that

$$\forall p \in \mathcal{F} : \mathrm{prem}(p, \mathcal{A}) = 0.$$

By assumption, $\mathcal{A}$ is an irreducible ascending chain. So $\mathcal{A}$ has a generic point

$$\tilde{\eta} = (u_1, \ldots, u_d, \eta_1, \ldots, \eta_r).$$

Let $p \in \mathcal{F}$. Then p can be written as

$$p = \sum_{i=1}^{r} \alpha_i(y_1, \ldots, y_r) f_i(u_1, \ldots, u_d, y_1, \ldots, y_i),$$

where $\alpha_i(y_1, \ldots, y_r) \in \mathcal{K}_0[y_1, \ldots, y_r]$ are such that $p \in \mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$. Since $\tilde{\eta}$ is an (extended) zero of all f_i, and $u_1, \ldots, u_d$ are all transcendental over $\mathcal{K}$, we must have

$$p(u_1, \ldots, u_d, \eta_1, \ldots, \eta_r) = 0.$$

So $\tilde{\eta}$ is an extended zero of p, and application of Proposition 4.2.31 yields

$$\mathrm{prem}(p, \mathcal{A}) = 0.$$

Since $p \in \mathcal{F}$ was arbitrary, this proves that $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$.

(iii) $\Rightarrow$ (ii) Assume that $\mathcal{F}$ is a prime ideal in $\mathcal{K}[u_1, \ldots, u_d, y_1, \ldots, y_r]$, and that $\mathcal{A}$ is a (Ritt-) characteristic set of $\mathcal{F}$. Since $\mathcal{F}$ is prime, it follows immediately from Lemma 4.2.33 that $\tilde{\mathcal{F}}$ is a prime ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$. So we only have to prove that $\mathcal{A}$ is a (Ritt-) characteristic set of $\tilde{\mathcal{F}}$.

Let $p \in \tilde{\mathcal{F}}$ and assume that p is reduced w.r.t. $\mathcal{A}$. Since $p \in \tilde{\mathcal{F}}$ there exists a nonzero polynomial $\alpha \in \mathcal{K}[u_1, \ldots, u_d]$ such that $\alpha \cdot p \in \mathcal{F}$. However, the polynomial $\alpha \cdot p$ remains reduced w.r.t. $\mathcal{A}$. Application of Lemma 4.2.14, and using the assumption that $\mathcal{A}$ is a Ritt-characteristic set of $\mathcal{F}$, we conclude that $\alpha \cdot p = 0$. Since α is nonzero, we must have $p = 0$. According to Lemma 4.2.14, this implies that $\mathcal{A}$ is a Ritt-characteristic set of $\tilde{\mathcal{F}}$.

(ii) $\Rightarrow$ (i) Let $\mathcal{A} = (f_1, \ldots, f_r)$ be a (Ritt-) characteristic set of the prime polynomial ideal $\tilde{\mathcal{F}} = \langle f_1, \ldots, f_r \rangle$ in the ring $\mathcal{K}_0[y_1, \ldots, y_r]$. Suppose that $\mathcal{A}$ is reducible. We have to derive a contradiction.

Since $\mathcal{A}$ is reducible, there exists an integer $j < r$ such that

$$\mathcal{A}_j = (f_1, \ldots, f_j)$$

is an irreducible ascending chain in $\mathcal{K}_0[y_1, \ldots, y_j]$, with generic point $\tilde{\eta} = (\eta_1, \ldots, \eta_j)$, and (using the notation of Definition 4.2.29)

$$f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}) \text{ is reducible in } \mathcal{K}_j[y_{j+1}].$$

So there exist polynomials p and q in$\mathcal{K}_j[y_{j+1}]$ such that

$$f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}) = p(\eta_1, \ldots, \eta_j, y_{j+1}) q(\eta_1, \ldots, \eta_j, y_{j+1}), \tag{C.8}$$

with

$$0 < \deg_{y_{j+1}}(p(\eta_1, \ldots, \eta_j, y_{j+1})) < \deg_{y_{j+1}}(f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1})), \tag{C.9}$$

$$0 < \deg_{y_{j+1}}(q(\eta_1, \ldots, \eta_j, y_{j+1})) < \deg_{y_{j+1}}(f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1})). \tag{C.10}$$

Substitution of $(y_1, \ldots, y_j)$ for $(\eta_1, \ldots, \eta_j)$ yields:

$$f_{j+1}(y_1, \ldots, y_{j+1}) = p(y_1, \ldots, y_{j+1}) q(y_1, \ldots, y_{j+1}) + r(y_1, \ldots, y_{j+1}). \tag{C.11}$$

Because of (C.8) we clearly have $r(\eta_1, \ldots, \eta_j, y_{j+1}) = 0$.

