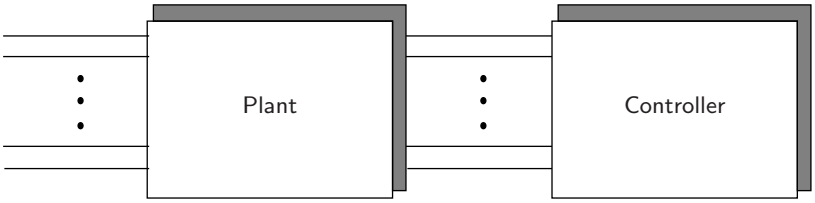
Introduction to the Mathematical Theory of Systems and Control



Jan Willem Polderman
Jan C. Willems

Contents

Preface **ix**

1 Dynamical Systems **1**
1.1 Introduction 1
1.2 Models 3
1.2.1 The universum and the behavior 3
1.2.2 Behavioral equations 4
1.2.3 Latent variables 5
1.3 Dynamical Systems 8
1.3.1 The basic concept 9
1.3.2 Latent variables in dynamical systems 10
1.4 Linearity and Time-Invariance 15
1.5 Dynamical Behavioral Equations 16
1.6 Recapitulation 19
1.7 Notes and References 20
1.8 Exercises 20

2 Systems Defined by Linear Differential Equations **27**
2.1 Introduction 27
2.2 Notation 28

2.3 Constant-Coefficient Differential Equations 31
2.3.1 Linear constant-coefficient differential equations . . . 31
2.3.2 Weak solutions of differential equations 33
2.4 Behaviors Defined by Differential Equations 37
2.4.1 Topological properties of the behavior 38
2.4.2 Linearity and time-invariance 43
2.5 The Calculus of Equations 44
2.5.1 Polynomial rings and polynomial matrices 44
2.5.2 Equivalent representations 45
2.5.3 Elementary row operations and unimodular polynomial matrices 49
2.5.4 The Bezout identity 52
2.5.5 Left and right unimodular transformations 54
2.5.6 Minimal and full row rank representations 57
2.6 Recapitulation 60
2.7 Notes and References 61
2.8 Exercises 61
2.8.1 Analytical problems 63
2.8.2 Algebraic problems 64

3 Time Domain Description of Linear Systems 67
3.1 Introduction 67
3.2 Autonomous Systems 68
3.2.1 The scalar case 71
3.2.2 The multivariable case 79
3.3 Systems in Input/Output Form 83
3.4 Systems Defined by an Input/Output Map 98
3.5 Relation Between Differential Systems and Convolution Systems 101
3.6 When Are Two Representations Equivalent? 103
3.7 Recapitulation 106
3.8 Notes and References 107
3.9 Exercises 107

4 State Space Models 119
4.1 Introduction 119
4.2 Differential Systems with Latent Variables 120

4.3 State Space Models 120
4.4 Input/State/Output Models 126
4.5 The Behavior of i/s/o Models 127
4.5.1 The zero input case 128
4.5.2 The nonzero input case: The variation of the constants formula 129
4.5.3 The input/state/output behavior 131
4.5.4 How to calculate e^{At}? 133
4.5.4.1 Via the Jordan form 134
4.5.4.2 Using the theory of autonomous behaviors 137
4.5.4.3 Using the partial fraction expansion of $(I\xi - A)^{-1}$ 140
4.6 State Space Transformations 142
4.7 Linearization of Nonlinear i/s/o Systems 143
4.8 Recapitulation 148
4.9 Notes and References 149
4.10 Exercises 149

5 Controllability and Observability 155
5.1 Introduction 155
5.2 Controllability 156
5.2.1 Controllability of input/state/output systems 167
5.2.1.1 Controllability of i/s systems 167
5.2.1.2 Controllability of i/s/o systems 174
5.2.2 Stabilizability 175
5.3 Observability 177
5.3.1 Observability of i/s/o systems 181
5.3.2 Detectability 187
5.4 The Kalman Decomposition 188
5.5 Polynomial Tests for Controllability and Observability . . . 192
5.6 Recapitulation 193
5.7 Notes and References 194
5.8 Exercises 195

6 Elimination of Latent Variables and State Space Representations 205
6.1 Introduction 205

6.2 Elimination of Latent Variables 206
6.2.1 Modeling from first principles 206
6.2.2 Elimination procedure 210
6.2.3 Elimination of latent variables in interconnections . 214
6.3 Elimination of State Variables 216
6.4 From i/o to i/s/o Model 220
6.4.1 The observer canonical form 221
6.4.2 The controller canonical form 225
6.5 Canonical Forms and Minimal State Space Representations 229
6.5.1 Canonical forms 230
6.5.2 Equivalent state representations 232
6.5.3 Minimal state space representations 233
6.6 Image Representations 234
6.7 Recapitulation 236
6.8 Notes and References 237
6.9 Exercises 237

7 Stability Theory 247
7.1 Introduction 247
7.2 Stability of Autonomous Systems 250
7.3 The Routh–Hurwitz Conditions 254
7.3.1 The Routh test 255
7.3.2 The Hurwitz test 257
7.4 The Lyapunov Equation 259
7.5 Stability by Linearization 268
7.6 Input/Output Stability 271
7.7 Recapitulation 276
7.8 Notes and References 277
7.9 Exercises 277

8 Time- and Frequency-Domain Characteristics of Linear Time-Invariant Systems 287
8.1 Introduction 287
8.2 The Transfer Function and the Frequency Response 288
8.2.1 Convolution systems 289
8.2.2 Differential systems 291

8.2.3 The transfer function represents the controllable part of the behavior . . . 295
8.2.4 The transfer function of interconnected systems . . . 295
8.3 Time-Domain Characteristics . . . 297
8.4 Frequency-Domain Response Characteristics . . . 300
8.4.1 The Bode plot . . . 302
8.4.2 The Nyquist plot . . . 303
8.5 First- and Second-Order Systems . . . 304
8.5.1 First-order systems . . . 304
8.5.2 Second-order systems . . . 304
8.6 Rational Transfer Functions . . . 307
8.6.1 Pole/zero diagram . . . 308
8.6.2 The transfer function of i/s/o representations . . . 308
8.6.3 The Bode plot of rational transfer functions . . . 310
8.7 Recapitulation . . . 313
8.8 Notes and References . . . 313
8.9 Exercises . . . 314

9 Pole Placement by State Feedback 317
9.1 Open Loop and Feedback Control . . . 317
9.2 Linear State Feedback . . . 323
9.3 The Pole Placement Problem . . . 324
9.4 Proof of the Pole Placement Theorem . . . 325
9.4.1 System similarity and pole placement . . . 326
9.4.2 Controllability is necessary for pole placement . . . 327
9.4.3 Pole placement for controllable single-input systems 327
9.4.4 Pole placement for controllable multi-input systems 329
9.5 Algorithms for Pole Placement . . . 331
9.6 Stabilization . . . 333
9.7 Stabilization of Nonlinear Systems . . . 335
9.8 Recapitulation . . . 339
9.9 Notes and References . . . 339
9.10 Exercises . . . 340

10 Observers and Dynamic Compensators 347
10.1 Introduction . . . 347
10.2 State Observers . . . 350

10.3 Pole Placement in Observers 352
10.4 Unobservable Systems 355
10.5 Feedback Compensators 356
10.6 Reduced Order Observers and Compensators 364
10.7 Stabilization of Nonlinear Systems 368
10.8 Control in a Behavioral Setting 370
10.8.1 Motivation 370
10.8.2 Control as interconnection 373
10.8.3 Pole placement 375
10.8.4 An algorithm for pole placement 377
10.9 Recapitulation 382
10.10Notes and References 383
10.11Exercises 383

A Simulation Exercises **391**
A.1 Stabilization of a Cart 391
A.2 Temperature Control of a Container 393
A.3 Autonomous Dynamics of Coupled Masses 396
A.4 Satellite Dynamics 397
A.4.1 Motivation 398
A.4.2 Mathematical modeling 398
A.4.3 Equilibrium Analysis 401
A.4.4 Linearization 401
A.4.5 Analysis of the model 402
A.4.6 Simulation 402
A.5 Dynamics of a Motorbike 402
A.6 Stabilization of a Double Pendulum 404
A.6.1 Modeling 404
A.6.2 Linearization 406
A.6.3 Analysis 407
A.6.4 Stabilization 408
A.7 Notes and References 409

B Background Material **411**
B.1 Polynomial Matrices 411
B.2 Partial Fraction Expansion 417
B.3 Fourier and Laplace Transforms 418

B.3.1 Fourier transform 419
B.3.2 Laplace transform 421
B.4 Notes and References 421
B.5 Exercises 422

Notation **423**

References **425**

Index **429**

Preface

The purpose of this preface is twofold. Firstly, to give an informal historical introduction to the subject area of this book, *Systems and Control*, and secondly, to explain the philosophy of the approach to this subject taken in this book and to outline the topics that will be covered.

A brief history of systems and control

Control theory has two main roots: *regulation* and *trajectory optimization*. The first, regulation, is the more important and engineering oriented one. The second, trajectory optimization, is mathematics based. However, as we shall see, these roots have to a large extent merged in the second half of the twentieth century.

The problem of regulation is to design mechanisms that keep certain to-be-controlled variables at constant values against external disturbances that act on the plant that is being regulated, or changes in its properties. The system that is being controlled is usually referred to as the *plant*, a passe-partout term that can mean a physical or a chemical system, for example. It could also be an economic or a biological system, but one would not use the engineering term "plant" in that case.

Examples of regulation problems from our immediate environment abound. Houses are regulated by thermostats so that the inside temperature remains constant, notwithstanding variations in the outside weather conditions or changes in the situation in the house: doors that may be open or closed, the

number of persons present in a room, activity in the kitchen, etc. Motors in washing machines, in dryers, and in many other household appliances are controlled to run at a fixed speed, independent of the load. Modern automobiles have dozens of devices that regulate various variables. It is, in fact, possible to view also the suspension of an automobile as a regulatory device that absorbs the irregularities of the road so as to improve the comfort and safety of the passengers. Regulation is indeed a very important aspect of modern technology. For many reasons, such as efficiency, quality control, safety, and reliability, industrial production processes require regulation in order to guarantee that certain key variables (temperatures, mixtures, pressures, etc.) be kept at appropriate values. Factors that inhibit these desired values from being achieved are external disturbances, as for example the properties of raw materials and loading levels or changes in the properties of the plant, for example due to aging of the equipment or to failure of some devices. Regulation problems also occur in other areas, such as economics and biology.

One of the central concepts in control is *feedback:* the value of one variable in the plant is measured and used (*fed back*) in order to take appropriate action through a control variable at another point in the plant. A good example of a feedback regulator is a thermostat: it senses the room temperature, compares it with the set point (the desired temperature), and feeds back the result to the boiler, which then starts or shuts off depending on whether the temperature is too low or too high.

Man has been devising control devices ever since the beginning of civilization, as can be expected from the prevalence of regulation problems. Control historians attribute the first conscious design of a regulatory feedback mechanism in the West to the Dutch inventor Cornelis Drebbel (1572–1633). Drebbel designed a clever contraption combining thermal and mechanical effects in order to keep the temperature of an oven at a constant temperature. Being an alchemist as well as an inventor, Drebbel believed that his oven, the *Athanor,* would turn lead into gold. Needless to say, he did not meet with much success in this endeavor, notwithstanding the inventiveness of his temperature control mechanism. Later in the seventeenth century, Christiaan Huygens (1629–1695) invented a flywheel device for speed control of windmills. This idea was the basis of the centrifugal fly-ball governor (see Figure P.1) used by James Watt (1736–1819), the inventor of the steam engine. The centrifugal governor regulated the speed of a steam engine. It was a very successful device used in all steam engines during the industrial revolution, and it became the first mass-produced control mechanism in existence. Many control laboratories have therefore taken Watt's fly-ball governor as their favorite icon. The control problem for steam engine speed occurred in a very natural way. During the nineteenth century, prime movers driven by steam engines were running throughout the grim factories of the industrial revolution. It was clearly important to avoid the

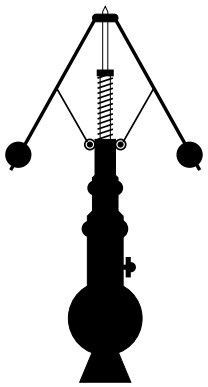


FIGURE P.1. Fly ball governor.

speed changes that would naturally occur in the prime mover when there was a change in the load, which occurred, for example, when a machine was disconnected from the prime mover. Watt's fly-ball governor achieved this goal by letting more steam into the engine when the speed decreased and less steam when the speed increased, thus achieving a speed that tends to be insensitive to load variations. It was soon realized that this adjustment should be done cautiously, since by overreacting (called *overcompensation*), an all too enthusiastic governor could bring the steam engine into oscillatory motion. Because of the characteristic sound that accompanied it, this phenomenon was called *hunting*. Nowadays, we recognize this as an instability due to high gain control. The problem of tuning centrifugal governors that achieved fast regulation but avoided hunting was propounded to James Clerk Maxwell (1831–1870) (the discoverer of the equations for electromagnetic fields) who reduced the question to one about the stability of differential equations. His paper *"On Governors,"* published in 1868 in the *Proceedings of the Royal Society of London*, can be viewed as the first mathematical paper on control theory viewed from the perspective of regulation. Maxwell's problem and its solution are discussed in Chapter 7 of this book, under the heading of the Routh-Hurwitz problem.

The field of control viewed as regulation remained mainly technology driven during the first half of the twentieth century. There were two very important developments in this period, both of which had a lasting influence on the field. First, there was the invention of the *Proportional–Integral–Differential (PID)* controller. The PID controller produces a control signal that consists of the weighted sum of three terms (a PID controller is therefore often called a *three-term* controller). The P-term produces a signal that is proportional to the error between the actual and the desired value of the to-be-controlled variable. It achieves the basic feedback compensation control, leading to a control input whose purpose is to make the to-be-controlled variable in-

crease when it is too low and decrease when it is too high. The I-term feeds back the integral of the error. This term results in a very large correction signal whenever this error does not converge to zero. For the error there hence holds, *Go to zero or bust!* When properly tuned, this term achieves *robustness*, good performance not only for the nominal plant but also for plants that are close to it, since the I-term tends to force the error to zero for a wide range of the plant parameters. The D-term acts on the derivative of the error. It results in a control correction signal as soon as the error starts increasing or decreasing, and it can thus be expected that this anticipatory action results in a fast response. The PID controller had, and still has, a very large technological impact, particularly in the area of chemical process control. A second important event that stimulated the development

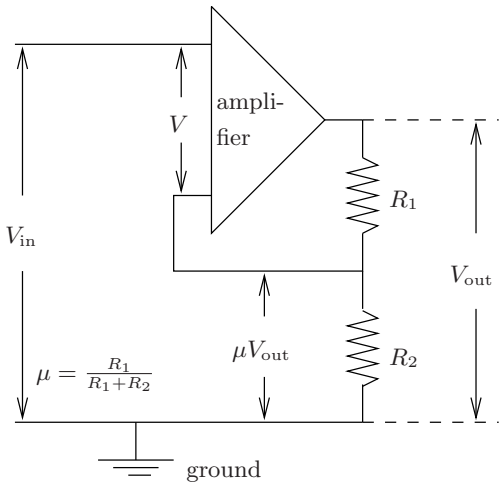


FIGURE P.2. Feedback amplifier.

of regulation in the first half of the twentieth century was the invention in the 1930s of the feedback amplifier by Black. The feedback amplifier (see Figure P.2) was an impressive technological development: it permitted signals to be amplified in a reliable way, insensitive to the parameter changes inherent in vacuum-tube (and also solid-state) amplifiers. (See also Exercise 9.3.) The key idea of Black's negative feedback amplifier is subtle but simple. Assume that we have an electronic amplifier that amplifies its input voltage V to $V_{\text{out}} = KV$. Now use a voltage divider and feed back μV_{out} to the amplifier input, so that when subtracted (whence the term *negative* feedback amplifier) from the input voltage V_{in} to the feedback amplifier, the input voltage to the amplifier itself equals $V = V_{\text{in}} - \mu V_{\text{out}}$. Combining

these two relations yields the crucial formula

$$V_{\text{out}} = \frac{1}{\mu + \frac{1}{K}} V_{\text{in}}.$$

This equation, simple as it may seem, carries an important message, see Exercise 9.3. *What's the big deal with this formula?* Well, the value of the gain K of an electronic amplifier is typically large, but also very unstable, as a consequence of sensitivity to aging, temperature, loading, etc. The voltage divider, on the other hand, can be implemented by means of passive resistors, which results in a very stable value for μ. Now, for large (although uncertain) Ks, there holds $\frac{1}{\mu+\frac{1}{K}} \approx \frac{1}{\mu}$, and so somehow Black's magic circuitry results in an amplifier with a stable amplification gain $\frac{1}{\mu}$ based on an amplifier that has an inherent uncertain gain K.

The invention of the negative feedback amplifier had far-reaching applications to telephone technology and other areas of communication, since long-distance communication was very hampered by the annoying drifting of the gains of the amplifiers used in repeater stations. Pursuing the above analysis in more detail shows also that the larger the amplifier gain K, the more insensitive the overall gain $\frac{1}{\mu+\frac{1}{K}}$ of the feedback amplifier becomes. However, at high gains, the above circuit could become dynamically unstable because of dynamic effects in the amplifier. For amplifiers, this phenomenon is called *singing*, again because of the characteristic noise produced by the resistors that accompanies this instability. Nyquist, a colleague of Black at Bell Laboratories, analyzed this stability issue and came up with the celebrated *Nyquist stability criterion*. By pursuing these ideas further, various techniques were developed for setting the gains of feedback controllers. The sum total of these design methods was termed *classical control theory* and comprised such things as the Nyquist stability test, Bode plots, gain and phase margins, techniques for tuning PID regulators, lead–lag compensation, and root–locus methods.

This account of the history of control brings us to the 1950s. We will now backtrack and follow the other historical root of control, *trajectory optimization*. The problem of trajectory transfer is the question of determining the paths of a dynamical system that transfer the system from a given initial to a prescribed terminal state. Often paths are sought that are optimal in some sense. A beautiful example of such a problem is the *brachystochrone problem* that was posed by Johann Bernoulli in 1696, very soon after the discovery of differential calculus. At that time he was professor at the University of Groningen, where he taught from 1695 to 1705. The brachystochrone problem consists in finding the path between two given points A and B along which a body falling under its own weight moves in the shortest possible time. In 1696 Johann Bernoulli posed this problem as a public challenge to his contemporaries. Six eminent mathematicians (and not just any six!) solved the problem: Johann himself, his elder brother

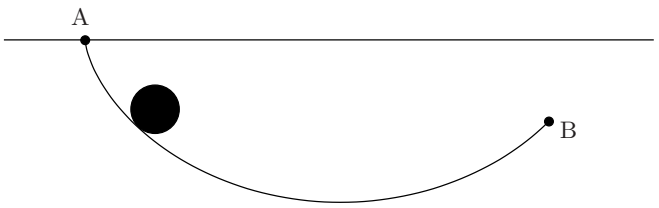


FIGURE P.3. Brachystochrone.

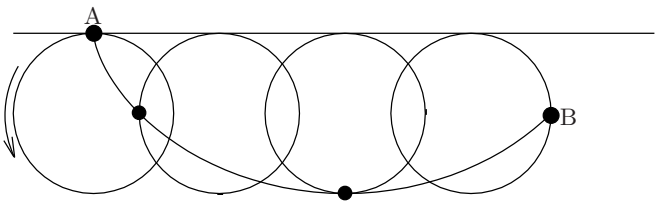


FIGURE P.4. Cycloid.

Jakob, Leibniz, de l'Hôpital, Tschirnhaus, and Newton. Newton submitted his solution anonymously, but Johann Bernoulli recognized the culprit, since, as he put it, *ex ungue leonem: you can tell the lion by its claws.* The brachystochrone turned out to be the cycloid traced by a point on the circle that rolls without slipping on the horizontal line through A and passes through A and B. It is easy to see that this defines the cycloid uniquely (see Figures P.3 and P.4).

The brachystochrone problem led to the development of the *Calculus of Variations*, of crucial importance in a number of areas of applied mathematics, above all in the attempts to express the laws of mechanics in terms of variational principles. Indeed, to the amazement of its discoverers, it was observed that the possible trajectories of a mechanical system are precisely those that minimize a suitable *action integral*. In the words of Legendre, *Ours is the best of all possible worlds.* Thus the calculus of variations had far-reaching applications beyond that of finding optimal paths: in certain applications, it could also tell us what paths are physically possible. Out of these developments came the Euler–Lagrange and Hamilton equations as conditions for the vanishing of the first variation. Later, Legendre and Weierstrass added conditions for the nonpositivity of the second variation, thus obtaining conditions for trajectories to be local minima.

The problem of finding optimal trajectories in the above sense, while extremely important for the development of mathematics and mathematical physics, was not viewed as a control problem until the second half of the twentieth century. However, this changed in 1956 with the publication of Pontryagin's *maximum principle*. The maximum principle consists of a very

general set of necessary conditions that a control input that generates an optimal path has to satisfy. This result is an important generalization of the classical problems in the calculus of variations. Not only does it allow a much larger class of problems to be tackled, but importantly, it brought forward the problem of *optimal input selection* (in contrast to optimal *path* selection) as the central issue of trajectory optimization.

Around the same time that the maximum principle appeared, it was realized that the (optimal) input could also be implemented as a function of the state. That is, rather than looking for a control input as a function of time, it is possible to choose the (optimal) input as a *feedback* function of the state. This idea is the basis for *dynamic programming*, which was formulated by Bellman in the late 1950s and which was promptly published in many of the applied mathematics journals in existence. With the insight obtained by dynamic programming, the distinction between (feedback based) *regulation* and the (input selection based) *trajectory optimization* became blurred. Of course, the distinction is more subtle than the above suggests, particularly because it may not be possible to measure the whole state accurately; but we do not enter into this issue here. Out of all these developments, both in

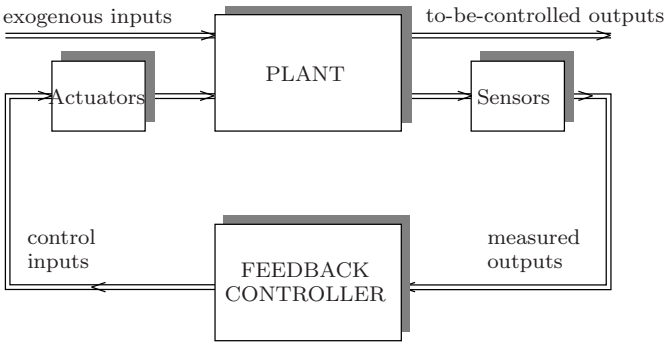


FIGURE P.5. Intelligent control.

the areas of regulation and of trajectory planning, the picture of Figure P.5 emerged as the central one in control theory. The basic aim of control as it is generally perceived is the design of the feedback processor in Figure P.5. It emphasizes *feedback* as the basic principle of control: the controller accepts the measured outputs of the plant as its own inputs, and from there, it computes the desired control inputs to the plant. In this setup, we consider the plant as a *black box* that is driven by *inputs* and that produces *outputs*. The controller functions as follows. From the sensor outputs, information is obtained about the disturbances, about the actual dynamics of the plant if these are poorly understood, of unknown parameters, and of the internal state of the plant. Based on these sensor observations, and on the control

objectives, the feedback processor computes what control input to apply. Via the actuators, appropriate influence is thus exerted on the plant.

Often, the aim of the control action is to steer the to-be-controlled outputs back to their desired equilibria. This is called *stabilization*, and will be studied in Chapters 9 and 10 of this book. However, the goal of the controller may also be *disturbance attenuation*: making sure that the disturbance inputs have limited effect on the to-be-controlled outputs; or it may be *tracking*: making sure that the plant can follow exogenous inputs. Or the design question may be *robustness*: the controller should be so designed that the controlled system should meet its *specs* (that is, that it should achieve the design specifications, as stability, tracking, or a degree of disturbance attenuation) for a wide range of plant parameters.

The mathematical techniques used to model the plant, to analyze it, and to synthesize controllers took a major shift in the late 1950s and early 1960s with the introduction of state space ideas. The classical way of viewing a system is in terms of the transfer function from inputs to outputs. By specifying the way in which exponential inputs transform into exponential outputs, one obtains (at least for linear time-invariant systems) an insightful specification of a dynamical system. The mathematics underlying these ideas are Fourier and Laplace transforms, and these very much dominated control theory until the early 1960s. In the early sixties, the prevalent models used shifted from transfer function to state space models. Instead of viewing a system simply as a relation between inputs and outputs, state space models consider this transformation as taking place via the transformation of the internal state of the system. When state models came into vogue, differential equations became the dominant mathematical framework needed. State space models have many advantages indeed. They are more akin to the classical mathematical models used in physics, chemistry, and economics. They provide a more versatile language, especially because it is much easier to incorporate nonlinear effects. They are also more adapted to computations. Under the impetus of this new way of looking at systems, the field expanded enormously. Important new concepts were introduced, notably (among many others) those of *controllability* and *observability*, which became of central importance in control theory. These concepts are discussed in Chapter 5.

Three important theoretical developments in control, all using state space models, characterized the late 1950s: the *maximum principle, dynamic programming*, and the *Linear–Quadratic–Gaussian (LQG) problem* . As already mentioned, the maximum principle can be seen as the culmination of a long, 300-year historical development related to trajectory optimization. Dynamic programming provided algorithms for computing optimal trajectories in feedback form, and it merged the feedback control picture of Figure P.5 with the optimal path selection problems of the calculus of variations. The LQG problem, finally, was a true feedback control result:

it showed how to compute the feedback control processor of Figure P.5 in order to achieve optimal disturbance attenuation. In this result the plant is assumed to be *linear*, the optimality criterion involves an integral of a *quadratic* expression in the system variables, and the disturbances are modeled as *Gaussian* stochastic processes. Whence the terminology LQG problem. The LQG problem, unfortunately, falls beyond the scope of this introductory book. In addition to being impressive theoretical results in their own right, these developments had a deep and lasting influence on the mathematical outlook taken in control theory. In order to emphasize this, it is customary to refer to the state space theory as *modern control theory* to distinguish it from the *classical control theory* described earlier.

Unfortunately, this paradigm shift had its downsides as well. Rather than aiming for a good balance between mathematics and engineering, the field of systems and control became mainly mathematics driven. In particular, mathematical modeling was not given the central place in systems theory that it deserves. Robustness, i.e., the integrity of the control action against plant variations, was not given the central place in control theory that it deserved. Fortunately, this situation changed with the recent formulation and the solution of what is called the H_∞ problem. The H_∞ problem gives a method for designing a feedback processor as in Figure P.5 that is optimally robust in some well-defined sense. Unfortunately, the H_∞ problem also falls beyond the scope of this introductory book.

A short description of the contents of this book

Both the transfer function and the state space approaches view a system as a signal processor that accepts inputs and transforms them into outputs. In the transfer function approach, this processor is described through the way in which exponential inputs are transformed into exponential outputs. In the state space approach, this processor involves the state as intermediate variable, but the ultimate aim remains to describe how inputs lead to outputs. This input/output point of view plays an important role in this book, particularly in the later chapters. However, our starting point is different, more general, and, we claim, more adapted to modeling and more suitable for applications.

As a paradigm for control, input/output or input/state/output models are often very suitable. Many control problems can be viewed in terms of plants that are driven by control inputs through actuators and feedback mechanisms that compute the control action on the basis of the outputs of sensors, as depicted in Figure P.5. However, as a tool for modeling dynamical systems, the input/output point of view is unnecessarily restrictive. Most physical systems do not have a preferred signal flow direction, and it is important to let the mathematical structures reflect this. This is the approach taken in this book: we view systems as defined by any relation among dy-

namic variables, and it is only when turning to control in Chapters 9 and 10, that we adopt the input/state/output point of view. The general model structures that we develop in the first half of the book are referred to as the *behavioral approach*. We now briefly explain the main underlying ideas.

We view a mathematical model as a subset of a universum of possibilities. Before we accept a mathematical model as a description of reality, all outcomes in the universum are in principle possible. After we accept the mathematical model as a convenient description of reality, we declare that only outcomes in a certain subset are possible. Thus a mathematical model is an exclusion law: it excludes all outcomes except those in a given subset. This subset is called the *behavior* of the mathematical model. Proceeding from this perspective, we arrive at the notion of a dynamical system as simply a subset of time-trajectories, as a family of time signals taking on values in a suitable signal space. This will be the starting point taken in this book. Thus the input/output signal flow graph emerges in general as a construct, sometimes a purely mathematical one, not necessarily implying a physical structure.

We take the description of a dynamical system in terms of its behavior, thus in terms of the time trajectories that it permits, as the vantage point from which the concepts put forward in this book unfolds. We are especially interested in *linear time-invariant differential systems:* "linearity" means that these systems obey the *superposition principle*, "time-invariance" that the laws of the system do not depend explicitly on time, and "differential" that they can be described by differential equations. Specific examples of such systems abound: linear electrical circuits, linear (or linearized) mechanical systems, linearized chemical reactions, the majority of the models used in econometrics, many examples from biology, etc.

Understanding linear time-invariant differential systems requires first of all an accurate mathematical description of the behavior, i.e., of the solution set of a system of differential equations. This issue—how one wants to define a solution of a system of differential equations—turns out to be more subtle than it may at first appear and is discussed in detail in Chapter 2. Linear time-invariant differential systems have a very nice structure. When we have a set of variables that can be described by such a system, then there is a transparent way of describing how trajectories in the behavior are generated. Some of the variables, it turns out, are free, unconstrained. They can thus be viewed as unexplained by the model and imposed on the system by the environment. These variables are called *inputs*. However, once these free variables are chosen, the remaining variables (called the *outputs*) are not yet completely specified. Indeed, the internal dynamics of the system generates many possible trajectories depending on the past history of the system, i.e., on the initial conditions inside the system. The formalization of these initial conditions is done by the concept of *state*. Discovering this

structure of the behavior with free inputs, bound outputs, and the memory, the state variables, is the program of Chapters 3, 4, and 5.

When one models an (interconnected) physical system from first principles, then unavoidably auxiliary variables, in addition to the variables modeled, will appear in the model. Those auxiliary variables are called *latent* variables, in order to distinguish them from the *manifest* variables, which are the variables whose behavior the model aims at describing. The interaction between manifest and latent variables is one of the recurring themes in this book.

We use this behavioral definition in order to study some important features of dynamical systems. Two important properties that play a central role are *controllability and observability*. Controllability refers to the question of whether or not one trajectory of a dynamical system can be steered towards another one. Observability refers to the question of what one can deduce from the observation of one set of system variables about the behavior of another set. Controllability and observability are classical concepts in control theory. The novel feature of the approach taken in this book is to cast these properties in the context of behaviors.

The book uses the behavioral approach in order to present a systematic view for constructing and analyzing mathematical models. The book also aims at explaining some synthesis problems, notably the design of control algorithms. We treat control from a classical, input/output point of view. It is also possible to approach control problems from a behavioral point of view. But, while this offers some important advantages, it is still a relatively undeveloped area of research, and it is not ready for exposition in an introductory text. We will touch on these developments briefly in Section 10.8.

We now proceed to give a chapter-by-chapter overview of the topics covered in this book.

In the first chapter we discuss the mathematical definition of a dynamical system that we use and the rationale underlying this concept. The basic ingredients of this definition are the behavior of a dynamical system as the central object of study and the notions of manifest and latent variables. The manifest variables are what the model aims at describing. Latent variables are introduced as auxiliary variables in the modeling process but are often also introduced for mathematical reasons, for purposes of analysis, or in order to exhibit a special property.

In the second chapter, we introduce linear time-invariant differential systems. It is this model class that we shall be mainly concerned with in this book. The crucial concept discussed is the notion of a solution - more specifically, of a *weak* solution of a system of differential equations. As we shall see, systems of linear time-invariant differential equations are parametrized by polynomial matrices. An important part of this chapter is devoted to

the study of properties of polynomial matrices and their interplay with differential equations.

In the third chapter we study the behavior of linear differential systems in detail. We prove that the variables in such systems may be divided into two sets: one set contains the variables that are free (we call them *inputs*), the other set contains the variables that are bound (we call them *outputs*). We also study how the relation between inputs and outputs can be expressed as a convolution integral.

The fourth chapter is devoted to *state models*. The state of a dynamical system parametrizes its memory, the extent to which the past influences the future. State equations, that is, the equations linking the manifest variables to the state, turn out to be first-order differential equations. The output of a system is determined only after the input and the initial conditions have been specified.

Chapter 5 deals with *controllability* and *observability*. A controllable system is one in which an arbitrary past trajectory can be steered so as to be concatenated with an arbitrary future trajectory. An observable system is one in which the latent variables can be deduced from the manifest variables. These properties play a central role in control theory.

In the sixth chapter we take another look at latent variable and state space systems. In particular, we show how to eliminate latent variables and how to introduce state variables. Thus a system of linear differential equations containing latent variables can be transformed in an equivalent system in which these latent variables have been eliminated.

Stability is the topic of Chapter 7. We give the classical stability conditions of systems of differential equations in terms of the roots of the associated polynomial or of the eigenvalue locations of the system matrix. We also discuss the Routh–Hurwitz tests, which provide conditions for polynomials to have only roots with negative real part.

Up to Chapter 7, we have treated systems in their natural, time-domain setting. However, linear time-invariant systems can also be described by the way in which they process sinusoidal or, more generally, exponential signals. The resulting frequency domain description of systems is explained in Chapter 8. In addition, we discuss some characteristic features and nomenclature for system responses related to the step response and the frequency domain properties.

The remainder of the book is concerned with control theory. Chapter 9 starts with an explanation of the difference between open-loop and feedback control. We subsequently prove the pole placement theorem. This theorem states that for a controllable system, there exists, for any desired monic polynomial, a state feedback gain matrix such that the eigenvalues of the closed loop system are the roots of the desired polynomial. This result,

called the *pole placement theorem*, is one of the central achievements of modern control theory.

The tenth chapter is devoted to observers: algorithms for deducing the system state from measured inputs and outputs. The design of observers is very similar to the stabilization and pole placement procedures. Observers are subsequently used in the construction of output feedback compensators. Three important cybernetic principles underpin our construction of observers and feedback compensators. The first principle is *error feedback:* The estimate of the state is updated through the error between the actual and the expected observations. The second is *certainty equivalence*. This principle suggest that when one needs the value of an unobserved variable, for example for determining the suitable control action, it is reasonable to use the estimated value of that variable, as if it were the exact value. The third cybernetic principle used is the *separation principle*. This implies that we will separate the design of the observer and the controller. Thus the observer is not designed with its use for control in mind, and the controller is not adapted to the fact that the observer produces only estimates of the state.

Notes and references

There are a number of books on the history of control. The origins of control, going back all the way to the Babylonians, are described in [40]. Two other history books on the subject, spanning the period from the industrial revolution to the postwar era, are [10, 11]. The second of these books has a detailed account of the invention of the PID regulator and the negative feedback amplifier. A collection of historically important papers, including original articles by Maxwell, Hurwitz, Black, Nyquist, Bode, Pontryagin, and Bellman, among others, have been reprinted in [9]. The history of the brachystochrone problem has been recounted in most books on the history of mathematics. Its relation to the maximum principle is described in [53]. The book [19] contains the history of the calculus of variations.

There are numerous books that explain classical control. Take any textbook on control written before 1960. The state space approach to systems, and the development of the LQG problem happened very much under the impetus of the work of Kalman. An inspiring early book that explains some of the main ideas is [15]. The special issue [5] of the *IEEE Transactions on Automatic Control* contains a collection of papers devoted to the *Linear–Quadratic–Gaussian problem*, up-to-date at the time of publication. Texts devoted to this problem are, for example, [33, 3, 4]. Classical control theory emphasizes simple, but nevertheless often very effective and robust, controllers. Optimal control à la Pontryagin and LQ control aims at trajectory transfer and at shaping the transient response; LQG techniques center on disturbance attenuation; while H_∞ control emphasizes regulation against both disturbances and plant uncertainties. The latter, H_∞ control, is an important recent development that originated with the ideas of Zames [66]. This theory culminated in the remarkable double-Riccati-equation

paper [16]. The behavioral approach originated in [55] and was further developed, for example, in [56, 57, 58, 59, 60] and in this book. In [61] some control synthesis ideas are put forward from this vantage point.

1
Dynamical Systems

1.1 Introduction

We start this book at the very beginning, by asking ourselves the question, *What is a dynamical system?*

Disregarding for a moment the dynamical aspects—forgetting about time—we are immediately led to ponder the more basic issue, *What is a mathematical model?* What does it tell us? What is its mathematical nature? Mind you, we are not asking a philosophical question: we will not engage in an erudite discourse about the relation between reality and its mathematical description. Neither are we going to elucidate the methodology involved in actually deriving, setting up, postulating mathematical models. What we are asking is the simple question, *When we accept a mathematical expression, a formula, as an adequate description of a phenomenon, what mathematical structure have we obtained?*

We view a mathematical model as an *exclusion law*. A mathematical model expresses the opinion that some things can happen, are possible, while others cannot, are declared impossible. Thus Kepler claims that planetary orbits that do not satisfy his three famous laws are impossible. In particular, he judges nonelliptical orbits as unphysical. The second law of thermodynamics limits the transformation of heat into mechanical work. Certain combinations of heat, work, and temperature histories are declared to be impossible. Economic production functions tell us that certain amounts of raw materials, capital, and labor are needed in order to manufacture a

finished product: it prohibits the creation of finished products unless the required resources are available.

We formalize these ideas by stating that a mathematical model selects a certain subset from a universum of possibilities. This subset consists of the occurrences that the model allows, that it declares possible. We call the subset in question the *behavior* of the mathematical model.

True, we have been trained to think of mathematical models in terms of equations. *How do equations enter this picture?* Simply, an equation can be viewed as a law excluding the occurrence of certain outcomes, namely, those combinations of variables for which the equations are not satisfied. This way, equations define a behavior. We therefore speak of *behavioral equations* when mathematical equations are intended to model a phenomenon. It is important to emphasize already at this point that behavioral equations provide an effective, but at the same time highly nonunique, way of specifying a behavior. Different equations can define the same mathematical model. One should therefore not exaggerate the intrinsic significance of a specific set of behavioral equations.