Let $\bar{\mathcal{K}}_0$ denote the algebraic closure of $\mathcal{K}_0$. In completely the same way as in the proof of Proposition 4.2.30 we can show that:

$$\forall (\xi_1, \ldots, \xi_j) \in \bar{\mathcal{K}}_0^j : \left.\begin{array}{c} f_1(\xi_1) = 0 \\ \vdots \\ f_j(\xi_1, \ldots, \xi_j) = 0 \end{array}\right\} \Longrightarrow r(\xi_1, \ldots, \xi_j, y_{j+1}) = 0. \tag{C.12}$$

We repeat the same argument as in the proof of Proposition 4.2.30. Regard $f_1, \ldots, f_j$ as polynomials in $\mathcal{K}_0[y_1, \ldots, y_{j+1}]$ and apply the Hilbert Nullstellensatz on (C.12). Then we find an integer $\rho \in \mathbf{N}$ and polynomials $\beta_i \in \mathcal{K}_0[y_1, \ldots, y_{j+1}]$ such that

$$r^\rho(y_1, \ldots, y_{j+1}) = \sum_{i=1}^{j} \beta_i(y_1, \ldots, y_{j+1}) f_i(y_1, \ldots, y_i).$$

So $r^\rho \in \tilde{\mathcal{F}}$. Now $\tilde{\mathcal{F}}$ is a prime ideal, and thus it follows that $r \in \tilde{\mathcal{F}}$. Recalling (C.11) we conclude that

$$p(y_1, \ldots, y_{j+1}) q(y_1, \ldots, y_{j+1}) \in \tilde{\mathcal{F}}. \tag{C.13}$$

Since $\tilde{\mathcal{F}}$ is a prime ideal in $\mathcal{K}_0[y_1, \ldots, y_r]$, either p or q (or both) has to be an element of $\tilde{\mathcal{F}}$. Without loss of generality we assume that $p \in \tilde{\mathcal{F}}$.

Next, recall that at the moment $\mathcal{A}$ is considered as an ascending chain in $\mathcal{K}_0[y_1, \ldots, y_r]$ and $\mathcal{A}_j$ as an ascending chain in $\mathcal{K}_0[y_1, \ldots, y_j]$. So all polynomials we are considering, are polynomials in the indeterminates $y_1, \ldots, y_r$ with coefficients in $\mathcal{K}_0$.

Since p is a polynomial in the indeterminates $y_1, \ldots, y_{j+1}$, p is reduced with respect to $f_{j+2}, \ldots, f_r$. Moreover, $p \in \tilde{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ is a prime ideal with Ritt-characteristic set $\mathcal{A}$, so $\mathrm{prem}(p, \mathcal{A}) = 0$. Hence there exist integers $\nu_1, \ldots, \nu_{j+1} \in \mathbf{N}$ and polynomials $\beta_1, \ldots, \beta_{j+1} \in \mathcal{K}_0[y_1, \ldots, y_{j+1}]$ such that

$$I_1^{\nu_1} \cdots I_{j+1}^{\nu_{j+1}} p = \sum_{i=1}^{j+1} \beta_i(y_1, \ldots, y_{j+1}) f_i(y_1, \ldots, y_i). \tag{C.14}$$

By assumption, $\mathcal{A}_j$ is an irreducible ascending chain, with generic point $\tilde{\eta}_j = (\eta_1, \ldots, \eta_j)$. The initials $I_1, \ldots, I_{j+1}$ are polynomials in $K_0[y_1, \ldots, y_j]$ that are reduced with respect to $\mathcal{A}$, so in particular they are reduced with respect to $\mathcal{A}_j$. According to Proposition 4.2.31, $\tilde{\eta}_j$ is not an extended zero of $I_1, \ldots, I_{j+1}$:

$$\forall i \in \{1, \ldots, j+1\} : C_i := I_i(\eta_1, \ldots, \eta_{i-1}) \neq 0.$$

Next substitute $(\eta_1, \ldots, \eta_j)$ for $(y_1, \ldots, y_j)$ in formula (C.14). Since $(\eta_1, \ldots, \eta_j)$ is a common zero of $f_1, \ldots, f_j$, we obtain

$$C_1^{\nu_1} \cdots C_{j+1}^{\nu_{j+1}} \cdot p(\eta_1, \ldots, \eta_j, y_{j+1}) = \beta_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}) f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1}).$$

Therefore

$$\deg_{y_{j+1}}(p(\eta_1, \ldots, \eta_j, y_{j+1})) \geq \deg_{y_{j+1}}(f_{j+1}(\eta_1, \ldots, \eta_j, y_{j+1})),$$

and this contradicts (C.9). We conclude that $\mathcal{A}$ must be irreducible.

This completes the proof. ∎

Bibliography

[1] M.F. Atiyah and I.G. MacDonald, *Introduction to Commutative Algebra.* Addison-Wesley, Reading, Massachusetts, 1969.

[2] T. Becker and V. Weispfenning, *Gröbner Bases: a computational approach to commutative algebra*, volume 141 of *Graduate texts in mathematics.* Springer-Verlag, Berlin, 1993. Written in cooperation with H. Kredel.

[3] R. Bellman and K.L. Cooke, *Differential-Difference Equations*, volume 6 of *Mathematics in science and engineering.* Academic Press, New York, 1963.

[4] W. Böge, R. Gebauer, and H. Kredel, Some examples for solving systems of algebraic equations by calculating Gröbner bases. *J. Symbolic Computation*, 1:83–98, 1986.

[5] J.W. Brewer, J.W. Bunce, and F.S. Van Vleck, *Linear systems over commutative rings*, volume 104 of *Lecture notes in pure and applied mathematics.* Marcel Dekker, New York, 1986.

[6] I.N. Bronstein and K.A. Semendjajew, *Taschenbuch der Mathematik.* B.G. Teubner, Leipzig, 1989. 24. Auflage.

[7] B. Buchberger, Gröbner bases: an algorithmic method in polynomial ideal theory. In N.K. Bose, editor, *Multidimensional Systems Theory*, Mathematics and Its Applications, chapter 6, pages 184–232. D. Reidel, Dordrecht, 1985.

[8] C. Byrnes, M. Hazewinkel, C. Martin, and Y. Rouchaleau, Introduction to geometrical methods for the theory of linear systems. In C.I. Byrnes and C.F. Martin, editors, *Geometrical Methods for the Theory of Linear Systems*, pages 1–84, Dordrecht, 1980. D. Reidel.

[9] B.W. Char, K.O. Geddes, G.H. Gonnet, B.L. Leong, M.B. Monagan, and S.M. Watt, *Maple V Library Reference Manual.* Springer-Verlag, New York, 1991.

[10] S.C. Chou, *Mechanical Geometry Theorem Proving.* D. Reidel, Dordrecht, 1988.

[11] R.V. Churchill and J.W. Brown, *Complex Variables and Applications.* McGraw-Hill, London, fourth edition, 1984.

[12] G. Conte and A.M. Perdon, Systems over a principal ideal domain. A polynomial model approach. *SIAM J. Contr. & Opt.*, 20:112–124, 1982.

[13] J.B. Conway, *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1985.

[14] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1992.

[15] M.A. Cruz and J.K. Hale, Stability of functional differential equations of neutral type. *J. Diff. Eqns.*, 7:334–355, 1970.

[16] R.F. Curtain and K. Glover, Robust stabilization of infinite dimensional systems by finite dimensional controllers. *Syst. & Contr. Letters*, 7:41–47, 1986.

[17] R.F. Curtain and A.J. Pritchard, *Infinite-Dimensional Linear Systems Theory*, volume 8 of *Lecture Notes in Control and Information Sciences*. Springer Verlag, Berlin, 1978.

[18] K.B. Datta and M.L.J. Hautus, Decoupling of multivariable control systems over unique factorization domains. *SIAM J. Contr. & Opt.*, 22:28–39, 1984.

[19] S. Eilenberg, *Automata, languages, and machines, Volume A*, volume 59-A of *Pure and Applied Mathematics*. Academic Press, New York, 1974.

[20] R. Eising, Pole assignment, a new proof and algorithm. *Syst. & Contr. Letters*, 2:6–12, 1982.

[21] R. Eising, Pole assignment for systems over rings. *Syst. & Contr. Letters*, 2:225–229, 1982.

[22] R. Eising and M.L.J. Hautus, Realization algorithms for systems over a principal ideal domain. *Math. Systems Theory*, 14:353–366, 1981.

[23] E. Emre, On necessary and sufficient conditions for regulation of linear systems over rings. *SIAM J. Contr. & Opt.*, 20:155–160, 1982.

[24] E. Emre and P.P. Khargonekar, Regulation of split linear systems over rings: coefficient-assignment and observers. *IEEE Trans. Aut. Contr.*, AC-27:104–113, 1982.

[25] E. Emre and G.J. Knowles, Control of linear systems with fixed noncommensurate point delays. *IEEE Trans. Aut. Contr.*, AC-29:1083–1090, 1984.

[26] K. Forsman, Applications of constructive algebra to control problems. Tekn. lic. dissertation, Linköping University, Linköping, 1990. Linköping Studies in Science and Technology, Thesis No. 231.

[27] K. Forsman, *Constructive Commutative Algebra in Nonlinear Control Theory*. PhD thesis, Linköping University, Linköping, 1991. Linköping Studies in Science and Technology. Dissertations No. 261.

[28] G. Gallo and B. Mishra, Efficient algorithms and bounds for Wu-Ritt characteristic sets. In T. Mora and C. Traverso, editors, *Effective methods in algebraic geometry*, pages 119–142, Boston, 1991. Birkhäuser.

[29] F.R. Gantmacher, *The Theory of Matrices, Volume I.* Chelsea, New York, 1959.

[30] P. Gianni, B. Trager, and G. Zacharias, Gröbner bases and primary decomposition of polynomial ideals. In L. Robbiano, editor, *Computational Aspects of Commutative Algebra*, pages 15–33. Academic Press, London, 1989.

[31] K. Glover, Robust stabilization of linear multivariable systems: relations to approximation. *Int. J. Contr.*, 43:741–766, 1986.

[32] W. Gröbner, *Moderne algebraische Geometrie, die idealtheoretischen Grundlagen.* Springer-Verlag, Wien, 1949.

[33] G. Gu, P.P. Khargonekar, and E.B. Lee, Approximation of infinite-dimensional systems. *IEEE Trans. Aut. Contr.*, 34:610–618, 1989.

[34] G. Gu, P.P. Khargonekar, E.B. Lee, and P. Misra, Finite-dimensional approximations of unstable infinite-dimensional systems. *SIAM J. Contr. & Opt.*, 30:704–716, 1992.

[35] L.C.G.J.M. Habets, Characteristic sets in commutative algebra: an overview. Memorandum COSOR 92-24, Eindhoven University of Technology, Eindhoven, 1992.

[36] L.C.G.J.M. Habets, A reliable stability test for exponential polynomials. Memorandum COSOR 92-48, Eindhoven University of Technology, Eindhoven, 1992.

[37] L.C.G.J.M. Habets, Stabilization of time-delay systems: an overview of the algebraic approach. EUT Report 92-WSK-02, Eindhoven University of Technology, Eindhoven, 1992.