In addition to behavioral equations and the behavior of a mathematical model, there is a third concept that enters our modeling language *ab initio: latent variables*. We think of the variables that we try to model as *manifest* variables: they are the attributes on which the modeler in principle focuses attention. However, in order to come up with a mathematical model for a phenomenon, one invariably has to consider other, *auxiliary*, variables. We call them *latent* variables. These may be introduced for no other reason than in order to express in a convenient way the laws governing a model. For example, when modeling the behavior of a complex system, it may be convenient to view it as an interconnection of component subsystems. Of course, the variables describing these subsystems are, in general, different from those describing the original system. When modeling the external terminal behavior of an electrical circuit, we usually need to introduce the currents and voltages in the internal branches as auxiliary variables. When expressing the first and second laws of thermodynamics, it has been proven convenient to introduce the internal energy and entropy as latent variables. When discussing the synthesis of feedback control laws, it is often imperative to consider models that display their internal state explicitly. We think of these internal variables as latent variables. Thus in first principles modeling, we distinguish two types of variables. The terminology first principles modeling refers to the fact that the physical laws that play a role in the system at hand are the elementary laws from physics, mechanics, electrical circuits, etc.

This triptych—*behavior/behavioral equations/manifest and latent variables*—is the essential structure of our modeling language. The fact that we take the behavior, and not the behavioral equations, as the central object specifying a mathematical model has the consequence that basic system

properties (such as time-invariance, linearity, stability, controllability, observability) will also refer to the behavior. The subsequent problem then always arises how to deduce these properties from the behavioral equations.

1.2 Models

1.2.1 The universum and the behavior

Assume that we have a phenomenon that we want to model. To start with, we cast the situation in the language of mathematics by assuming that the *phenomenon* produces outcomes in a set $\mathbb{U}$, which we call the *universum*. Often $\mathbb{U}$ consists of a product space, for example a finite dimensional vector space. Now, a (deterministic) mathematical model for the phenomenon (viewed purely from the black-box point of view, that is, by looking at the phenomenon only from its terminals, by looking at the model as descriptive but not explanatory) claims that certain outcomes are possible, while others are not. Hence a model recognizes a certain subset $\mathfrak{B}$ of $\mathbb{U}$. This subset is called the *behavior* (of the model). Formally:

Definition 1.2.1 A *mathematical model* is a pair $(\mathbb{U}, \mathfrak{B})$ with $\mathbb{U}$ a set, called the *universum*—its elements are called *outcomes*—and $\mathfrak{B}$ a subset of $\mathbb{U}$, called the *behavior*. □

Example 1.2.2 During the ice age, shortly after Prometheus stole fire from the gods, man realized that H_2O could appear, depending on the temperature, as liquid water, steam, or ice. It took a while longer before this situation was captured in a mathematical model. The generally accepted model, with the temperature in degrees Celsius, is $\mathbb{U} = \{\text{ice, water, steam}\} \times [-273, \infty)$ and $\mathfrak{B} = ((\{\text{ice}\} \times [-273, 0]) \cup (\{\text{water }\} \times [0, 100]) \cup (\{\text{steam}\} \times [100, \infty))$. □

Example 1.2.3 Economists believe that there exists a relation between the amount P produced of a particular economic resource, the capital K invested in the necessary infrastructure, and the labor L expended towards its production. A typical model looks like $\mathbb{U} = \mathbb{R}^3_+$ and $\mathfrak{B} = \{(P, K, L) \in \mathbb{R}^3_+ \mid P = F(K, L)\}$, where $F : \mathbb{R}^2_+ \to \mathbb{R}_+$ is the *production function*. Typically, $F : (K, L) \mapsto \alpha K^\beta L^\gamma$, with $\alpha, \beta, \gamma \in \mathbb{R}_+$, $0 \leq \beta \leq 1$, $0 \leq \gamma \leq 1$, constant parameters depending on the production process, for example the type of technology used. Before we modeled the situation, we were ready to believe that every triple $(P, K, L) \in \mathbb{R}^3_+$ could occur. After introduction of the production function, we limit these possibilities to the triples satisfying $P = \alpha K^\beta L^\gamma$. The subset of $\mathbb{R}^3_+$ obtained this way is the behavior in the example under consideration.

□

1.2.2 Behavioral equations

In applications, models are often described by equations (see Example 1.2.3). Thus the behavior consists of those elements in the universum for which "balance" equations are satisfied.

Definition 1.2.4 Let $\mathbb{U}$ be a universum, $\mathbb{E}$ a set, and $f_1, f_2 : \mathbb{U} \to \mathbb{E}$. The mathematical model $(\mathbb{U}, \mathfrak{B})$ with $\mathfrak{B} = \{u \in \mathbb{U} \mid f_1(u) = f_2(u)\}$ is said to be described by *behavioral equations* and is denoted by $(\mathbb{U}, \mathbb{E}, f_1, f_2)$. The set $\mathbb{E}$ is called the *equating space*. We also call $(\mathbb{U}, \mathbb{E}, f_1, f_2)$ a *behavioral equation representation* of $(\mathbb{U}, \mathfrak{B})$. □

Often, an appropriate way of looking at $f_1(u) = f_2(u)$ is as *equilibrium conditions:* the behavior $\mathfrak{B}$ consists of those outcomes for which two (sets of) quantities are in balance.

Example 1.2.5 Consider an electrical resistor. We may view this as imposing a relation between the voltage V across the resistor and the current I through it. Ohm recognized more than a century ago that (for metal wires) the voltage is proportional to the current: $V = RI$, with the proportionality factor R called the resistance. This yields a mathematical model with universum $\mathbb{U} = \mathbb{R}^2$ and behavior $\mathfrak{B}$, induced by the behavioral equation $V = RI$. Here $\mathbb{E} = \mathbb{R}$, $f_1 : (V, I) \mapsto V$, and $f_2(V, I) : I \mapsto RI$. Thus $\mathfrak{B} = \{(I, V) \in \mathbb{R}^2 \mid V = RI\}$.
Of course, nowadays we know many devices imposing much more complicated relations between V and I, which we nevertheless choose to call (non-Ohmic) resistors. An example is an (ideal) diode, given by the (I, V) characteristic $\mathfrak{B} = \{(I, V) \in \mathbb{R}^2 \mid (V \geq 0 \text{ and } I = 0) \text{ or } (V = 0 \text{ and } I \leq 0)\}$. Other resistors may exhibit even more complex behavior, due to hysteresis, for example.

□

Example 1.2.6 Three hundred years ago, Sir Isaac Newton discovered (better: deduced from Kepler's laws since, as he put it, *Hypotheses non fingo*) that masses attract each other according to the inverse square law. Let us formalize what this says about the relation between the force F and the position vector q of the mass m. We assume that the other mass M is located at the origin of $\mathbb{R}^3$. The universum $\mathbb{U}$ consists of all conceivable force/position vectors, yielding $\mathbb{U} = \mathbb{R}^3 \times \mathbb{R}^3$. After Newton told us the behavioral equations $F = -k\frac{mMq}{\|q\|^3}$, we knew more: $\mathfrak{B} = \{(F, q) \in \mathbb{R}^3 \times$

$\mathbb{R}^3 \mid F = -k\frac{mMq}{\|q\|^3}\}$, with k the gravitational constant, $k = 6.67 \times 10^{-8}$ cm^3/g.sec^2. Note that $\mathfrak{B}$ has three degrees of freedom–down three from the six degrees of freedom in $\mathbb{U}$. □

In many applications models are described by *behavioral inequalities*. It is easy to accommodate this situation in our setup. Simply take in the above definition $\mathbb{E}$ to be an ordered space and consider the behavioral inequality $f_1(u) \leq f_2(u)$. Many models in operations research (e.g., in linear programming) and in economics are of this nature. In this book we will not pursue models described by inequalities.

Note further that whereas behavioral equations specify the behavior uniquely, the converse is obviously not true. Clearly, if $f_1(u) = f_2(u)$ is a set of behavioral equations for a certain phenomenon and if $f : \mathbb{E} \to \mathbb{E}'$ is any bijection, then the set of behavioral equations $(f \circ f_1)(u) = (f \circ f_2)(u)$ form another set of behavioral equations yielding the same mathematical model. Since we have a tendency to think of mathematical models in terms of behavioral equations, most models being presented in this form, it is important to emphasize their ancillary role: *it is the behavior, the solution set of the behavioral equations, not the behavioral equations themselves, that is the essential result of a modeling procedure.*

1.2.3 Latent variables

Our view of a mathematical model as expressed in Definition 1.2.1 is as follows: identify the outcomes of the phenomenon that we want to model (specify the universum $\mathbb{U}$) and identify the behavior (specify $\mathfrak{B} \subseteq \mathbb{U}$). However, in most modeling exercises we need to introduce other variables in addition to the attributes in $\mathbb{U}$ that we try to model. We call these other, auxiliary, variables *latent variables*. In a bit, we will give a series of instances where latent variables appear. Let us start with two concrete examples.

Example 1.2.7 Consider a one-port resistive electrical circuit. This consists of a graph with nodes and branches. Each of the branches contains a resistor, except one, which is an external port. An example is shown in Figure 1.1. Assume that we want to model the port behavior, the relation between the voltage drop across and the current through the external port. Introduce as auxiliary variables the voltages $(V_1, \ldots, V_5)$ across and the currents $(I_1, \ldots, I_5)$ through the internal branches, numbered in the obvious way as indicated in Figure 1.1. The following relations must be satisfied:

- *Kirchhoff's current law:* the sum of the currents entering each node must be zero;

- *Kirchhoff's voltage law:* the sum of the voltage drops across the branches of any loop must be zero;
- The *constitutive laws of the resistors* in the branches.

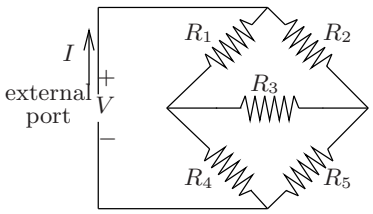


FIGURE 1.1. Electrical circuit with resistors only.

These yield:

Constitution laws *Kirchhoff's current laws* *Kirchhoff's voltage laws*

$$\begin{array}{lll} R_1I_1 = V_1, & I = I_1 + I_2, & V_1 + V_4 = V, \\ R_2I_2 = V_2, & I_1 = I_3 + I_4, & V_2 + V_5 = V, \\ R_3I_3 = V_3, & I_5 = I_2 + I_3, & V_1 + V_4 = V_2 + V_5, \\ R_4I_4 = V_4, & I = I_4 + I_5, & V_1 + V_3 = V_2, \\ R_5I_5 = V_5, & & V_3 + V_5 = V_4. \end{array}$$

Our basic purpose is to express the relation between the voltage across and current into the external port. In the above example, this is a relation of the form $V = RI$ (where R can be calculated from R_1, R_2, R_3, R_4, and R_5), obtained by eliminating $(V_1, \ldots, V_5, I_1, \ldots, I_5)$ from the above equations. However, the basic model, the one obtained from *first principles*, involves the variables $(V_1, \ldots, V_5, I_1, \ldots, I_5)$ in addition to the variables (V, I) whose behavior we are trying to describe. The node voltages and the currents through the internal branches (the variables $(V_1, \ldots, V_5, I_1, \ldots, I_5)$ in the above example) are thus latent variables. The port variables (V, I) are the manifest variables. The relation between I and V is obtained by eliminating the latent variables. How to do that in a systematic way is explained in Chapter 6. See also Exercise 6.1. □

Example 1.2.8 An economist is trying to figure out how much of a package of n economic goods will be produced. As a firm believer in equilibrium theory, our economist assumes that the production volumes consist of those points where, product for product, the supply equals the demand. This equilibrium set is a subset of $\mathbb{R}^n_+$. It is the behavior that we are looking for. In order to specify this set, we can proceed as follows. Introduce as latent variables the price, the supply, and the demand of each of the

n products. Next determine, using economic theory or experimentation, the supply and demand functions $S_i : \mathbb{R}^n_+ \to \mathbb{R}_+$ and $D_i : \mathbb{R}^n_+ \to \mathbb{R}_+$. Thus $S_i(p_1, p_2, \ldots, p_n)$ and $D_i(p_1, p_2, \ldots, p_n)$ are equal to the amount of product i that is bought and produced when the going market prices are $p_1, p_2, \ldots, p_n$. This yields the behavioral equations

$$\begin{aligned} s_i &= S_i(p_1, p_2, \ldots, p_n), \\ d_i &= D_i(p_1, p_2, \ldots, p_n), \\ s_i &= d_i = P_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

These behavioral equations describe the relation between the prices p_i, the supplies s_i, the demands d_i, and the production volumes P_i. The P_is for which these equations are solvable yield the desired behavior. Clearly, this behavior is most conveniently specified in terms of the above equations, that is, in terms of the behavior of the variables p_i, s_i, d_i, and $P_i (i = 1, 2, \ldots, n)$ jointly. The manifest behavioral equations would consist of an equation involving $P_1, P_2, \ldots, P_n$ only. □

These examples illustrate the following definition.

Definition 1.2.9 A *mathematical model with latent variables* is defined as a triple $(\mathbb{U}, \mathbb{U}_\ell, \mathfrak{B}_\mathrm{f})$ with $\mathbb{U}$ the *universum* of manifest variables, $\mathbb{U}_\ell$ the universum of *latent variables*, and $\mathfrak{B}_\mathrm{f} \subseteq \mathbb{U} \times \mathbb{U}_\ell$ the *full behavior*. It defines the *manifest mathematical model* $(\mathbb{U}, \mathfrak{B})$ with $\mathfrak{B} := \{u \in \mathbb{U} \mid \exists \ell \in \mathbb{U}_\ell \text{ such that } (u, \ell) \in \mathfrak{B}_\mathrm{f}\}$; $\mathfrak{B}$ is called the *manifest behavior* (or the *external behavior*) or simply the *behavior*. We call $(\mathbb{U}, \mathbb{U}_\ell, \mathfrak{B}_\mathrm{f})$ *a latent variable representation* of $(\mathbb{U}, \mathfrak{B})$. □

Note that in our framework we view the attributes in $\mathbb{U}$ as those variables that the model aims at describing. We think of these variables as *manifest*, as *external*. We think of the latent variables as auxiliary variables, as *internal*. In pondering about the difference between manifest variables and latent variables it is helpful *in the first instance* to think of the signal variables being directly *measurable;* they are *explicit*, while the latent variables are not: they are *implicit*, unobservable, or—better—only indirectly observable through the manifest variables. Examples: in pedagogy, scores of tests can be viewed as manifest, and native or emotional intelligence can be viewed as a latent variable aimed at explaining these scores. In thermodynamics, pressure, temperature, and volume can be viewed as manifest variables, while the internal energy and entropy can be viewed as latent variables. In economics, sales can be viewed as manifest, while consumer demand could be considered as a latent variable. We emphasize, however, that which variables are observed and measured through sensors, and which are not, is something that is really part of the instrumentation and the technological setup of a system. Particularly, in control applications

one should not be nonchalant about declaring certain variables measurable and observed. Therefore, we will not further encourage the point of view that identifies *manifest* with *observable*, and *latent* with *unobservable*.

Situations in which basic models use latent variables either for mathematical reasons or in order to express the basic laws occur very frequently. Let us mention a few: *internal voltages* and *currents* in electrical circuits in order to express the external port behavior; *momentum* in Hamiltonian mechanics in order to describe the evolution of the position; *internal energy* and *entropy* in thermodynamics in order to formulate laws restricting the evolution of the temperature and the exchange of heat and mechanical work; *prices* in economics in order to explain the production and exchange of economic goods; *state variables* in system theory in order to express the memory of a dynamical system; the *wave function* in quantum mechanics underlying observables; and finally, the *basic probability space* Ω in probability theory: the big latent variable space in the sky, our example of a latent variable space *par excellence*.

Latent variables invariably appear whenever we model a system by the method of *tearing* and *zooming*. The system is viewed as an interconnection of subsystems, and the modeling process is carried out by *zooming* in on the individual subsystems. The overall model is then obtained by combining the models of the subsystems with the interconnection constraints. This ultimate model invariably contains latent variables: the auxiliary variables introduced in order to express the interconnections play this role.

Of course, equations can also be used to express the full behavior $\mathfrak{B}_{\mathrm{f}}$ of a latent variable model (see Examples 1.2.7 and 1.2.8). We then speak of *full behavioral equations*.

1.3 Dynamical Systems

We now apply the ideas of Section 1.2 in order to set up a language for dynamical systems. The adjective *dynamical* refers to phenomena with a *delayed reaction*, phenomena with an *aftereffect*, with *transients*, *oscillations*, and, perhaps, an approach to *equilibrium*. In short, phenomena in which the *time evolution* is one of the crucial features. We view a dynamical system in the logical context of Definition 1.2.1 simply as a mathematical model, but a mathematical model in which the objects of interest are functions of time: the universum is a function space. We take the point of view that a dynamical system constrains the time signals that the system can conceivably produce. The collection of all the signals compatible with these laws defines what we call the *behavior* of the dynamical system. This yields the following definition.

1.3.1 The basic concept

Definition 1.3.1 A *dynamical system* Σ is defined as a triple

$$\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B}),$$

with $\mathbb{T}$ a subset of $\mathbb{R}$, called the *time axis*, $\mathbb{W}$ a set called the *signal space*, and $\mathfrak{B}$ a subset of $\mathbb{W}^{\mathbb{T}}$ called the *behavior* ($\mathbb{W}^{\mathbb{T}}$ is standard mathematical notation for the collection of all maps from $\mathbb{T}$ to $\mathbb{W}$). □

The above definition will be used as a *leitmotiv* throughout this book. The set $\mathbb{T}$ specifies the set of time instances relevant to our problem. Usually $\mathbb{T}$ equals $\mathbb{R}$ or $\mathbb{R}_+$ (in *continuous-time systems*), $\mathbb{Z}$ or $\mathbb{Z}_+$ (in *discrete-time systems*), or, more generally, an interval in $\mathbb{R}$ or $\mathbb{Z}$.

The set $\mathbb{W}$ specifies the way in which the outcomes of the signals produced by the dynamical system are formalized as elements of a set. These outcomes are the variables whose evolution in time we are describing. In what are called *lumped systems*, systems with a few well-defined simple components each with a finite number of degrees of freedom, $\mathbb{W}$ is usually a finite-dimensional vector space. Typical examples are electrical circuits and mass–spring–damper mechanical systems. In this book we consider almost exclusively lumped systems. They are of paramount importance in engineering, physics, and economics. In *distributed systems*, $\mathbb{W}$ is often an infinite-dimensional vector space. For example, the deformation of flexible bodies or the evolution of heat in media are typically described by partial differential equations that lead to an infinite-dimensional function space $\mathbb{W}$. In areas such as digital communication and computer science, signal spaces $\mathbb{W}$ that are finite sets play an important role. When $\mathbb{W}$ is a finite set, the term *discrete-event systems* is often used.

In Definition 1.3.1 the behavior $\mathfrak{B}$ is simply a family of time trajectories taking their values in the signal space. Thus elements of $\mathfrak{B}$ constitute precisely the trajectories compatible with the laws that govern the system: $\mathfrak{B}$ consists of all time signals which—according to the model—can conceivably occur, are compatible with the laws governing Σ, while those outside $\mathfrak{B}$ cannot occur, are prohibited. The behavior is hence the essential feature of a dynamical system.

Example 1.3.2 According to Kepler, the motion of planets in the solar system obeys three laws:

(K.1) planets move in elliptical orbits with the sun at one of the foci;

(K.2) the radius vector from the sun to the planet sweeps out equal areas in equal times;

(K.3) the square of the period of revolution is proportional to the third power of the major axis of the ellipse.

If a definition is to show proper respect and do justice to history, Kepler's laws should provide the very first example of a dynamical system. They do. Take $\mathbb{T} = \mathbb{R}$ (disregarding biblical considerations and modern cosmology: we assume that the planets have always been there, rotating, and will always rotate), $\mathbb{W} = \mathbb{R}^3$ (the position space of the planets), and $\mathfrak{B} = \{w : \mathbb{R} \to \mathbb{R}^3 \mid$ Kepler's laws are satisfied$\}$. Thus the behavior $\mathfrak{B}$ in this example consists of the *planetary motions* that, according to Kepler, are possible, all trajectories mapping the time-axis $\mathbb{R}$ into $\mathbb{R}^3$ that satisfy his three famous laws. Since for a given trajectory $w : \mathbb{R} \to \mathbb{R}^3$ one can unambiguously decide whether or not it satisfies Kepler's laws, $\mathfrak{B}$ is indeed well-defined. Kepler's laws form a beautiful example of a dynamical system in the sense of our definition, since it is one of the few instances in which $\mathfrak{B}$ can be described explicitly, and not indirectly through differential equations. It took no lesser man than Newton to think up appropriate behavioral differential equations for this dynamical system. □

Example 1.3.3 Let us consider the motion of a particle in a *potential field* subject to an external force. The purpose of the model is to relate the position q of the particle in $\mathbb{R}^3$ to the external force F. Thus $\mathbb{W}$, the signal space, equals $\mathbb{R}^3 \times \mathbb{R}^3$: three components for the position q, three for the force F. Let V: $\mathbb{R}^3 \to \mathbb{R}$ denote the potential field. Then the trajectories (q, F), which, according to the laws of mechanics, are possible, are those that satisfy the differential equation

$$m\frac{d^2q}{dt^2} + V'(q) = F,$$

where m denotes the mass of the particle and V' the gradient of V. Formalizing this model as a dynamical system yields $\mathbb{T} = \mathbb{R}$, $\mathbb{W} = \mathbb{R}^3 \times \mathbb{R}^3$, and $\mathfrak{B} = \{(q, F) \mid \mathbb{R} \to \mathbb{R}^3 \times \mathbb{R}^3 \mid m\frac{d^2q}{dt^2} + V'(q) = F\}$. □

1.3.2 Latent variables in dynamical systems

The definition of a *latent variable model* is easily generalized to dynamical systems.

Definition 1.3.4 A *dynamical system with latent variables* is defined as $\Sigma_L = (\mathbb{T}, \mathbb{W}, \mathbb{L}, \mathfrak{B}_\mathrm{f})$ with $\mathbb{T} \subseteq \mathbb{R}$ the *time-axis*, $\mathbb{W}$ the *(manifest) signal space*, $\mathbb{L}$ the *latent variable space*, and $\mathfrak{B}_\mathrm{f} \subseteq (\mathbb{W} \times \mathbb{L})^\mathbb{T}$ the *full behavior*. It defines a *latent variable representation* of the *manifest dynamical system* $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ with *(manifest) behavior* $\mathfrak{B} := \{w : \mathbb{T} \to \mathbb{W} \mid \exists\ \ell : \mathbb{T} \to \mathbb{L}$ such that $(w, \ell) \in \mathfrak{B}_\mathrm{f}\}$. □

Sometimes we will refer to the full behavior as the *internal* behavior and to the manifest behavior as the *external* behavior. Note that in a dynamical

system with latent variables each trajectory in the full behavior $\mathfrak{B}_{\mathrm{f}}$ consists of a pair (w, ℓ) with $w : \mathbb{T} \to \mathbb{W}$ and $\ell : \mathbb{T} \to \mathbb{L}$. The *manifest* signal w is the one that we are really interested in. The *latent variable* signal ℓ in a sense "supports" w. If $(w, \ell) \in \mathfrak{B}_{\mathrm{f}}$, then w is a possible manifest variable trajectory since ℓ can occur simultaneously with w.

Let us now look at two typical examples of how dynamical models are constructed from first principles. We will see that latent variables are unavoidably introduced in the process. Thus, whereas Definition 1.3.1 is a good concept as a basic notion of a dynamical system, typical models will involve additional variables to those whose behavior we wish to model.

Example 1.3.5 Our first example considers the port behavior of the electrical circuit shown in Figure 1.2. We assume that the elements R_C, R_L, L,

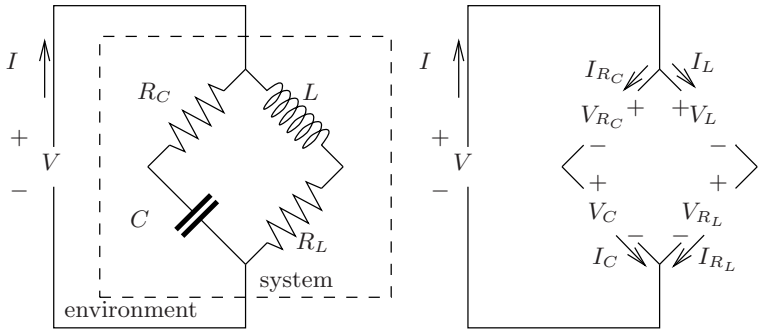


FIGURE 1.2. Electrical circuit.

and C all have positive values. The circuit interacts with its environment through the external port. The variables that describe this interaction are the current I into the circuit and the voltage V across its external terminals. These are the *manifest* variables. Hence $\mathbb{W} = \mathbb{R}^2$. As time- axis in this example we take $\mathbb{T} = \mathbb{R}$. In order to specify the port behavior, we introduce as auxiliary variables the currents through and the voltages across the internal branches of the circuit, as shown in Figure 1.2. These are the *latent variables*. Hence $\mathbb{L} = \mathbb{R}^8$.

The following equations specify the laws governing the dynamics of this circuit. They define the relations between the manifest variables (the port current and voltage) and the latent variables (the branch voltages and currents). These equations constitute the *full behavioral equations*.

Constitutive equations:

$$V_{R_C} = R_C I_{R_C}, \qquad V_{R_L} = R_L I_{R_L}, \qquad C\frac{dV_C}{dt} = I_C, \qquad L\frac{dI_L}{dt} = V_L; \tag{1.1}$$

Kirchhoff's current laws:

$$I = I_{R_C} + I_L, \qquad I_{R_C} = I_C, \qquad I_L = I_{R_L}, \qquad I_C + I_{R_L} = I; \tag{1.2}$$

Kirchhoff's voltage laws:

$$V = V_{R_C} + V_C, \qquad V = V_L + V_{R_L}, \qquad V_{R_C} + V_C = V_L + V_{R_L}. \tag{1.3}$$

In what sense do these equations specify a manifest behavior? In principle this is clear from Definition 1.3.4. But is there a more explicit way of describing the manifest behavior other than through (1.1, 1.2, 1.3)? Let us attempt to eliminate the latent variables in order to come up with an explicit relation involving V and I only. In the example at hand we will do this elimination in an ad hoc fashion. In Chapter 6, we will learn how to do it in a systematic way.

Note first that the constitutive equations (1.1) allow us to eliminate V_{R_C}, V_{R_L}, I_C, and V_L from equations (1.2, 1.3). These may hence be replaced by

$$I = I_{R_C} + I_L, \qquad I_{R_C} = C\frac{dV_C}{dt}, \qquad I_L = I_{R_L}, \tag{1.4}$$

$$V = R_C I_{R_C} + V_C, \qquad V = L\frac{dI_L}{dt} + R_L I_{R_L}. \tag{1.5}$$

Note that we have also dropped the equations $I_C + I_{R_L} = I$ and $V_{R_C} + V_C = V_L + V_{R_L}$, since these are obviously redundant. Next, use $I_{R_L} = I_L$ and $I_{R_C} = \frac{V - V_C}{R_C}$ to eliminate I_{R_L} and I_{R_C} from (1.4) and (1.5) to obtain

$$R_L I_L + L\frac{dI_L}{dt} = V, \tag{1.6}$$

$$V_C + CR_C\frac{dV_C}{dt} = V, \tag{1.7}$$

$$I = \frac{V - V_C}{R_C} + I_L. \tag{1.8}$$

We should still eliminate I_L and V_C from equations (1.6, 1.7, 1.8) in order to come up with an equation that contains only the variables V and I. Use equation (1.8) in (1.6) to obtain

$$V_C + \frac{L}{R_L}\frac{dV_C}{dt} = (1 + \frac{R_C}{R_L})V + \frac{L}{R_L}\frac{dV}{dt} - R_c I - \frac{LR_C}{R_L}\frac{dI}{dt}, \tag{1.9}$$

$$V_C + CR_C\frac{dV_C}{dt} = V. \tag{1.10}$$

Next, divide (1.9) by $\frac{L}{R_L}$ and (1.10) by CR_C, and subtract. This yields

$$(\frac{R_L}{L} - \frac{1}{CR_C})V_C = (\frac{R_C}{L} + \frac{R_L}{L} - \frac{1}{CR_C})V + \frac{dV}{dt} - \frac{R_C R_L}{L}I - R_C\frac{dI}{dt}. \tag{1.11}$$

Now it becomes necessary to consider two cases:

Case 1: $CR_C \neq \frac{L}{R_L}$. Solve (1.11) for V_C and substitute into (1.10). This yields, after some rearranging,

$$(\frac{R_C}{R_L} + (1 + \frac{R_C}{R_L})CR_C\frac{d}{dt} + CR_C\frac{L}{R_L}\frac{d^2}{dt^2})V = (1 + CR_C\frac{d}{dt})(1 + \frac{L}{R_L}\frac{d}{dt})R_C I \tag{1.12}$$

as the relation between V and I.

Case 2: $CR_C = \frac{L}{R_L}$. Then (1.11) immediately yields

$$(\frac{R_C}{R_L} + CR_C\frac{d}{dt})V = (1 + CR_C\frac{d}{dt})R_C I \tag{1.13}$$

as the relation between V and I. We claim that equations (1.12, 1.13) specify the manifest behavior defined by the full behavioral equations (1.1, 1.2, 1.3). Indeed, our derivation shows that (1.1, 1.2, 1.3) imply (1.12, 1.13). But we should also show the converse. We do not enter into the details here, although in the case at hand it is easy to prove that (1.12, 1.13) imply (1.1, 1.2, 1.3). This issue will be discussed in full generality in Chapter 6.

This example illustrates a number of issues that are important in the sequel. In particular:

1. The full behavioral equations (1.1, 1.2, 1.3) are all linear differential equations. (Note: we consider algebraic relations as differential equations of order zero). The manifest behavior, it turns out, is also described by a linear differential equation, (1.12) or (1.13). A coincidence? Not really: in Chapter 6 we will learn that this is the case in general.

2. The differential equation describing the manifest behavior is (1.12) when $CR_C \neq \frac{L}{R_L}$. This is an equation of order two. When $CR_C = \frac{L}{R_L}$, however, it is given by (1.13), which is of order one. Thus the order of the differential equation describing the manifest behavior turns out to be a sensitive function of the values of the circuit elements.

3. We need to give an interpretation to the anomalous case $CR_C = \frac{L}{R_L}$, in the sense that for these values a discontinuity appears in the manifest behavioral equations. This interpretation, it turns out, is *observability*, which will be discussed in Chapter 5. □

Example 1.3.6 As a second example for the occurrence of latent variables, let us consider a Leontieff model for an economy in which several economic goods are transformed by means of a number of production processes. We are interested in describing the evolution in time of the total utility of the goods in the economy. Assume that there are N production processes in which n economic goods are transformed into goods of the same kind, and that in order to produce one unit of good j by means of the k th production process, we need at least a_{ij}^k units of good i. The real numbers a_{ij}^k, $k \in \underline{N} :=$

$\{1, 2, \ldots, N\}$, $i, j \in \underline{n} := \{1, 2, \ldots, n\}$, are called the *technology coefficients*. We assume that in each time unit one production cycle takes place.

Denote by

$q_i(t)$	the quantity of product i available at time t
$u_i^k(t)$	the quantity of product i assigned to the production process k at time t,
$y_i^k(t)$	the quantity of product i acquired from the production process k at time t.

Then the following hold:

$$\begin{aligned} \sum_{k=1}^{n} u_i^k(t) &\leq q_i(t) \qquad \forall i \in \underline{n}, \\ \sum_{i=1}^{n} a_{ij}^k u_i^k(t) &\geq y_j^k(t+1) \quad \forall k \in \underline{N}, i \in \underline{n}, \\ q_i(t) &\leq \sum_{k=1}^{n} y_i^k(t) \quad \forall i \in \underline{n}. \end{aligned} \tag{1.14}$$

The underlying structure of the economy is shown in Figure 1.3. The differences between the right-hand and left-hand sides of the above inequalities are due to such things as inefficient production, imbalance of the available products, consumption, and other forms of waste. Now assume that the to-

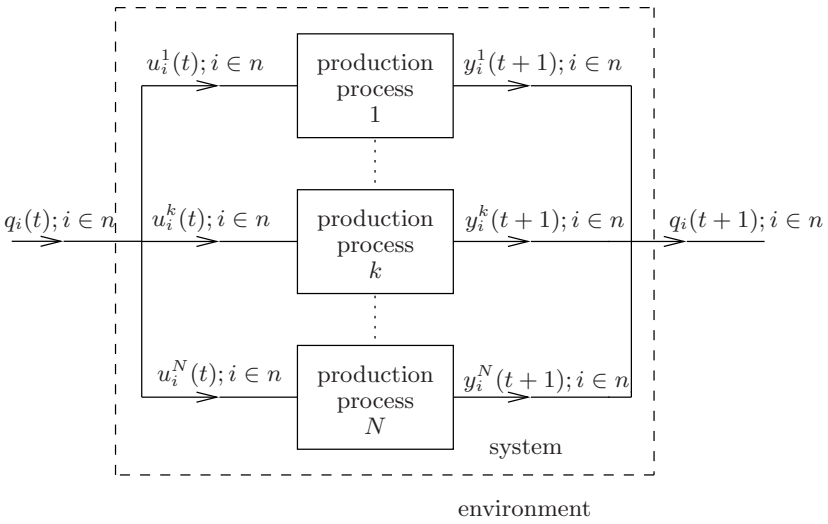


FIGURE 1.3. Leontieff economy.

tal utility of the goods in the economy is a function of the available amount of goods $q_1, q_2, \ldots, q_n$; i.e., $J : \mathbb{Z} \to \mathbb{R}_+$ is given by

$$J(t) \;=\; \eta(q_1(t), \ldots, q_n(t)),$$

with $\eta : \mathbb{R}_+^n \to \mathbb{R}_+$ a given function, the *utility*. For example, if we identify utility with resale value (in dollars, say), then $\eta(q_1, q_2, \ldots, q_n)$ is equal to $\sum_{k=1}^n p_i q_i$ with p_i the per unit selling price of good i.

How does this example fit in our modeling philosophy?

The first question to ask is, *What is the time-set?* It is reasonable to take $\mathbb{T} = \mathbb{Z}$. This does not mean that we believe that the products have always existed and that the factories in question are blessed with life eternal. What instead it says is that for the purposes of our analysis it is reasonable to assume that the production cycles have already taken place very many times before and that we expect very many more production cycles to come.

The second question is, What are we trying to model? *What is our signal space?* As die-hard utilitarians we decide that all we care about is the total utility J, whence $\mathbb{W} = \mathbb{R}_+$.

The third question is, *How is the behavior defined?* This is done by inequalities (1.14). Observe that these inequalities involve, in addition to the manifest variable J, as latent variables the us, qs, and ys. Hence $\mathbb{L} = \mathbb{R}_+^n \times \mathbb{R}_+^{n\times m} \times \mathbb{R}_+^{n\times p}$.

The full behavior is now defined as consisting of those trajectories satisfying the behavioral difference inequalities (1.14). These relations define the intrinsic dynamical system with $\mathbb{T} = \mathbb{Z}$, $\mathbb{W} = \mathbb{R}_+$, and the manifest behavior $\mathfrak{B} = \{J : \mathbb{Z} \to \mathbb{R}_+ \mid \exists\; q_i : \mathbb{Z} \to \mathbb{R}_+,\; u_i^k : \mathbb{Z} \to \mathbb{R}_+,\; y_i^k : \mathbb{Z} \to \mathbb{R}_+,\; i \in \underline{n}, k \in \underline{N}$, such that the inequalities (1.14) are satisfied for all $t \in \mathbb{Z}\}$. Note that in contrast to the previous example, where it was reasonably easy to obtain behavioral equations (1.12) or (1.13) explicitly in terms of the external attributes V and I, it appears impossible in the present example to eliminate the qs, us, and ys and obtain an explicit behavioral equation (or, more likely, inequality) describing $\mathfrak{B}$ entirely in terms of the J, the variables of interest in this example. □

1.4 Linearity and Time-Invariance

Until now we have discussed dynamical systems purely on a set-theoretic level. In order to obtain a workable theory it is necessary to impose more structure. Of particular importance in applications are linearity and time-invariance. These notions are now introduced.

Definition 1.4.1 A dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ is said to be *linear* if $\mathbb{W}$ is a vector space (over a field $\mathbb{F}$: for the purposes of this book, think of it as $\mathbb{R}$ or $\mathbb{C}$), and $\mathfrak{B}$ is a linear subspace of $\mathbb{W}^{\mathbb{T}}$ (which is a vector space in the obvious way by pointwise addition and multiplication by a scalar). □

Thus linear systems obey the *superposition principle* in its ultimate and very simplest form: $\{w_1(\cdot), w_2(\cdot) \in \mathfrak{B}; \alpha, \beta \in \mathbb{F}\} \Rightarrow \{\alpha w_1(\cdot) + \beta w_2(\cdot) \in \mathfrak{B}\}$.

Time-invariance is a property of dynamical systems governed by laws that do not explicitly depend on time: if one trajectory is *legal* (that is, in the behavior), then the shifted trajectory is also *legal*.

Definition 1.4.2 A dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ with $\mathbb{T} = \mathbb{Z}$ or $\mathbb{R}$ is said to be *time-invariant* if $\sigma^t\mathfrak{B} = \mathfrak{B}$ for all $t \in \mathbb{T}$ (σ^t denotes the *backward t-shift*: $(\sigma^t f)(t') := f(t' + t)$). If $\mathbb{T} = \mathbb{Z}$, then this condition is equivalent to $\sigma\mathfrak{B} = \mathfrak{B}$. If $\mathbb{T} = \mathbb{Z}_+$ or $\mathbb{R}_+$, then time-invariance requires $\sigma^t\mathfrak{B} \subseteq \mathfrak{B}$ for all $t \in \mathbb{T}$. In this book we will almost exclusively deal with $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$, and therefore we may as well think of time-invariance as $\sigma^t\mathfrak{B} = \mathfrak{B}$. The condition $\sigma^t\mathfrak{B} = \mathfrak{B}$ is called *shift-invariance* of $\mathfrak{B}$. □

Essentially all the examples that we have seen up to now are examples of time-invariant systems.