[38] L.C.G.J.M. Habets, On the genericity of stabilizability for time-delay systems. Memorandum COSOR 93-14, Eindhoven University of Technology, Eindhoven, 1993.

[39] L.C.G.J.M. Habets, A reachability test for systems over polynomial rings using Gröbner bases. In *Proc. ACC '93*, volume 1, pages 226–230, San Francisco, 1993.

[40] L.C.G.J.M. Habets, Testing reachability and stabilizability of systems over polynomial rings using Gröbner bases. Memorandum COSOR 93-29, Eindhoven University of Technology, Eindhoven, 1993.

[41] J. Hale, *Theory of Functional Differential Equations*, volume 3 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 1977.

[42] M.L.J. Hautus, Linear equations over rings. Personal communication.

[43] M.L.J. Hautus, Controllability and observability conditions of linear autonomous systems. *Indag. Math.*, 31:443–448, 1969.

[44] M.L.J. Hautus and E.D. Sontag, An approach to detectability and observers. *Lectures in Applied Mathematics*, 18:99–135, 1980.

[45] H. Hironaka, Resolution of singularities of an algebraic variety over a field of characteristic zero: I, II. *Annals of Math.*, 79:109–326, 1964.

[46] K. Hoffman, *Banach spaces of analytic functions.* Prentice-Hall, London, 1962.

[47] T. Kailath, *Linear systems.* Prentice-Hall, Englewood Cliffs, N.J., 1980.

[48] R.E. Kalman, P.L. Falb, and M.A. Arbib, *Topics in mathematical system theory.* McGraw-Hill, New York, 1969.

[49] E.W. Kamen, Lectures on Algebraic System Theory: Linear Systems Over Rings. Nasa contractor report 3016, NASA, 1978.

[50] E.W. Kamen, On the relationship between zero criteria for two-variable polynomials and asymptotic stability of delay differential equations. *IEEE Trans. Aut. Contr.*, AC-25:983–984, 1980.

[51] E.W. Kamen, Linear systems with commensurate time delays: stability and stabilization independent of delay. *IEEE Trans. Aut. Contr.*, AC-27:367–375, 1982.

[52] E.W. Kamen, Correction to "Linear systems with commensurate time delays: stability and stabilization independent of delay". *IEEE Trans. Aut. Contr.*, AC-28:248–249, 1983.

[53] E.W. Kamen, P.P. Khargonekar, and A. Tannenbaum, Pointwise stability and feedback control of linear systems with noncommensurate time delays. *Acta Appl. Math.*, 2:159–184, 1984.

[54] E.W. Kamen, P.P. Khargonekar, and A. Tannenbaum, Stabilization of time-delay systems using finite-dimensional compensators. *IEEE Trans. Aut. Contr.*, AC-30:75–78, 1985.

[55] E.W. Kamen, P.P. Khargonekar, and A. Tannenbaum, New techniques for the control of linear infinite-dimensional systems. In C.I. Byrnes and A. Lindquist, editors, *Frequency Domain and State Space Methods for Linear Systems*, pages 355–365. Elseviers Science, North-Holland, 1986.

[56] E.W. Kamen, P.P. Khargonekar, and A. Tannenbaum, Proper stable Bezout factorizations and feedback control of linear time-delay systems. *Int. J. Contr.*, 43:837–857, 1986.

[57] P.P. Khargonekar, On matrix fraction representations for linear systems over commutative rings. *SIAM J. Contr. & Opt.*, 20:172–197, 1982.

[58] P.P. Khargonekar and E.D. Sontag, On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings. *IEEE Trans. Aut. Contr.*, AC-27:627–638, 1982.

[59] E.R. Kolchin, *Differential Algebra and Algebraic Groups*, volume 54 of *Pure and applied mathematics*. Academic Press, New York, 1973.

[60] V.B. Kolmanovskii and V.R. Nosov, *Stability of Functional Differential Equations*, volume 180 of *Mathematics in science and engineering*. Academic Press, London, 1986.

[61] E. Kreyszig, *Introductory functional analysis with applications*. Wiley, London, 1978.

[62] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. Wiley, New York, 1972.

[63] Y.N. Lakshman, A single exponential bound on the complexity of computing Gröbner bases of zero dimensional ideals. In T. Mora and C. Traverso, editors, *Effective methods in algebraic geometry*, pages 227–234, Boston, 1991. Birkhäuser.

[64] Y.N. Lakshman and D. Lazard, On the complexity of zero-dimensional algebraic systems. In T. Mora and C. Traverso, editors, *Effective methods in algebraic geometry*, pages 217–225, Boston, 1991. Birkhäuser.

[65] S. Lang, *Introduction to algebraic geometry*. Addison-Wesley, Reading, Massachusetts, 1972. Third printing.

[66] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*. Wiley, New York, 1967.

[67] E.B. Lee and A.W. Olbrot, On reachability over polynomial rings and a related genericity problem. *Int. J. Systems Sci.*, 13:109–113, 1982.

[68] A. Manitius, Necessary and sufficient conditions of approximate controllability for general linear retarded systems. *SIAM J. Contr. & Opt.*, 19:516–532, 1981.

[69] A. Manitius, F-Controllability and observability of linear retarded systems. *Appl. Math. Optim.*, 9:73–95, 1982.