Example 1.4.3 As an example of a time-varying system, consider the motion of a point-mass with a time-varying mass $m(\cdot)$, for example, a burning rocket. The differential equation describing this motion is given by

$$\frac{d}{dt}(m(t)\frac{d}{dt}q) = F.$$

If we view this as a model for the manifest variables $(q, F) \in \mathbb{R}^3 \times \mathbb{R}^3$, then the resulting dynamical system is linear but time-varying. If we view this as a model for the manifest variables $(q, F, m) \in \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}_+$, then the resulting dynamical system is time-invariant but nonlinear (see Exercise 1.5). □

The notion of linearity and time-invariance can in an obvious way be extended to latent variable dynamical systems. We will not explicitly write down these formal definitions.

1.5 Dynamical Behavioral Equations

In most models encountered in applications, the behavior is described through equations. The behavior, a subset of a universum, is then simply

defined as those elements of this universum satisfying a set of equations, called *behavioral equations*. In dynamical systems these behavioral equations often take the form of differential equations or of integral equations. All of our examples have been of this type. Correction: all except Kepler's laws, Example 1.3.2, where the behavior was described explicitly, although even there one could associate equations to K.1, K.2, and K.3.

We now formalize this. We describe first the ideas in terms of difference equations, since they involve the fewest difficulties of a technical mathematical nature. A *behavioral difference equation* representation of a discrete-time dynamical system with time-axis $\mathbb{T} = \mathbb{Z}$ and signal space $\mathbb{W}$ is defined by a nonnegative integer L (called the *lag*, or the *order* of the difference equation), a set $\mathbb{E}$ (called the *equating space*), and two maps $f_1, f_2 : \mathbb{W}^{L+1} \to \mathbb{E}$, yielding the difference equations

$$f_1(w, \sigma w, \ldots, \sigma^{L-1} w, \sigma^L w) = f_2(w, \sigma w, \ldots, \sigma^{L-1} w, \sigma^L w).$$

Note that this is nothing more than a compact way of writing the difference equation

$$f_1(w(t), w(t+1), \ldots, w(t+L)) = f_2(w(t), w(t+1), \ldots, w(t+L)).$$

These equations define the behavior by

$$\mathfrak{B}=\{w : \mathbb{Z} \to \mathbb{W} \mid f_1(w, \sigma w, \ldots, \sigma^{L-1} w, \sigma^L w) = f_2(w, \sigma w, \ldots, \sigma^{L-1} w, \sigma^L w)\}.$$

In many applications it is logical to consider difference equations that have both positive and negative lags, yielding the behavioral equations

$$f_1(\sigma^{L_{\min}} w, \sigma^{L_{\min}+1} w, \ldots, \sigma^{L_{\max}} w) = f_2(\sigma^{L_{\min}} w, \sigma^{L_{\min}+1} w, \ldots, \sigma^{L_{\max}} w). \tag{1.15}$$

We call $L_{\max} - L_{\min}$ the *lag* of this difference equation. We assume $L_{\max} \geq L_{\min}$, but either or both of them could be negative. Whether forward lags (powers of σ) or backward lags (powers of σ^{-1}) are used is much a matter of tradition. In control theory, forward lags are common, but econometricians like backward lags. The behavior defined by (1.15) is

$$\mathfrak{B} = \{w : \mathbb{Z} \to \mathbb{W} \mid f_1(\sigma^{L_{\min}} w, \ldots, \sigma^{L_{\max}} w) = f_2(\sigma^{L_{\min}} w, \ldots, \sigma^{L_{\max}} w)\}.$$

It is clear that the system obtained this way defines a time-invariant dynamical system.

Example 1.5.1 As a simple example, consider the following algorithm for computing the *moving average* of a time-series:

$$a(t) = \sum_{k=-T}^{T} \alpha_k s(t+k),$$

where the α_ks are nonnegative weighting coefficients that sum to 1. In the above equation $a(t)$ denotes the (real-valued) moving average at time t, and s denotes the time-series of which the moving average is taken. This equation can easily be cast in the form (1.15) by taking $L_{\min} = -T, L_{\max} = T$, and defining f_1 and f_2 in the obvious way. This example shows that it is often convenient to use both negative and positive lags. □

The continuous-time analogue of a behavioral difference equation is a *behavioral differential equation*. Let $\mathbb{T} = \mathbb{R}$, and assume that the signal space $\mathbb{W}$ is $\mathbb{R}^q$. Let L be a nonnegative integer (called the *order* of the differential equation), $\mathbb{E}$ a set (called the *equating space*), and $f_1, f_2 : (\mathbb{R}^q)^{L+1} \to \mathbb{E}$ two maps. Consider the differential equation

$$f_1(w, \frac{dw}{dt}, \ldots, \frac{d^{L-1}w}{dt^{L-1}}, \frac{d^Lw}{dt^L}) = f_2(w, \frac{dw}{dt}, \ldots, \frac{d^{L-1}}{dt^{L-1}}, \frac{d^Lw}{dt^L}). \qquad (1.16)$$

This differential equation intuitively describes the dynamical system $\Sigma = (\mathbb{T}, \mathbb{R}^q, \mathfrak{B})$ with $\mathbb{T} = \mathbb{R}$ and $\mathfrak{B}$ the collection of all time-functions $w(\cdot) : \mathbb{T} \to \mathbb{R}^q$ such that this differential equation is satisfied. Intuitively, it is clear what this means. But what is the precise mathematical significance? *What does it mean that $w(\cdot)$ satisfies this differential equation?* It turns out that we must let the precise meaning depend to some extent on the context, and hence we will not enter into details now. We will belabor this point in Chapter 2 in the context of linear differential equations.

If one looks around at the mathematical models used in areas such as physics, engineering, economics, and biology, then one is struck by the prevalence of models that use the language of differential (and difference) equations. Indeed, all the examples of continuous-time dynamical systems that we have seen up to now were in the form of behavioral differential equations, and as Newton showed, even Kepler's laws can be cast as the solution set of an appropriate second-order differential equation involving the inverse square law of gravitation. So we are led to ponder the question, *What is so special about differential equation models? Why are they so common?* It is not easy to give a brief and convincing answer to this. An important property is that the behavior defined by differential equations is *locally specified*. This means the following. Let $\Sigma = (\mathbb{R}, \mathbb{W}, \mathfrak{B})$ be a time-invariant dynamical system. Define, from $\mathfrak{B}$, the behavior restricted to a small time interval $(-\epsilon, \epsilon)$ as follows:

$$\mathfrak{B}_\epsilon := \{\tilde{w} : (-\epsilon, \epsilon) \to \mathbb{W} \mid \exists w \in \mathfrak{B} \text{ such that } \tilde{w}(t) = w(t) \text{for} - \epsilon < t < \epsilon\}.$$

We call Σ *locally specified* if for all $\epsilon > 0$,

$$(w \in \mathfrak{B}) \Leftrightarrow ((\sigma^t w) \mid_{(-\epsilon,\epsilon)} \in \mathfrak{B}_\epsilon \text{ for all } t \in \mathbb{R}).$$

It is easy to see that a system described by behavioral differential equations is locally specified. In other words, w is *legal* if and only if all its restrictions

to any arbitrarily small time interval look *legal*. This is a crucial property of behaviors described by differential equations. In our context, it holds for systems described by ordinary differential equations with time as the independent variable, but more generally, a similar property of "locally specified" holds for partial differential equation models. The fact that the behavior of models described by differential equations has this property of being *locally specified* explains their prevalence: in time, there is no *action at a distance*. In order to verify that the trajectory w belongs to the behavior, it suffices to examine what the trajectory looks like in the immediate neighborhood of each point. Of course, many useful models do not exhibit this property: witness systems described by differential-delay equations. As an example of a nonlocally-specified behavior, consider the simplistic growth model described by the following equation:

$$\frac{dw}{dt} = \alpha\sigma^{-\Delta}w \qquad (\text{i.e.,} \frac{dw}{dt}(t) = \alpha w(t-\Delta))$$

expressing that the growth of a population is proportional to the size of the population Δ time units ago.

Of course, latent variable models are also often described by differential or difference equations. In the case of differential equations this leads to behavioral equations of the form

$$f_1(w, \ldots, \frac{d^L w}{dt^L}, \ell, \ldots, \frac{d^L \ell}{dt^L}) = f_2(w, \ldots, \frac{d^L w}{dt^L}, \ell, \ldots, \frac{d^L \ell}{dt^L}). \tag{1.17}$$

In the next chapters we will study *linear* systems described by differential equations such as (1.16) and (1.17) in much detail. One of the first issues that need to be considered, of course, is, *What exactly is meant by a solution?*

1.6 Recapitulation

In this chapter we have introduced some basic mathematical language and concepts that are used throughout this book. We started by discussing completely general models, but soon specialized to dynamical systems, that is, to phenomena in which the *time evolution* is of central importance.

The basic notions introduced were the following:

- A *mathematical model*, which we viewed as being defined by a subset $\mathfrak{B}$, called the *behavior*, of a universum $\mathbb{U}$ (Definition 1.2.1).
- *Behavioral equations*, which serve to specify $\mathfrak{B}$ as the set of solutions of a system of equations (Definition 1.2.4).
- *Manifest* and *latent variables*. The manifest variables are those whose behavior the model aims at describing. The latent variables are auxiliary

variables introduced in the modeling process. First principle models are typically given by equations involving *both* manifest *and* latent variables. We call these equations *full behavioral equations* (Definition 1.2.9).

- A *dynamical system* is a mathematical model for a phenomenon that evolves over time. A *dynamical system* is defined by three sets: the *time-axis* $\mathbb{T}$, a subset of $\mathbb{R}$ consisting of the relevant time instances; the *signal space* $\mathbb{W}$, the set in which the time trajectories take on their values; and the *behavior* $\mathfrak{B}$, a subset of $\mathbb{W}^{\mathbb{T}}$ consisting of all trajectories $\mathbb{W} : \mathbb{T} \to \mathbb{W}$ that according to the model can occur. Thus a dynamical system is defined as a triple $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$ (Definition 1.3.1).
- Just as was the case for general models, first principle dynamical models typically involve *latent*, in addition to *manifest*, variables (Definition 1.3.4).
- Important properties of dynamical systems are *linearity* and *time-invariance*. A linear dynamical system is one for which the *superposition principle* holds. In a time-invariant dynamical system the laws do not depend explicitly on time. Its behavior is shift-invariant (Section 1.4).
- Dynamical systems are often described by behavioral equations that are *differential* or *difference* equations. The behavior consists of the solution set of these equations. Systems described by differential equations are locally specified (Section 1.5).

1.7 Notes and References

The modeling language described in this chapter has been developed in [56, 57, 55, 58, 59, 60]. There are numerous books on mathematical modeling, but none of them seem to have come to the elegant mathematical formalization and notions that we put forward here. However, models have been and will be used very effectively without a formalized mathematical setting, and most books on modeling use a learn-while-you-do philosophy. A book with nice examples of mathematical models from a variety of disciplines is [38].

1.8 Exercises

As simulation exercises illustrating the material covered in this chapter we suggest A.1 and A.2.

1.1 Model the external port behavior of the resistive circuit shown in Figure 1.4 using latent variables and Kirchhoff's laws. Eliminate the latent variables and obtain a behavioral equation for the manifest behavior. Call two resistive circuits *equivalent* if they have the same manifest behavior. For what values of R_1, R_2, and R_3 are the circuits (1), (2), and (3) shown in Figure 1.4 equivalent?

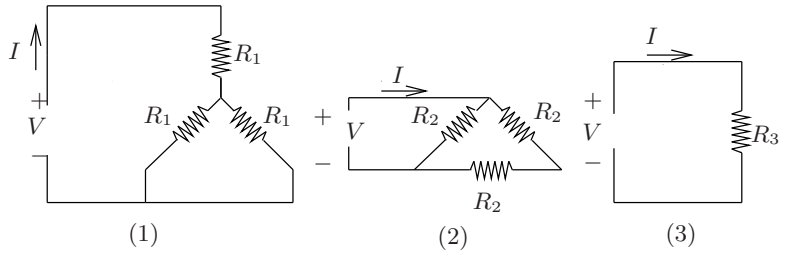


FIGURE 1.4. Resistive circuits.

1.2 Consider a linear mathematical model $(\mathbb{R}^q, \mathfrak{B})$. Let $p := q - \dim\ \mathfrak{B}$. Prove that $\mathfrak{B}$ is the behavior of a linear model if and only if there exists a full row rank matrix $R \in \mathbb{R}^{p\times q}$ such that $\mathfrak{B}$ is described by the behavioral equations

$$Rw = 0. \tag{1.18}$$

Similarly, prove that $\mathfrak{B}$ is a linear model if and only if there exists a full column rank matrix $M \in \mathbb{R}^{q\times(q-p)}$ such that $\mathfrak{B}$ is the manifest behavior of the latent variable model

$$w = M\ell \tag{1.19}$$

with $\ell \in \mathbb{R}^{q-p}$ latent variables. It is natural to call (1.18) a *kernel representation* of the mathematical model $(\mathbb{R}^n, \mathfrak{B})$ and (1.19) an *image representation*. Why?

1.3 Consider the pinned elastic beam shown below in Figure 1.5. We want to

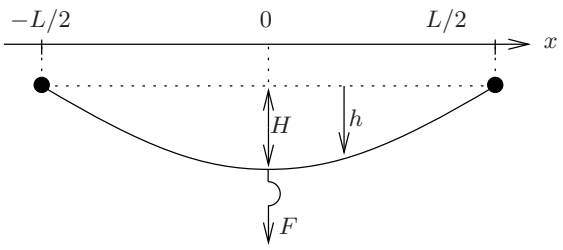


FIGURE 1.5. Elastic beam.

describe the static relation between the force F applied at the center of the beam and its displacement H. We expect this relationship to be like that of a spring $H = \alpha F$. How do we demonstrate that, and how do we determine α?

Elasticity theory yields the following relations describing the deflection h of the beam:

$$\frac{d^2}{dx^2}\left(EI(x)\frac{d^2h}{dx^2}(x)\right) = 0, \qquad 0 < |x| \le \frac{L}{2}, \tag{1.20}$$

with the boundary conditions

$$h(-\frac{L}{2}) = h(\frac{L}{2}) = 0, \qquad \frac{d^2h}{dx^2}(-\frac{L}{2}) = \frac{d^2h}{dx^2}(\frac{L}{2}) = 0 \tag{1.21}$$

and the loading conditions

$$h, \frac{dh}{dx}, \frac{d^2h}{dx^2} \quad \text{continuous at } x = 0, \tag{1.22}$$

$$\frac{d^3h}{dx^3}(0^+) = \frac{d^3h}{dx^3}(0^-) + \frac{F}{EI(0)}, \tag{1.23}$$

where E denotes the modulus of elasticity of the material of the beam—it is a parameter depending on the material—and $I(x)$ is the area moment of the cross section of the beam at x. This is a function of the geometry, in particular of the thickness of the beam at x. It is not essential for our purposes to understand how these equations are arrived at.

How can we view equations (1.20–1.23) *as defining a mathematical model for the relation between F and H?* The universum is $\mathbb{R}^2$: a priori all pairs (F, H) are conceivable. In order to specify the behavior, introduce as latent variables the deflection $h(x)$ for $-\frac{L}{2} \leq x \leq \frac{L}{2}$. Now define the full behavior

$$\begin{aligned}\mathfrak{B}_{\mathrm{f}} = \{(F, H, h) \mid (F, H) \in \mathbb{R}^2,\ h \in \mathcal{C}^2([-\tfrac{L}{2}, \tfrac{L}{2}],\ \mathbb{R}),\\ \text{(1.20–1.23) are satisfied and } H = h(0)\},\end{aligned}$$

from which we can derive the manifest behavior

$$\mathfrak{B} = \{(F, H) \in \mathbb{R}^2 \mid \exists\, h : \left[-\frac{L}{2}, \frac{L}{2}\right] \to \mathbb{R} \text{ such that } (F, H, h) \in \mathfrak{B}_{\mathrm{f}}\}.$$

For many purposes the model $(\mathbb{W}, \mathfrak{B})$ in this implicit form (with $h(\cdot)$ not eliminated from the equations) is an adequate one. In fact, when the area moment I depends on x, this implicit model may be the most explicit expression for the relation between F and H that we can hope to derive. Hence elimination of the latent variables may be next to impossible to achieve. It is possible to prove that $\mathfrak{B}$ is given by $\mathfrak{B} = \{(F, H) \in \mathbb{R} \mid H = \alpha F\}$ for some suitable constant $\alpha \in \mathbb{R}$.

Prove that this defines a linear latent variable model. Prove that there exists an $\alpha \in \mathbb{R}$ such that $\mathfrak{B} = \{(F, H) \in \mathbb{R}^2 \mid H = \alpha F\}$. Assume that $I(x)$ is independent of x. Prove that $\alpha = \frac{L^3}{48EI}$.

1.4 Consider a continuous-time dynamical system with $\mathbb{T} = \mathbb{R}$, signal space $\mathbb{W} = \mathbb{R}$, and behavior consisting of all sinusoidal signals with period 2π. In other words, $w : \mathbb{R} \to \mathbb{R}$ is assumed to belong to $\mathfrak{B}$ if and only if there exist $A \in \mathbb{R}_+$ and $\varphi \in [0, 2\pi)$ such that $w(t) = A \sin(t + \varphi)$.

(i) Is this dynamical system linear?

(ii) Time-invariant?

(iii) Is the differential equation $w + \frac{d^2w}{dt^2} = 0$ a behavioral equation for it?

1.5 Consider the dynamical system relating the position $q \in \mathbb{R}^3$ of a burning rocket with mass $m : \mathbb{R} \to \mathbb{R}_+$ under the influence of an external force $F \in \mathbb{R}^3$. The equation of motion is given by

$$\frac{d}{dt}(m(t)\frac{dq}{dt}) = F. \tag{1.24}$$

Prove that this system, viewed as relating q and F (with $m(.)$ as a "parameter") is linear but time-varying, whereas if you view it as relating q, F, and m, then it is time-invariant but nonlinear. Complete this model with an equation explaining the relation between F, q, and m, for example,

$$\frac{dm}{dt} = \alpha F \frac{dq}{dt}. \tag{1.25}$$

Here $\alpha \in \mathbb{R}^+$ is a parameter. Give a reasonable physical explanation of (1.25) in terms of power and energy. View (1.24,1.25) as a model in the variables q, F, and m. Is it time-invariant? Linear?

1.6 Consider the pendulum shown in Figure 1.6. Assume that we want to model the relationship between the position w_1 of the mass and the position w_2 of the tip of the pendulum (say with the ultimate goal of designing a controller that stabilizes w_1 at a fixed value by using w_2 as control, as we do when we balance an inverted broom on our hand). In order to obtain such a model, introduce as auxiliary variables the force F in the bar and the real-valued proportionality factor α of F and $w_1 - w_2$. We obtain the

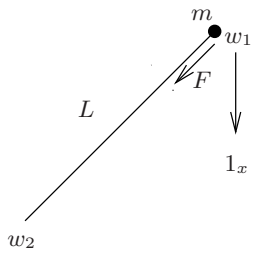


FIGURE 1.6. A pendulum.

behavioral equations

$$\begin{aligned} m\frac{d^2 w_1}{dt^2} &= mgw_z + F, \\ \|w_1 - w_2\| &= L, \\ F &= \alpha(w_1 - w_2). \end{aligned}$$

Here m denotes the mass of the pendulum, L its length, g the gravitational constant, and $\vec{1}_z$ the unit vector in the z-direction.

Define formally the full and the manifest behavior.

1.7 Let $\Sigma_L = (\mathbb{Z}, \mathbb{W}, \mathbb{L}, \mathfrak{B}_{\mathrm{f}})$ be a latent variable dynamical system and $\Sigma = (\mathbb{Z}, \mathbb{W}, \mathfrak{B})$ the manifest dynamical system induced by it. Prove that Σ is linear if Σ_L is. Prove that Σ is time-invariant if Σ_L is.

1.8 Consider the time-invariant dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$. How would you define an *equilibrium*, that is, a *static motion*? Formalize the family of static motions as a dynamical system Σ^{stat}. Assume that $\mathfrak{B}$ is described by a difference or a differential equation. Give the equations describing $\mathfrak{B}^{\mathrm{stat}}$.

1.9 Consider the electrical circuit shown in Figure 1.7. Assume that the values

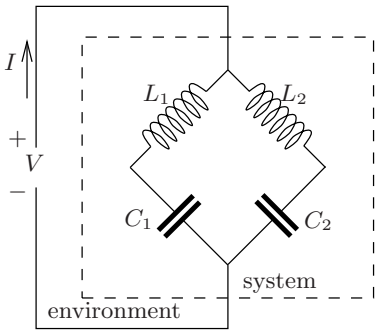


FIGURE 1.7. Electrical circuit.

of the elements L_1, L_2, C_1, and C_2 are all positive. The circuit interacts with its environment through the external port. The variables that describe this interaction are the current I into the circuit and the voltage V across its external terminals. These are the *manifest* variables. In order to specify the terminal behavior, introduce as auxiliary variables the currents through and the voltages across the internal branches of the circuit. These are the *latent variables*. Follow the ideas set forth in Example 1.3.5 in order to come up with a differential equation describing the manifest behavior.

1.10 Consider the electrical circuit shown in Figure 1.8 Assume that we want to model the relation between the switch position and the voltage across the capacitor C. The voltage source gives a constant voltage V. Assume that the switch position is modeled as follows:

$$s(t) = \begin{cases} 1 & \text{if the switch is closed,} \\ 0 & \text{if the switch is open.} \end{cases}$$

Formalize this as a mathematical model. Specify clearly the sets $\mathbb{T}, \mathbb{W}, \mathfrak{B}$ and, if needed, $\mathbb{L}$ and $\mathfrak{B}_f$.

1.11 Consider the mechanical system shown in Figure 1.9.

A spring exerts a force that is a function of its extension. A damper exerts a force that is a function of the velocity of the piston. Assume that the spring and the damper are both linear. We want to describe the relation

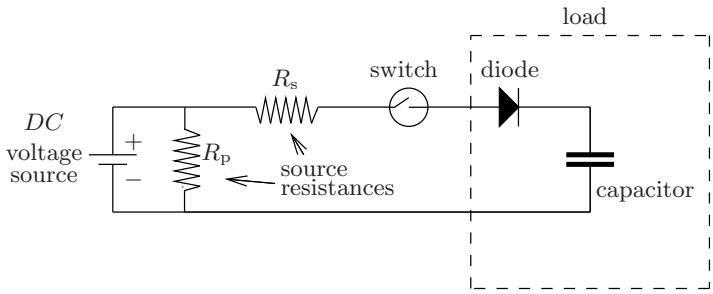


FIGURE 1.8. Electrical circuit with nonlinear elements.

between the external force F and the position q of the mass. Give the differential equation relating F and q. Define this carefully as a dynamical system. Assume instead that you want to study the relation between the force and the internal energy of this mechanical system. How would you now formalize this as a dynamical system? Repeat this for the relation between the force and the heat produced in the damper. Are these latter two dynamical systems linear? Time-invariant?

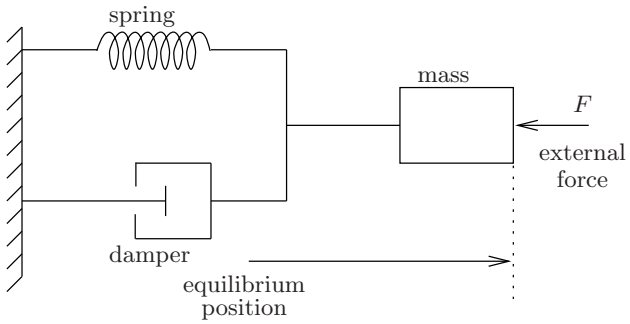


FIGURE 1.9. Mass–spring–damper system.

1.12 Consider the linear mechanical system shown in Figure 1.10. Use the equations derived in the previous exercise to model the relation between F_1 and q_1, and F_2 and q_2. Now hook the two masses together. Argue that this comes down to imposing the additional equations $F_1 = F_2$ and $q_1 + q_2 = \Delta$, with $\Delta > 0$ a fixed constant. Define a new equilibrium position for the first mass. Write a differential equation describing the behavior of q_1', the position of the first mass measured from this new equilibrium.

1.13 Consider the time-invariant dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}, \mathfrak{B})$. Define it to be *time-reversible* if $(w \in \mathfrak{B}) \Leftrightarrow (\text{rev}(w) \in \mathfrak{B})$ with the map rev defined by

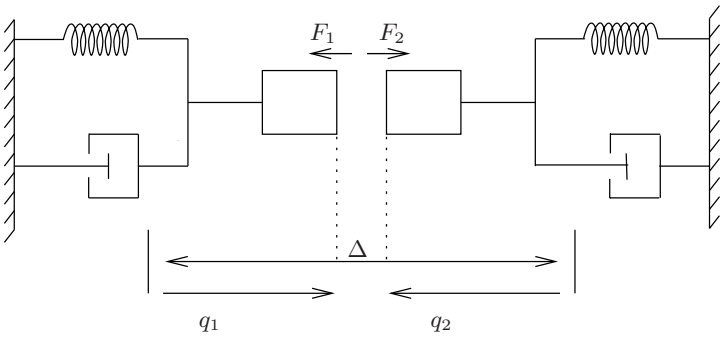


FIGURE 1.10. Mass–spring–damper system.

$(\mathrm{rev}(w))(t) := w(-t)$. Which is of the following scalar differential equations defines a time-reversible system?

(i) $\frac{d^2w}{dt^2} + w = 0$.

(ii) $\frac{dw}{dt} + \alpha w = 0$ (the answer may depend on the parameter α).

(iii) $\frac{d^2w}{dt^2} - w = 0$.

(iv) $\frac{d^nw}{dt^n} = 0$ (the answer may depend on the parameter n).

2

Systems Defined by Linear Differential Equations

2.1 Introduction

In this chapter we discuss a very common class of dynamical systems. It consists of the systems that are:

- linear
- time-invariant
- described by differential (or, in discrete time, difference) equations.

The importance of such dynamical systems stems from at least two aspects. Firstly, their prevalence in applications. Indeed, many models used in science and (electrical, mechanical, chemical) engineering are by their very nature linear and time-invariant. Secondly, the small signal behavior of a nonlinear time-invariant dynamical system in the neighborhood of an equilibrium point is time-invariant and approximately linear. The process of substituting the nonlinear model by the linear one is called *linearization* and is discussed in Chapter 4.

Linear systems lend themselves much better to analysis and synthesis techniques than nonlinear systems do. Much more is known about them. As such, the theory of linear systems not only plays an exemplary role for the nonlinear case, but has also reached a much higher degree of perfection.

The organization of this chapter is as follows. Some of the notational conventions are discussed in Section 2.2. The systems under consideration are

those described by *linear constant-coefficient differential equations*. What do we mean by a solution to such an equation? The seemingly natural answer to this question, demanding sufficient differentiability, is not quite adequate for our purposes. In particular, we want to be able to talk about solutions that are not differentiable. Therefore, in Section 2.3, the concept of weak solution is introduced. It is an extension of the classical notion, in which solutions are required to be sufficiently differentiable functions. The behavior is then defined as the set of weak solutions of the particular system of differential equations. In Section 2.4 some topological properties of the behavior are derived, and it is proved that the resulting dynamical system is linear and time-invariant.

A dynamical system is determined by its behavior, as introduced in Chapter 1. The behaviors studied in this chapter are described by systems of behavioral differential equations. Obviously, *different* behaviors are described by *different* equations. However, different equations do not necessarily describe different behaviors. In Section 2.5 it is explained that systems of differential equations that can be transformed into each other by premultiplication by a *unimodular* matrix represent the same behavior. Conversely, we will investigate the relation between representations that define the same behavior. It turns out that under a certain condition such differential equation representations can be transformed into each other by means of premultiplication by a suitable unimodular matrix.

Some of the mathematical background is provided in Appendix B. Appropriate references to this appendix are given whenever needed.

2.2 Notation

The class of dynamical systems that are studied in this chapter consists of those that can be described by the following type of behavioral differential equation:

$$R(\frac{d}{dt})w = 0, \tag{2.1}$$

or more explicitly,

$$R_0 w + R_1 \frac{d}{dt} w + \cdots + R_L \frac{d^L}{dt^L} w = 0, \tag{2.2}$$

with $R_0, R_1, \ldots, R_L \in \mathbb{R}^{g\times q}$ given coefficient matrices. Written in terms of the polynomial matrix (we will discuss the notation in more detail soon),

$$R(\xi) = R_0 + R_1\xi + \cdots + R_L\xi^L \in \mathbb{R}^{g\times q}[\xi],$$

this leads to (2.1). Thus $R(\xi)$ is a matrix of polynomials, $\frac{d}{dt}$ is the differentiation operator, and $w : \mathbb{R} \to \mathbb{R}^q$ is the signal that is being modeled by

the behavioral equation (2.1). We shall see that this defines a dynamical system with time axis $\mathbb{T} = \mathbb{R}$, with signal space $\mathbb{R}^q$, and with behavior $\mathfrak{B}$ consisting of those ws for which equation (2.1) is satisfied. However, in order to make this precise, we have more explaining to do; in particular, we should clarify the notation and specify what it means that w satisfies the differential equation (2.1).

Let us first backtrack in order to explain the notation in detail. Consider the following system of g linear constant-coefficient differential equations in the q real-valued signals $w_1, w_2, \ldots, w_q$:

$$\begin{array}{ccccccccc} r_{110}w_1 + \cdots + r_{11n_{11}}\dfrac{d^{n_{11}}}{dt^{n_{11}}}w_1 + \cdots + r_{1q0}w_q + \cdots + r_{1qn_{1q}}\dfrac{d^{n_{1q}}}{dt^{n_{1q}}}w_q = 0, \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad \vdots\ \vdots \\ r_{g10}w_1 + \cdots + r_{g1n_{g1}}\dfrac{d^{n_{g1}}}{dt^{n_{g1}}}w_1 + \cdots + r_{gq0}w_q + \cdots + r_{gqn_{gq}}\dfrac{d^{n_{gq}}}{dt^{n_{gq}}}w_q = 0. \end{array} \tag{2.3}$$

There are g scalar differential equations in (2.3). Each of these differential equations involves the scalar signals $w_1, w_2, \ldots, w_q$ (to save space we have only written how the first variable w_1 and the last variable w_q enter in the differential equations). Further, every one of these differential equations involves a certain number of derivatives of each of the variables $w_1, w_2, \cdots, w_q$. It is a linear constant-coefficient differential equation, meaning that the coefficients, the $r_{k\ell j}$s, multiplying these derivatives are real numbers. In the notation used in equation (2.3), the kth of these differential equations involves up to the $n_{k\ell}$th derivative of the variable w_ℓ, and the coefficient of the j-th derivative of w_ℓ, $\frac{d^j}{dt^j}w_\ell$, in the kth equation is $r_{k\ell j}$. Of course, in a concrete example sparsity will always be on our side, meaning that the great majority of coefficients $r_{k\ell j}$ turns out to be zero.

Nobody would wish to proceed to set up a general theory with the above cumbersome notation. Polynomial matrices are the appropriate tool to achieve the desired compactification of the notation of (2.3). Polynomial matrices play a very important role in this book. The notation and some salient facts are explained in Section 2.5 and in Appendix B. In particular, $\mathbb{R}[\xi]$ denotes the real polynomials in ξ. The symbol ξ in a polynomial is usually called the *indeterminate*. $\mathbb{R}^{n_1\times n_2}[\xi]$ denotes the set of real polynomial matrices with n_1 rows and n_2 columns, and $\mathbb{R}^{\bullet\times n}[\xi]$ the real polynomial matrices with n columns and any (finite) number of rows. Let $r(\xi) \in \mathbb{R}[\xi]$ be a polynomial with real coefficients. Thus $r(\xi)$ is an expression of the form

$$r(\xi) = \alpha_0 + \alpha_1\xi + \cdots + \alpha_{n-1}\xi^{n-1} + \alpha_n\xi^n,$$

with $\alpha_0, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and where ξ is the indeterminate. Now replace in this polynomial the indeterminate by the differentiation operator $\frac{d}{dt}$. This

yields the differential operator

$$r(\frac{d}{dt}) = \alpha_0 + \alpha_1 \frac{d}{dt} + \cdots + \alpha_{n-1} \frac{d^{n-1}}{dt^{n-1}} + \alpha_n \frac{d^n}{dt^n}.$$

We can let this differential operator act on an n-times differentiable function $f : \mathbb{R} \to \mathbb{R}$, yielding

$$r(\frac{d}{dt})f = \alpha_0 f + \alpha_1 \frac{d}{dt} f + \cdots + \alpha_{n-1} \frac{d^{n-1}}{dt^{n-1}} f + \alpha_n \frac{d^n}{dt^n} f.$$

Let us now generalize this to the multivariable case. Construct from (2.3) the polynomials

$$r_{k\ell}(\xi) = r_{k\ell 0} + r_{k\ell 1}\xi + \cdots + r_{k\ell n_{k\ell}}\xi^{n_{k\ell}}, \qquad k = 1, 2, \cdots, g, \quad \ell = 1, 2, \cdots, q$$

and organize them into the $g \times q$ polynomial matrix

$$R(\xi) := \begin{bmatrix} r_{11}(\xi) & \cdots & r_{1q}(\xi) \\ \vdots & & \vdots \\ r_{g1}(\xi) & \cdots & r_{gq}(\xi) \end{bmatrix}.$$

Note that we may as well write

$$R(\xi) = R_0 + R_1\xi + \cdots + R_{L-1}\xi^{L-1} + R_L\xi^L \tag{2.4}$$

with L the maximum of the orders $n_{k\ell}$ and with $R_j \in \mathbb{R}^{g \times q}$ the matrix

$$R_j := \begin{bmatrix} r_{11j} & \cdots & r_{gqj} \\ \vdots & & \vdots \\ r_{g1j} & \cdots & r_{gqj} \end{bmatrix}$$

(assume that $r_{k\ell j}$ is defined to be zero if $j > n_{k\ell}$). Now replace ξ in (2.4), as in the scalar case, by $\frac{d}{dt}$. This yields the differential operator

$$R(\frac{d}{dt}) = R_0 + R_1 \frac{d}{dt} + \cdots + R_{L-1} \frac{d^{L-1}}{dt^{L-1}} + R_L \frac{d^L}{dt^L}.$$

Acting on a sufficiently differentiable, in this case at least L times differentiable, time function $w : \mathbb{R} \to \mathbb{R}^q$, this yields the time function $e : \mathbb{R} \to \mathbb{R}^g$ defined by

$$e = R(\frac{d}{dt})w.$$

Next organize in (2.2) the time functions $w_1, w_2, \ldots, w_q$ into the column vector $w = \mathrm{col}[w_1, w_2, \ldots, w_q]$, $w : \mathbb{R} \to \mathbb{R}^q$, and verify that (2.1) $R(\frac{d}{dt})w = 0$, is nothing more than a mercifully compact version of the unwieldy system of differential equations (2.3). The discussion above is illustrated in the following example.

Example 2.2.1 Let $R(\xi) \in \mathbb{R}^{2\times 3}[\xi]$ be given by

$$R(\xi) = \begin{bmatrix} \xi^3 & -2+\xi & 3 \\ -1+\xi^2 & 1+\xi+\xi^2 & \xi \end{bmatrix}.$$

The multivariable differential equation $R(\frac{d}{dt})w = 0$ is

$$\frac{d^3}{dt^3}w_1 - 2w_2 + \frac{d}{dt}w_2 + 3w_3 = 0,$$

$$-w_1 + \frac{d^2}{dt^2}w_1 + w_2 + \frac{d}{dt}w_2 + \frac{d^2}{dt^2}w_2 + \frac{d}{dt}w_3 = 0.$$

□

2.3 Constant-Coefficient Differential Equations

In this section we study linear constant-coefficient ordinary differential equations as behavioral equations. Our aim in this section is to formalize (2.1) as a representation of a dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$. As the time axis $\mathbb{T}$ we take $\mathbb{R}$, and the signal space $\mathbb{W}$ is $\mathbb{R}^q$. In order to explain what the behavior is, we need to discuss the notions of strong and weak solutions to (2.1).

2.3.1 Linear constant-coefficient differential equations

The main object of study in this chapter, and to a certain extent of this book, is the behavior defined by the system of differential equations

$$R(\frac{d}{dt})w = 0, \qquad R(\xi) \in \mathbb{R}^{g\times q}[\xi]. \tag{2.5}$$

Equation (2.5) represents a system of g linear differential equations in the q scalar variables $w_1, \ldots, w_q$. In order to let (2.5) define a behavior, we have to specify the space of time functions of which the behavior is a subspace, and also we have to be precise about when we want to consider a function w to be a *solution* of this system of differential equations. A first attempt could be to restrict the attention to functions that are sufficiently smooth so that all derivatives appearing in (2.5) exist. This would have the advantage that the notion of *solution* is quite clear. However, from a system-theoretic point of view this choice is not satisfactory. For typically, the vector-valued function w contains components, called *inputs,* that can be freely manipulated, and one wants to be able to instantaneously change the value of these input variables, as in the case of a step input. As an example, think of instantaneously switching on a voltage source in an electrical network, or suddenly applying a force in a mechanical system.

Example 2.3.1 Consider the electrical circuit shown in Figure 2.1. The variable V_C is the voltage across the capacitor. From Kirchhoff's laws we

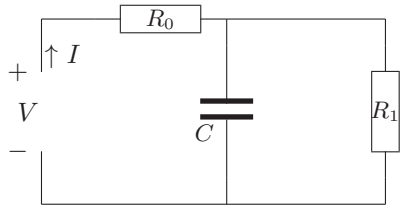


FIGURE 2.1. Electrical circuit.

obtain

$$\begin{aligned} I &= \frac{1}{R_1} V_C + C \tfrac{d}{dt} V_C, \\ V &= R_0 I + V_C. \end{aligned}$$

After eliminating V_C by substituting the second equation in the first, we obtain

$$V + CR_1 \frac{d}{dt} V = (R_0 + R_1) I + CR_0 R_1 \frac{d}{dt} I \tag{2.6}$$

as the differential equation relating the port voltage V to the port current I.

Now assume that for $t < 0$ this circuit was shorted ($V = 0$), and that at $t = 0$ a 1 volt battery is attached to it, see Figure 2.2. What is the

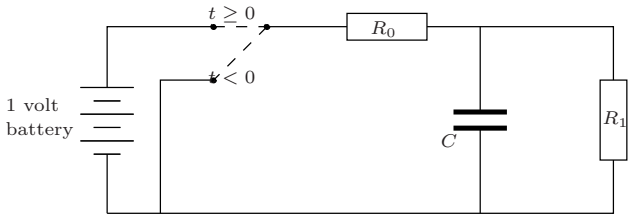


FIGURE 2.2. A voltage source attached to the circuit.

behavior of the terminal variables V and I? For V this is clear, it is the step function shown in Figure 2.3. In Chapter 3 we will see how to compute the corresponding current I. No doubt many readers see that I has the exponential-type response shown in Figure 2.4. From the above graphs and from physical considerations it appears that the differential equation (2.6) has this pair (V, I) as a solution. Note, however, that both V and

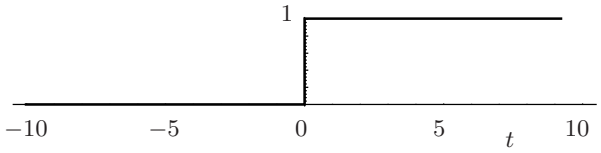


FIGURE 2.3. The voltage can be a step function.

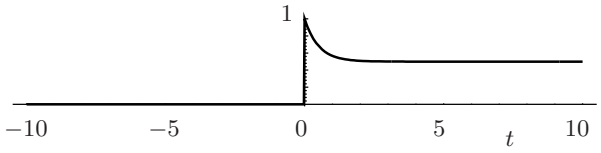


FIGURE 2.4. The response of the current.

I are discontinuous and hence not differentiable at $t = 0$. So, in order to interpret this voltage/current pair as a solution of (2.6) we have to extend, for good physical and engineering reasons, the notion of a solution beyond that which would require both V and I to be differentiable functions of time. □

Example 2.3.1 shows that confining the solution of a differential equation to the class of functions that are sufficiently smooth is too restrictive, and that we should look for a meaningful notion of solution that includes solutions that are discontinuous. We will therefore allow a much larger class of admissible trajectories, namely the set of *locally integrable functions* (see Definition 2.3.4). The advantage of this choice is the flexibility of the trajectories, which is attractive for applications. A drawback is that now it is not clear anymore when we want to consider a function $w : \mathbb{R} \to \mathbb{R}^q$ to be a solution of the equation (2.5). We therefore explain first what it means for a locally integrable function to be a solution of (2.5).

2.3.2 Weak solutions of differential equations

For a sufficiently smooth function it is obvious whether or not it is a solution of (2.5). For convenience we devote a definition to this.

Definition 2.3.2 (Strong solution) A function $w : \mathbb{R} \to \mathbb{R}^q$ is called a *strong solution* of (2.5) if the components of w are as often differentiable as required by the equation (2.5), and if it is a solution in the ordinary sense, that is, if $(R(\frac{d}{dt})w)(t) = 0$ for all $t \in \mathbb{R}$. □

In order to avoid having to specify how many times a function is differentiable, we frequently use the set of infinitely differentiable functions.

Definition 2.3.3 (Infinitely differentiable function) A function $w : \mathbb{R} \to \mathbb{R}^q$ is called *infinitely differentiable* if w is k times differentiable for all $k \in \mathbb{N}$. The space of infinitely differentiable functions $w : \mathbb{R} \to \mathbb{R}^q$ is denoted by $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. □

Definition 2.3.4 (Locally integrable function) A function $w : \mathbb{R} \to \mathbb{R}^q$ is called *locally integrable* if for all $a, b \in \mathbb{R}$,

$$\int_a^b \|w(t)\| dt < \infty.$$

Here $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^q$: if $v \in \mathbb{R}^q$, then $\|v\| = \sqrt{\sum_{i=1}^q v_i^2}$. The space of locally integrable functions $w : \mathbb{R} \to \mathbb{R}^q$ is denoted by $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. □

If a function w is only locally integrable, then plugging w into equation (2.5) makes, in general, no sense, since it may be that not all the required derivatives exist. Therefore the concept of *weak* solution is introduced. It extends the notion of solution to $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. Before we explain this generally, we present an illustrative example.

Example 2.3.5 Consider the simple differential equation

$$w_1 + \frac{d^2}{dt^2} w_1 + w_2 - \frac{d}{dt} w_2 = 0. \tag{2.7}$$

If we integrate (2.7) twice, we obtain

$$w_1(t) + \int_0^t \int_0^\tau w_1(s) ds d\tau - \int_0^t w_2(\tau) d\tau + \int_0^t \int_0^\tau w_2(s) ds d\tau = c_0 + c_1 t, \quad c_1, c_0 \in \mathbb{R}. \tag{2.8}$$

It should be clear that for every strong solution of (2.7), there exist constants c_1, c_0 such that (2.8) is satisfied. Conversely, if (w_1, w_2) satisfies (2.8) for some c_0, c_1, and if (w_1, w_2) is sufficiently smooth, then (w_1, w_2) satisfies (2.7). The interesting feature of (2.8) is that this equation does not impose any smoothness conditions on the pair of locally integrable functions (w_1, w_2), whereas (2.7) has no meaning in the classical sense if we want to consider functions that are not twice differentiable. This observation suggests that we could call w a *weak solution* of (2.7) if it satisfies (2.8) for some constants c_0, c_1. Although appealing, calling (w_1, w_2) a weak solution if (2.8) is satisfied for *all* t is not quite natural. For if, for example, we

were to change the value of w_2 at an arbitrary time instant, (2.8) remains true. Of course, w_1 could *not* be changed without making (2.8) false, but it seems reasonable to treat all components of w equally. Therefore we call (w_1, w_2) a weak solution of (2.7) if (2.8) is satisfied for *almost all* t, that is, except for t in a "small set", for example a finite set. To turn this informal discussion into a rigorous definition, we have to explain what exactly we mean by a "small set". □

Definition 2.3.6 (Set of zero measure) A set $N \subset \mathbb{R}$ is said to have *measure zero*, if $\int_N dt = 0$. Equivalently, N has measure zero if for every $\epsilon > 0$, there exist intervals I_k such that $N \subset \bigcup_{k=0}^{\infty} I_k$ and the sum of the lengths of the intervals does not exceed ϵ. □

Typical examples of sets of measure zero are finite sets and countable sets, for example the set of rational numbers. Two functions $f, g \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ are equal except on a set of measure zero if for all $a, b \in \mathbb{R}$ there holds $\int_a^b \|f(t) - g(t)\| dt = 0$. To streamline the terminology and the notation, we write $f = g$ almost everywhere ($f = g$ a.e.), or $f(t) = g(t)$ for almost all t.

Definition 2.3.7 (Weak solution) Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ and consider

$$R(\frac{d}{dt})w = 0. \tag{2.9}$$

Define the *integral operator* acting on $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ by

$$(\int w)(t) := \int_0^t w(\tau) d\tau, \qquad (\int^{k+1} w)(t) := \int_0^t (\int^k w)(\tau) d\tau, \quad k \geq 1. \tag{2.10}$$

Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ be given. To $R(\xi)$ we associate the polynomial matrix $R^*(\xi)$ defined as follows. Let L be the maximal power of ξ occurring in $R(\xi)$, say

$$R(\xi) = R_0 + R_1 \xi + \cdots + R_L \xi^L, \quad R_L \neq 0. \tag{2.11}$$

Define

$$R^*(\xi) := \xi^L R(\frac{1}{\xi}) = R_L + R_{L-1}\xi + \cdots + R_1 \xi^{L-1} + R_0 \xi^L.$$

Now, consider the *integral equation*

$$((R_0(\int)^L + R_1(\int)^{L-1} + \cdots + R_{L-1} \int + R_L)w)(t) = c_0 + c_1 t + \cdots + c_{L-1} t^{L-1},$$

with $c_i \in \mathbb{R}^g$, or, in compact notation

$$(R^*(\int)w)(t) = c_0 + c_1 t + \cdots + c_{L-1} t^{L-1}, \quad c_i \in \mathbb{R}^g. \tag{2.12}$$

We call $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ a *weak* solution of (2.9) if there exist constant vectors $c_i \in \mathbb{R}^g$ such that (2.12) is satisfied for almost all $t \in \mathbb{R}$. □

Remark 2.3.8 The lower limit in (2.10), which we took to be zero, is immaterial for the definition of weak solution: any other choice for the lower limit leads to an equivalent notion of weak solution. In fact, even within expressions of the form $(\int)^k$, the lower limits in the multiple integral need not be identical. This is shown in the following lemma. □

Lemma 2.3.9 *Let $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. For any $t_i \in \mathbb{R}$, $i = 0, \ldots, n-1$, there exist $c_i \in \mathbb{R}^q$, $i = 0, \ldots, n-1$ such that for all $t \in \mathbb{R}$*

$$\int_{t_0}^{t}\int_{t_1}^{\tau_1}\cdots\int_{t_{n-1}}^{\tau_{n-1}} w(\tau_n)d\tau_n\cdots d\tau_1 = \int_{0}^{t}\int_{0}^{\tau_1}\cdots\int_{0}^{\tau_{n-1}} w(\tau_n)d\tau_n\cdots d\tau_1$$
$$+c_0 + c_1 t + \cdots + c_{n-1}t^{n-1}.$$

Proof The proof is not difficult, and the reader is referred to Exercise 2.12. □

Example 2.3.10 Let $g = 1,\ q = 2$. Consider the differential equation

$$\frac{d}{dt}w_2 = w_2 + w_1. \tag{2.13}$$

A *strong* solution of (2.13) is obtained by taking

$$w_2(t) = \int_0^t e^{(t-\tau)}w_1(\tau)d\tau, \tag{2.14}$$

where w_1 is any continuous function. This follows by a simple calculation. In Chapter 3 we explain how to find solutions of the form (2.14) is in a systematic way. An example of a *weak* solution that is not a *strong* solution is

$$(w_1(t), w_2(t)) = \begin{cases} (0,0) & t < 0, \\ (1, e^t - 1) & t \geq 0. \end{cases} \tag{2.15}$$

For the proof the reader is referred to Exercise 2.1. □

The following result shows that the definition of *weak* solution indeed provides an extension of the notion of *strong* solution.

Theorem 2.3.11 *Consider the behavior defined by*

$$R(\frac{d}{dt})w = 0. \tag{2.16}$$

1. *Every strong solution of (2.16) is also a weak solution.*

2. *Every weak solution that is sufficiently smooth (in the notation of (2.11) L times differentiable) is also a strong solution.*

Proof See Exercise 2.21. □

Remark 2.3.12 We have introduced the concept of weak solution only to give a mathematically sound basis to nondifferentiable solutions of differential equations. We will not use it to actually *solve* differential equations via the "weak approach". However, we will sometimes check that proposed functions are indeed weak solutions. For instance, in Chapter 3 we will verify that the pair (w_1, w_2) related by

$$w_2(t) = \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} e^{\lambda(t-\tau)} w_1(\tau) d\tau$$

satisfies the differential equation

$$(\frac{d}{dt} - \lambda)^k w_2 = w_1$$

strongly if w_1 is continuous and *weakly* for every locally integrable, possibly discontinuous w_1. This is proven in Section 3.3.

Another, perhaps more elegant, way of choosing the set of admissible trajectories is to include *generalized functions,* or *distributions*. Distribution theory is beyond the scope of this book. However, it should be remarked that the approach that is chosen here is quite close to what distribution theory would give. □

2.4 Behaviors Defined by Differential Equations

In this section we formally define behaviors represented by $R(\frac{d}{dt})w = 0$. We prove two fundamental structural properties, *linearity* and *time invariance*. Furthermore, we consider some *topological* properties of the behavior.

With the aid of the notion of weak solution, we are now able to define the behavior corresponding to the behavioral equation (2.5).

Definition 2.4.1 Equation (2.5) defines the dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$, where $\mathfrak{B}$ is defined as

$$\mathfrak{B} := \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid w \text{ is a weak solution of } R(\tfrac{d}{dt})w = 0.\}$$

□

2.4.1 Topological properties of the behavior

In this subsection we discuss some important *topological* properties of the behavior. Topological properties are related to such notions as *open* and *closed*, as well as *dense* and *convergence*. Before we can make any statements concerning these matters, we need to define convergence in $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$. It turns out that the behavior defined by $R(\frac{d}{dt})w = 0$ is *closed* with respect to that notion of convergence. That is, if a sequence of trajectories in the behavior converges to some trajectory w, then this trajectory w also belongs to the behavior. A second, equally important, result is that every weak solution may be obtained as the limit of a sequence of strong solutions. In other words the subbehavior of strong solutions is a dense subset of the behavior.

Definition 2.4.2 (Convergence in the sense of $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$) A sequence $\{w_k\}$ in $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$ converges to $w \in \mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$ in the sense of $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$ if for all $a, b \in \mathbb{R}$,

$$\lim_{k\to\infty} \int_a^b \|w(t) - w_k(t)\| dt = 0.$$

Here, $\|.\|$ denotes the Euclidean norm in $\mathbb{R}^q$. □

Example 2.4.3 Consider the sequence of functions $\{w_k\}$ with $w_k(t)$ defined by:

$$w_k(t) = \begin{cases} 0 & \text{for } |t| > \dfrac{1}{k}, \\ \\ 1 & \text{for } |t| \leq \dfrac{1}{k}. \end{cases}$$

It is not difficult to check that w_k converges (in the sense of $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$) to the zero function, even though $w_k(0) = 1$ for all k. On the other hand, if we define $w_k(t) = k$ for $|t| < \frac{1}{k}$ and zero otherwise, then w_k does *not* converge in the sense of $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R})$ at all, let alone to zero, although $w_k(t)$ converges pointwise to zero for all $t \neq 0$. □

Theorem 2.4.4 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ be given and let $\mathfrak{B}$ be the behavior defined by $R(\frac{d}{dt})w = 0$. If $w_k \in \mathfrak{B}$ converges to $w \in \mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$ in the sense of $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R}^q)$, then $w \in \mathfrak{B}$.*

Proof Since $w_k \in \mathfrak{B}$, there exist vectors $c_{0,k}, \ldots, c_{L-1,k}$, such that

$$R^*(\textstyle\int) w_k = c_{0,k} + \cdots + c_{L-1,k} t^{L-1}, \qquad k = 0, 1, 2, \ldots. \tag{2.17}$$

Since $w_k \to w$ as $k \to \infty$, and since integration is a continuous operation on $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ ($\lim_{k\to\infty} \int_0^t w_k(\tau)d\tau = \int_0^t \lim_{k\to\infty} w_k(\tau)d\tau$, in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$), see Exercise 2.19, it follows that

$$\lim_{k\to\infty} R^*(\textstyle\int) w_k = R^*(\textstyle\int) w \quad \text{in the sense of } \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q).$$

In other words, the right-hand side of (2.17) converges, to some function, in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. It is not difficult to check (see Exercise 2.20) that this can be the case only if the coefficients $c_{i,k}$ converge, say $\lim_{k\to\infty} c_{i,k} =: c_i$. This implies that the sequence of polynomials on the right-hand side of (2.17) converges to the polynomial $c_0 + \cdots + c_{L-1}t^{L-1}$. As a consequence

$$R^*(\textstyle\int) w = c_0 + \ldots + c_{L-1}t^{L-1} \quad \text{for almost all } t,$$

and hence $w \in \mathfrak{B}$. □

We now prove that every trajectory in the behavior can be seen as a limit of smooth trajectories in the behavior, i.e., as a limit of strong solutions of $R(\frac{d}{dt})w = 0$, see Theorem 2.3.11, Part 2. For the construction of the limiting sequence we use a very special function, namely a bell-shaped function that is zero outside the interval $[-1, 1]$, nonzero inside this interval, but nevertheless infinitely differentiable.

Definition 2.4.5 The function ϕ is defined by

$$\phi(t) = \begin{cases} 0 & \text{if } |t| \geq 1, \\ e^{-\frac{1}{1-t^2}} & \text{for } |t| < 1. \end{cases} \tag{2.18}$$

□

Remark 2.4.6 Notice that ϕ is indeed infinitely differentiable; see Figure 2.5 and Exercise 2.17. Functions that are infinitely differentiable and have compact support are called *flat* functions. □

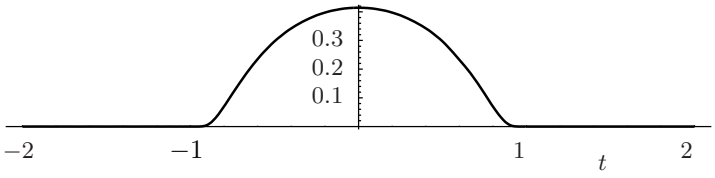


FIGURE 2.5. The graph of the function ϕ defined in (2.18).

We will use the following:

Lemma 2.4.7 *Let $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ and let ϕ be given by (2.18). Define the function v by*

$$v(t) := \int_{-\infty}^{\infty} \phi(\tau) w(t-\tau) d\tau. \tag{2.19}$$

Then v is infinitely differentiable.

Proof Verify that

$$(\frac{d^n}{dt^n} v)(t) = \int_{-\infty}^{\infty} \phi^{(n)}(\tau) w(t-\tau) d\tau.$$

□

Remark 2.4.8 The function v defined by (2.19) is known as the *convolution product* of ϕ and w, and is usually denoted by

$$\phi * w.$$

□

Lemma 2.4.9 *Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ and let $\mathfrak{B}$ be the behavior defined by $R(\frac{d}{dt})w = 0$. Let ϕ be the function defined in Definition 2.4.5. Then for every $w \in \mathfrak{B}$ we have that $\phi * w \in \mathfrak{B}$.*

Proof Let $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. By interchanging the order of integration (see Exercise 2.18), it follows that

$$\int (\phi * w) = \phi * (\int w). \tag{2.20}$$

Repeated application of (2.20) yields

$$R^*(\int)(\phi * w) = \phi * (R^*(\int) w). \tag{2.21}$$

From (2.21) we conclude that since $w \in \mathfrak{B}$:

$$\begin{aligned} R^*(\int)(\phi * w) &= \phi * (R^*(\int) w) \\ &= \phi * (c_0 + \cdots + c_{L-1} t^{L-1}). \end{aligned} \tag{2.22}$$

It remains to show that the convolution product of a polynomial with ϕ is again a polynomial of at most the same degree. To that end, observe that by L-fold differentiation, we obtain zero:

$$\frac{d^L}{dt^L}[\phi * (c_0 + \cdots + c_{L-1} t^{L-1})] = \phi * [\frac{d^L}{dt^L}(c_0 + \cdots + c_{L-1} t^{L-1})] = 0.$$

Combining (2.21) and (2.22) yields

$$R^*(\textstyle\int)(w * \phi) = e_0 + \cdots + e_{L-1}t^{L-1},$$

so that indeed $w * \phi \in \mathfrak{B}$. □

The following theorem shows that every locally integrable function can be approximated arbitrarily well by $\mathcal{C}^\infty$ functions, or otherwise stated, $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ is dense in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$.

Theorem 2.4.10 *Let $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. There exists a sequence $\{w_k\}$ in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ that converges to w in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$.*

Proof [sketch, $g = q = 1$] Define ψ as the normalized version of ϕ (defined by (2.18)),

$$\psi(t) := \frac{\phi(t)}{\int_{-\infty}^{\infty} \phi(\tau)d\tau},$$

and define ψ_k by

$$\psi_k(t) := k\psi(kt).$$

Define the functions w_k by

$$w_k(t) := w * \psi_k. \tag{2.23}$$

From Lemma 2.4.7 we know that w_k is in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. We claim that w_k converges to w in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$. The proof of this claim is worked out in Exercise 2.15.

□

Remark 2.4.11 In Figure 2.6 we have depicted the functions ψ_k. The sequence $\{\psi_k\}$ does not converge in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$, for the zero function, obviously the only reasonable candidate limit, is *not* the limit, yet $w * \psi_k$ converges for every $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$. In that sense one could say that ψ_k converges to the unity element with respect to $*$. Readers familiar with distribution theory recognize the Dirac delta as the limit of the the ψ_ks. As an illustration, Figure 2.7 shows the first four approximations of the step function.

□

Theorem 2.4.10 and Lemma 2.4.9 are interesting in their own right. For us, the importance is reflected by the following consequences.

Corollary 2.4.12 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and let $\mathfrak{B}$ be the behavior defined by $R(\frac{d}{dt})w = 0$. For every $w \in \mathfrak{B}$ there exists a sequence $w_k \in \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ such that w_k converges to w in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$.*

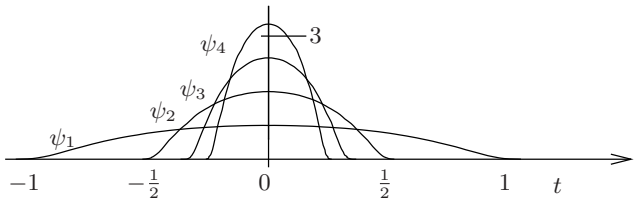


FIGURE 2.6. The graph of the functions $\psi_1, \ldots, \psi_4$.

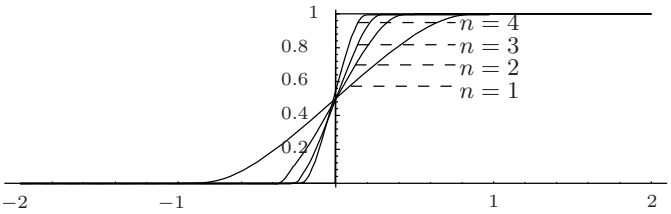


FIGURE 2.7. Approximation of the step function.

Proof Define $w_k \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ as in (2.23). By Lemma 2.4.9, $w_k \in \mathfrak{B}$, and by Theorem 2.4.10, w_k converges to w in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. □

Corollary 2.4.12 implies in particular that every weak solution can be approximated arbitrarily closely by a strong one.

The last theorem in this section is of crucial importance in the remainder of the book. It states that the behavior corresponding to $R(\frac{d}{dt})w = 0$ is completely determined by its $\mathcal{C}^\infty$ part and hence by the strong solutions.

Theorem 2.4.13 *Let $R_1(\xi) \in \mathbb{R}^{g_1 \times q}[\xi]$ and $R_2(\xi) \in \mathbb{R}^{g_2 \times q}[\xi]$. Denote the corresponding behaviors by $\mathfrak{B}_1$ and $\mathfrak{B}_2$. If $\mathfrak{B}_1 \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) = \mathfrak{B}_2 \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, then $\mathfrak{B}_1 = \mathfrak{B}_2$.*

Proof Choose $w \in \mathfrak{B}_1$. By Corollary 2.4.12 there exists a sequence $w_k \in \mathfrak{B}_1 \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) = \mathfrak{B}_2 \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ converging to w. By Theorem 2.4.4 it follows that $w \in \mathfrak{B}_2$. This shows that $\mathfrak{B}_1 \subset \mathfrak{B}_2$. In the same way one proves that $\mathfrak{B}_2 \subset \mathfrak{B}_1$. □

Remark 2.4.14 The results derived in this section play an important role in the sequel. At several places in the theory, properties of the behavior will first be derived for the $\mathcal{C}^\infty$ part of the behavior and will then be proved for the whole behavior using a denseness argument and/or the fact that the behavior is completely determined by its $\mathcal{C}^\infty$ part. To illustrate this we now show how the time-invariance could be proven along these lines. □

2.4.2 Linearity and time-invariance

In the previous subsection we derived some topological properties of the behavior. Next we study two structural properties of the behavior: linearity and time-invariance. The proof of linearity is straightforward; the proof of time-invariance relies on the property that the set of strong solutions is dense in the behavior. We refer the reader to Section 1.4 for the definitions of linearity and time-invariance.

Theorem 2.4.15 *The behavior $\mathfrak{B}$ as defined in Definition 2.4.1 is linear and time-invariant.*

Proof *Linearity.* Let $w_1, w_2 \in \mathfrak{B}$ and let $\lambda \in \mathbb{C}$. Define $w := w_1 + \lambda w_2$. We have to prove that $w \in \mathfrak{B}$. Since $w_1, w_2 \in \mathfrak{B}$, there exist vectors $c'_0, \ldots, c'_{L-1}$ and $c''_0, \ldots, c''_{L-1}$ such that

$$R^*(\textstyle\int) w_1 = c'_0 + \cdots + c'_{L-1} t^{L-1} \quad \text{and} \quad R^*(\textstyle\int) w_2 = c''_0 + \cdots + c''_{L-1} t^{L-1}.$$

Define $c_i := c'_i + \lambda c''_i$. Then

$$\begin{aligned} R^*(\textstyle\int) w &= R^*(\textstyle\int)(w_1 + \lambda w_2) \\ &= R^*(\textstyle\int) w_1 + \lambda R^*(\textstyle\int) w_2 \\ &= c'_0 + \cdots + c'_{L-1} t^{L-1} + \lambda (c''_0 + \cdots + c''_{L-1} t^{L-1}) \\ &= c_0 + \cdots + c_{L-1} t^{L-1}. \end{aligned}$$

This shows the linearity of $\mathfrak{B}$.

Time-invariance. Let $w \in \mathfrak{B}$ and let $\tilde{w}$ be defined as $\tilde{w}(t) := w(t - t_1)$. By Theorem 2.4.10 there exists a sequence $w_k \in \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ such that w_k converges to w in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. Since w_k is smooth, $R(\frac{d}{dt}) w_k = 0$ strongly, i.e., in the classical sense. Define the time-shifted versions of w_k as $\tilde{w}_k(t) := w_k(t - t_1)$. Since $\frac{d}{dt}(w(t - t_1)) = (\frac{d}{dt} w)(t - t_1)$, in other words the differentiation operator commutes with the shift operator, it is clear that also $R(\frac{d}{dt}) \tilde{w}_k = 0$ in the classical sense. Since w_k converges to w, $\tilde{w}_k$ converges to $\tilde{w}$. By Theorem 2.4.4 we conclude that $\tilde{w} \in \mathfrak{B}$. □

Remark 2.4.16 With the aid of Lemma 2.3.9 it is possible to give a direct proof of time-invariance without using the results from Section 2.4.1. The present proof, however, is nicer, since it avoids cumbersome calculations and shows how properties of the strong part of the behavior carry over to the whole behavior. □

2.5 The Calculus of Equations

2.5.1 Polynomial rings and polynomial matrices

The main objects of study in this book are linear time-invariant behaviors that can be described by

$$R(\frac{d}{dt})w = 0, \tag{2.24}$$

where $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ is a *polynomial matrix*. As was explained in Section 2.2, (2.24) provides a convenient and compact notation. However, it is not just for notational convenience that we have introduced this. We will see that we can develop a *calculus of representations* based on the structure of the set $\mathbb{R}[\xi]$. By this we mean that certain algebraic properties of polynomials and polynomial matrices are relevant to the study of behavioral representations of the form (2.24). To take full advantage of these properties, we briefly introduce some elements from abstract algebra. Undoubtedly the most important fact that we use is that the set of polynomials with real or complex coefficients forms a *ring*. A ring is a nonempty set on which two binary operations addition, $+$, and multiplication, $\bullet$, are defined[1]. Addition and multiplication in rings are, in general, different in nature from what we are used to in $\mathbb{R}$. However, they are required to satisfy the usual properties such as, for example, associativity and distributivity, except that multiplication is not required to be commutative. Addition of two polynomials $a(\xi) = a_0 + a_1\xi + \cdots + a_n\xi^n$, $b(\xi) = b_0 + b_1\xi + \cdots + b_m\xi^m$ is defined as $a(\xi) + b(\xi) := (a_0 + b_0) + (a_1 + b_1)\xi + \cdots$, multiplication as $a(\xi) \bullet b(\xi) := a_0b_0 + (a_1b_0 + a_0b_1)\xi + (a_0b_2 + a_1b_1 + a_2b_0)\xi^2 + \cdots$. Usually, the multiplication symbol $\bullet$ is dropped, and we will henceforth follow this convention. With these definitions of addition and multiplication, the set of polynomials with coefficients in $\mathbb{R}$ or $\mathbb{C}$, denoted by $\mathbb{R}[\xi]$ or $\mathbb{C}[\xi]$, forms a ring. The polynomial rings $\mathbb{R}[\xi]$ and $\mathbb{C}[\xi]$ have the property that *division with remainder* is possible, i.e., for every two elements $a(\xi)$ and $b(\xi)$, with $a(\xi)$ nonzero, there exist polynomials $q(\xi), r(\xi)$ in the ring such that

$$b(\xi) = q(\xi)a(\xi) + r(\xi) \text{ with } \deg r(\xi) < \deg a(\xi). \tag{2.25}$$

Notice that (2.25) is the polynomial analogue of a similar property of the integers: for every $a, b \in \mathbb{Z}$, $a \neq 0$, there exist $q, r \in \mathbb{Z}$ such that $b = q\,a + r$ and $|r| < |a|$. Just as for the integer case, the polynomials $q(\xi)$ (the *quotient* of $b(\xi)$ and $a(\xi)$) and $r(\xi)$ (the *remainder*), can be computed by *long division*.

---

[1] A *binary operation* on a set A is a map from $A \times A$ to A. For example, addition is a binary operation on $\mathbb{R}$. It assigns to every pair of real numbers their sum. Another binary operation on $\mathbb{R}$ is multiplication.

Example 2.5.1 Take $a(\xi) = 2 - \xi + \xi^2$ and $b(\xi) = 4 - 2\xi + 3\xi^2 + 3\xi^3 + \xi^4$. Then $b(\xi) = (5 + 4\xi + \xi^2)a(\xi) - 6 - 5\xi$. Thus, $q(\xi) = 5 + 4\xi + \xi^2$ and $r(\xi) = -6 - 5\xi$. □

Rings in which division with remainder is possible are called *Euclidean rings*. The reader is referred to any introductory algebra textbook, e.g., [34], for more details on rings and for a precise definition of Euclidean rings. All the details that we use are provided in the text.

The definitions of addition and multiplication of polynomials induce in a natural way addition and multiplication of *polynomial matrices*, provided, of course, that these matrices are of compatible sizes. Also, we can speak about the *determinant* of a square polynomial matrix. Notice that this determinant is a scalar polynomial. Finally, if $R(\xi)$ is a square matrix with $\det R(\xi) \neq 0$ (by that we mean that the determinant is not the zero polynomial), then we can speak about the *inverse*, $R^{-1}(\xi)$, which is a matrix of *rational* functions, i.e., a matrix whose entries are fractions of polynomials.

2.5.2 Equivalent representations

In Section 2.3 we have introduced dynamical systems described by systems of differential equations $R(\frac{d}{dt})w = 0$. We have seen in particular what it means that a map $w : \mathbb{R} \to \mathbb{R}^q$ belongs to the behavior of the resulting dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$. This system is said to be *represented*, or *parametrized*, by the polynomial matrix $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Note that $R(\xi)$ obviously defines Σ. However, there are many polynomial matrices that represent the *same* dynamical system. This leads to the following definition.

Definition 2.5.2 (Equivalent differential equations) Let $R_i(\xi) \in \mathbb{R}^{g_i\times q}[\xi]$, $i = 1, 2$. The differential equations

$$R_1(\frac{d}{dt})w = 0 \quad \text{and} \quad R_2(\frac{d}{dt})w = 0$$

are said to be *equivalent* if they define the same dynamical system. In other words, equivalence means that w is a weak solution of $R_1(\frac{d}{dt})w = 0$ *if and only if* it is also a weak solution of $R_2(\frac{d}{dt})w = 0$. □

Example 2.5.3 Consider the system of differential equations

$$\begin{aligned} w_1 + \frac{d^2}{dt^2}w_1 &= 0, \\ -w_2 + \frac{d^2}{dt^2}w_2 &= 0 \end{aligned} \tag{2.26}$$

and the seemingly different one:

$$\begin{array}{rcl} w_1 + \dfrac{d^2}{dt^2}w_1 & = & 0, \\ \dfrac{d^2}{dt^2}w_1 + \dfrac{d^4}{dt^4}w_1 - w_2 + \dfrac{d^2}{dt^2}w_2 & = & 0. \end{array} \tag{2.27}$$

These systems of equations define the same dynamical system. To see this, observe that the first differential equation implies that

$$\frac{d^2}{dt^2}w_1 + \frac{d^4}{dt^4}w_1 = 0.$$

Adding this to the second differential equation in (2.26) or subtracting it from the second in (2.27) demonstrates the equivalence. In fact, in this case the solutions can be displayed explicitly. The solution set of both (2.26) and (2.27) consists of the functions of the form

$$\begin{array}{rcl} w_1(t) & = & A_1 \cos t + A_2 \sin t, \\ w_2(t) & = & B_1 e^t + B_2 e^{-t}, \end{array} \tag{2.28}$$

where A_1, A_2, B_1, B_2 range over the set of real numbers. Why this is precisely the solution set is explained in Section 3.2. Anyway, it shows that (2.26) and (2.27) are equivalent.

Let us now see how we can reformulate the operations above in an algebraic fashion. In polynomial notation, equations (2.26) and (2.27) read as $R_1(\frac{d}{dt})w = 0$ and $R_2(\frac{d}{dt})w = 0$ with $R_1(\xi)$ and $R_2(\xi)$ given by

$$R_1(\xi) = \begin{bmatrix} 1+\xi^2 & 0 \\ 0 & -1+\xi^2 \end{bmatrix}, \qquad R_2(\xi) = \begin{bmatrix} 1+\xi^2 & 0 \\ \xi^2+\xi^4 & -1+\xi^2 \end{bmatrix}.$$

In algebraic terms, these operations that transformed the representation (2.26) into (2.27) are: *multiply the first row of $R_1(\xi)$ by ξ^2 and add it to the second.* In polynomial notation,

$$U(\xi)R_1(\xi) = R_2(\xi) \quad \text{with} \quad U(\xi) = \begin{bmatrix} 1 & 0 \\ \xi^2 & 1 \end{bmatrix}.$$

If we restrict attention to $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ solutions, then it is clear that $R_1(\frac{d}{dt})w = 0$ implies that $U(\frac{d}{dt})R_1(\frac{d}{dt})w = 0$ and hence $R_2(\frac{d}{dt})w = 0$. In this simple example it also clear how to prove the converse, since

$$V(\xi)R_2(\xi) = R_1(\xi) \quad \text{with} \quad V(\xi) = \begin{bmatrix} 1 & 0 \\ -\xi^2 & 1 \end{bmatrix}. \tag{2.29}$$

The operation (2.29) is nothing but replacing the second row of $R_2(\xi)$ by the difference of the second row and ξ^2 times the first row. Notice that $V(\xi)U(\xi) = I$. The polynomial matrices $U(\xi)$ and $V(\xi)$ appear to be

each other's inverses. The matrix $U(\xi)$, transforming $R_1(\xi)$ into $R_2(\xi)$ by $R_2(\xi) = U(\xi)R_1(\xi)$, thus has the special property that its inverse is also a polynomial matrix (rather than merely a matrix of rational functions, which one would have expected). This special property, called *unimodularity* of $U(\xi)$, turns out to be the key in classifying equivalent representations. □

The first thing that we learn from Example 2.5.3 is that for any polynomial matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ and $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$, we have that $R(\frac{d}{dt})w = 0$ implies that $U(\frac{d}{dt})R(\frac{d}{dt})w = 0$. How about the converse implication? It is very tempting to conjecture that $U(\frac{d}{dt})R(\frac{d}{dt})w = 0$ implies that $U^{-1}(\frac{d}{dt})U(\frac{d}{dt})R(\frac{d}{dt})w = 0$ and therefore $R(\frac{d}{dt})w = 0$. This would, however, be an abuse of analogy. For $U^{-1}(\xi)$ may have a proper meaning as a *matrix of rational functions*, but it need not be polynomial, and therefore $U^{-1}(\frac{d}{dt})$ does have no meaning in general. (What is the meaning of $\frac{1+\frac{d}{dt}}{2+(\frac{d}{dt})^2}$?) On the other hand, if there happens to exist a polynomial matrix $V(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that $V(\xi)U(\xi) = I$, in other words, if the inverse of $U(\xi)$ is again a polynomial matrix, then the converse *is* true: $R(\frac{d}{dt})w = 0$ and $U(\frac{d}{dt})R(\frac{d}{dt})w = 0$ define the same behavior. This observation leads us to the following result.

Theorem 2.5.4 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$. Define $R'(\xi) := U(\xi)R(\xi)$. Denote the behaviors corresponding to $R(\xi)$ and $R'(\xi)$ by $\mathfrak{B}$ and $\mathfrak{B}'$ respectively. Then:*

1. $\mathfrak{B} \subset \mathfrak{B}'$.
2. *If in addition, $U^{-1}(\xi)$ exists and if $U^{-1}(\xi) \in \mathbb{R}^{g\times g}[\xi]$, then $\mathfrak{B} = \mathfrak{B}'$.*

Proof 1. Choose $w \in \mathfrak{B}$. By Corollary 2.4.12, there exists a sequence $\{w_k\} \in \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ converging to w. Since $R(\frac{d}{dt})w_k = 0$ in the usual sense, i.e., strongly, $U(\frac{d}{dt})R(\frac{d}{dt})w_k = 0$, implying that $w_k \in \mathfrak{B}'$. By Theorem 2.4.4, it follows that then also $w \in \mathfrak{B}'$.

2. By 1, it suffices to prove that $\mathfrak{B}' \subset \mathfrak{B}$. Since $U^{-1}(\xi)$ is a well-defined polynomial matrix, this follows by just applying Part 1 to $U^{-1}(\xi)R(\xi)$ and $R(\xi)$. □

Polynomial matrices with a polynomial inverse play a very important role. Hence the following definition.

Definition 2.5.5 (Unimodular matrix) Let $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$. Then $U(\xi)$ is said to be a *unimodular* polynomial matrix if there exists a polynomial matrix $V(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that $V(\xi)U(\xi) = I$. Equivalently, if $\det U(\xi)$ is equal to a nonzero constant (see Exercise 2.7). □

Remark 2.5.6 Theorem 2.5.4 is a first example of the power of the polynomial approach. It raises some interesting questions:

1. Can we characterize all unimodular matrices in some simple constructive way?
2. How can we use Theorem 2.5.4?
3. What about the converse of Theorem 2.5.4: Is it true that if two polynomial matrices $R_1(\xi)$ and $R_2(\xi)$ define the same behavior, then there exists a unimodular matrix $U(\xi)$ such that $U(\xi)R_1(\xi) = R_2(\xi)$?