[70] E. Mayr and A. Meyer, The complexity of the word problem for commutative semigroups and polynomial ideals. *Adv. Math.*, 46:305–329, 1982.

[71] A.S. Morse, Ring models for delay-differential systems. *Automatica*, 12:529–531, 1976.

[72] J.W. Nieuwenhuis and J.C. Willems, Continuity of dynamical systems: a system theoretical approach. *Math. Contr. Sign. & Syst.*, 1:147–165, 1988.

[73] L. Pandolfi, On feedback stabilization of functional differential equations. *Boll. Un. Mat. Ital.*, 11:626–635, 1975.

[74] L. Pandolfi, Controllability properties of perturbed distributed parameter systems. *Lin. Alg. Appl.*, 122/123/124:525–538, 1989.

[75] J.R. Partington, K. Glover, H.J. Zwart, and R.F. Curtain, L_∞ approximation and nuclearity of delay systems. *Syst. & Contr. Letters*, 10:59–65, 1988.

[76] F. Pauer and M. Pfeifhofer, The theory of Gröbner bases. *L'Enseignement Mathématique*, 34:215–232, 1988.

[77] L.S. Pontryagin, On the zeros of some elementary transcendental functions. *Izv. Akad. Nauk SSR, Ser. Mat.*, 6:115–134, 1942. The English translition is given in *Amer. Math. Soc. Transl.*, Ser. 2, Vol. 1, pp. 95-110, 1955.

[78] J.F. Ritt, *Differential Equations from the Algebraic Standpoint*, volume 14 of *Amer. Math. Soc. Colloq. Publ.* American Mathematical Society, New York, 1932.

[79] J.F. Ritt, *Differential Algebra*, volume 33 of *Amer. Math. Soc. Colloq. Publ.* American Mathematical Society, New York, 1950. Also published in 1966 by Dover Publications, New York.

[80] Y. Rouchaleau, Régulation statique et dynamique d'un système héréditaire. In A. Bensoussan and J.L. Lions, editors, *Analysis and Optimization of Systems*, volume 44 of *Lecture Notes in Control and Information Sciences*, pages 532–547, Berlin, 1982. Springer-Verlag.

[81] Y. Rouchaleau and B.F. Wyman, Linear dynamical systems over integral domains. *J. Comput. System Sci.*, 9:129–142, 1974.

[82] W. Rudin, The closed ideals in an algebra of analytic functions. *Canadian J. Math.*, 9:426–434, 1957.

[83] W. Rudin, *Real and complex analysis.* McGraw-Hill, New York, third edition, 1987.

[84] J.M. Schumacher, A direct approach to compensator design for distributed parameter systems. *SIAM J. Contr. & Opt.*, 21:823–836, 1983.

[85] E.D. Sontag, Linear systems over commutative rings: a survey. *Ricerche di Automatica*, 7:1–34, 1976.

[86] E.D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, volume 6 of *Texts in Applied Mathematics.* Springer-Verlag, New York, 1990.

[87] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis.* Springer-Verlag, New York, 1980.

[88] D.J. Struik, *Lectures on Classical Differential Geometry.* Addison-Wesley, Reading, MA, 2nd. edition, 1961. Also published by Dover Publications, New York, 1988.

[89] A. Tannenbaum, On pole assignability over polynomial rings. *Syst. & Contr. Letters*, 2:13–16, 1982.

[90] A. Tannenbaum, Polynomial rings over arbitrary fields in two or more variables are not pole assignable. *Syst. & Contr. Letters*, 2:222–224, 1982.

[91] G.P. Tolstov, *Fourier Series.* Prentice-Hall, Englewood Cliffs, NJ, 1962. Also published by Dover Publications, New York, 1976.

[92] N.G. Čebotarev and N.N. Meĭman, The Routh-Hurwitz problem for polynomials and entire functions. *Trudy Mat. Inst. Steklov*, 26, 1949.

[93] B.L. van der Waerden, *Algebra Volume I.* Springer-Verlag, New York, seventh edition, 1991.

[94] B.L. van der Waerden, *Algebra Volume II.* Springer Verlag, New York, fifth edition, 1991.

[95] M. Vidyasagar, *Control System Synthesis: A Factorization Approach.* MIT Press, Cambridge, MA, 1985.

[96] D.M. Wang, The generalization of characteristic sets algorithm. RISC-Linz Series 89-51.0, Johannes Kepler University, Linz, 1989.

[97] D.M. Wang. An implementation of charactersitic sets method in Maple. Available by anonymous ftp: 129.132.101.33, 1991.

[98] D.M. Wang, Characteristic sets and zero structure of polynomial sets. Preprint, Research Institute for Symbolic Computation, Johannes Kepler University, Linz, 1992.

[99] D.M. Wang, A method for factorizing multivariate polynomials over successive algebraic extension fields. Preprint, Research Institute for Symbolic Computation, Johannes Kepler University, Linz, 1992.

[100] P.J. Wangersky and W.J. Cunningham, Time lag in prey-predation population models. *Ecology*, 38:136–139, 1957.

[101] Wu-Wen-Tsun, Basic principles of mechanical theorem proving in elementary geometries. *J. Sys. Sci. & Math. Sciences*, 4:207–235, 1984. Also published in: *J. of Automated Reasoning*, vol. 2, pp. 221-252, 1986.