As for the answers to these questions, we will see that unimodular matrices can be characterized in that they can all be factorized as products of matrices of a very simple form.

The usefulness of Theorem 2.5.4 is that since premultiplication of $R(\xi)$ by a matrix $U(\xi)$ of which the inverse is again polynomial does not change the behavior, this property can be used to bring $R(\xi)$ into a convenient form, such as a triangular form.

The last question is more involved. The answer is affirmative, provided that the matrices $R_1(\xi)$ and $R_2(\xi)$ have the same number of rows. The proof of this statement involves elements that will be treated only in Chapter 3 and is therefore postponed until the end of that chapter. □

It is worthwhile to pause a bit on the idea of unimodularity. Consider the polynomial matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$. Its determinant is defined by the same formula as the determinant of an ordinary matrix, so that $\det U(\xi)$ is a scalar polynomial. Of course, $\det U(\xi)$ may be the zero polynomial However, if $\det U(\xi)$ is not equal to the zero polynomial, we can define the inverse of $U(\xi)$ as

$$U^{-1}(\xi) = \frac{1}{\det U(\xi)}\tilde{U}(\xi),$$

where $\tilde{U}(\xi)$ denotes the matrix of cofactors of $U(\xi)$, i.e., the (i,j)th element of $\tilde{U}(\xi)$ is equal to $(-1)^{i+j}$ times the determinant of the matrix obtained by deleting the jth row and the ith column in $U(\xi)$ (this is Cramer's rule). Note that $U^{-1}(\xi)$ is a matrix of *rational functions*: each element is the ratio of two polynomials. However, in some special cases $U^{-1}(\xi)$ is itself a polynomial matrix. This occurs if and only if $\det U(\xi)$ is a nonzero constant: a nonzero polynomial of degree zero. This special class of polynomial matrices is the class of the unimodular matrices.

Examples of unimodular matrices are:

1. Nonsingular square matrices with constant coefficients.
2. Upper triangular square polynomial matrices with nonzero constants on the diagonal.

3. Lower triangular square polynomial matrices with nonzero constants on the diagonal.

Observe further that:

1. I is unimodular.
2. If $U_1(\xi)$ and $U_2(\xi)$ are unimodular, so is $U_1(\xi)U_2(\xi)$.
3. If $U(\xi)$ is unimodular, so is $U^{-1}(\xi)$.

This gives the set of unimodular matrices the structure of a *group* (see Exercise 2.25).

2.5.3 Elementary row operations and unimodular polynomial matrices

It follows from Theorem 2.5.4 that if $\mathfrak{B}$ is represented by a matrix $R(\xi)$, and if $U(\xi)$ is unimodular, then $\mathfrak{B}$ is also represented by $U(\xi)R(\xi)$. This is a rather abstract result that seems to rely very much on algebraic properties of polynomial matrices. We now give an interpretation in terms of simple operations on the differential equation $R(\frac{d}{dt})w = 0$ itself.

There are a number of *elementary operations* on the equations $R(\frac{d}{dt})w = 0$ by means of which equivalent representations may be generated.

Denote the rows of $R(\xi)$ by $r_1(\xi), r_2(\xi), \ldots, r_g(\xi)$. Hence $r_i(\xi) \in \mathbb{R}^{1\times q}[\xi]$ for $i = 1, 2, \ldots, g$. Now consider the following *elementary row* operations on $R(\xi)$, and, accordingly, on the differential equation $R(\frac{d}{dt})w = 0$. There are three types of elementary row operations:

1. *Interchange row i and row j*

$$\tilde{R}(\xi) = \begin{bmatrix} r_1(\xi) \\ \vdots \\ r_{i-1}(\xi) \\ r_j(\xi) \\ r_{i+1}(\xi) \\ \vdots \\ r_{j-1}(\xi) \\ r_i(\xi) \\ r_{j+1}(\xi) \\ \vdots \\ r_g(\xi) \end{bmatrix}.$$

2. *Multiply a row by a nonzero constant.* Let $0 \neq \alpha \in \mathbb{R}, 1 \leq i \leq g$, and define

$$\tilde{R}(\xi) = \begin{bmatrix} r_1(\xi) \\ \vdots \\ r_{i-1}(\xi) \\ \alpha r_i(\xi) \\ r_{i+1}(\xi) \\ \vdots \\ r_g(\xi) \end{bmatrix} \longleftarrow i\text{th row.}$$

3. *Replace a row by the sum of that row and the product of ξ^d and another row.* In terms of the equations, this means differentiate the jth equation d times and add the result to the ith equation. Let $d \in \mathbb{Z}_+, 1 \leq i \leq g, 1 \leq j \leq g, i \neq j$, and define

$$\tilde{R}(\xi) = \begin{bmatrix} r_1(\xi) \\ \vdots \\ r_{i-1}(\xi) \\ r_i(\xi) + \xi^d r_j(\xi) \\ r_{i+1}(\xi) \\ \vdots \\ r_g(\xi) \end{bmatrix} \longleftarrow i\text{th row.}$$

For all of these three types of elementary row operations it is clear that

$$R(\frac{d}{dt})w = 0 \text{ and } \tilde{R}(\frac{d}{dt})w = 0$$

have the same *strong* solutions, and therefore, by Corollary 2.4.12, they define the same behavior.

Each of the three elementary operations corresponds to premultiplication by a unimodular matrix. The first operation, interchange of two rows, corresponds to premultiplication by the matrix M that is obtained by interchanging the ith and jth columns in the identity matrix. Such a matrix is called a *permutation matrix*. The second elementary operation corresponds to replacing $R(\xi)$ by $DR(\xi)$, where D is the diagonal matrix

$$D = \operatorname{diag}\,(1, \ldots, 1, \underset{\substack{\uparrow \\ i\text{th place}}}{\alpha}, 1, \ldots, 1)\,, \tag{2.27}$$

while the third elementary operation corresponds to replacing $R(\xi)$ by $N(\xi)R(\xi)$, where $N(\xi)$ is the polynomial matrix

$$N(\xi) = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & \xi^d & 0 & \dots & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \longleftarrow i\text{th row.} \\ \\ \\ \\ \end{matrix}$$

$$\uparrow$$
$$j\text{th column}$$

(2.28)

The matrices $M, D, N(\xi)$ are called *elementary unimodular* matrices. Of course, applying a finite sequence of elementary row operations to $R(\xi)$ also generates equivalent representations. Now, by combining elementary row operations (that is, premultiplying $R(\xi)$ by a finite product of matrices like M, D, and $N(\xi)$), we obtain a very rich and flexible way of obtaining equivalent differential equations.

The question now arises of how general the class of unimodular matrices is that can be written as a finite product of elementary ones. It turns out that *every* unimodular matrix may be written as a product of elementary unimodular matrices.

Theorem 2.5.7 *$U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ is unimodular if and only if it is a finite product of elementary unimodular matrices.*

Proof The proof is not difficult, and the reader is encouraged to check it; see Appendix B, Theorem B.1.3. □

Remark 2.5.8 Combining Theorem 2.5.4 and 2.5.7 we conclude that two polynomial matrices $R(\xi)$ and $R'(\xi)$ represent the same behavior if $R'(\xi)$ can be obtained from $R(\xi)$ by a sequence of elementary row operations on $R(\xi)$. There are two other operations by means of which we obtain equivalent representations, namely adding and deleting zero-rows. Indeed, since rows in $R(\xi)$ that consist of zeros only do not contribute to the specification of the behavior we might as well delete or add such rows. Notice that adding and deleting zero-rows differ from elementary row operations in that elementary row operations leave the number of rows unchanged.

If $R(\xi)$ does not contain zero-rows, then we may be able to create a zero-row by applying elementary row operations. Subsequently we can then delete this zero-row. You may wonder if we can always create zero-rows or if we can check beforehand whether or not $R(\xi)$ contains "hidden" zero-rows. These questions lead to the notions of full row rank and minimal representation. We come back to this issue in Sections 2.5.6 and 3.6. See also Corollary 3.6.3. □

2.5.4 The Bezout identity

Unimodular matrices play a crucial role in the characterization of equivalent representations. We have seen that every unimodular matrix can be written as the product of elementary unimodular matrices. In terms of manipulations of equations, this means that $R_1(\frac{d}{dt})w = 0$ is equivalent to $R_2(\frac{d}{dt})w = 0$ if $R_1(\xi)$ can be transformed into $R_2(\xi)$ by means of elementary row operations. In practice, carrying out a whole sequence of elementary row operations can be quite cumbersome (and boring), and therefore we would like to know whether or not an arbitrary row operation can be replaced by a succession of elementary row operations. Let us illustrate this with an example.

Example 2.5.9 Consider the square matrix $R(\xi)$ given by

$$R(\xi) = \begin{bmatrix} -1+\xi^2 & 1+\xi^3 \\ 2+\xi^3 & -4+\xi^3 \end{bmatrix}.$$

Suppose we want to transform this matrix into upper triangular form by elementary row operations. To see that this can be done, first subtract ξ times the first row from the second: in more abstract notation, $r_2(\xi) := r_2(\xi) - \xi r_1(\xi)$. Then, replace in the resulting matrix $r_1(\xi)$ by $r_1(\xi) - \xi r_2(\xi)$, subsequently $r_1(\xi) := r_1(\xi) + 2r_2(\xi)$, then $r_2(\xi) := r_2(\xi) - \frac{1}{3}\xi r_1(\xi)$, and finally $r_2(\xi) := r_2(\xi) - \frac{2}{3}r_1(\xi)$. The result is that the first column of $R(\xi)$ has been transformed into

$$\begin{bmatrix} 3 \\ 0 \end{bmatrix}.$$

By applying these elementary row operations to the second column, the desired triangular form is obtained.

A much faster way to obtain an upper triangular form works as follows. Just replace $r_2(\xi)$ by $2+\xi^3$ times the first row plus $1-\xi^2$ times the second row. It is obvious that this row operation creates the desired zero at the lower left corner. However, it is not at all clear that the latter row operation corresponds to a sequence of elementary row operations. How can we check that this is the case? Recall that a row operation corresponds to premultiplication by a polynomial matrix. If this polynomial matrix is unimodular,

then this row operation is equivalent to a sequence of elementary ones. The row operation that we just described corresponds to premultiplication by a polynomial matrix of the form

$$U(\xi) = \begin{bmatrix} * & * \\ 2+\xi^3 & 1-\xi^2 \end{bmatrix}. \tag{2.29}$$

We have used the notation $*$ for entries that for the moment are left unspecified. If we assume that the row operation did not change the first row, then the first row of $U(\xi)$ has to be [1 0], which definitely does not yield a unimodular matrix. The question now is, Can we find polynomial entries for the $*$s in $U(\xi)$ such that the resulting matrix is unimodular? The answer to this question is *yes*, as shown by the following choice:

$$U(\xi) = \begin{bmatrix} -1+2\xi-\xi^2 & -2+\xi \\ 2+\xi^3 & 1-\xi^2 \end{bmatrix}. \tag{2.30}$$

It is readily verified that $U(\xi)$ in (2.30) is unimodular.

From this we may conclude that replacing $r_2(\xi)$ by $2+\xi^3$ times the first row plus $1-\xi^2$ times the second row is indeed the result of a sequence of elementary row operations. Two new questions arise. Firstly, is there a criterion purely in terms of the second row of (2.29) that allows us to conclude that there exist $*$s such that (2.29) is unimodular? Secondly, how can the $*$s be determined from the second row?

The answer to the first question is that since $2+\xi^3$ and $1-\xi^2$ have no nonconstant common factors, there exists a unimodular matrix of the form (2.29). This is shown in Theorem 2.5.10. Once it has been established that the desired unimodular matrix indeed exists, it remains to determine the entries explicitly. One way to do this is by means of elementary *column operations*: perform elementary column operations on the last row of (2.29) to obtain the vector [0 1]. The sequence of elementary column operations that achieves this is postmultiplication by a unimodular matrix $V(\xi)$ such that

$$\begin{bmatrix} 2+\xi^3 & 1-\xi^2 \end{bmatrix} V(\xi) = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

So take $U(\xi) = V^{-1}(\xi)$.

A second way of finding the $*$s is by substituting polynomials with unknown coefficients in the first row of (2.29) and determining the coefficients such that the determinant is a nonzero constant. For the particular case of a two-by-two matrix this can readily be done, since it yields linear equations in the unknown coefficients.

The determination of the unspecified entries in the unimodular matrix by means of elementary column operations is not very effective. For what is the advantage of elementary column operations applied to $U(\xi)$ as opposed to elementary row operations applied to $R(\xi)$ itself?

The answer to the second question, how to determine the $*$s, is that it suffices to know that a unimodular matrix $U(\xi)$ exists. But we don't have to calculate the $*$s to guarantee their existence. This is the content of the next result. □

Theorem 2.5.10 *Let $r_1(\xi), \ldots, r_k(\xi) \in \mathbb{R}[\xi]$ and assume that $r_1(\xi), \ldots, r_k(\xi)$ have no common[2] factor. Then there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{k\times k}[\xi]$ such that the last row of $U(\xi)$ equals $[r_1(\xi), \ldots, r_k(\xi)]$.*

Proof See Appendix B, Theorem B.1.6. □

Remark 2.5.11 As already indicated in Example 2.5.9, Theorem 2.5.10 shows that there exists a unimodular matrix of the form (2.29). This implies that there exist polynomials $a(\xi), b(\xi)$ such that when plugged into the first row of $U(\xi)$, $\det U(\xi) = 1$. In other words

$$a(\xi)(2 + \xi^3) + b(\xi)(1 - \xi^2) = 1. \tag{2.31}$$

Equation (2.31) is known as the *Bezout identity* or the *Bezout equation*. It plays an important role in algebra, and it is also very useful in our context. The more general Bezout identity for more than two polynomials follows easily from the proof of Theorem 2.5.10. □

A polynomial with real coefficients is called *monic* if the coefficient of its leading term is unity. The greatest common divisor of a set of polynomials $a_1(\xi), \ldots, a_k(\xi)$ is defined as the unique *monic* polynomial $g(\xi)$ that divides all the $a_j(\xi)$s and is such that *any* other polynomial with that property divides $g(\xi)$.

Corollary 2.5.12 (Bezout) *Let $r_1(\xi), \ldots, r_k(\xi) \in \mathbb{R}[\xi]$. Assume that the greatest common divisor of $r_1(\xi), \ldots, r_k(\xi)$ is 1, i.e., the k polynomials have no common factor, i.e., they are coprime. Then there exist polynomials $a_1(\xi), \ldots, a_k(\xi) \in \mathbb{R}[\xi]$ such that*

$$r_1(\xi)a_1(\xi) + \cdots + r_k(\xi)a_k(\xi) = 1.$$

Proof See Appendix B, Corollary B.1.7. □

2.5.5 Left and right unimodular transformations

We have seen that by multiplying a polynomial matrix from the left by a unimodular matrix, we transform given behavioral equations into equiva-

[2]By "no common factor" we always mean "no *nonconstant* common factor".

lent equations. We refer to these transformations as *left unimodular transformations*. If the behavior is represented by the matrix $R(\xi)$, then the transformed equations are represented by $U(\xi)R(\xi)$.

It is sometimes convenient to use *right unimodular transformations*, that is, to multiply $R(\xi)$ by a unimodular matrix from the right. Left transformations do not change the behavior. Right transformations, however, do change the behavior. The usefulness of applying right transformations lies in the fact that the behavior, although altered, remains structurally the same, since right transformations represent *isomorphisms*[3] of vector spaces.

Before we can apply right transformations, there is a technical difficulty to overcome. Consider the behavior $\mathfrak{B}$ defined by $R(\frac{d}{dt})w = 0$, where $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Let $V(\xi) \in \mathbb{R}^{q\times q}[\xi]$ be unimodular. We can postmultiply $R(\xi)$ by $V(\xi)$, resulting in $R(\xi)V(\xi)$. Of course, in general, the behavior $\mathfrak{B}'$ defined by $R(\frac{d}{dt})V(\frac{d}{dt})w = 0$ differs from the original behavior $\mathfrak{B}$. However, there is an obvious candidate for an isomorphism between $\mathfrak{B}$ and $\mathfrak{B}'$. To see this, choose $w \in \mathfrak{B}'$. Then $R(\frac{d}{dt})V(\frac{d}{dt})w = 0$, and this implies that $V(\frac{d}{dt})w$ belongs to $\mathfrak{B}$. Conversely, if $w \in \mathfrak{B}$, then $V^{-1}(\frac{d}{dt})w \in \mathfrak{B}'$. Notice that since $V(\xi)$ is unimodular, $V^{-1}(\xi)$ is again a polynomial matrix. It seems that $V(\frac{d}{dt})$ indeed defines an isomorphism between $\mathfrak{B}$ and $\mathfrak{B}'$. But there is a snag in this reasoning. Indeed, the operator $V(\frac{d}{dt})$ is only defined for those elements of $\mathfrak{B}$ that are sufficiently smooth. Since by Theorem 2.4.10 a behavior is completely determined by its $\mathcal{C}^\infty$ part, on which $V(\frac{d}{dt})$ is perfectly well defined, there is a way out: we just restrict our attention to the $\mathcal{C}^\infty$ parts of $\mathfrak{B}$ and $\mathfrak{B}'$.

For future reference we formulate these observations in a theorem.

Theorem 2.5.13 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and $V(\xi) \in \mathbb{R}^{q\times q}[\xi]$. Denote the behaviors defined by $R(\frac{d}{dt})w = 0$ and $R(\frac{d}{dt})V(\frac{d}{dt})w = 0$ by $\mathfrak{B}$ and $\mathfrak{B}'$ respectively. If $V(\xi)$ is unimodular, then $\mathfrak{B}\cap\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^q)$ and $\mathfrak{B}'\cap\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^q)$ are isomorphic as vector spaces.*

There exists a subclass of right unimodular transformations for which we do not have to restrict the isomorphism to the $\mathcal{C}^\infty$ part of the behavior, namely those that correspond to unimodular matrices that do not depend on ξ, in other words, matrices that are invertible in $\mathbb{R}^{q\times q}$. These transformations replace the variable w by Vw, where V is a nonsingular matrix in $\mathbb{R}^{q\times q}$, and they are called *static right unimodular transformations* , since they do not involve differentiations of the trajectories in the behavior. Static right unimodular transformations are useful in the context of input/output systems, as will be shown in Theorem 3.3.24.

---

[3] An isomorphism of two vector spaces is a bijective linear map from the first to the second.

Left unimodular transformations can be used to bring the polynomial matrix in a more suitable form, like the *upper triangular form.*

Theorem 2.5.14 (Upper triangular form) *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. There exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that $U(\xi)R(\xi) = T(\xi)$ and $T_{ij}(\xi) = 0$ for $i = 1, \ldots, g,\ j < i$.*

Proof See Appendix B, Theorem B.1.1. □

By applying both left and right unimodular transformations we can bring a polynomial matrix into an even more convenient form, called the *Smith form.*

Theorem 2.5.15 (Smith form, square case) *Let $R(\xi) \in \mathbb{R}^{g\times g}[\xi]$. There exist unimodular matrices $U(\xi), V(\xi) \in \mathbb{R}^{g\times g}$ such that*

1. *$U(\xi)R(\xi)V(\xi) = \operatorname{diag}(d_1(\xi), \ldots, d_g(\xi))$.*

2. *There exist (scalar) polynomials $q_i(\xi)$ such that $d_{i+1}(\xi) = q_i(\xi)d_i(\xi)$, $i = 1, \ldots, g-1$.*

Proof See Appendix B, Theorem B.1.4. □

Remark 2.5.16 The first part of Theorem 2.5.15 states that every square polynomial matrix may be transformed into a diagonal matrix by means of pre- and postmultiplication by suitable unimodular matrices. The second part claims that in addition, the elements down the diagonal divide each other. In fact, the diagonal form, i.e., without the division property, suffices for the applications that we will encounter. The main reasons why we also give the full result are convenience of reference, that the proof is not more difficult, and that it is useful in the analysis in Chapter 3; see Remark 3.2.19.

It is not difficult to see that if we require the nonzero diagonal elements to be monic, then the Smith form is unique: that is, for a given matrix $R(\xi)$ there exists exactly one matrix that satisfies the properties 1 and 2 in Theorem 2.5.15. □

Remark 2.5.17 If $R(\xi)$ is nonsquare, then the Smith form is also defined and is obtained via the same algorithm. If $R(\xi)$ is *wide* ($g < q$) or *tall*

$(g > q)$, the Smith forms are given by

$$\begin{bmatrix} d_1(\xi) & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & d_g(\xi) & 0 & \cdots & 0 \end{bmatrix}, \quad \begin{bmatrix} d_1(\xi) & & & \\ & \ddots & & \\ & & & d_q(\xi) \\ 0 & \cdots & 0 & \\ \vdots & & \vdots & \\ 0 & \cdots & 0 & \end{bmatrix}$$

respectively. □

2.5.6 Minimal and full row rank representations

We have seen that different sets of equations may define the same behavior. Given this fact, it is natural to look for representations that are in some sense as simple as possible. In this section we concentrate on one particular feature of a representation, namely the parameter g, the number of equations needed to describe a given behavior. The first step is to start from a representation $R(\xi)$ and try to reduce the number of equations by creating zero-rows by means of left unimodular transformations. This leads to the notion of *full row rank* representation. Loosely speaking, this is a representation in which the number of rows cannot be reduced any further. The precise definition follows shortly.

A representation is called *minimal* if the number of rows is minimal among all possible equivalent representations. At this stage it is not clear whether the reduction of a representation to a full row rank representation leads to a minimal one. Yet this turns out to be true. Before we discuss the notions that we use in the development, we present a simple example.

Example 2.5.18 Let $R(\xi)$ be given by

$$R(\xi) = \begin{bmatrix} -1+\xi & 1+\xi \\ -2+\xi+\xi^2 & 2+3\xi+\xi^2 \end{bmatrix}.$$

Let $\mathfrak{B}$ denote the behavior represented by $R(\frac{d}{dt})w = 0$. Subtracting $\xi + 2$ times the first row from the second yields

$$\begin{bmatrix} -1+\xi & 1+\xi \\ 0 & 0 \end{bmatrix}. \tag{2.32}$$

$\mathfrak{B}$ is thus also represented by (2.32). Because a zero-row does not impose any restriction on w, we could as well delete it. Therefore, $\mathfrak{B}$ is also represented by

$$\begin{bmatrix} -1+\xi & 1+\xi \end{bmatrix}.$$

Intuitively, it is clear that no further reduction is possible. In more complicated situations, we would like to have a precise criterion on the basis of which we can conclude that we are dealing with the minimal number of equations. □

Definition 2.5.19 (Independence) The polynomial *vectors* $r_1(\xi), \ldots, r_k(\xi)$ are said to be *independent* over $\mathbb{R}[\xi]$ if

$$\sum_{j=1}^{k} a_j(\xi) r_j(\xi) = 0 \quad \Leftrightarrow \quad a_1(\xi) = \cdots = a_k(\xi) = 0.$$

Here $a_1(\xi), \ldots, a_k(\xi)$ are scalar polynomials. □

Remark 2.5.20 Notice the analogy of Definition 2.5.19 with the notion of independence of vectors over a field. See, however, Exercise 2.26. □

Definition 2.5.21 (Row rank and column rank) Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The *row rank* (*column rank*) of $R(\xi)$ is defined as the maximal number of independent rows (columns). We say that $R(\xi)$ has *full row rank* (*full column rank*) if the row rank (column rank) equals the number of rows (columns) in $R(\xi)$. □

Theorem 2.5.22 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Denote the row rank and the column rank of $R(\xi)$ by k_r and k_c respectively.*

1. *The integers k_r, k_c are invariant with respect to pre- and postmultiplication by unimodular matrices.*
2. *$k_r = k_c$, and hence we can speak about the* rank *of $R(\xi)$.*
3. *There exists a $k_r \times k_r$ submatrix of $R(\xi)$ with nonzero determinant.*

Proof (1) The statements are easily verified for pre- and postmultiplication by *elementary* unimodular matrices, see Exercise 2.25. Since every unimodular matrix is the product of elementary unimodular matrices (Theorem 2.5.7), the result follows.

(2) Choose unimodular matrices $U(\xi), V(\xi)$ such that $U(\xi)R(\xi)V(\xi)$ is in Smith form. By part 1, we know that the row and column ranks of $R(\xi)$ and $U(\xi)R(\xi)V(\xi)$ agree. Obviously, both the row and column ranks of $U(\xi)R(\xi)V(\xi)$ equal the size of the nonzero diagonal part of that matrix. This implies that $k_r = k_c$.

(3) Choose k_r independent rows to form the matrix $R_r(\xi)$. The row rank, and hence the column rank, of $R_r(\xi)$ is equal to k_r, and hence we can select k_r independent columns of $R_r(\xi)$. From these columns we form the square

matrix $R_{rc}(\xi)$. Since $R_{rc}(\xi)$ has full row rank, it follows from its Smith form that its determinant is not the zero polynomial. □

Theorem 2.5.23 *Every behavior $\mathfrak{B}$ defined by $R(\frac{d}{dt})w = 0$, $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ admits an equivalent full row rank representation, that is, there exists a representation $\tilde{R}(\frac{d}{dt})w = 0$ of $\mathfrak{B}$ with $\tilde{R}(\xi) \in \mathbb{R}^{\tilde{g}\times q}$ of full row rank.*

Proof If $R(\xi)$ has full row rank, there is nothing to prove. Suppose that $R(\xi)$ does not have full row rank. Denote the rows of $R(\xi)$ by $r_1(\xi), \ldots, r_g(\xi)$. By definition of dependence there exist nonzero scalar polynomials $a_1(\xi), \ldots, a_g(\xi)$ such that

$$\sum_{j=1}^{g} a_j(\xi) r_j(\xi) = 0. \tag{2.33}$$

We may assume that $a_1(\xi), \ldots, a_g(\xi)$ have no common nonconstant factor (otherwise we could divide the $a_j(\xi)$s by this factor without changing (2.33)). By Theorem 2.5.10 there exists a unimodular matrix $U(\xi)$ with last row $[a_1(\xi), \ldots, a_g(\xi)]$. From (2.33) it follows that the last row of $U(\xi)R(\xi)$ is the zero-row. So as long as a matrix does not consist of zero-rows and independent rows, we can create at least one more zero-row by premultiplication by a suitable unimodular matrix. This implies that there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that

$$U(\xi)R(\xi) = \begin{bmatrix} \tilde{R}(\xi) \\ 0 \end{bmatrix}$$

with $\tilde{R}(\xi) \in \mathbb{R}^{\tilde{g}\times q}[\xi]$ of full row rank. Of course, the behavior defined by $R(\frac{d}{dt})w = 0$ equals the behavior defined by $\tilde{R}(\frac{d}{dt})w = 0$, since deletion of zero-rows does not change the behavior. This completes the proof. □

Definition 2.5.24 (Minimality) Let the behavior $\mathfrak{B}$ be defined by $R(\frac{d}{dt})w = 0$, $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The representation $R(\frac{d}{dt})w = 0$ is called *minimal* if every other representation has at least g rows, that is, if $w \in \mathfrak{B}$ if and only if $R'(\frac{d}{dt})w = 0$ for some $R'(\xi) \in \mathbb{R}^{g'\times q}[\xi]$ implies $g' \geq g$. □

The following result is almost immediate.

Theorem 2.5.25 *If $R(\xi)$ is a minimal representation of $\mathfrak{B}$, then $R(\xi)$ has full row rank.*

Proof Suppose that $R(\xi)$ does not have full row rank. Then there exists a unimodular matrix $U(\xi)$ such that

$$U(\xi)R(\xi) = \begin{bmatrix} R'(\xi) \\ 0 \end{bmatrix}. \tag{2.34}$$

Of course, $\mathfrak{B}$ is also represented by (2.34) and hence by $R'(\frac{d}{dt})w = 0$. Since the number of rows of $R'(\xi)$ is strictly smaller than the number of rows of $R(\xi)$, the latter cannot be minimal. □

As stated in Theorem 2.5.25, minimality implies full row rank, but the converse is also true: *full row rank implies minimality*. Moreover all minimal representations may be transformed into each other by means of left unimodular multiplication. This implies in particular that all minimal representations have the same number of rows. The proofs of these statements have to be postponed until the end of Chapter 3, Theorem 3.6.4, because we need to develop some other ingredients of the theory of linear dynamical systems first.

2.6 Recapitulation

In this chapter we introduced the main class of dynamical systems that we will deal with in this book: systems described by linear constant-coefficient differential equations of the form $R(\frac{d}{dt})w = 0$, with $R(\xi)$ a polynomial matrix. We explained why the classical notion of a solution is inadequate for (engineering) applications. In order to accommodate this difficulty, the notion of a *weak solution*, defined in terms of an integral equation associated with the differential equation, was introduced.

The main points of Chapter 2 are:

- Every strong solution is a weak solution, and every weak solution that is sufficiently smooth is a strong solution (Theorem 2.3.11).
- Every weak solution can be approximated by a sequence of infinitely differentiable ones. In other words, the C^∞ part of the behavior is dense in the behavior (Corollary 2.4.12).
- The systems under consideration, i.e., those described by behavioral differential equations of the form $R(\frac{d}{dt})w = 0$, are linear and time-invariant (Theorem 2.4.15).
- If there exists a unimodular polynomial matrix $U(\xi)$ such that $R_2(\xi) = U(\xi)R_1(\xi)$, then $R_1(\frac{d}{dt})w = 0$ and $R_2(\frac{d}{dt})w = 0$ represent the same behavior (Theorem 2.5.4). Such representations are called equivalent. In Chapter 3 we will show that the converse is also true: if $R_1(\xi)$ and $R_2(\xi)$ represent the same behavior and if they have the same number of rows, then there exists a unimodular matrix $U(\xi)$ such that $U(\xi)R_2(\xi) = R_1(\xi)$.
- We introduced the concept of *minimal* and *full row rank* representation. Each system of differential equations $R(\frac{d}{dt})w = 0$ is equivalent to one in which the corresponding polynomial matrix has a minimal number of rows among all possible equivalent representations. Such a minimal representation is also of full row rank. Thus "minimal" implies "full row rank". The converse is also true. This will be proved in Chapter 3.

2.7 Notes and References

The material of this chapter requires a mathematical background that goes beyond what is offered in standard calculus courses. Introductory books that may serve as background are [51] for the analysis part and [34] for the algebra part. A more advanced book on matrices with entries in a Euclidean domain is [43]. Pioneering books for the use of polynomial matrices in system theory are [8, 48, 63]. The results on equivalent and minimal representations were first brought forward in [60].

2.8 Exercises

2.1 Consider Example 2.3.10. Prove that the function (w_1, w_2) given by (2.15) is a weak solution of (2.13).

2.2 Assume that a mass M at rest is hit by a unit force F at $t = 0$:

$$F(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0. \end{cases}$$

Assume that the motion of the mass obeys Newton's law:

$$M \frac{d^2}{dt^2} q = F. \tag{2.35}$$

Compute the resulting displacement q as a function of time. Do you obtain a weak or a strong solution of (2.35)? Repeat the question for

$$F(t) = \begin{cases} 0 & t < 0, \\ t & t \geq 0. \end{cases}$$

2.3 Consider the RC-circuit of Example 2.3.1. Assume for ease of calculation that $R_0 = R_1 = 1$, $C = 1$. Prove that with the port voltage

$$V(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0 \end{cases}$$

the current

$$I(t) = \begin{cases} 0 & t < 0, \\ \frac{1}{2} + \frac{1}{2} e^{-2t} & t \geq 0 \end{cases}$$

yields a weak solution of (2.6).

Assume instead that a current source

$$I(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0 \end{cases}$$

is switched on at $t = 0$. Compute, analogously to the previous situation, a corresponding weak solution of (2.6).

2.4 A permutation matrix is a matrix in which each row and each column contains exactly one entry that is equal to one and in which all other entries are equal to zero. Consider the permutation matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

(a) Show that it is a unimodular matrix.

(b) Write it as the product of elementary factors.

(c) Prove that every permutation matrix can be written as the product of matrices of the type (2.27) and (2.28).

2.5 Let the polynomial matrix $U(\xi)$ be given by

$$\begin{bmatrix} -1+\xi^2+\xi^3 & \xi \\ \xi+\xi^2 & 1 \end{bmatrix}$$

Is $U(\xi)$ unimodular? Write it as a product of elementary matrices.

2.6 Let $r_1(\xi) = 2 - 3\xi + \xi^2$, $r_2(\xi) = 6 - 5\xi + \xi^2$, and $r_3(\xi) = 12 - 7\xi + \xi^2$.

(a) Find a unimodular matrix $U(\xi) \in \mathbb{R}^{3\times 3}[\xi]$ such that the last row of $U(\xi)$ equals $[r_1(\xi), r_2(\xi), r_3(\xi)]$.

(b) Find polynomials $a_1(\xi), a_2(\xi), a_3(\xi)$ such that

$$r_1(\xi)a_1(\xi) + r_2(\xi)a_2(\xi) + r_3(\xi)a_3(\xi) = 1.$$

2.7 In the definition of unimodular matrix, Definition 2.5.5, it is stated that a polynomial matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ has a polynomial inverse $V(\xi) \in \mathbb{R}^{g\times g}[\xi]$ if and only if $\det U(\xi)$ equals a nonzero constant. Prove this statement. Hint: for the sufficiency part use the discussion just below Definition 2.5.5.

2.8 Consider the differential system

$$\begin{aligned} -w_1 + \frac{d^2}{dt^2}w_1 + w_2 + \frac{d}{dt}w_2 &= 0, \\ -\frac{d}{dt}w_1 + \frac{d^2}{dt^2}w_1 + \frac{d}{dt}w_2 &= 0. \end{aligned}$$

Is it a full row rank representation? If not, construct an equivalent full row rank representation.

2.9 Consider the electrical circuit studied in Example 1.3.5. Write the equations (1.1, 1.2, 1.3) in Section 1.3 in polynomial matrix form

$$R(\frac{d}{dt})w = 0$$

with $w = \mathrm{col}(V, I, V_{R_C}, I_{R_C}, V_{R_L}, I_{R_L}, V_C, I_C, V_L, I_L)$.

2.10 Consider the polynomial matrix obtained in Exercise 2.9. Is the resulting system minimal in the sense of Definition 2.5.24? Prove that in this case you can obtain an equivalent minimal system by simply dropping two equations. (We have seen in Example 2.5.18 that in general, reduction to minimal form requires more involved manipulations than simply dropping redundant equations.)

2.11 Assume that $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and that

$$R(\frac{d}{dt})w = 0$$

is a minimal representation. Prove that $g \leq q$. Conclude that every system can hence be represented by a number of equations that is less than or equal to the number of components of w.

2.12 Prove Lemma 2.3.9.

2.8.1 Analytical problems

Below we have listed some exercises that fill in the gaps that were left in the proofs of some of the analytical results. As they do not really have any system-theoretic significance, we list these exercises separately.

2.13 Define the functions $w_n \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ as

$$w_n(t) = \begin{cases} 0 & |t| < n, \\ n & |t| \geq n. \end{cases}$$

Prove that w_n converges to the zero function in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$.

2.14 (a) Does *pointwise* convergence ($\lim_{k\to\infty} w_k(t) = w(t)$ for all t) imply convergence in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$? Provide a counterexample.

(b) The sequence of functions $\{w_k\}$ is said to converge to w uniformly in t if for all $\epsilon > 0$ there exists an N such that for all t and for all $k \geq N$, $\|w_k(t) - w(t)\| < \epsilon$. The difference with pointwise convergence is that N is not allowed to depend on t. Equivalently, $\lim_{k\to\infty}(\sup_t \|w_k(t) - w(t)\|) = 0$. Does *uniform* convergence imply convergence in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$?

2.15 Complete the proof of Theorem 2.4.10.

2.16 Let $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ be a weak solution of $R(\frac{d}{dt})w = 0$, where $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$.

(a) Let $\psi_i \in \mathbb{R}$ and $\tau_i \in \mathbb{R}$, $i = 1,\ldots,N$. Prove that $\sum_{i=1}^N \psi_i w(t-\tau_i)$ is a weak solution of $R(\frac{d}{dt})w = 0$.

(b) Assume that $\psi \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ is such that

$$(\psi * w)(t) := \int_{-\infty}^{\infty} \psi(\tau)w(t-\tau)d\tau \tag{2.36}$$

is a well-defined integral with $\psi * w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$. Prove that $\psi * w$ is also a weak solution of $R(\frac{d}{dt})w = 0$.

Note: In Section 2.4.1, Lemma 2.4.9, we showed that the integral (2.36) is always well-defined and belongs to the behavior for a special choice of ψ. In particular, ψ had compact support. Other conditions under which (2.36) is obviously well-defined are, for example, that there exists $t' \in \mathbb{R}$ such that $\psi(t) = 0$ for $t > t'$ and $w(t) = 0$ for $t < t'$, or if $w \in \mathfrak{L}_1(\mathbb{R},\mathbb{R}^q)$ and $\psi \in \mathfrak{L}_1(\mathbb{R},\mathbb{R})$.

2.17 Show that the function ϕ defined in (2.18) is infinitely differentiable.

2.18 Let $\psi \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ and $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$. Define v by

$$v(t) := \int_0^t w(s)ds.$$

Show that

$$\int_0^t (\psi * w)(\tau)d\tau = (\psi * v)(t).$$

2.19 Prove that integration is a continuous operation on $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$; i.e., show that if $\lim_{k\to\infty} w_k = w$ in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$, see Definition 2.4.2, then

$$\lim_{k\to\infty} \int_a^b w_k(t)dt = \int_a^b w(t)dt.$$

This fact is used in the proof of Theorem 2.4.4.

2.20 Let $w_k(t) = c_{0,k} + \cdots + c_{n,k}t^n$, $c_{i,k} \in \mathbb{R}^q$. Prove that if the sequence $\{w_k\}$ converges in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$, then the sequences $c_{i,k}$ converge in $\mathbb{R}^q$ for all i. For simplicity you may confine yourself to the case $q = 1, n = 1$.