[102] Y. Yamamoto, Pseudo-rational input/output maps and their realizations: a fractional representation approach to infinite-dimensional systems. *SIAM J. Contr. & Opt.*, 26:1415–1430, 1988.

[103] J. Zabczyk, *Mathematical Control Theory: An Introduction.* Systems & Control: Foundations & Applications. Birkhäuser, Boston, 1992.

[104] O. Zariski and P. Samuel, *Commutative Algebra, Volume I.* Van Nostrand, Princeton, New Jersey, 1958.

[105] O. Zariski and P. Samuel, *Commutative Algebra, Volume II.* Van Nostrand, Princeton, New Jersey, 1960.

Samenvatting

Systemen met tijdvertraging kunnen worden gezien als een vrij eenvoudige generalisatie van lineaire tijdsinvariante systemen. Ze worden beschreven door vergelijkingen van de vorm

$$\Sigma \begin{cases} \dot{x}(t) &= \sum_{i=1}^{k} (A_i x(t-\tau_i) + B_i u(t-\tau_i)), \\ y(t) &= \sum_{i=1}^{k} (C_i x(t-\tau_i) + D_i u(t-\tau_i)), \end{cases}$$

waarbij $x \in \mathbb{R}^n$ een evolutie-variabele is, $u \in \mathbb{R}^m$ een ingangsvariabele, en $y \in \mathbb{R}^p$ een uitgangsvariabele. De parameters $\tau_i > 0$ $(i = 1, \ldots, k)$ beschrijven de tijdvertragingen die in het systeem voorkomen. In tegenstelling tot systemen zonder tijdvertraging zijn $\dot{x}$ en y op het tijdstip t niet alleen afhankelijk van x en u op het tijdstip t, maar ook van de waarden van x en u op specifieke tijdstippen in het verleden. Daarom wordt Σ ook vaak een systeem met puntvertragingen genoemd, om het verschil met systemen met zogenaamde gedistribueerde tijdvertragingen beter aan te duiden.

In de literatuur wordt een systeem met tijdvertraging vaak beschreven als een oneindig-dimensionaal systeem. Bij de bestudering van deze systemen maakt men dan voornamelijk gebruik van functionaal-analytische methoden. In dit proefschrift wordt echter gekozen voor een andere, meer algebraïsche aanpak, die ook in de literatuur wordt voorgesteld. Na de invoering van een aantal vertragingsoperatoren kan men een systeem met tijdvertraging beschrijven als een lineair systeem over een polynoomring. De vertragingsoperatoren worden dan beschouwd als onbepaalden, en daarmee wordt het tijdvertragingskarakter van het systeem (tijdelijk) geëlimineerd. Deze algebraïsche aanpak heeft een belangrijk voordeel: verschillende methoden uit de constructieve commutatieve algebra kunnen worden toegepast om systeemtheoretische problemen op te lossen. Soms biedt deze aanpak echter onvoldoende soelaas. In dat geval kan men vaak nog resultaat boeken door het tijdvertragingskarakter van het systeem expliciet te gebruiken. Dit verdoezelen en weer oprakelen van informatie over het tijdvertragingskarakter van het systeem loopt als een rode draad door dit proefschrift.

In Hoofdstuk 2 wordt gestart met een inleiding over systemen over ringen in het algemeen, waarbij voornamelijk wordt ingegaan op de begrippen bereikbaarheid en stabiliseerbaarheid door dynamische terugkoppeling. Deze eigenschappen kunnen worden gekarakteriseerd met behulp van rechts-inverteerbaarheidscondities op een polynoommatrix die aan het gegeven systeem gerelateerd is. Voor systemen met

tijdvertraging kan de conditie voor stabiliseerbaarheid nog verder worden gespecialiseerd. Door expliciet van het tijdvertragingskarakter gebruik te maken, verkrijgt men een rangconditie die kan worden gezien als een specialisatie van de Hautustest naar systemen met tijdvertraging. Het blijkt dat deze rangconditie voor stabiliseerbaarheid erg zwak is. Nadat er een natuurlijke topologie gedefinieerd is op de ruimte die alle systemen met puntvertragingen beschrijft, wordt bewezen dat de verzameling van alle stabiliseerbare systemen een deelverzameling bevat die open is en dicht ligt in de ruimte van alle tijdvertraagde systemen. Dit betekent dat stabiliseerbaarheid in deze topologie een generieke eigenschap is.

Het tweede deel van het proefschrift is meer algoritmisch van aard. Eerst wordt een overzicht gegeven van twee methoden uit de constructieve commutatieve algebra voor de manipulatie van polynoomidealen: Gröbnerbases en karakteristieke verzamelingen. Vervolgens wordt met name de Gröbnerbasismethode gebruikt om de rechts-inverteerbaarheidscondities voor bereikbaarheid en stabiliseerbaarheid expliciet te verifiëren. Hiertoe worden enkele polynoomidealen ingevoerd die de bereikbaarheid en stabiliseerbaarheid van een systeem op een eenvoudige wijze karakteriseren. De berekening van een Gröbnerbasis van ieder van deze idealen leidt tot een algoritme om de bereikbaarheid van een systeem over een polynoomring te testen. Dezelfde methoden kunnen worden gebruikt om na te gaan of een willekeurige niet-vierkante polynoommatrix rechts-inverteerbaar is. Met één van de voorgestelde algoritmen is het ook mogelijk een polynomiale rechter-inverse te bepalen. Het stabiliseerbaarheidsprobleem ligt een stuk moeilijker en wordt alleen voor systemen met tijdvertraging opgelost. Door gebruik te maken van het tijdvertragingskarakter van deze systemen kan men verschillende algoritmen verkrijgen om stabiliseerbaarheid te testen. Naast de berekening van Gröbnerbases is hier ook de (numerieke) bepaling van nulpunten van univariabele polynomen voor nodig.