2.21 Prove Theorem 2.3.11.

2.8.2 Algebraic problems

Below we have listed some exercises that fill in the gaps that were left in the proofs of some of the algebraic results. As they do not really have any system-theoretic significance, we have listed them separately.

2.22 Prove Theorem 2.5.22, part 1, for elementary unimodular matrices.

2.23 Let $a(\xi), b(\xi) \in \mathbb{R}[\xi]$ be polynomials. Prove that $a(\xi)$ and $b(\xi)$ are coprime if and only if there exist polynomials $p(\xi)$ and $q(\xi)$ such that $a(\xi)p(\xi) + b(\xi)q(\xi) = 1$.

2.24 Let $a(\xi), b(\xi) \in \mathbb{R}[\xi]$ be polynomials of degree n and m respectively. Assume that $a(\xi)$ and $b(\xi)$ are coprime, i.e., they have no nonconstant common factors.

(a) Use Corollary 2.5.12 to conclude that for every polynomial $c(\xi) \in \mathbb{R}[\xi]$ there exist polynomials $p(\xi)$ and $q(\xi)$ such that

$$a(\xi)p(\xi) + b(\xi)q(\xi) = c(\xi). \tag{2.37}$$

(b) Suppose that $\deg c(\xi) < n + m$. Prove that the previous statement remains true if we require that $\deg p(\xi) < m$ and $\deg q(\xi) < n$. Hint: Assume that $\deg p(\xi) \geq m$. Apply division with remainder of $p(\xi)$ by $b(\xi)$ ($p(\xi) = f(\xi)b(\xi) + \tilde{p}(\xi)$) to obtain a polynomial $\tilde{p}(\xi)$ of degree strictly smaller than m. Rewrite (2.37) as $a(\xi)\tilde{p}(\xi) + b(\xi)\tilde{q}(\xi) = c(\xi)$ for a suitable choice of $\tilde{q}(\xi)$ and argue by checking the degrees of the left- and right-hand sides that $\deg \tilde{q}(\xi) < n$.

(c) Assume that $\deg c(\xi) < n + m$. Prove that $p(\xi)$ with $\deg p(\xi) < m$ and $q(\xi)$ with $\deg q(\xi) < n$ such that (2.37) is satisfied are unique. Hint: In the previous part you have just proved that the linear map L that assigns to the pair $(p(\xi), q(\xi))$ the polynomial $a(\xi)p(\xi)+b(\xi)q(\xi)$ is *surjective*. Use the fact that a linear map between vector spaces of the same (finite) dimension is injective if and only if it is surjective to conclude that L is also *injective*.

(d) Now suppose that $a(\xi)$ and $b(\xi)$ are not coprime. Prove that there exist polynomials $c(\xi)$ such that (2.37) has no solution $p(\xi), q(\xi)$. Under what condition on $c(\xi)$ is (2.37) solvable?

2.25 Consider $\mathbb{R}^{g\times g}[\xi]$. Obviously, addition, $+$, and multiplication, $\bullet$, each define binary operations on $\mathbb{R}^{g\times g}[\xi]$. Prove that $(\mathbb{R}[\xi], \bullet, +)$ defines a ring. Let $\mathfrak{U} = \{U(\xi) \in \mathbb{R}^{g\times g}[\xi] \mid U(\xi) \text{ is unimodular}\}$. Prove that $(\mathfrak{U}, \bullet)$ forms a group.

2.26 If the *real* vectors $v_1, \ldots, v_k \in \mathbb{R}^n$ are linearly dependent over $\mathbb{R}$, then at least one of these vectors can be written as a linear combination (over $\mathbb{R}$) of the others. Show by means of an example that this is not true for *polynomial* vectors. See Definition 2.5.19 for independence of polynomials. Hint: Consider $v_1(\xi) := [\xi\ \xi^2]^T$, $v_2 := [1+\xi\ \ \xi+\xi^2]^T$.

3

Time Domain Description of Linear Systems

3.1 Introduction

In Chapter 2 we studied behaviors described by equations of the form $R(\frac{d}{dt})w = 0$. We obtained fundamental properties such as linearity, time-invariance, and the like, as well as the relation between the behavior and its representations. What we did not do, however, is pay attention to what the trajectories in the behavior, the weak solutions of $R(\frac{d}{dt})w = 0$, actually look like.

The first goal of this chapter is to give a complete and explicit characterization of all weak solutions of $R(\frac{d}{dt})w = 0$. This is done in two steps. In the first step we treat *autonomous* systems, that is, the case where $R(\xi)$ is a square polynomial matrix with nonzero determinant. In the second step the general case is covered. This leads to the notion of input/output representation. Loosely speaking, this means that we can split the trajectories w into two components, w_1 and w_2, one component that can be chosen freely, called the *input*, and the other, called the *output*, the future of which is completely determined by its past and the choice of the input.

The second goal of this chapter is to study an alternative representation of input/output systems, namely through convolution.

As an application of these results, we prove the claim made in Chapter 2 that two matrices $R_1(\xi)$ and $R_2(\xi)$ with the same number of rows represent the same behavior if and only if $R_2(\xi) = U(\xi)R_1(\xi)$ for some unimodular matrix $U(\xi)$. From that it follows that all minimal representations, equivalently all

full row rank representations, of a given behavior may be transformed into each other by multiplication from the left by a unimodular polynomial matrix.

The outline of the chapter is as follows. In Section 3.2 we consider the case where $R(\xi)$ is square and has nonzero determinant. The scalar case is treated first. The results obtained there are used to analyze the multivariable case. In Section 3.3 we define and study systems in *input/output* form. Also in this section the scalar case is treated first. By scalar we now mean that both the input and output variables are scalar. These systems are referred to as single-input/single-output (SISO) systems. Subsequently, we obtain the important result, valid for multivariable systems (i.e., $q \geq 2$), that every behavior can be written in input/output form. In Section 3.4 we study convolution systems, and in Section 3.5 we relate convolution systems to input/output systems described by differential equations. Section 3.6 contains the counterpart of Theorem 2.5.4. By their nature, the results of Section 3.6 belong to Chapter 2. Unfortunately, the results of Section 3.6 could not be given earlier, since they rely on elements of previous sections of the present chapter. In that respect, they also form nice applications of some of the results obtained in this chapter.

3.2 Autonomous Systems

In Chapter 2 we have defined behaviors described by systems of differential equations $R(\frac{d}{dt})w = 0$, where $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The $q\times q$ case is of special interest. For notational reasons that will become clear later, we prefer to denote $R(\xi)$ by $P(\xi)$ in this case. The corresponding system of differential equations that we study in this section is

$$P(\frac{d}{dt})w = 0 \tag{3.1}$$

with $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and $\det P(\xi) \neq 0$. With $\det P(\xi) \neq 0$, we mean that $\det P(\xi)$ is not equal to the zero polynomial. Of course, if $\deg \det P(\xi)$ is not zero, then $\det P(\xi)$ has roots. It turns out that in this case the behavior can be described quite explicitly. In fact, the roots of $\det P(\xi)$ play a leading role in this description. Thus the problem at hand is to describe the solution set of the system of differential equations (3.1), that is, the behavior of the dynamical system represented by it. The expression for the behavior becomes (notationally) much simpler if we consider differential equations with complex coefficients rather than with real coefficients. The reason for this is that in $\mathbb{C}$ every polynomial can be written as the product of first-order factors. In large parts of this chapter we hence assume that $P(\xi) \in \mathbb{C}^{q\times q}[\xi], \det P(\xi) \neq 0$, and of course we obtain an expression for all solutions $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{C}^q)$ of the differential equation $P(\frac{d}{dt})w = 0$. It is easy

to treat the real case from there. If $P(\xi)$ happens to have real coefficients, then it simply suffices to take the real part of the complex solutions w.

We will first determine the set of *strong solutions* of $P(\frac{d}{dt})w = 0$. Subsequently, we show that every *weak* solution is equivalent to a strong one, or stated otherwise, that for every weak solution there exists a strong solution such that they agree everywhere except on a set of measure zero (see Definition 2.3.6).

The main result of this section is that solutions of (3.1) are completely determined by their past. That is, if two solutions of $P(\frac{d}{dt})w = 0$ agree on the time interval $(-\infty, 0]$, then they agree on the whole time axis $\mathbb{R}$. Behaviors with this property are called *autonomous*.

Definition 3.2.1 A behavior $\mathfrak{B}$ is called *autonomous* if for all $w_1, w_2 \in \mathfrak{B}$

$$w_1(t) = w_2(t) \text{ for } t \leq 0 \quad \Rightarrow \qquad w_1(t) = w_2(t) \text{ for almost all } t.$$

In words: the future of every trajectory is completely determined by its past. □

The idea of Definition 3.2.1 is best illustrated by means of an example.

Example 3.2.2 Consider the mass–damper–spring system of Figure 3.1. The equation that describes the behavior of the displacement of the mass

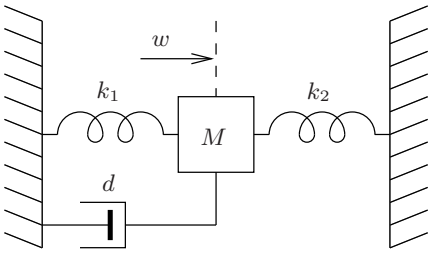


FIGURE 3.1. Autonomous mass–damper–spring system.

with respect to its equilibrium is

$$(k_1 + k_2)w + d\frac{d}{dt}w + M(\frac{d}{dt})^2 w = 0. \tag{3.2}$$

Mathematically speaking, it is clear that (3.2) defines an autonomous system. For suppose that we have two solutions w_1 and w_2 of (3.1) such that $w_1(t) = w_2(t)$ for $t \leq 0$. Since we are dealing with a linear system, $w := w_1 - w_2$ also belongs to the corresponding behavior. It follows from the theory of ordinary differential equations that since $w(t) = 0$ for $t \leq 0$, and w satisfies (3.2), w is identically zero. This implies that $w_1 = w_2$. In

fact, the solution of (3.2) for $t > 0$ is completely determined by $w(0)$ and $(\frac{d}{dt}w)(0)$. In turn, these initial conditions are obviously determined by $w(t)$ for $t \leq 0$. This provides a somewhat different, though related, explanation for the fact that (3.2) defines an autonomous behavior. Notice, however, that we have used arguments borrowed from theory that we did not provide in this book. In the sequel we derive that (3.2) defines an autonomous system by different means. Physically, the autonomous nature of the system is explained by the observation that once the mass has been in its equilibrium position in the past, it remains there forever. The only way to move the mass from its equilibrium position is to act on it with an external force. Such an action involves a corresponding external variable, which is not modeled by (3.2). In the next example we see that if we incorporate an external force in the model, then the system is no longer autonomous. □

Example 3.2.3 As an example of a non autonomous system, consider the mass–spring system in Figure 3.2. The difference with Example 3.2.2 is

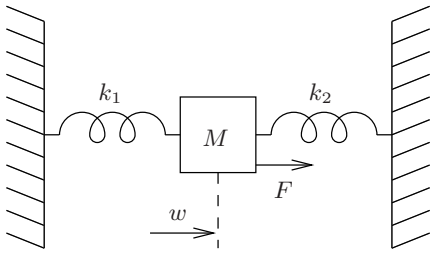


FIGURE 3.2. Non autonomous mass–spring system.

that now an external force can act on the mass. Denote this force by F. The equation describing the behavior becomes

$$(k_1 + k_2)w + M(\frac{d}{dt})^2 w = F. \tag{3.3}$$

To see that the behavior defined by (3.3) is not autonomous, it suffices to show that there exists a nonzero solution (w, F) of (3.3) that is zero on $(-\infty, 0]$. For convenience, assume that $M = 1$ and $k_1 = k_2 = \frac{1}{2}$. Take

$$w(t) = \begin{cases} 0 & t < 0 \\ 1 - \cos(t) & t \geq 0 \end{cases}, \qquad F(t) = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}. \tag{3.4}$$

It is not difficult to check that the pair (w, F) defined by (3.4) is a (weak) solution of (3.3). Because $(w, F) = (0, 0)$ is also a solution of (3.3), it follows that trajectories in the behavior are not completely determined by their past. Therefore, (3.4) viewed as a dynamical system in the variable (w, F) defines a non autonomous system. □

3.2.1 The scalar case

In order to get going, let us consider first the case $q = 1$. Let n denote the degree of $P(\xi)$, say $P(\xi) = P_0 + P_1\xi + \cdots + P_{n-1}\xi^{n-1} + P_n\xi^n$ with $P_n \neq 0$. The problem at hand is to describe the solution set of the scalar differential equation

$$P_0 w + P_1 \frac{d}{dt} w + \cdots + P_{n-1} \frac{d^{n-1}}{dt^{n-1}} w + P_n \frac{d^n}{dt^n} w = 0, \tag{3.5}$$

where $P_0, P_1, \ldots, P_{n-1}, P_n \in \mathbb{C}$. Note that (3.5) is indeed the scalar version of (3.1).

Before we characterize the behavior defined by (3.5), we derive that every *weak* solution of (3.5) is equal to an infinitely differentiable solution almost everywhere and therefore also to a *strong* solution almost everywhere. As a pleasant consequence, we may subsequently confine ourselves to strong solutions, which, as we will see, facilitates the analysis considerably.

Theorem 3.2.4 *Let $\mathfrak{B}$ be the behavior defined by $P(\frac{d}{dt})w = 0$, where $0 \neq P(\xi) \in \mathbb{R}[\xi]$. For every $w \in \mathfrak{B}$, there exists a $v \in \mathfrak{B}$ that is infinitely differentiable and such that $w = v$ almost everywhere.*

Proof Let $P(\xi)$ be given by

$$P(\xi) = P_0 + P_1\xi + \cdots + P_n\xi^n, \quad P_n \neq 0.$$

Every weak solution of $P(\frac{d}{dt})w = 0$ satisfies the associated integral equation

$$P_L w(t) + P_{L-1} \int_0^t w(\tau)d\tau + \cdots + P_0 \int_0^t \cdots \int_0^{\tau_{L-1}} w(\tau_L)d\tau_L \cdots d\tau_1 = r_{L-1}(t)$$

for some polynomial r_{L-1} of degree at most $L - 1$. Equivalently,

$$P_L w(t) = -P_{L-1} \int_0^t w(\tau)d\tau - \cdots - P_0 \int_0^t \cdots \int_0^{\tau_{L-1}} w(\tau_L)d\tau_L \cdots d\tau_1 + r_{L-1}(t). \tag{3.6}$$

Since $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$, it follows that the right-hand side of (3.6) is $\mathcal{C}^0(\mathbb{R}, \mathbb{R})$ (continuous), and hence, since $P_L \neq 0$, we conclude that w equals a function in $\mathcal{C}^0(\mathbb{R}, \mathbb{R})$ almost everywhere. But this implies that the right-hand side of (3.6) is $\mathcal{C}^1(\mathbb{R}, \mathbb{R})$ (continuously differentiable), and it follows that w equals a function in $\mathcal{C}^1(\mathbb{R}, \mathbb{R})$ almost everywhere. Repeating this argument yields that w equals a function $v \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ almost everywhere. Finally, Theorem 2.3.11, part 2, implies that v is a strong solution of $P(\frac{d}{dt})w = 0$. □

Theorem 3.2.4 allows us to restrict attention to strong solutions of (3.5). The following theorem, which can be considered as the main result of the

theory of differential equations, gives an explicit description of the behavior represented by (3.5).

Theorem 3.2.5 *Let $P(\xi) \in \mathbb{R}[\xi]$ be a monic polynomial and let $\lambda_i \in \mathbb{C}$, $i = 1, \ldots, N$, be the distinct roots of $P(\xi)$ of multiplicity n_i: $P(\xi) = \prod_{k=1}^{N}(\xi - \lambda_k)^{n_k}$. The corresponding behavior $\mathfrak{B}$ is autonomous and is a finite-dimensional subspace of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{C})$ of dimension $n = \deg P(\xi)$. Moreover, $w \in \mathfrak{B}$ if and only if it is of the form*

$$w(t) = \sum_{k=1}^{N} \sum_{\ell=0}^{n_k-1} r_{k\ell} t^\ell e^{\lambda_k t} \tag{3.7}$$

with $r_{k\ell}, k = 1, 2, \ldots, N, \ell = 0, 1, \ldots, n_k - 1$, arbitrary complex numbers. Equivalently,

$$w(t) = \sum_{k=1}^{N} r_k(t) e^{\lambda_k t} \tag{3.8}$$

with $r_k(\xi) \in \mathbb{C}[\xi]$ an arbitrary polynomial of degree less than n_k.

The polynomial $P(\xi)$ is called the *characteristic polynomial* of $\mathfrak{B}$, and the roots of $P(\xi)$ are called the *characteristic values.*

Before we can prove this theorem, we state and prove some preliminary results. The first one is a theorem that gives an explicit expression for the action of a polynomial differential operator applied to a function of the form $t^k e^{\lambda t}$.

Notation: We use the notation $P^{(k)}(\xi)$ to denote the kth derivative of the polynomial matrix $P(\xi)$. So if $P(\xi) = P_0 + P_1\xi + \cdots + P_{L-1}\xi^{L-1} + P_L\xi^L$, then

$$P^{(1)} = P_1 + \cdots + (L-1)P_{L-1}\xi^{L-2} + LP_L\xi^{L-1} \text{ and } P^{(k+1)}(\xi) = (P^{(k)})^{(1)}(\xi). \tag{3.9}$$

In the next theorem we use *binomial coefficients*. Recall that for integers $j \geq \ell$ these are defined as:

$$\begin{pmatrix} j \\ \ell \end{pmatrix} = \frac{j!}{(j-\ell)!\ \ell!}.$$

Lemma 3.2.6 *Let $P(\xi) = \sum_{k=0}^{n} P_k \xi^k$ and $w(t) = t^m e^{\lambda t}$. Then*

$$(P(\frac{d}{dt})w)(t) = \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} P^{(m-k)}(\lambda) t^k e^{\lambda t}.$$

Proof Since $\frac{d^k}{dt^k} e^{\lambda t} = \lambda^k e^{\lambda t}$, it follows that $P(\frac{d}{dt})e^{\lambda t} = P(\lambda)e^{\lambda t}$. Differentiating the left- and right-hand side of this equality m times with respect

to λ yields

$$P(\frac{d}{dt})t^m e^{\lambda t} = \sum_{k=0}^{m} \left(\begin{array}{c} m \\ k \end{array} \right) P^{(m-k)}(\lambda) t^k e^{\lambda t}.$$

□

Corollary 3.2.7

(i) For all $k \geq 0$, $l \leq k$, *and* $\lambda \in \mathbb{C}$, $(\frac{d}{dt} - \lambda)^{k+1}(t^l e^{\lambda t}) = 0$.

(ii) For all $k \geq 0$ *and* $\lambda \in \mathbb{C}$, $(\frac{d}{dt} - \lambda)^k(t^k e^{\lambda t}) = k! e^{\lambda t}$.

Proof The proof is left as an exercise to the reader. □

The following theorem is an important step in the proof of Theorem 3.2.5.

Theorem 3.2.8 *Let* $N, M \in \mathbb{N}$ *and let* $I \subset \mathbb{R}$ *be an interval of positive length. Let* $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ *be mutually distinct. Define the functions* $b_{ij} : I \to \mathbb{C}$ *for* $i = 1, \ldots, N,\ j = 1, \ldots, M$ *as follows:*

$$b_{ij}(t) = t^{j-1} e^{\lambda_i t}.$$

The functions b_{ij} *are linearly independent on* I *over* $\mathbb{C}$;. *By that we mean that if* $\sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_{ij} b_{ij}(t) = 0$ *for all* $t \in I$, *where the* $\alpha_{ij} \in \mathbb{C}$, *then* $\alpha_{ij} = 0$ *for all* i *and* j.

Proof Suppose that

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_{ij} b_{ij}(t) = 0 \text{ for all } t \in I. \tag{3.10}$$

Define functions $b_1, \ldots, b_N$ as $b_i := \sum_{j=1}^{M} \alpha_{ij} b_{ij}$. We first show that all the b_is are identically zero. To that end define $P(\xi) := \prod_{i=2}^{N} (\xi - \lambda_i)^M$ and $Q(\xi) := (\xi - \lambda_1)^M$. Since the λ_is are mutually distinct, $P(\xi)$ and $Q(\xi)$ have no common factors, and hence by Corollary B.1.7 there exist polynomials $a(\xi)$ and $b(\xi)$ such that $a(\xi)P(\xi) + b(\xi)Q(\xi) = 1$. By (3.10) and Corollary 3.2.7(i) it follows that $P(\frac{d}{dt})b_1 = 0$ and $Q(\frac{d}{dt})b_1 = 0$. This implies that

$$b_1 = \left(a(\frac{d}{dt})P(\frac{d}{dt}) + b(\frac{d}{dt})Q(\frac{d}{dt}) \right) b_1 = 0.$$

In the same way one proves that $b_2 = 0, \ldots, b_N = 0$.

Consider b_i. By Corollary 3.2.7(i, ii) we conclude that $(\frac{d}{dt} - \lambda_i)^{M-1} b_i(t) = \alpha_{i,M}(M-1)! e^{\lambda_i t} = 0$. This implies that $\alpha_{i,M} = 0$. Operating sequentially by $(\frac{d}{dt} - \lambda_i)^{M-2}$, $(\frac{d}{dt} - \lambda_i)^{M-3}$, ... proves that the remaining coefficients of b_i are zero. This shows that the functions b_{ij} are indeed linearly independent. □

We are now ready to prove Theorem 3.2.5.

Proof of Theorem 3.2.5 From Corollary 3.2.7 we know that every function of the form $t^j e^{\lambda_i t}$, $i = 1, \ldots, N,\ j = 0, \ldots, n_i - 1$, is a solution of (3.5). By the linearity of the differential equation it follows that the linear space spanned by these functions is a subspace of the behavior $\mathfrak{B}$ associated with (3.5). This implies that the dimension of the behavior $\mathfrak{B}$ is at least n. If we can prove that the dimension is also at most n, we are done. This is proven by induction on n.

For $n = 0$ the statement that $\mathfrak{B}$ has dimension zero follows trivially. Suppose that for every polynomial $\tilde{P}(\xi)$ of degree n, the behavior $\mathfrak{B}$ corresponding to that polynomial has dimension n.

Let $P(\xi)$ be a polynomial of degree $n+1$ and let λ be a root of $P(\xi)$. Then $P(\xi) = \tilde{P}(\xi)(\xi - \lambda)$, where $\tilde{P}(\xi)$ is a polynomial of degree n. Let $w(t)$ be a solution of $P(\frac{d}{dt})w(t) = 0$. Note that

$$P(\frac{d}{dt})w = 0 \Longleftrightarrow \tilde{P}(\frac{d}{dt})(\frac{d}{dt} - \lambda)w = 0$$

and

$$\tilde{P}(\frac{d}{dt})(\frac{d}{dt} - \lambda)w = 0 \Longleftrightarrow \tilde{P}(\frac{d}{dt})w' = 0 \text{ with } w' := (\frac{d}{dt} - \lambda)w.$$

Notice that since w is a solution of $P(\frac{d}{dt})w = 0$, we have that $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ (by Theorem 3.2.4), so that w' is well-defined. By the induction hypothesis the behavior defined by $\tilde{P}(\xi)$ has an n-dimensional basis b_i, $i = 1, \ldots, n$, so that w' can be written as

$$w'(t) = \sum_{i=1}^{n} r_i b_i(t),$$

By variation of constants, every strong solution of $(\frac{d}{dt} - \lambda)w = w'$ is given by:

$$\begin{aligned} w(t) &= w(0)e^{\lambda t} + \int_0^t e^{\lambda(t-\tau)} w'(\tau) d\tau \\ &= w(0)e^{\lambda t} + \int_0^t e^{\lambda(t-\tau)} \sum_{i=1}^{n} r_i b_i(\tau) d\tau \\ &= w(0)e^{\lambda t} + \sum_{i=1}^{n} r_i \int_0^t e^{\lambda(t-\tau)} b_i(\tau) d\tau. \end{aligned}$$

Now, $w(t)$ has been written as a linear combination of $n + 1$ functions, namely $e^{\lambda t}$ and

$$\int_0^t e^{\lambda(t-\tau)} b_i(\tau) d\tau, \qquad i = 1, \ldots, n.$$

Notice that these $n+1$ functions all belong to the behavior defined by $P(\xi)$. This implies that the behavior corresponding to $P(\frac{d}{dt})w = 0$ has dimension at most equal to $n + 1$.

Finally, we show that $\mathfrak{B}$ is autonomous. Suppose that $w_1, w_2 \in \mathfrak{B}$ and that $w_1(t) = w_2(t)$ for all $t \leq 0$. Define $w := w_1 - w_2$. Then by linearity of $\mathfrak{B}$, also $w \in \mathfrak{B}$. We have to prove that $w(t) = 0$ for all $t > 0$. We know that w can be expressed as a linear combination of the form (3.7), so that in particular,

$$0 = \sum_{k=1}^{N} \sum_{\ell=0}^{n_k-1} r_{k\ell} t^\ell e^{\lambda_k t}, \quad t \leq 0.$$

By Theorem 3.2.8 the functions $t^\ell e^{\lambda_k t}$ are linearly independent, so that all coefficients are zero. It follows that $w(t) = 0$ for $t > 0$.

This completes the proof. □

Remark 3.2.9 Theorem 3.2.5 provides, despite its somewhat lengthy proof, an elegant way of characterizing all trajectories in a scalar autonomous behavior. The result may not be very easy to prove, but it is easy to recall and to use. □

Before we proceed, we present some examples.

Example 3.2.10 Consider the equation

$$2y - 2\frac{d}{dt}y + \frac{d^2}{dt^2}y = 0. \tag{3.11}$$

The corresponding polynomial is $P(\xi) = 2 - 2\xi + \xi^2$, and it factorizes as

$$(\xi - 1 - i)(\xi - 1 + i).$$

Again, according to Theorem 3.2.5, every solution of (3.11) can be written as

$$y(t) = r_1 e^{(1+i)t} + r_2 e^{(1-i)t}.$$

Here r_1 and r_2 are complex coefficients. Suppose we are interested in real solutions only. One way to obtain real solutions is to take r_2 as the complex conjugate of r_1. Another way is as follows. Write $y = y_r + y_i$, with y_r and y_i the real and imaginary parts of y respectively. From $P(\frac{d}{dt})y = 0$, it follows that $P(\frac{d}{dt})y_r = 0$ and $P(\frac{d}{dt})y_i = 0$. Let us determine y_r and y_i. Write $r_1 = r_{1\mathrm{r}} + r_{1\mathrm{i}} i$ and $r_2 = r_{2\mathrm{r}} + r_{2\mathrm{i}} i$. Then

$$\begin{aligned} y_r(t) &= \mathrm{Re}[(r_{1\mathrm{r}} + ir_{1\mathrm{i}})e^{(1+i)t} + (r_{2\mathrm{r}} + ir_{2\mathrm{i}})e^{(1-i)t}] \\ &= e^t \,\mathrm{Re}[(r_{1\mathrm{r}} + ir_{1\mathrm{i}})(\cos t + i \sin t) + (r_{2\mathrm{r}} + ir_{2\mathrm{i}})(\cos t - i \sin t)] \\ &= e^t((r_{1\mathrm{r}} + r_{2\mathrm{r}}) \cos t + (r_{2\mathrm{i}} - r_{1\mathrm{i}}) \sin t). \end{aligned} \tag{3.12}$$

Similarly we find

$$y_i(t) = e^t((r_{1\mathrm{i}} + r_{2\mathrm{i}}) \cos t + (r_{1\mathrm{r}} - r_{2\mathrm{r}}) \sin t).$$

By defining $\alpha := r_{1\mathrm{r}} + r_{2\mathrm{r}}$ and $\beta := r_{2\mathrm{i}} - r_{1\mathrm{i}}$, we obtain from (3.12) that the real solutions of $P(\frac{d}{dt})y = 0$ consist of the functions of the form

$$y(t) = \alpha e^t \cos t + \beta e^t \sin t, \quad \alpha, \beta \in \mathbb{R}.$$

□

Example 3.2.11 In Example 3.2.2 we discussed an autonomous mass–damper–spring system. Take $d = 0$. The equation describing the position of the mass with respect to its equilibrium is

$$(k_1 + k_2)w + M\frac{d^2}{dt^2}w = 0.$$

The corresponding polynomial is $P(\xi) = M\xi^2 + (k_1 + k_2)$. The roots of $P(\xi)$ are

$$\lambda_1 = i\sqrt{\frac{k_1 + k_2}{M}}, \quad \lambda_2 = -i\sqrt{\frac{k_1 + k_2}{M}}.$$

According to Theorem 3.2.5 every possible motion of the mass is of the form

$$w(t) = c_1 e^{i\sqrt{\frac{k_1+k_2}{M}}t} + c_2 e^{-i\sqrt{\frac{k_1+k_2}{M}}t}, \qquad c_1, c_2 \in \mathbb{C}. \tag{3.13}$$

Complex trajectories have no physical meaning, so we take the real part of (3.13) to obtain a general expression for the real trajectories. This gives

$$w(t) = \alpha \sin(\sqrt{\frac{k_1 + k_2}{M}}t) + \beta \cos(\sqrt{\frac{k_1 + k_2}{M}}t), \quad \alpha, \beta \in \mathbb{R},$$

or equivalently,

$$w(t) = A\cos(\sqrt{\frac{k_1 + k_2}{M}}t + \phi), \quad A \in \mathbb{R}, \phi \in [0, 2\pi).$$

□

Example 3.2.12 As a third example, consider the behavior defined by

$$-6y - 11\frac{d}{dt}y - 3\frac{d^2}{dt^2}y + 3\frac{d^3}{dt^3}y + \frac{d^4}{dt^4}y = 0. \tag{3.14}$$

The corresponding polynomial is

$$P(\xi) = -6 - 11\xi - 3\xi^2 + 3\xi^3 + \xi^4 = (\xi + 1)^2(\xi - 2)(\xi + 3),$$

and according to Theorem 3.2.5, the solutions of (3.14) can be written as

$$y(t) = r_{10}e^{-t} + r_{11}te^{-t} + r_{20}e^{2t} + r_{30}e^{-3t}. \tag{3.15}$$

The coefficients r_{ij} in (3.15) are completely free, but if we are interested only in real solutions, then they should be real. □

Taking real parts in (3.8) and proceeding as in Example 3.2.10, yields the following immediate consequence of Theorem 3.2.8.

Corollary 3.2.13 *Consider the dynamical system* $(\mathbb{R}, \mathbb{R}, \mathfrak{B})$ *represented by the differential equation (3.5) with real coefficients. Then* $w \in \mathfrak{B}$ *if and only if it is of the form*

$$w(t) = \sum_{k=1}^{N'} r_k(t) e^{\lambda_k t} + \sum_{k=1}^{N''} (r'_k(t) \cos \omega_k t + r''_k(t) \sin \omega_k t) e^{\lambda'_k t}, \qquad (3.16)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{N'}$ *are the real roots of* $P(\xi), n_1, n_2, \ldots, n_{N'}$ *their multiplicities;* $\lambda'_1 \pm i\omega_1, \lambda'_2 \pm i\omega_2, \ldots, \lambda'_{N''} \pm i\omega_{N''}$ *the roots with nonzero imaginary part,* $n'_1, n'_2, \ldots, n'_{N''}$ *their multiplicities and* $r_k(t), r'_k(t), r''_k(t) \in \mathbb{R}[t]$ *arbitrary polynomials of degrees at most,* $n_k - 1, n'_k - 1$, *and* $n'_k - 1$ *respectively.*

Proof If $\lambda \in \mathbb{R}$, then obviously $e^{\lambda t}$ is real-valued. This explains the first part of (3.16).

If $\lambda = \mu + i\omega$ is a root of multiplicity m of $P(\xi)$, then $\bar{\lambda} = \mu - i\omega$ is also a root of $P(\xi)$ and has the same multiplicity. The roots λ and $\bar{\lambda}$ give rise to (complex-valued) solutions of the form

$$w(t) = \sum_{j=0}^{m-1} (a_j + ia'_j) t^j e^{\lambda t} + \sum_{j=0}^{m-1} (b_j + ib'_j) t^j e^{\bar{\lambda} t}, \quad a_j, a'_j, b_j, b'_j \in \mathbb{R}. \quad (3.17)$$

Taking real and imaginary parts in (3.17) yields the result. □

Remark 3.2.14 Some remarks related to Theorem 3.2.5 and Corollary 3.2.13:

1. From Theorem 3.2.5 it follows that $\mathfrak{B}$ is a finite-dimensional vector space. Its dimension is equal to the number of free coefficients in (3.7), namely n, the degree of $P(\xi)$. Note also that

   $$e^{\lambda_1 t}, te^{\lambda_1 t}, \ldots, t^{n_1 - 1} e^{\lambda_1 t}, \ldots, e^{\lambda_N t}, te^{\lambda_N t}, \ldots, t^{n_N - 1} e^{\lambda_N t}$$

   form a basis for $\mathfrak{B}$ in the complex case, while in the real case this basis consists of

   $$e^{\lambda_1 t}, \ldots, t^{n_1 - 1} e^{\lambda_1 t}, \ldots, e^{\lambda_{N'} t}, \ldots, t^{n_{N'} - 1} e^{\lambda_{N'} t}$$
   $$e^{\lambda'_1 t} \cos \omega_1 t, e^{\lambda'_1 t} \sin \omega_1 t, \ldots, t^{n'_1 - 1} e^{\lambda'_1 t} \cos \omega_1 t, t^{n'_1 - 1} e^{\lambda'_1 t} \sin \omega_1 t, \ldots,$$
   $$e^{\lambda'_{N''} t} \cos \omega_{N''} t, \ldots, t^{n'_{N''} - 1} e^{\lambda'_{N''} t} \cos \omega_{N''} t, t^{n'_{N''} - 1} e^{\lambda'_{N''} t} \sin \omega_{N''} t.$$

2. The constants $r_{k\ell}$ appearing in the polynomials in the expression of a general solution w, (3.7), have an interpretation in terms of the value

of w and its first $n-1$ derivatives at $t=0$. We derive this relation for the case that all roots of $P(\xi)$ have multiplicity one. Indeed

$$(\frac{d}{dt})^k(\sum_{i=1}^{n} r_i e^{\lambda_i t})(0) = \sum_{i=1}^{n} r_i \lambda_i^k.$$

This shows that there is a simple relation between $w(0)$, $w^{(1)}(0)$,..., $w^{(n-1)}(0)$ and the coefficients $\{r_i\}$. Written in matrix notation

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} w(0) \\ w^{(1)}(0) \\ \vdots \\ w^{(n-1)}(0) \end{bmatrix}. \tag{3.18}$$

The matrix in (3.18) has a simple structure and is called a *Vandermonde matrix*. It is not difficult to prove that this matrix is nonsingular if and only if the λ_is are mutually distinct. See Exercise 3.16. Hence the linear relation between the initial conditions and the coefficients r_i is bijective.

The reader is encouraged to treat the general case, multiplicities larger than one, by doing Exercises 3.3 and 3.16.

3. Let $w \in \mathfrak{B}$. Now consider its *past*, w^-, and its *future*, w^+. Formally

$$w^- : (-\infty, 0) \to \mathbb{R} \text{ is defined by } w^-(t) := w(t),\ t < 0;$$

$$w^+ : [0, \infty) \to \mathbb{R} \text{ is defined by } w^+(t) := w(t),\ t \geq 0.$$

Each element $w \in \mathfrak{B}$ is of the form (3.7). By (3.18) the coefficients r_{kl} are uniquely determined by $w(0), \ldots, (\frac{d^{n-1}}{dt^{n-1}}w)(0)$, which in turn are determined by w^-. It follows that w^+ is uniquely determined by w^-, so that indeed the corresponding behavior is autonomous. Thus the dynamical system represented by the differential equation (3.5) has the special property that the *past* of an element in its behavior *uniquely specifies its future*. This explains the title of Section 3.2. Mathematically inclined readers understand that this follows from the fact that all functions of the form (3.7) are *analytic*.

□

3.2.2 The multivariable case

We now consider the general case, $g = q \geq 1$. Recall that we are interested in the characterization of the behavior $\mathfrak{B}$ corresponding to the linear differential equation

$$P(\frac{d}{dt})w = 0, \qquad P(\xi) \in \mathbb{R}^{q \times q}[\xi], \qquad \det P(\xi) \neq 0.$$

Recall that for the scalar case, every weak solution of $P(\frac{d}{dt})w = 0$ is equal to a strong one almost everywhere. This result is also true for the multivariable case, but the proof is less straightforward. The reader is referred to Exercise 3.15 for a suggested proof. For the sake of completeness we state the result in the form of a theorem (see also Theorem 3.2.4).

Theorem 3.2.15 *Let $\mathfrak{B}$ be the behavior defined by $P(\frac{d}{dt})w = 0$, where $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and $\det P(\xi) \neq 0$. For every $w \in \mathfrak{B}$ there exists a $v \in \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ such that $w(t) = v(t)$ for almost all t.*

Proof See Exercise 3.15. □

Because of Theorem 3.2.15, the object of interest in this section is

$$\mathfrak{B} := \left\{ w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid w \text{ is a strong solution of } P(\frac{d}{dt})w = 0 \right\}.$$

Refer to (3.9) for the notation for the higher-order derivatives of polynomial matrices.

Theorem 3.2.16 *Let $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and let $\lambda_i \in \mathbb{C}$, $i = 1, \ldots, N$, be the distinct roots of $\det P(\xi)$ of multiplicity n_i: $\det P(\xi) = c\prod_{k=1}^{N}(\xi - \lambda_k)^{n_k}$ for some nonzero constant c. The corresponding behavior $\mathfrak{B}$ is autonomous and is a finite-dimensional subspace of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{C}^q)$ of dimension $n = \deg\det P(\xi)$. Moreover, $w \in \mathfrak{B}$ if and only if it is of the form*

$$w(t) = \sum_{i=1}^{N}\sum_{j=0}^{n_i-1} B_{ij}t^j e^{\lambda_i t}, \tag{3.19}$$

where the vectors $B_{ij} \in \mathbb{C}^q$ satisfy the relations

$$\sum_{j=\ell}^{n_i-1} \binom{j}{\ell} P^{(j-\ell)}(\lambda_i)B_{ij} = 0, \qquad i = 1, \ldots, N;\ \ell = 0, \ldots, n_i - 1. \tag{3.20}$$

In matrix notation we get for $i = 1, \ldots, N$:

$$\begin{bmatrix} \binom{0}{0}P^{(0)}(\lambda_i) & \cdots & \cdots & \binom{n_i-1}{0}P^{(n_i-1)}(\lambda_i) \\ 0 & \ddots & & \binom{n_i-1}{1}P^{(n_i-2)}(\lambda_i) \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \binom{n_i-1}{n_i-1}P^{(0)}(\lambda_i) \end{bmatrix} \begin{bmatrix} B_{i,0} \\ B_{i,1} \\ \vdots \\ B_{i,n_{i-1}} \end{bmatrix} = 0.$$

The polynomial $\prod_{k=1}^{N}(\xi-\lambda_k)^{n_k}$ is called the *characteristic polynomial* of the autonomous behavior $\mathfrak{B}$. The roots of $\det P(\xi)$, $\lambda_1,\cdots,\lambda_N$, are called the characteristic values of the behavior. For the case that $P(\xi)=I\xi-A$ for some matrix A, the characteristic values are just the eigenvalues of A.

Proof For the proof of Theorem 3.2.16 we use a lemma, which we state first. □

Remark 3.2.17 If all the roots of $\det P(\xi)$ have multiplicity one, say $\det P(\xi)=\prod_{k=1}^{n}(\xi-\lambda_k)$, then (3.19, 3.20) reduce to $w(t)=\sum_{k=1}^{n}B_k e^{\lambda_k t}$, $P(\lambda_k)B_k=0$. □

Before we give the proof of Theorem 3.2.16, we state the multivariable analogue of Lemma 3.2.6.

Lemma 3.2.18 *Let $P(\xi)\in\mathbb{R}^{q\times q}[\xi]$, m a nonnegative integer, $\lambda\in\mathbb{C}, A\in\mathbb{C}^q$, and $w(t)=At^m e^{\lambda t}$. Then*

$$(P(\frac{d}{dt})w)(t)=\sum_{k=0}^{m}\begin{pmatrix} m \\ k \end{pmatrix}P^{(m-k)}(\lambda)At^k e^{\lambda t}.$$

Proof See Exercise 3.11. □

Proof of Theorem 3.2.16 The proof is divided into four parts. In the first part we show that the dimension of $\mathfrak{B}$ is equal to the degree of the determinant of $P(\xi)$. In the second part we show that every $w\in\mathfrak{B}$ is of the form (3.19), and in the third part we show that the coefficients B_{ij} should satisfy (3.20). Finally, we show that $\mathfrak{B}$ is autonomous.

Choose unimodular matrices $U(\xi),V(\xi)$ such that $D(\xi):=U(\xi)P(\xi)V(\xi)=\operatorname{diag}(d_1(\xi),\ldots,d_q(\xi))$ is in Smith form. Notice that because $0\neq\det P(\xi)=c\det D(\xi)$ for some nonzero constant c, the polynomials $d_i(\xi)$ are nonzero for $i=1,\ldots,q$.

(i) Since $D(\xi)$ is diagonal, the behavior $\mathfrak{B}_D$ defined by $D(\frac{d}{dt})w=0$ can be obtained as follows. Every component w_i of a solution of $D(\frac{d}{dt})w=0$ satisfies $d_i(\frac{d}{dt})w_i=0$, and conversely, every w for which the components satisfy $d_i(\frac{d}{dt})w_i=0$, $i=1,\ldots q$, is a solution of $D(\frac{d}{dt})w=0$. By Theorem 3.2.5, the dimension of the (scalar) behavior of $d_i(\frac{d}{dt})w_i=0$ equals the degree of $d_i(\xi)$. This implies that the dimension of $\mathfrak{B}_D$ is equal to the sum of the degrees of the diagonal elements, which is the degree of the determinant of $D(\xi)$. After left and right unimodular transformations, the determinant has changed only by a multiplicative nonzero constant, and hence the dimension of $\mathfrak{B}_D$ equals $\deg\det P(\xi)$. Finally, by Theorem 2.5.13, the $\mathcal{C}^\infty$ parts of the behaviors defined by $P(\xi)$ and $D(\xi)$ are isomorphic, and hence they have the same dimension.

An alternative proof that does not rely on the Smith form is suggested in Exercise 3.10.

(ii) Since $D(\xi)$ is diagonal, it follows from Theorem 3.2.5 that every solution of $D(\frac{d}{dt})\tilde{w} = 0$ is of the form

$$\tilde{w}(t) = \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \tilde{B}_{ij} t^j e^{\lambda_i t}. \tag{3.21}$$

Since $D(\xi) = U(\xi)P(\xi)V(\xi)$, (3.21) implies that every $w \in \mathfrak{B}$ can be written as

$$w(t) = V(\frac{d}{dt})\tilde{w}(t) = \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} V(\frac{d}{dt}) \left(\tilde{B}_{ij} t^j e^{\lambda_i t} \right) =: \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} B_{ij} t^j e^{\lambda_i t}.$$

The last equality follows from Lemma 3.2.18.

(iii) Next we prove that every function of the form (3.19) belongs to $\mathfrak{B}$ if and only if the relations (3.20) hold.

Suppose that w is given by (3.19). Then it follows by Lemma 3.2.18 that

$$\begin{aligned} P(\frac{d}{dt})w &= \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} P(\frac{d}{dt}) B_{ij} t^j e^{\lambda_i t} = \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \sum_{\ell=0}^{j} \binom{j}{\ell} P^{(j-\ell)}(\lambda_i) B_{ij} t^\ell e^{\lambda_i t} \\ &= \sum_{i=1}^{N} \sum_{\ell=0}^{n_i-1} \left[\sum_{j=\ell}^{n_i-1} \binom{j}{\ell} P^{(j-\ell)}(\lambda_i) B_{ij} \right] t^\ell e^{\lambda_i t}. \end{aligned} \tag{3.22}$$

Now, $w \in \mathfrak{B}$ if and only if the last line of (3.22) is identically zero. By Theorem 3.2.8, the functions $t^\ell e^{\lambda_i t}$ are linearly independent, and hence the (vector-valued) coefficients in the last line of (3.22) should be zero. This implies that w belongs to $\mathfrak{B}$ if and only if

$$\sum_{j=\ell}^{n_i-1} \binom{j}{\ell} P^{(j-\ell)}(\lambda_i) B_{ij} = 0 \text{ for } i = 1, \ldots, N, \ \ell = 0, \ldots, n_i - 1.$$

The fact that $\mathfrak{B}$ is autonomous follows in the same way as for the scalar case, and therefore we omit the details. □

Remark 3.2.19 The Smith form of the matrix $P(\xi)$ gives some useful information about the structure of the corresponding behavior. To see this, let $D(\xi)$ be the Smith form of $P(\xi)$. Let λ_i be a root of $\det P(\xi)$ of multiplicity n_i. In principle, we could expect elements $w \in \mathfrak{B}$ of the form $B_{ij} t^j e^{\lambda_i t}$ for $j = 0, \ldots, n_i - 1$. If, however, the factor $(\xi - \lambda_i)$ appears in $D(\xi)$ at most with the power m_i, for some $m_i \leq n_i - 1$, then we can conclude that the coefficients B_{ij} with $j \geq m_i$ are zero. An example clarifies this point.

Let $P(\xi)$ be such that $\det P(\xi) = \xi^3$. In principle, we can expect solutions of $P(\frac{d}{dt})w = 0$ of the form

$$B_0 + B_1 t + B_2 t^2.$$

Suppose, however, that the Smith form of $P(\xi)$ is

$$D(\xi) = \begin{bmatrix} \xi & 0 \\ 0 & \xi^2 \end{bmatrix}.$$

Then the solutions of $D(\frac{d}{dt})w = 0$ are of the form

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} 0 \\ c_3 \end{bmatrix} t. \tag{3.23}$$

From (3.23) we conclude that quadratic terms do not appear, and hence B_2 should be zero.

If, on the other hand, the Smith form of $P(\xi)$ is

$$D(\xi) = \begin{bmatrix} 1 & 0 \\ 0 & \xi^3 \end{bmatrix},$$

then the solutions of $D(\frac{d}{dt})w = 0$ are of the form

$$\begin{bmatrix} 0 \\ c_0 \end{bmatrix} + \begin{bmatrix} 0 \\ c_1 \end{bmatrix} t + \begin{bmatrix} 0 \\ c_2 \end{bmatrix} t^2.$$

Note that for this case quadratic terms do occur. □

3.3 Systems in Input/Output Form

For autonomous behaviors, the future of a trajectory is completely determined by its past; two trajectories with the same past are necessarily identical. It follows from Section 3.2 that behaviors defined by *square* full row rank polynomial matrices are autonomous. In this section we study the case where the number of rows of $R(\xi)$ is less than the number of columns, more generally, when the rank of the polynomial matrix $R(\xi)$ is less than the number of columns. It turns out that in this case the trajectories contain *free components*, parts of w that are not uniquely determined by their past. The reader may find it convenient to refer to the analogy with underdetermined systems of linear equations. When there are more variables than equations, it can be expected that some of the variables are not restricted by the equations. The analogy can be carried even further, since it turns out that the number of free components is actually equal to the number of

variables minus the number of independent equations. However, it requires some work to derive this appealing result.

Before we concentrate on representations, we give a behavioral definition of an *input/output* system that is in the same spirit as Definition 3.2.1.

Definition 3.3.1 Let $\mathfrak{B}$ be a behavior with signal space $\mathbb{R}^q$. Partition the signal space as $\mathbb{R}^q = \mathbb{R}^m \times \mathbb{R}^p$ and $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ correspondingly as $w = \mathrm{col}(w_1, w_2)$ ($w_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ and $w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^p)$). This partition is called an *input/output partition* if:

1. w_1 is *free*; i.e., for all $w_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$, there exists a $w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^p)$ such that $\mathrm{col}(w_1, w_2) \in \mathfrak{B}$.
2. w_2 does not contain any further free components; i.e., given w_1, none of the components of w_2 can be chosen freely. Stated differently, w_1 is *maximally free*.

If 1 and 2 hold, then w_1 is called an *input variable* and w_2 is called an *output variable*. □

To illustrate Definition 3.3.1, consider the following examples.

Example 3.3.2 This is a continuation of Example 3.2.3. We have already seen that the mass–spring system is not autonomous. In fact, from physical considerations, it should be clear that the force acting on the mass can be any time function and can thus be seen as a free variable. Also, given the force, the position of the mass as a function of time is completely determined by the past (in fact, by the position and velocity at $t = 0$). So from an intuitive point of view, the mass–spring system can be considered as an input/output system with the force as input and the position as output. □

A more mathematically oriented example is the following.

Example 3.3.3 Let w_1, w_2 be scalar variables and let $\mathfrak{B}$ be the behavior defined by

$$-w_2 + \frac{d}{dt} w_2 = w_1. \tag{3.24}$$

For a given $w_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$, define w_2 by

$$w_2(t) := \int_0^t e^{t-\tau} w_1(\tau) d\tau, \quad t \in \mathbb{R}. \tag{3.25}$$

We should emphasize that w_2 is defined by (3.25) for *all* t, also for $t < 0$. It follows by substitution in (3.24) that if w_1 is continuous, then $(w_1, w_2) \in \mathfrak{B}$.

Actually, in this case, it is a *strong* solution. Later in this section we will see that it is a *weak* solution if w_1 is merely in $\mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R})$. This implies that w_1 is a free variable. Once w_1 is given, w_2 cannot be chosen to be completely free, for if (w_1, w_2) and (w_1, w_2') satisfy (3.24), then

$$(-1 + \frac{d}{dt})(w_2 - w_2') = 0.$$

This means that $w_2 - w_2'$ should satisfy an equation of the type studied in Section 3.2. In other words, w_2 is completely determined by its past and w_1. The conclusion is that w_1 is maximally free. □

Remark 3.3.4 The partition of w into input and output is in general not unique. However, in Examples 3.3.2 and 3.3.3 there is no choice. In the latter example, given w_2, there will not always be a w_1 such that $(w_1, w_2) \in \mathfrak{B}$. For example, if w_2 is not continuous, there does not exist a $w_1 \in \mathfrak{L}_1^{\rm loc}(\mathbb{R},\mathbb{R})$ such that (3.24) is satisfied. So w_2 is not free and can therefore not be viewed as input. See Exercise 3.26.

A trivial example illustrates that there are cases in which the choice of input is indeed not unique. Consider the behavior defined by

$$w_1 = w_2.$$

It is clear that we can either take w_1 as input and w_2 as output, or vice versa. Trivial as this example may be, it has some consequences for modeling physical systems as input/output systems. For instance, when modeling the voltage/current behavior of a resistor, either of the two variables can act as an input. □

We now focus on a special class of behaviors defined by equations of the form $R(\frac{d}{dt})w = 0$, for which the polynomial matrix $R(\xi)$ has a special form. Before we can specify that form, we need the notion of a matrix of proper rational functions.

Definition 3.3.5 A matrix of rational functions (i.e., each of the entries is the ratio of two polynomials) is called *proper* if in each entry the degree of the numerator does not exceed the degree of the denominator. It is called *strictly proper* if in each entry the degree of the numerator is strictly smaller than the degree of the denominator. □

In what follows, $R(\xi)$ is assumed to be of the form

$$R(\xi) = \begin{bmatrix} -Q(\xi) & P(\xi) \end{bmatrix},$$

where $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$ and $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ satisfy:

- $\det P(\xi) \neq 0$.

- $P^{-1}(\xi)Q(\xi)$ is a matrix of proper rational functions. By Cramer's rule, the entries of $P^{-1}(\xi)Q(\xi)$ are always rational functions. The condition that they are proper, however, imposes a restriction.

For notational reasons we partition w conformably as

$$w = \begin{bmatrix} u \\ y \end{bmatrix},$$

so that the behavioral equations $R(\frac{d}{dt})w = 0$ may be written as $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$. The corresponding behavior is

$$\mathfrak{B} = \left\{ w = \mathrm{col}(u, y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^p) \mid P(\frac{d}{dt})y = Q(\frac{d}{dt})u, \text{ weakly} \right\}. \tag{3.26}$$

Remark 3.3.6 The matrix $P^{-1}(\xi)Q(\xi)$ is often referred to as the *transfer matrix* of the behavior defined by (3.26). The transfer matrix plays an important role in applications. We come back to it in Chapter 8. □

We will show that behaviors of the form (3.26) are indeed input/output systems in the sense of Definition 3.3.1, with u as input and y as output. In particular, we show that for every $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ there exists a $y \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^p)$ such that $(u, y) \in \mathfrak{B}$. Moreover, we also show that every behavior defined by $R(\frac{d}{dt})w = 0$, of which (3.26) is a special case, admits a representation of the form (3.26).

We first give a complete characterization of the behavior $\mathfrak{B}$ defined by (3.26). We do this in terms of the so-called *partial fraction expansion* of the quotient of two polynomials.

Theorem 3.3.7 (Partial fraction expansion, scalar case)
Let $P(\xi), Q(\xi) \in \mathbb{R}[\xi]$, and $\deg Q(\xi) = m \leq n = \deg P(\xi)$. Suppose $P(\xi) = \prod_{i=1}^{N}(\xi - \lambda_i)^{n_i}$, $\lambda_i \neq \lambda_j$, $i \neq j$. Then there exist a_0 and $a_{ij} \in \mathbb{C}$ such that

$$P^{-1}(\xi)Q(\xi) = a_0 + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{a_{ij}}{(\xi - \lambda_i)^j}. \tag{3.27}$$

Proof The proof is given in Appendix B, Theorem B.2.1. □

Corollary 3.3.8 *Let the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$ be given by (3.27). Then*

$$Q(\xi) = a_0 P(\xi) + \sum_{i=1}^{N} \sum_{j=1}^{n_i} a_{ij} [\prod_{k \neq i}^{N} (\xi - \lambda_k)^{n_k}](\xi - \lambda_i)^{n_i - j}.$$

Proof See Exercise 3.30. □

Remark 3.3.9 The coefficients a_{ij} can be calculated as follows:

$$\begin{aligned} a_0 &= \lim_{\lambda\to\infty} \frac{Q(\lambda)}{P(\lambda)}, \\ a_{in_i} &= \lim_{\lambda\to\lambda_i} (\lambda-\lambda_i)^{n_i} \frac{Q(\lambda)}{P(\lambda)}, && i=1,\ldots,N, \\ a_{ij} &= \lim_{\lambda\to\lambda_i} (\lambda-\lambda_i)^{j} \left[\frac{Q(\lambda)}{P(\lambda)} - \sum_{k=j+1}^{n_i} \frac{a_{ik}}{(\lambda-\lambda_i)^k}\right], && i=1,\ldots,N, \\ & && j=1,\ldots,n_i-1. \end{aligned}$$

□

Example 3.3.10 Take $Q(\xi) = 8 - 17\xi + 8\xi^2 + 3\xi^3$ and $P(\xi) = 1 - 2\xi + 2\xi^2 - 2\xi^3 + \xi^4$. The polynomial $P(\xi)$ factors as $P(\xi) = (\xi-1)^2(\xi-i)(\xi+i)$. If we take $\lambda_1 = 1,\ \lambda_2 = i$, and $\lambda_3 = -i$, then according to Theorem 3.3.7, $\frac{Q(\xi)}{P(\xi)}$ can be written as

$$P^{-1}(\xi)Q(\xi) = \frac{a_{11}}{(\xi-1)} + \frac{a_{12}}{(\xi-1)^2} + \frac{a_{21}}{(\xi-i)} + \frac{a_{31}}{(\xi+i)}. \tag{3.28}$$

The coefficients of (3.28) are calculated as follows:

$$a_{12} = \lim_{\lambda\to 1}(\lambda-1)^2\frac{Q(\lambda)}{P(\lambda)} = \lim_{\lambda\to 1}\frac{8-17\lambda+8\lambda^2+3\lambda^3}{1+\lambda^2} = 1,$$

$$a_{11} = \lim_{\lambda\to 1}(\lambda-1)\left[\frac{Q(\lambda)}{P(\lambda)} - \frac{1}{(\lambda-1)^2}\right] = \lim_{\lambda\to 1}\frac{-7+10\lambda+3\lambda^2}{1+\lambda^2} = 3.$$

The other coefficients are calculated accordingly. This yields that $a_{21} = -5i$ and $a_{31} = 5i$. Hence the partial fraction expansion is

$$P^{-1}(\xi)Q(\xi) = \frac{1}{(\xi-1)^2} + \frac{3}{(\xi-1)} - \frac{5i}{(\xi-i)} + \frac{5i}{(\xi+i)}.$$

□

Remark 3.3.11 (Partial fraction expansion, matrix case) Formula (3.27) is called the *partial fraction expansion of* $P^{-1}(\xi)Q(\xi)$. If $P(\xi)$ and $Q(\xi)$ are polynomial matrices, with $\det P(\xi) \neq 0$, then the partial fraction expansion of the matrix of rational functions $P^{-1}(\xi)Q(\xi)$ is defined entry-wise. The complex numbers λ_i are now the roots of the determinant of $P(\xi)$. □

We frequently use the following result.

Lemma 3.3.12 *Let* $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ *and* $k \geq 1$. *Define*

$$y_k(t) := \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} e^{\lambda(t-\tau)} u(\tau) d\tau. \tag{3.29}$$

Then $\mathrm{col}(u, y_k)$ *satisfies*

$$(\frac{d}{dt} - \lambda)^k y_k = u \tag{3.30}$$

weakly

Proof The proof consists of two parts. In the first part we make the assumption that u is infinitely differentiable, so that (u, y_k) is a strong solution. In the second part we show that (u, y_k) is a weak solution by approximating u by a sequence of $\mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ functions. (i) Suppose that u is infinitely differentiable.

(Induction) For $k = 1$ the statement that (u, y_k) is a strong solution of (3.29) is clearly true. Suppose that $n > 1$ and that for all $k \leq n$, (3.29) defines a solution of (3.30). Consider

$$y_{n+1}(t) = \int_0^t \frac{(t-\tau)^n}{n!} e^{\lambda(t-\tau)} u(\tau) d\tau.$$

Since u is smooth, so is y_{n+1}. The derivative of y_{n+1} satisfies

$$\begin{aligned} \frac{d}{dt} y_{n+1}(t) &= \int_0^t \frac{(t-\tau)^{n-1}}{(n-1)!} e^{\lambda(t-\tau)} u(\tau) d\tau + \lambda \int_0^t \frac{(t-\tau)^n}{(n)!} e^{\lambda(t-\tau)} u(\tau) d\tau \\ &= y_n(t) + \lambda y_{n+1}(t), \end{aligned}$$

from which we conclude that

$$(\frac{d}{dt} - \lambda) y_{n+1} = y_n,$$

which in turn implies

$$(\frac{d}{dt} - \lambda)^{n+1} y_{n+1} = (\frac{d}{dt} - \lambda)^n y_n = u.$$

The last equality follows from the induction hypothesis.

(ii) Now assume that $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ and let y_k be defined by (3.29). Choose a sequence $u_n \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ such that u_n converges to u in the sense of $\mathfrak{L}_1^{\mathrm{loc}}$. Define $y_{k,n}$ by

$$y_{k,n}(t) := \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} e^{\lambda(t-\tau)} u_n(\tau) d\tau.$$

Then $(u_n, y_{k,n})$ converges to (u, y_k) (y_k defined by (3.29)) in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$ as n tends to infinity; see Exercise 3.32. By Theorem 2.4.4 it follows that (u, y_k) is a weak solution of (3.30). □

Lemma 3.3.12 and the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$ allow us to provide an explicit expression of a particular solution of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$.

Theorem 3.3.13 *Let $\mathfrak{B}$ be the behavior defined by (3.26), and let $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$. Let the partial fraction expansion of the transfer matrix $P^{-1}(\xi)Q(\xi)$ be given by*

$$P^{-1}(\xi)Q(\xi) = A_0 + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{A_{ij}}{(\xi - \lambda_i)^j}.$$

Define y by

$$y(t) := A_0 u(t) + \sum_{i=1}^{N} \sum_{j=1}^{n_i} A_{ij} \int_0^t \frac{(t-\tau)^{j-1}}{(j-1)!} e^{\lambda_i (t-\tau)} u(\tau) d\tau, \qquad t \in \mathbb{R}. \quad (3.31)$$

Then $(u, y) \in \mathfrak{B}$.

Proof For simplicity, we treat the single-input/single-output case $p = 1$, $m = 1$ only. The multivariable case is proven analogously but is technically more involved. A proof for the multivariable case is suggested in Exercise 3.24.

Let $\{\lambda_i\}$ be the distinct roots of multiplicity n_i of $P(\xi)$, and let a_0 and $a_{ij}, i = 1, \ldots, N,\ j = 1, \ldots, n_i$ be the coefficients of the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$, as in (3.27). First assume that $u \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$.

$$y_0(t) := a_0 u(t),$$

Define

$$y_{ij}(t) := a_{ij} \int_0^t \frac{(t-\tau)^{j-1}}{(j-1)!} e^{\lambda_i (t-\tau)} u(\tau) d\tau.$$

Then

$$y(t) = y_0(t) + \sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij}(t).$$

Let us calculate $P(\frac{d}{dt})y(t)$:

$$\begin{aligned} P(\frac{d}{dt})y(t) &= \prod_{k=1}^{N}(\frac{d}{dt}-\lambda_k)^{n_k}\left(y_0(t)+\sum_{i=1}^{N}\sum_{j=1}^{n_i}y_{ij}(t)\right) \\ &= P(\frac{d}{dt})a_0u(t)+\sum_{i=1}^{N}\sum_{j=1}^{n_i}\prod_{k=1}^{N}(\frac{d}{dt}-\lambda_k)^{n_k}y_{ij}(t) \\ &= P(\frac{d}{dt})a_0u(t)+\sum_{i=1}^{N}\sum_{j=1}^{n_i}[\prod_{k\neq i}^{N}(\frac{d}{dt}-\lambda_k)^{n_k}](\frac{d}{dt}-\lambda_i)^{n_i}y_{ij}(t) \\ &= P(\frac{d}{dt})a_0u(t)+\sum_{i=1}^{N}\sum_{j=1}^{n_i}[\prod_{k\neq i}^{N}(\frac{d}{dt}-\lambda_k)^{n_k}](\frac{d}{dt}-\lambda_i)^{n_i-j}a_{ij}u(t) \\ &= Q(\frac{d}{dt})u(t). \end{aligned} \tag{3.32}$$

The third equality follows from Lemma 3.3.12 and the last equality from Corollary 3.3.8.

The case where $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ goes along the same lines as the second part of the proof of Lemma 3.3.12. □

Corollary 3.3.14 *Let $\mathfrak{B}$ be the behavior defined by (3.26). Then (u,y) defines an input/output partition in the sense of Definition 3.3.1.*

Proof It follows from Theorem 3.3.13 that u is free: (3.31) shows that for any $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^m)$ there exists a $y \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^p)$ such that $(u,y) \in \mathfrak{B}$. Next we show that u is *maximally free*; i.e., y does not contain any other free components. Let $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^m)$ be given, and suppose $(u,y_1),(u,y_2) \in \mathfrak{B}$. By linearity it follows that $(0,y_1-y_2) \in \mathfrak{B}$. This implies that $P(\frac{d}{dt})(y_1-y_2)=0$. Since by assumption $\det P(\xi) \neq 0$, we conclude from Section 3.2 that $P(\frac{d}{dt})y=0$ defines an autonomous behavior, and therefore y_1-y_2 is uniquely determined by its past, whence y does not contain any further free components.

The conclusion is that for given $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^m)$, the set of ys such that $(u,y) \in \mathfrak{B}$ is a finite-dimensional *affine*[1] subspace of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$. □

Corollary 3.3.14 justifies the following definition.

---

[1] An *affine* subspace of a linear space is a shifted linear subspace. In other words, a subset S of a linear space $\mathfrak{X}$ is affine if it is of the form $S=a+\mathfrak{V}$ with $a \in \mathfrak{X}$ and $\mathfrak{V}$ a linear subspace of $\mathfrak{X}$. Its dimension is defined as the dimension of the linear subspace $\mathfrak{V}$. An example of a one-dimensional affine subspace in $\mathbb{R}^2$ is a line that does not pass through the origin. In input/output behaviors, the set of outputs y for a given input u such that $(u,y) \in \mathfrak{B}$ is affine. If $(u,y) \in \mathfrak{B}$, then all possible outputs y' can be written as $y'=y+y_{\mathrm{hom}}$ where y_{hom} satisfies $P(\frac{d}{dt})y_{\mathrm{hom}}=0$. Therefore, the dimension of the set of possible outputs corresponding to this input u equals $\deg\det P(\xi)$.

Definition 3.3.15 A dynamical system represented by $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ with $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$ and $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ is said to be in *input/output form* if it satisfies:

- $\det P(\xi) \neq 0$.
- $P^{-1}(\xi)Q(\xi)$ is a matrix of proper rational functions. (By Cramer's rule, the entries of $P^{-1}(\xi)Q(\xi)$ are rational functions.)

□

Example 3.3.16 Consider the behavior defined in Example 2.3.10. There $P(\xi) = -1 + \xi$ and $Q(\xi) = 1$; hence $P^{-1}(\xi)Q(\xi)$ is proper, and thus it is a system in input/output form. One may check that the pair $(u, y) := (w_1, w_2)$, defined by (2.14), belongs to $\mathfrak{B}$ for every $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$. In other words, the dynamical system defined by (2.13) does not put any restriction on the function u. Also, once u is given, y is completely determined by its past. □

Example 3.3.17 Consider the spring–mass system of Example 3.2.3. Take the force on the mass as u and the position of the mass as y. The equation relating u and y becomes

$$(k_1 + k_2)y + M\frac{d^2}{dt^2}y = u, \tag{3.33}$$

so that $P(\xi) = M\xi^2 + k_1 + k_2$ and $Q(\xi) = 1$. Obviously, in this case $P^{-1}(\xi)Q(\xi) = \frac{1}{k_2+k_1\xi+M\xi^2}$ is proper, and hence (3.33) is in input/output form, with the force as input and the displacement as output. □

Remark 3.3.18 It is the condition that $P^{-1}(\xi)Q(\xi)$ is proper that guarantees that w_2 is a free variable. For instance, in an equation like

$$w_1 = \frac{d}{dt}w_2, \tag{3.34}$$

w_2 is *not* a free variable. To see this, take $w_2(t) = 1$ for $t \geq 0, w_2(t) = 0$ for $t < 0$, and check that there is no $w_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$ such that (3.34) holds; see Exercise 3.25. However, if we had allowed *distributions* (extension of the set of admissible trajectories) or if we had confined ourselves to $\mathcal{C}^\infty$ functions (restriction of admissible trajectories), then w_2 would have been a free variable for the system (3.34). Let us elaborate on the latter case. Consider a system of the form (3.26), where $\det P(\xi) \neq 0$, but where we do not assume that $P^{-1}(\xi)Q(\xi)$ is proper. Choose $k \in \mathbb{N}$ such that $\xi^{-k}P^{-1}(\xi)Q(\xi)$ is proper (convince yourself that such a k always exists), and define $\tilde{P}(\xi) := \xi^k P(\xi)$. Choose $w_2 \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. By Theorem 3.3.13

there exists $\tilde{w_1}$ such that $\tilde{P}(\frac{d}{dt})\tilde{w_1} = Q(\frac{d}{dt})w_2$. Moreover, since w_2 is infinitely differentiable, we conclude from (3.31) that $\tilde{w_1}$ is also infinitely differentiable. In particular, $w_1 := (\frac{d}{dt})^k\tilde{w_1}$ is well-defined and satisfies $P(\frac{d}{dt})w_1 = Q(\frac{d}{dt})w_2$. □

Theorem 3.3.13 combined with Theorem 3.2.16 allows us to to characterize the *complete* behavior of the system (3.26).

Theorem 3.3.19 *Consider the behavior*

$$\mathfrak{B} = \left\{(u,y) \in \mathfrak{L}_1^{\rm loc}(\mathbb{R}, (\mathbb{R}^m \times \mathbb{R}^p)) \mid P(\frac{d}{dt})y = Q(\frac{d}{dt})u, \text{ weakly}\right\}.$$

Then $\mathfrak{B} = \mathfrak{B}_{\rm i/o} + \mathfrak{B}_{\rm hom}$,
with $\mathfrak{B}_{\rm i/o} = \left\{(u, y_{\rm i/o}) \mid u \in \mathfrak{L}_1^{\rm loc}(\mathbb{R}, \mathbb{R}^m) \text{ and } y_{\rm i/o} \text{ is given by (3.31)}\right\}$
and $\mathfrak{B}_{\rm hom} = \{(0, y_{\rm hom}) \mid y_{\rm hom} \text{ is of the form (3.19) }\}$.

Proof This follows from the observation that *every* solution of a linear time-invariant differential equation can be written as the sum of a *particular* solution and a solution of the associated *homogeneous* equation. More precisely, let $(u, y) \in \mathfrak{B}$, and let $y_{\rm i/o}$ be defined by (3.31). By Theorem 3.3.13 we know that $(u, y_{\rm i/o}) \in \mathfrak{B}$. Define $y_{\rm hom} = y - y_{\rm i/o}$. By linearity of $\mathfrak{B}$, we have that $(u, y) - (u, y_{\rm i/o}) \in \mathfrak{B}$, and therefore $(0, y_{\rm hom}) \in \mathfrak{B}$. This implies that $P(\frac{d}{dt})y_{\rm hom} = 0$, and hence $y_{\rm hom}$ is of the form (3.19). This shows that $\mathfrak{B} \subset \mathfrak{B}_{\rm i/o} + \mathfrak{B}_{\rm hom}$.

Conversely, by Theorem 3.3.13, we have that $\mathfrak{B}_{\rm i/o} \subset \mathfrak{B}$. Further, $P(\frac{d}{dt})y_{\rm hom} = 0$ implies that $(0, y_{\rm hom}) \in \mathfrak{B}$, so that also $\mathfrak{B}_{\rm hom} \subset \mathfrak{B}$. Again by linearity, it follows that $\mathfrak{B}_{\rm i/o} + \mathfrak{B}_{\rm hom} \subset \mathfrak{B}$. Hence $\mathfrak{B} = \mathfrak{B}_{\rm i/o} + \mathfrak{B}_{\rm hom}$. □

The following corollary expresses that the past of the output does not restrict the future of the input.

Corollary 3.3.20 *Let* $(u, y) \in \mathfrak{B}$ *as in Theorem 3.3.19. Let* $\tilde{u} \in \mathfrak{L}_1^{\rm loc}(\mathbb{R}, \mathbb{R}^m)$ *be such that* $\tilde{u}(t) = u(t)$, $t \le t_0$, *for some* $t_0 \in \mathbb{R}$. *Then there exists* $\tilde{y}$ *such that* $(\tilde{u}, \tilde{y}) \in \mathfrak{B}$ *and* $\tilde{y}(t) = y(t)$, $t \le t_0$.

Proof By time invariance we may assume that $t_0 = 0$. Since $(u, y) \in \mathfrak{B}$, it follows from Theorem 3.3.19 that

$$y(t) = y_h(t) + A_0u(t) + \sum_{i=1}^{N}\sum_{j=1}^{n_i} A_{ij}\int_0^t \frac{(t-\tau)^{(j-1)}}{(j-1)!}e^{\lambda_i(t-\tau)}u(\tau)d\tau, \quad (3.35)$$

where y_h satisfies $P(\frac{d}{dt})y_h = 0$. Now simply define $\tilde{y}$ as in (3.35), but with u replaced by $\tilde{u}$. Since $\tilde{u}(t) = u(t)$ for $t \le 0$, it follows that $\tilde{y}(t) = y(t)$ for $t \le 0$. □

Remark 3.3.21 Corollary 3.3.20 implies *nonanticipation*. Indeed, the past of the output is not restricted by the future of the input. Nor is the future of the input restricted by the past of the output. This implies that y does *not anticipate* u, or simply that the relation between the input and the output is *nonanticipating*. We say that y does not anticipate u *strictly* if for all $(u, y) \in \mathfrak{B}$ and $\tilde{u} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ such that $\tilde{u}(t) = u(t)$, $t < t_0$ (notice the difference; the inequality $t \leq t_0$ has been replaced by a strict inequality) for some $t_0 \in \mathbb{R}$, there exists $\tilde{y}$ such that $(\tilde{u}, \tilde{y}) \in \mathfrak{B}$ and $\tilde{y}(t) = y(t)$, $t \leq t_0$. In other words, inputs that are equal in the strict past generate outputs that are equal up to and including the present.

Consider, for instance, the mass–spring system of Example 3.2.3. Suppose the system is in its equilibrium position for $t < 0$. At $t = 0$ the force on the mass could be changed abruptly, causing the mass to leave its equilibrium. It is clear that in principle the force that can be applied from $t = 0$ on is not restricted by the fact that the mass was in its equilibrium position before that time instant. And of course the input force after $t = 0$ has no influence on the position of the mass before $t = 0$, so that indeed the position of the mass does not anticipate the force. Thus in this system the force is the input, the position is the output, and the system is strictly nonanticipating. □

Let us now return to the differential equation $R(\frac{d}{dt})w = 0$ given by (2.5). We have seen in Section 2.5 that there always exists an equivalent system with the corresponding $R(\xi)$ of full row rank. We may thus assume without loss of generality that $R(\xi)$ has indeed full row rank. The question that we now want to consider is the following. Is there a sense in which this system can be viewed as an input/output system? In other words, is there a partition of w, possibly after permutation of the components, as $w = (u, y)$ such that (u, y) satisfies (3.26)? The answer to this question is *yes*. We now explain in what sense this is so.

Theorem 3.3.22 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ be of full row rank. If $g < q$, then there exists a choice of columns of $R(\xi)$: $c_{i_1}(\xi), \ldots, c_{i_g}(\xi) \in \mathbb{R}^{g\times q}[\xi]$ such that*

1. $\det \begin{bmatrix} c_{i_1}(\xi) & \cdots & c_{i_g}(\xi) \end{bmatrix} \neq 0.$
2. $\begin{bmatrix} c_{i_1}(\xi) & \cdots & c_{i_g}(\xi) \end{bmatrix}^{-1} \begin{bmatrix} c_{i_{g+1}}(\xi) & \cdots & c_{i_q}(\xi) \end{bmatrix}$ *is a proper rational matrix.*

Proof Choose $R_1(\xi)$ as a $g \times g$ nonsingular submatrix of $R(\xi)$ such that the degree of its determinant is maximal among the $g \times g$ submatrices of $R(\xi)$. Since $R(\xi)$ has full row rank, we know that $\det R_1(\xi)$ is not the zero polynomial. Denote the matrix formed by the remaining columns of $R(\xi)$

by $R_2(\xi)$. We claim that $R_1^{-1}(\xi)R_2(\xi)$ is proper. To see that, notice that by Cramer's rule, the ijth entry of $R_1^{-1}(\xi)R_2(\xi)$ is given by

$$\big[R_1^{-1}(\xi)R_2(\xi)\big]_{ij} = \frac{\det R_{1ij}(\xi)}{\det R_1(\xi)},$$

where the matrix $R_{1ij}(\xi)$ is obtained by replacing the ith column of $R_1(\xi)$ by the jth column of $R_2(\xi)$. Since the determinant of $R_1(\xi)$ is maximal among the $g\times g$ submatrices of $R(\xi)$, it follows that $\deg\det R_{1ij}(\xi) \leq \deg\det R_1(\xi)$. This implies that $R_1^{-1}(\xi)R_2(\xi)$ is proper. □

Corollary 3.3.23 *Let $R(\frac{d}{dt})w = 0$, $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$, be a full row rank representation of the behavior $\mathfrak{B}$. $\mathfrak{B}$ admits an i/o representation in the sense of Definition 3.3.1 with, in the notation of Theorem 3.3.22, input $u = \mathrm{col}(w_{i_{g+1}}, \ldots, w_{i_q})$ and output $y = \mathrm{col}(w_{i_1}, \ldots, w_{i_g})$.*

Proof If $g = q$, then $\det R(\xi) \neq 0$, so that by Corollary 3.3.23 the behavior is autonomous. In other words, w does not contain any free components at all. This is a special case of an input/output representation, a behavior with outputs only.