Tenslotte wordt nog ingegaan op de vraag hoe een systeem met tijdvertraging daadwerkelijk gestabiliseerd kan worden door middel van dynamische uitgangsterugkoppeling. Daartoe wordt eerst een numeriek algoritme ontwikkeld om de stabiliteit van een tijdvertraagd systeem te onderzoeken. Vervolgens worden enkele, in de literatuur reeds bekende, methoden behandeld om het stabiliseerbaarheidsprobleem voor systemen met tijdvertraging constructief op te lossen. Sommige van deze methoden hebben een algebraïsch karakter, andere zijn gebaseerd op de theorie van oneindig-dimensionale systemen. Dit illustreert dat zowel de algebraïsche als de functionaal-analytische aanpak van systemen met tijdvertraging hun eigen merites hebben. Wellicht kan dit proefschrift ertoe bijdragen om de enigszins onderbelichte algebraïsche aanpak wat meer onder de aandacht te brengen.

Curriculum Vitae

7 september 1966	Geboren te Eindhoven.
1978-1984	Gymnasium B, Van Maerlantlyceum, Eindhoven.
1984-1989	Technische Universiteit Eindhoven, 1986 Propadeuse Wiskunde, 1989 Doctoraalexamen Wiskunde (met lof), Afstudeerrichting Systeemtheorie.
1989-1990	Toegevoegd onderzoeker, groep Systeemtheorie, Technische Universiteit Eindhoven.
1989-1991	Postdoctorale opleiding in het kader van het Netwerk Systeem- en Regeltheorie.
1990-1994	Onderzoeker in opleiding bij de Faculteit Wiskunde en Informatica van de Technische Universiteit Eindhoven, op een onderzoeksproject financieel ondersteund door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

STELLINGEN

behorende bij het proefschrift

Algebraic and computational aspects of time-delay systems

van

Luc Habets

1. Zij $T \in \mathbb{R}(s)^{p\times m}$ een eigenlijke overdrachtsmatrix, en veronderstel dat $\Sigma = (A,B,C,D)$ een minimale realisatie is van T. We beschouwen het probleem van optimale robuuste stabilisatie in de zogenaamde gap-metriek (zie [2]). We zijn geïnteresseerd in een dynamische compensator C die niet alleen T stabiliseert, maar ook systemen in een zo groot mogelijke omgeving van T. Men wil deze compensator C daarom zo kiezen dat, in termen van de gap-metriek, de straal van de bol rond T waarbinnen alle systemen door C worden gestabiliseerd, gemaximaliseerd wordt.

Laat X en Y de unieke positief definiete oplossingen zijn van de algebraïsche Riccativergelijkingen

$$\begin{aligned}(A-BH^{-1}D^TC)^TX+X(A-BH^{-1}D^TC)-XBH^{-1}B^TX+C^TL^{-1}C &= 0,\\ (A-BD^TL^{-1}C)Y+Y(A-BD^TL^{-1}C)^T-YC^TL^{-1}CY+BH^{-1}B^T &= 0,\end{aligned}$$

met $H := I + D^TD$ en $L := I + DD^T$. Dan wordt de maximale stabiliteitsstraal $r_{\max}$ in de gap-metriek gegeven door (zie [3], [4]):

$$r_{\max} = \frac{1}{\sqrt{1+\lambda_{\max}(XY)}},$$

waarbij $\lambda_{\max}(XY)$ de grootste eigenwaarde van de matrix XY aanduidt.

Dit resultaat kan als volgt in twee stappen bewezen worden. Zij V de overdrachtsmatrix van het aan Σ gerelateerde systeem Υ dat volledig beschreven wordt door de minimale anti-stabiele realisatie

$$\Upsilon = \left(-(A-BF)^T, (I+XY)C^TL^{-\frac{1}{2}}, H^{-\frac{1}{2}}B^T, -H^{-\frac{1}{2}}D^TL^{\frac{1}{2}}\right),$$

met $F := H^{-1}(D^TC + B^TX)$, en definieer $\mathbb{R}H_\infty := \mathbb{R}(s)\cap H_\infty$. Dan geldt

$$r_{\max} = \frac{1}{\sqrt{1+\left(\inf_{R\in\mathbb{R}H_\infty{}^{m\times p}}\|V+R\|_\infty\right)^2}}.$$

Het infimum in de bovenstaande formule is te bepalen als de norm van de Hankeloperator Γ_V met symbool V (zie [1]), en op die manier volgt dat

$$\inf_{R\in\mathbb{R}H_\infty{}^{m\times p}}\|V+R\|_\infty = \|\Gamma_V\| = \sqrt{\lambda_{\max}(XY)}.$$

Referenties

[1] B.A. Francis, *A Course in H_∞ Control Theory*, volume 88 of *Lecture Notes in Control and Information Sciences*. Springer Verlag, Berlin, 1987.

[2] T.T. Georgiou and M.C. Smith, Optimal robustness in the gap metric. *IEEE Trans. Aut. Contr.*, 35:673-686, 1990.

[3] K. Glover and D. McFarlane, Robust stabilization of normalized coprime factor plant descriptions with H_∞-bounded uncertainty. *IEEE Trans. Aut. Contr.*, 34:821-830, 1989.

[4] L.C.G.J.M. Habets, *Robust Stabilization in the Gap-topology*, volume 150 of *Lecture Notes in Control and Information Sciences*. Springer Verlag, Berlin, 1991.

2. Beschouw een polynomiaal systeem met ingang u en uitgang y dat één commensurabele tijdvertraging τ bevat. Zij σ de met de tijdvertraging τ corresponderende vertragingsoperator. Veronderstel dat het systeem gegeven wordt door een n-tal gekoppelde differentie-differentiaal vergelijkingen van differentiaal-orde 1 in de interne variabelen $\bar{x} = (x_1, \ldots, x_n)$:

$$\begin{aligned} \dot{x}_1(t) &= f_1(\bar{x}, u), \\ &\vdots \qquad\qquad \vdots \\ \dot{x}_n(t) &= f_n(\bar{x}, u), \end{aligned}$$

en een uitgangsvergelijking

$$y = h(\bar{x}, u),$$

waarin de functies $f_1, \ldots, f_n$ en h polynomen zijn in de variabelen

$$\bar{x}(t), \sigma\bar{x}(t), \ldots, \sigma^k\bar{x}(t),$$
$$u(t), \sigma u(t), \ldots, \sigma^k u(t),$$

voor zekere $k \in \mathbb{N}$.

Dan is het mogelijk de interne variabelen $\bar{x}$ te elimineren: er bestaat een differentie-differentiaal vergelijking in u en y, waaraan y, gegeven u, voldoet. Bovendien kan de differentiaal-orde in y van deze differentie-differentiaal vergelijking $\leq n$ worden gekozen.

Referenties

[5] K. Forsman and L.C.G.J.M. Habets, Input-output equations and observability for polynomial delay systems. Memorandum COSOR 94-12, Eindhoven University of Technology, Eindhoven, 1994.

3. De transformatie die aan een rij vectoren $(x_i)_{i\in\mathbb{N}\cup\{0\}}$ in $\mathbb{R}^n$ de Laurentreeks

$$\sum_{i=0}^{\infty} x_i z^{-i}$$

toevoegt, wordt doorgaans met de naam *z-transformatie* aangeduid. Deze naam is onnodig suggestief. Evenals de Laplacetransformatie kan ook de z-transformatie met ieder willekeurig symbool worden uitgevoerd. De naam *discrete-tijd Laplacetransformatie* zou daarom veel minder verwarrend zijn.

4. In de computeralgebra blijken problemen die eenvoudig geformuleerd kunnen worden soms een verschrikkelijk gecompliceerde oplossing te hebben. In deze gevallen verkrijgt men als oplossing een object dat door een computer op eenvoudige wijze gemanipuleerd kan worden, maar dat voor de mens niet meer te overzien is. Het is dan twijfelachtig of het oorspronkelijke probleem, dan wel de door de computer berekende uitkomst, als oplossing beschouwd dient te worden. Weliswaar leveren exacte rekenmethoden in principe correcte antwoorden, maar deze methoden vereisen bovenal dat de juiste vraag wordt gesteld.

5. (a) In het euthanasiedebat worden de vragen naar de zin en naar het nut van het leven onterecht met elkaar verward. Terwijl de eerste, filosofische vraag doorgaans onbeantwoord blijft, kent men aan ieder mensenleven een economische waarde toe. Daarmee degradeert men het leven tot een wegwerpartikel.

(b) In een beschaafd land dient de overheid het onder (a) beschreven utiliteitsprincipe krachtdadig te verwerpen, en het leven, juist in zijn meest kwetsbare vorm, zo goed mogelijk te beschermen door middel van een strenge euthanasiewetgeving.

6. De toenemende ontkerkelijking en het groeiende materialisme zijn twee verschijnselen in onze huidige maatschappij die elkaar versterken. Enerzijds wordt getracht met een steeds grotere materiële rijkdom een verarmd geestelijk leven te camoufleren. Anderzijds lijkt menigeen zo zeer op te gaan in het eigen welvarend bestaan dat de diepere zin van het leven uit het oog wordt verloren.

7. Men moet niet proberen om gelukkig te *worden*, maar om gelukkig te *zijn*.

8. Politieke partijen die slechts de belangen van één bepaalde groep in de samenleving behartigen, vormen een bedreiging voor hun eigen achterban. Wanneer dit fenomeen algemeen navolging zou vinden, kan dit (in de meest extreme vorm) leiden tot een dictatuur van de grootste groepering binnen de samenleving.

9. Om te voorkomen dat de radiouitzendingen van de publieke omroep in Nederland Hilversumse eenheidsworst worden, dienen de regionale omroepen er voor te zorgen dat hun presentatoren beschikken over de tongval van de betreffende regio.

10. Het feit dat 25- en 50-jarige jubilea meestal op grootse wijze worden gevierd heeft meer te maken met het gegeven dat een mens aan iedere hand vijf vingers heeft, dan met het bijzondere karakter van deze zogenaamde kroonjaren.