Assume that $g < q$. In view of Theorem 2.5.23 we may assume that $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ has full row rank. According to Theorem 3.3.22 we can choose a $g \times g$ submatrix $R_1(\xi)$ of $R(\xi)$ such that $R_1^{-1}(\xi)R_2(\xi)$ is proper (of course, $R_2(\xi)$ is the submatrix of $R(\xi)$ consisting of the remaining columns). Partition w according to the choice of columns that led to $R_1(\xi)$ and $R_2(\xi)$ as (y, u). Thus if $R_1(\xi) = [c_{i_1}(\xi)\cdots c_{i_g}(\xi)]$, then $y = \mathrm{col}(w_{i_1}, \ldots, w_{i_g})$, and u contains the remaining components of w, $u = \mathrm{col}(w_{i_{g+1}}, \ldots, w_{i_q})$. The equation $R(\frac{d}{dt})w = 0$ can now equivalently be written as

$$R_1(\frac{d}{dt})y = -R_2(\frac{d}{dt})u. \tag{3.36}$$

Since by Theorem 3.3.22, $R_1^{-1}(\xi)R_2(\xi)$ is proper, it follows that (3.36) is indeed an i/o representation of $R(\frac{d}{dt})w = 0$. □

As a last result in this section we mention that a system in input/output form $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ with $P^{-1}(\xi)Q(\xi)$ proper may be transformed into a strictly proper input/output system by means of a static right unimodular transformation.

Theorem 3.3.24 *Let $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$ and $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ with $P^{-1}(\xi)Q(\xi)$ proper. Consider the input/output system $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$. Then there exists a matrix $M \in \mathbb{R}^{p\times m}$ such that the relation between u and $y' := y + Mu$ is a strictly proper input/output relation.*

Proof If $P^{-1}(\xi)Q(\xi)$ is strictly proper then there is nothing to prove. Assume that $P^{-1}(\xi)Q(\xi)$ is proper, but not strictly proper. Let the partial

fraction expansion of $P^{-1}(\xi)Q(\xi)$ be given by

$$P^{-1}(\xi)Q(\xi) = A_0 + \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{A_{ij}}{(\xi-\lambda_i)^j}. \tag{3.37}$$

By multiplying both sides of (3.37) from the left by $P(\xi)$ we get

$$Q(\xi) = P(\xi)A_0 + \tilde{Q}(\xi),$$

where $\tilde{Q}(\xi) \in \mathbb{R}^{p\times m}[\xi]$. The input/output equation $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ can now be rewritten as $P(\frac{d}{dt})y = (P(\frac{d}{dt})A_0 + \tilde{Q}(\frac{d}{dt}))u$, so that

$$P(\frac{d}{dt})[y - A_0u] = \tilde{Q}(\frac{d}{dt})u.$$

Define $M = -A_0$, and since

$$P^{-1}(\xi)\tilde{Q}(\xi) = P^{-1}(\xi)Q(\xi) - A_0 = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{A_{ij}}{(\xi-\lambda_i)^j},$$

it follows that $P^{-1}(\xi)\tilde{Q}(\xi)$ is strictly proper, and $P(\frac{d}{dt})y' = \tilde{Q}(\frac{d}{dt})u$. □

Example 3.3.25 Consider the i/o system defined by

$$y + 2\frac{d}{dt}y + \frac{d^2}{dt^2}y = u - 3\frac{d}{dt}u + 4\frac{d^2}{dt^2}u,$$

The corresponding polynomials are $p(\xi) = 1+2\xi+\xi^2$ and $q(\xi) = 1-3\xi+4\xi^2$. It is easily checked that $q(\xi) = 4p(\xi) - 3 - 11\xi$, from which it follows that

$$p(\frac{d}{dt})(y - 4u) = -3u - 11\frac{d}{dt}u.$$

Indeed, the relation between u and $y - 4u$ is a strictly proper i/o relation. □

Remark 3.3.26

1. The partition into input and output as given in Corollary 3.3.23 is in general *not* unique, since there may be more than just one choice of square submatrices with maximal determinant degree. As a trivial example, which we already discussed, consider the behavior defined by $w_1 = w_2$. It is obvious that either one of the two components of w can be viewed as the input variable. As another example, consider the manifest behavior of the RLC circuit of Example 1.3.5. The manifest behavior is described by (1.12) or (1.13). In both cases the input can be chosen to be the current I or the voltage V.

2. If $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ has full row rank and $g \leq q$, then for every i/o partition of w, it can be proven that the number of outputs (the number of components in the output vector) is g, and the number of inputs is $q - g$. See Exercise 3.27.

3. In our choice of the notions of input and output, we have opted to call those variables input that are *free* in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$. Thus in the system $\frac{d}{dt}w_1 + w_2 = 0$, we have chosen to call w_2 input and w_1 output but not vice versa. See Remark 3.3.18. This is in keeping with common usage in systems theory. If, however, we had considered only sufficiently smooth signals w_1, say in $\mathcal{C}^1(\mathbb{R}, \mathbb{R})$, then in the above equation w_1 would be free, and in this sense we could have considered (contrary to our chosen nomenclature) w_1 as an input and $w_2 = -\frac{d}{dt}w_1$ as the resulting output. There are, in fact, many useful devices that differentiate signals. For example, tachometers measure the position of a moving object and output the velocity, and it is certainly reasonable to call the position the input and the velocity the output of a tachometer.

□

Example 3.3.27 Consider the electrical circuit shown in Figure 3.3.

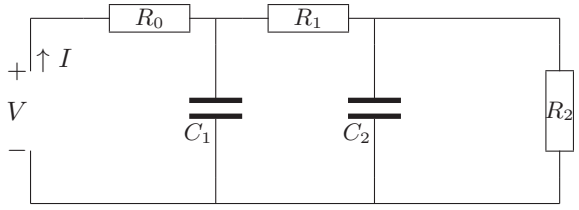


FIGURE 3.3. Electrical circuit.

Assume that C_1, C_2, R_1, R_2 are strictly positive and that $R_0 \geq 0$. After introducing as latent variables the voltages and currents in the internal branches and eliminating them (we omit these calculations), we arrive at the following behavioral equations,

$$\begin{aligned}&\left[1 + (C_1R_1 + C_1R_2 + C_2R_1 + C_2R_2)\tfrac{d}{dt} + C_1C_2R_1R_2\tfrac{d^2}{dt^2}\right]V = \\ &\left[R_0 + R_1 + R_2 + (C_1R_0R_1 + C_1R_0R_2 + C_2R_0R_2 + C_2R_1R_2)\tfrac{d}{dt}\right. \\ &\left.+(C_1C_2R_0R_1R_2)\tfrac{d^2}{dt^2}\right]I,\end{aligned}$$

as the differential equation describing the relation between the port voltage V and the port current I of this RC-circuit. In terms of our standard

notation, we have, with $w = \mathrm{col}(V, I)$,

$$R(\xi) = \begin{bmatrix} R_1(\xi) & R_2(\xi) \end{bmatrix},$$

where $R_1(\xi)$ and $R_2(\xi)$ are given by

$$\begin{aligned} R_1(\xi) = \quad & 1 + (C_1R_1 + C_1R_2 + C_2R_2)\xi + C_1C_2R_1R_2\xi^2 \\ R_2(\xi) = \quad & -(R_0 + R_1 + R_2) - (C_1R_0R_1 + C_1R_0R_2 + C_2R_0R_2 + C_2R_1R_2)\xi \\ & -(C_1C_2R_0R_1R_2)\xi^2 \end{aligned}$$

When $R_0 > 0$, it follows that V can be chosen as input and I as output or, conversely, that I can be seen as input and V as output. In the former case, the circuit is considered as an admittance (the voltage is input and the current is the output —the term *voltage controlled* is also used). In the latter case, the circuit is considered as an impedance (the current is input and the voltage is output—the circuit is *current controlled*).

When $R_0 = 0$, the input/output choice is restricted to I as input and V as output.

Above we have considered the input/output structure of the port variables (V, I). However, circuits as the one shown in Figure 3.3 can be used as filters relating the voltage V_{in} at the external port to the voltage V_{out} across the resistor R_2; see Figure 3.4.

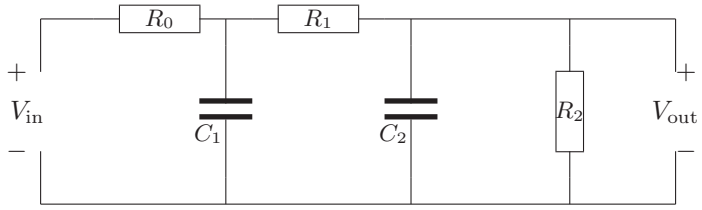


FIGURE 3.4. Electrical circuit.

The relation between V_{in} and V_{out} is (again we omit the details)

$$\begin{aligned} V_{\mathrm{in}} = & (1 + \frac{R_0 + R_1}{R_2}) + (C_2R_1 + C_1R_0 + C_2R_0 + \frac{C_1R_0R_1}{R_2})\tfrac{d}{dt} \\ & +(C_1C_2R_0R_1)\tfrac{d^2}{dt^2})V_{\mathrm{out}}. \end{aligned}$$

It follows from that in order to satisfy the requirements of Definition 3.3.1, we have to take V_{in} as input and V_{out} as output. □

3.4 Systems Defined by an Input/Output Map

We now briefly study dynamical systems that are defined by a *map* between the set of functions u and the set of functions y of the integral form

$$y(t) = \int_{-\infty}^{\infty} H(t,\tau)u(\tau)d\tau, \qquad H \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}^2, \mathbb{R}^{p\times m}). \tag{3.38}$$

The function H is called the *kernel* of the integral representation (3.38) (not to be confused with the kernel of a linear map). In order to define the behavior specified by (3.38), we have to define the set of admissible trajectories (u, y). A first attempt might be to allow every function u for which the integral (3.38) exists. This, however, has the effect that the set of admissible trajectories depends on the particular system, which is undesirable. Therefore, we take the set of all functions $(u, y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^p)$ for which u has *compact support.* We say that the function $u : \mathbb{R} \to \mathbb{R}^m$ has compact support if the set on which u is nonzero is bounded. The behavior $\mathfrak{B}$ corresponding to (3.38) is now defined as

$$\mathfrak{B} := \{(u, y) \mid u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m), u \text{ has compact support, } (u, y) \text{ satisfies (3.38)}\}. \tag{3.39}$$

Property 3.4.1 *The system defined by (3.39) has the following properties:*

- *It is linear.*
- *In general, it is time-varying. It is time-invariant if and only if for all t' and for all (t,τ), $H(t + t', \tau + t') = H(t,\tau)$; i.e., $H(t,\tau) = H(t - \tau, 0)$.*
- *u is a free variable, in the sense that for all $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ with compact support there corresponds a $y \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ such that $(u, y) \in \mathfrak{B}$.*

Remark 3.4.2 (3.38) defines a *map* from the set of input functions to the set of output functions, whereas the equation (2.5) only defines a *relation* on the Cartesian product of these two sets. Indeed, in (2.5) in Corollary 3.3.14 we saw that for every $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ there exists an n-dimensional (with $n = \deg \det P(\xi)$) affine subspace of corresponding outputs. □

Example 3.4.3 (Population dynamics) Let $y(t)$ denote the size of the population of a certain species at time t and let $P(t,\tau)$ denote the probability that an individual that is born at time $t - \tau$ is still alive at time t. Let $u(t)$ be the rate of births per time unit. A model that describes the relation between u and y is

$$y(t) = \int_{-\infty}^{\infty} P(t,\tau)u(t - \tau)d\tau. \tag{3.40}$$

For obvious reasons, $P(t,\tau) = 0$ for $t < \tau$, and hence, after a change of variables, (3.40) can be written as

$$y(t) = \int_{-\infty}^{t} P(t, t-\tau)u(\tau)d\tau.$$

A further refinement of the model is obtained when a maximum age is introduced:

$$y(t) = \int_{t-m}^{t} P(t, t-\tau)u(\tau)d\tau,$$

where m is the maximum age that the species can reach. Furthermore, time-invariance follows if $P(t,\tau)$ depends only on $t-\tau$, which is in many applications an acceptable assumption.

□

We are mainly concerned with systems of the form (3.39) that are time-invariant, i.e., for which $H(t,\tau)$ depends on (t,τ) through the difference of t and τ. With abuse of notation $H(t,\tau)$ is then written as a function of *one* argument: $H(t)$. Moreover, we assume that the system is *nonanticipating*: $H(t) = 0$ for $t < 0$. The system map then becomes

$$y(t) = \int_{-\infty}^{t} H(t-\tau)u(\tau)d\tau. \tag{3.41}$$

Since (3.41) is the convolution product of H and u, systems of the form (3.41) are called *nonanticipating convolution systems.*

For *convolution systems*, the kernel H is usually referred to as the *impulse response* of the system. The reason for this terminology is, loosely speaking, that if an input is applied that is zero everywhere and a pulse at time 0, then the output of the system is exactly the function H. This can be made precise by making use of the theory of distributions, but it can be explained intuitively by applying a sequence of inputs that approach a pulse. To that end, define

$$u_n(t) = \begin{cases} n & 0 \le t \le \dfrac{1}{n}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.42}$$

If H is continuous, then the response of the system to the input u_n behaves like $H(t)$ as n tends to infinity. Indeed,

$$\lim_{n\to\infty} \int_{-\infty}^{t} H(t-\tau)u_n(\tau)d\tau = \lim_{n\to\infty} n \int_{0}^{\frac{1}{n}} H(t-\tau)d\tau = H(t). \tag{3.43}$$

Of course, (3.43) is just an intuitive justification of the terminology used. Although the sequence of input functions u_n does not converge in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$, u_n can be seen as an approximation of a pulse at time $\tau = 0$. In the sense of distributions, u_n indeed converges to the celebrated Dirac δ function.

Example 3.4.4 Consider the convolution system with kernel,

$$H(t) = \begin{cases} e^{-t} & t \geq 0; \\ 0 & t < 0. \end{cases}$$

Thus

$$y(t) = \int_{-\infty}^{t} e^{-(t-\tau)} u(\tau) d\tau.$$

Let u_n be given by (3.42). Then

$$y_n(t) = \begin{cases} n\displaystyle\int_0^t H(t-\tau)d\tau = n\int_0^t e^{-(t-\tau)}d\tau = e^{-t}n(e^t-1), & 0 \leq t \leq \frac{1}{n}, \\ n\displaystyle\int_0^{\frac{1}{n}} H(t-\tau)d\tau = n\int_0^{\frac{1}{n}} e^{-(t-\tau)}d\tau = e^{-t}n(e^{\frac{1}{n}}-1), & t \geq \frac{1}{n}. \end{cases} \tag{3.44}$$

From (3.44) it follows that indeed

$$\lim_{n\to\infty} y_n(t) = e^{-t} = H(t) \quad \text{for all } t.$$

In Figure 3.5 we have depicted H and $y_1, y_2, y_3, \ldots, y_{10}$. □

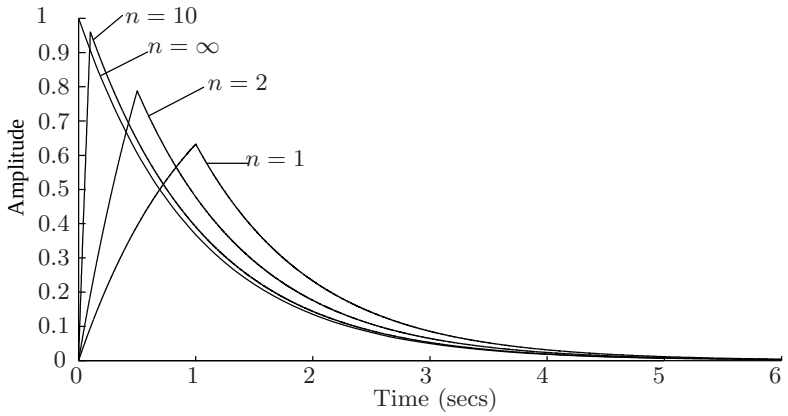


FIGURE 3.5. Approximation of the impulse response for $n = 1, 2, 10, \infty$.

3.5 Relation Between Differential Systems and Convolution Systems

In this section we study the relation between convolution systems and input/output systems described by differential equations of the form (3.26). Until now we have defined the set of admissible trajectories for systems described by differential equations to be $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. We now restrict the behavior to functions in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ that are zero at $-\infty$ and show that the resulting behavior can be described by a nonanticipating convolution system. By "zero at $-\infty$" we mean that the function is zero before some time. More precisely, for each such function w, there exists a $t' \in \mathbb{R}$ such that $w(t) = 0$ for all $t < t'$. For convenience we will distinguish the $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ case from the restricted case by referring to the latter as *systems initially at rest*.

Definition 3.5.1 Let the (linear differential) dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$ be given. With $\mathfrak{B}$ we associate the *initially at rest behavior* $\mathfrak{B}_0$:

$$\mathfrak{B}_0 := \{w \in \mathfrak{B} \mid \exists t_0 \text{ such that for all } t \leq t_0 : \ w(t) = 0\}.$$

Note that t_0 is not fixed; it may depend on w. □

Theorem 3.5.2

(i) Let $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$, $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ be such that $\det P(\xi) \neq 0$ and that $P^{-1}(\xi)Q(\xi)$ is strictly proper and assume that the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$ is given by

$$P^{-1}(\xi)Q(\xi) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{A_{ij}}{(\xi - \lambda_i)^j}.$$

Then the initially at rest behavior of the system defined by $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ is also described by the convolution system of the form (3.41) with H given by

$$H(t) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} A_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t} \qquad (t \geq 0).$$

(ii) Consider the convolution system described by $y(t) = \int_{-\infty}^{t} H(t - \tau)u(\tau)d\tau$. There exist polynomial matrices $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$, $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ such that the initially at rest behavior of the convolution system is also described by $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ if and only if H is of the

form

$$H(t) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} A_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t} \qquad (t \geq 0) \tag{3.45}$$

for some N, $n_i \in \mathbb{N}$, $A_{ij} \in \mathbb{C}^{p \times m}$, $\lambda_i \in \mathbb{C}$ *such that the complex* $\lambda_i s$ *come in complex conjugate pairs and the corresponding matrices* A_{ij} *also come in complex conjugate pairs.*

Proof (i) Define

$$H(t) = \begin{cases} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{n_i} A_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t} & \text{for } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, by Theorem 3.3.13, (3.31), and since we restrict our attention to the initially at rest part of the system, every solution (u, y) of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ satisfies

$$y(t) = \int_{-\infty}^{t} H(t-\tau)u(\tau)d\tau.$$

This proves part (i).

(ii) Define the *rational* matrix $T(\xi) \in \mathbb{R}^{p \times m}(\xi)$ by

$$T(\xi) := \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{A_{ij}}{(\xi - \lambda_i)^j}.$$

We want to find polynomial matrices $P(\xi)$ and $Q(\xi)$ such that $P^{-1}(\xi)Q(\xi) = T(\xi)$. This is easy. Define $d(\xi) \in \mathbb{R}[\xi]$ as

$$d(\xi) := \prod_{i=1}^{N} (\xi - \lambda_i)^{n_i}$$

and take $P(\xi)$ such that $(\xi - \lambda_i)^{n_i}$ divides $P(\xi)$ and such that $\det P(\xi) \neq 0$, e.g.,

$$P(\xi) := d(\xi) I_p \quad (I_p \text{ is the } p \times p \text{ identity-matrix}).$$

Finally, define $Q(\xi)$ as

$$Q(\xi) := P(\xi)T(\xi) \ (= d(\xi)T(\xi)).$$

For the the single-input/single-output case this comes down to

$$P(\xi) := \prod_{i=1}^{N} (\xi - \lambda)^{n_i} \quad \text{and} \quad Q(\xi) := P(\xi)T(\xi).$$

□

Remark 3.5.3 A function of the form (3.45) is called a (matrix of) *Bohl function*. A Bohl function is a finite sum of products of polynomials and exponentials. In the real case, a Bohl function is a finite sum of products of polynomials, real exponentials, sines, and cosines. □

3.6 When Are Two Representations Equivalent?

In Chapter 2, Theorem 2.5.4, we have seen that if $U(\xi)$ is unimodular, then $R(\xi)$ and $U(\xi)R(\xi)$ represent the same behavior. In this section we ask the converse question: *What is the relation between two matrices $R_1(\xi)$ and $R_2(\xi)$ that define the same behavior?* It turns out that if these matrices have the same number of rows, then $R_1(\xi)$ and $R_2(\xi)$ define the same behavior if and only if there exists a unimodular matrix $U(\xi)$ such that $R_2(\xi) = U(\xi)R_1(\xi)$.

Preparatory to this result, we prove the following lemma.

Lemma 3.6.1 *Let $P_1(\xi), P_2(\xi) \in \mathbb{R}^{q\times q}[\xi]$, with $\det P_1(\xi) \neq 0$. Denote the corresponding behaviors by $\mathfrak{B}_{P_1}$ and $\mathfrak{B}_{P_2}$ respectively. If $\mathfrak{B}_{P_1} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) = \mathfrak{B}_{P_2} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, then there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{q\times q}[\xi]$ such that $P_1(\xi) = U(\xi)P_2(\xi)$.*

Proof The proof goes by induction on q. Let $q = 1$. It follows from Theorem 3.2.5 that $\mathfrak{B}_{P_1} = \mathfrak{B}_{P_2}$ implies that the (scalar) polynomials $P_1(\xi)$ and $P_2(\xi)$ have the same roots. This can only be the case if $P_1(\xi) = uP_2(\xi)$ for some nonzero constant u. This yields the statement for the scalar case.

Assume now that the result is true for $q \leq n$, and let $P_i(\xi) \in \mathbb{R}^{(n+1)\times(n+1)}[\xi](i = 1, 2)$. By Theorem 2.5.14 (upper triangular form) it follows that by premultiplication by suitable unimodular matrices, both $P_1(\xi)$ and $P_2(\xi)$ can be transformed into the form

$$P_1(\xi) = \begin{bmatrix} P_{11}^{(1)} & P_{12}^{(1)} \\ 0 & P_{22}^{(1)} \end{bmatrix}, \qquad P_2(\xi) = \begin{bmatrix} P_{11}^{(2)} & P_{12}^{(2)} \\ 0 & P_{22}^{(2)} \end{bmatrix}$$

with $P_{11}^{(i)} \in \mathbb{R}^{n\times n}[\xi], i = 1, 2$. Partition w as $w = \mathrm{col}(w_1, w_2)$ with w_2 scalar. Choose w_1 such that $P_{11}^{(1)}(\frac{d}{dt})w_1 = 0$. Then $\mathrm{col}(w_1, 0) \in \mathfrak{B}_{P_1}$, and therefore also $\mathrm{col}(w_1, 0) \in \mathfrak{B}_{P_2}$, and hence $P_{11}^{(2)}(\frac{d}{dt})w_1 = 0$. The converse is, of course, also true, and by the induction hypothesis we conclude that there exists a unimodular matrix $U_{11}(\xi) \in \mathbb{R}^{n\times n}[\xi]$ such that $P_{11}^{(2)}(\xi) = U_{11}(\xi)P_{11}^{(1)}(\xi)$.

We now show that $P_{22}^{(1)}(\xi) = \alpha P_{22}^{(2)}(\xi)$ for some nonzero constant α. Choose w_2 such that $P_{22}^{(1)}(\frac{d}{dt})w_2 = 0$. Since by Theorem 3.2.5, w_2 is in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ and since $\det P_{11}^{(1)}(\xi) \neq 0$, it follows from Remark 3.3.18 that

there exists w_1 such that $P_{11}^{(1)}(\frac{d}{dt})w_1 + P_{12}^{(1)}(\frac{d}{dt})w_2 = 0$. In other words, $w \in \mathfrak{B}_{P_1}$. Then, by assumption, w also belongs to $\mathfrak{B}_{P_2}$, and therefore in particular, $P_{22}^{(2)}(\frac{d}{dt})w_2 = 0$. In the same way we conclude the converse, that $P_{22}^{(2)}(\frac{d}{dt})w_2 = 0$ implies $P_{22}^{(1)}(\frac{d}{dt})w_2 = 0$. This shows that indeed $P_{22}^{(1)}(\xi) = \alpha P_{11}^{(2)}(\xi)$ for some constant α.

What we have obtained thus far is that $P_1(\xi)$ and $P_2(\xi)$ are of the form

$$P_1(\xi) = \begin{bmatrix} P_{11}^{(1)}(\xi) & P_{12}^{(1)}(\xi) \\ 0 & P_{22}^{(1)}(\xi) \end{bmatrix}, \quad P_2(\xi) = \begin{bmatrix} U_{11}(\xi)P_{11}^{(1)}(\xi) & P_{12}^{(2)}(\xi) \\ 0 & \alpha P_{22}^{(1)}(\xi) \end{bmatrix}.$$

This is almost what we are after, except that the upper right corner of $P_2(\xi)$ still needs to be expressed in terms of $P_1(\xi)$. To derive such an expression, choose w_2 such that $P_{22}(1)(\frac{d}{dt})w_2 = 0$. As before, there exists w_1 such that

$$P_{11}^{(1)}(\frac{d}{dt})w_1 + P_{12}^{(1)}(\frac{d}{dt})w_2 = 0, \quad U_{11}(\frac{d}{dt})P_{11}^{(1)}(\frac{d}{dt})w_1 + P_{12}^{(2)}(\frac{d}{dt})w_2 = 0,$$

and therefore

$$\begin{aligned} &U_{11}(\tfrac{d}{dt})P_{11}^{(1)}(\frac{d}{dt})w_1 + U_{11}(\tfrac{d}{dt})P_{12}^{(1)}(\frac{d}{dt})w_2 = 0, \\ &U_{11}(\tfrac{d}{dt})P_{11}^{(1)}(\frac{d}{dt})w_1 + P_{12}^{(2)}(\frac{d}{dt})w_2 = 0. \end{aligned} \tag{3.46}$$

Subtraction of the two equations in (3.46) yields

$$\left(U_{11}(\frac{d}{dt})P_{12}^{(1)}(\frac{d}{dt}) - P_{12}^{(2)}(\frac{d}{dt})\right) w_2 = 0. \tag{3.47}$$

Hence $P_{22}(1)(\frac{d}{dt})w_2 = 0$ implies (3.47). It is not difficult to check (see Exercise 3.2) that therefore $P_{22}(1)(\xi)$ divides every entry of the polynomial vector $(U_{11}(\xi)P_{12}^{(1)}(\xi) - P_{12}^{(2)}(\xi))$, and hence there exists a polynomial vector $U_{12}(\xi) \in \mathbb{R}^{g\times 1}[\xi]$ such that $P_{12}^{(2)}(\xi) - U_{11}(\xi)P_{12}^{(1)}(\xi) = U_{12}(\xi)P_{22}(1)(\xi)$.

It follows that

$$P_2(\xi) = \underbrace{\begin{bmatrix} U_{11}(\xi) & U_{12}(\xi) \\ 0 & \alpha \end{bmatrix}}_{U(\xi)} P_1(\xi). \tag{3.48}$$

Since $U_{11}(\xi)$ is unimodular by the induction hypothesis and α is a nonzero constant, the matrix $U(\xi)$ in (3.48) is unimodular, and the lemma is proven. □

We are now ready to state and prove the converse of Theorem 2.5.4.

Theorem 3.6.2 *The polynomial matrices $R_1(\xi), R_2(\xi) \in \mathbb{R}^{g\times q}[\xi]$ define the same behavior $\mathfrak{B}$ if and only if there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that $R_1(\xi) = U(\xi)R_2(\xi)$.*

Proof The "if" part was already proven in Theorem 2.5.4. The "only if" part may be proved as follows. By elementary row operations, $R_1(\xi)$ and $R_2(\xi)$ may be transformed into

$$\begin{bmatrix} \tilde{R}_i(\xi) \\ 0 \end{bmatrix}, \quad i = 1, 2$$

with $\tilde{R}_i(\xi)$ of full row rank. According to Corollary 3.3.23, $\tilde{R}_1(\xi)$ may be written in the form

$$\tilde{R}_1(\xi) = \begin{bmatrix} \tilde{P}_1(\xi) & \tilde{Q}_1(\xi) \end{bmatrix} \tag{3.49}$$

with $\tilde{P}_1^{-1}(\xi)\tilde{Q}_1(\xi)$ proper. To obtain (3.49) could require a permutation of the columns of $\tilde{R}_1(\xi)$. It is, however, not a restriction to assume that $\tilde{R}_1(\xi)$ is already in the form (3.49). Partition $\tilde{R}_2(\xi)$ accordingly:

$$\tilde{R}_2(\xi) = \begin{bmatrix} \tilde{P}_2(\xi) & \tilde{Q}_2(\xi) \end{bmatrix}.$$

Choose w_1 such that $\tilde{P}_1(\frac{d}{dt})w_1 = 0$. Then $\mathrm{col}(w_1, 0) \in \mathfrak{B}$, and therefore also $\tilde{P}_2(\frac{d}{dt})w_1 = 0$. Conversely, $\tilde{P}_2(\frac{d}{dt})w_1 = 0$ implies that $\tilde{P}_1(\frac{d}{dt})w_1 = 0$. Since $\det \tilde{P}_1(\xi) \neq 0$, it follows from Lemma 3.6.1 that there exists a unimodular matrix $U(\xi)$ such that

$$\tilde{P}_2(\xi) = U(\xi)\tilde{P}_1(\xi). \tag{3.50}$$

Choose w_2 arbitrarily. Then, since $\tilde{P}_1^{-1}(\xi)\tilde{Q}_1(\xi)$ is proper, there exists w_1 such that

$$\tilde{P}_1(\frac{d}{dt})w_1 + \tilde{Q}_1(\frac{d}{dt})w_2 = 0,$$

and hence also $\tilde{P}_2(\frac{d}{dt})w_1 + \tilde{Q}_2(\frac{d}{dt})w_2 = 0$. By (3.50) it follows that

$$\tilde{P}_2(\frac{d}{dt})w_1 + U(\frac{d}{dt})\tilde{Q}_1(\frac{d}{dt})w_2 = 0 \quad \text{and} \quad \tilde{P}_2(\frac{d}{dt})w_1 + \tilde{Q}_2(\frac{d}{dt})w_2 = 0. \tag{3.51}$$

Subtracting the two equations in (3.51) yields that for all w_2

$$\left(U(\frac{d}{dt})\tilde{Q}_1(\frac{d}{dt}) - \tilde{Q}_2(\frac{d}{dt})\right) w_2 = 0. \tag{3.52}$$

Equation (3.52) can only hold if $\tilde{Q}_2(\xi) = U(\xi)\tilde{Q}_1(\xi)$, see Exercise 3.34. This shows that indeed $R_2(\xi) = U(\xi)R_1(\xi)$. □

Corollary 3.6.3 *Two matrices $R_1(\xi) \in \mathbb{R}^{g_1 \times q}[\xi]$ and $R_2(\xi) \in \mathbb{R}^{g_2 \times q}[\xi]$ define the same behavior if and only if $R_1(\xi)$ can be obtained from $R_2(\xi)$ by a series of operations of the following type:*

1. *Premultiplication by a unimodular matrix.*
2. *Addition or deletion of zero-rows.*

Proof See Exercise 3.35. □

In Chapter 2 we introduced the notion of *full row rank* and *minimal* of the system described by $R(\frac{d}{dt})w = 0$. We are now able to prove the claim made in Chapter 2 that these notions coincide.

Theorem 3.6.4 *Let the behavior $\mathfrak{B}$ be defined by $R(\frac{d}{dt})w = 0$, where, as usual, $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$.*

1. *The polynomial matrix $R(\xi)$ has full row rank if and only if it is a minimal representation.*
2. *All full row rank representations have the same number of rows.*
3. *Let $R_1(\xi)$ and $R_2(\xi)$ be two minimal representations of $\mathfrak{B}$. Then there exists a unimodular matrix $U(\xi)$ such that $U(\xi)R_1(\xi) = R_2(\xi)$.*

Proof 1. The "if" part was proven in Theorem 2.5.25. For the "only if" part we proceed as follows. Assume that $R(\xi)$ is of full row rank, and suppose that $R(\xi)$ is not minimal. Then there exists a representation $R'(\xi) \in \mathbb{R}^{g'\times q}[\xi]$ of $\mathfrak{B}$ with $g' < g$. It follows from Theorem 3.6.2 that there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that

$$U(\xi)R(\xi) = \begin{bmatrix} R'(\xi) \\ 0 \end{bmatrix}. \tag{3.53}$$

From Theorem 2.5.22 we know that the row rank of a matrix is invariant under premultiplication by a unimodular matrix, and therefore the row rank of the right-hand side of (3.53) is equal to g, the row rank of $R(\xi)$. Since zero-rows do not contribute to the row rank, $R'(\xi)$ should also have row rank equal to g. This is impossible, since the number of rows of $R'(\xi)$ was assumed to be strictly less than g'.

2. By definition of minimality, all minimal representations have the same number of rows. Since minimality and full row rank are the same, the statement follows.

3. This follows from Theorem 3.6.2 and part 2. □

3.7 Recapitulation

In this chapter we have discussed two important classes of systems: *autonomous* and *input/output* systems. The main points in Chapter 3 are:

- A system is autonomous if for any trajectory in the behavior, its future is completely determined by its past (Definition 3.2.1).

- Autonomous systems can be described by differential equations $R(\frac{d}{dt})w = 0$ with $R(\xi)$ a square polynomial matrix such that $\det R(\xi) \neq 0$. The behavior of such a system can be described quite explicitly through the roots of $\det R(\xi)$ and the vectors in the kernel of $R(\lambda)$ for the roots λ of $\det R(\xi)$ (Theorem 3.2.16). In the scalar case the behavior is completely determined by the roots of the scalar polynomial that defines the behavior (Theorem 3.2.5).
- For autonomous systems, every trajectory in the behavior is almost everywhere equal to an infinitely differentiable one (Theorem 3.2.4).
- Input/output systems can be described by $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ with the matrix of rational functions $P^{-1}(\xi)Q(\xi)$ proper. This matrix is referred to as the *transfer function* or *transfer matrix* and plays an important role in all of systems and control theory. The behavior of an input/output system can be expressed explicitly in terms of an integral expression involving the partial fraction expansion of the transfer function and the solutions of $P(\frac{d}{dt})y = 0$ (Theorems 3.3.13 and 3.3.19).
- Every behavior described by differential equations of the form $R(\frac{d}{dt})w = 0$ can be written in input/output form by selecting appropriate components of w as input and considering the remaining components as output. (Theorem 3.3.22 and Corollary 3.3.23).
- Using the above results we were able to complete two issues raised in Chapter 2:
  - The equivalence of all polynomial matrices that represent the same behavior (Corollary 3.6.3). Two polynomial matrices $R_1(\xi), R_2(\xi) \in \mathbb{R}^{g\times q}[\xi]$ define the same behavior if and only if there exists a unimodular matrix $U(\xi)$ such that $R_2(\xi)U(\xi) = R_1(\xi)$.
  - The characterization of *minimal* and *full row rank* representations (Theorem 3.6.4). The matrix $R(\xi)$ is minimal if and only if it has full row rank.

3.8 Notes and References

The material in this chapter is based on [59, 60], where, however, mainly the case of discrete-time systems, i.e., systems described by difference equations, is covered. Many of the proofs in this chapter appear here for the first time.

3.9 Exercises

As a simulation exercise illustrating the material covered in this chapter we suggest A.3.

3.1 Determine the behavior $\mathfrak{B}$ associated with the differential equation

$$-32w + 22\frac{d^2}{dt^2}w + 9\frac{d^3}{dt^3}w + \frac{d^4}{dt^4}w = 0.$$

3.2 Let $P_i(\xi) \in \mathbb{R}[\xi]$, $(i = 1, 2)$. Denote the corresponding behaviors by $\mathfrak{B}_i$. Assume that $\mathfrak{B}_1 \subset \mathfrak{B}_2$. Prove that the polynomial $P_1(\xi)$ divides $P_2(\xi)$.

3.3 Refer to Remark 3.2.14. Prove that:

(a) $\dfrac{d^k}{dt^k}(t^j e^{\lambda t})(0) = \begin{cases} 0 & j > k, \\ \dfrac{k!}{(k-j)!} & j \leq k. \end{cases}$

Hint: Use Leibniz's formula for the kth derivative of the product of two functions.

(b) $\dfrac{d^k}{dt^k}(\sum_{j=0}^{m-1} a_j t^j e^{\lambda t})(0) = \begin{cases} \sum_{j=0}^{k} a_j \dfrac{k!}{(k-j)!}\lambda^{k-j} & 1 \leq k \leq m-1, \\ \sum_{j=0}^{m-1} a_j \dfrac{k!}{(k-j)!}\lambda^{k-j} & k \geq m-1. \end{cases}$

(c) $\dfrac{d^k}{dt^k}(\sum_{j=0}^{m-1} a_j t^j e^{\lambda t})(0) = \sum_{j=0}^{m-1} a_j((\dfrac{d}{d\lambda})^j \lambda^k), \quad k \geq 0.$

(d) $\dfrac{d^k}{dt^k}(\sum_{i=1}^{N}\sum_{j=0}^{n_i-1} a_{ij} t^j e^{\lambda t})(0) = \sum_{i=1}^{N}\sum_{j=0}^{n_i-1} a_{ij}((\dfrac{d}{d\lambda})^j \lambda_i^k).$

(e) Derive a formula similar to (3.18) for the case that the multiplicities are allowed to be larger than one.

3.4 Prove Corollary 3.2.7.

3.5 Many differential equations occurring in physical applications, e.g., in mechanics, contain *even* derivatives only. Consider the behavioral equation

$$P(\frac{d^2}{dt^2})w = 0,$$

with $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$, $\det P(\xi) \neq 0$. Assume that the roots of $\det P(\xi)$ are real and simple (multiplicity one). Describe the *real* behavior of this system in terms of the roots λ_k of $\det P(\xi)$ and the kernel of $P(\lambda_k)$.

3.6 Consider the set of differential equations

$$\begin{aligned} w_1 + \frac{d^2}{dt^2}w_1 - 3w_2 - \frac{d}{dt}w_2 + \frac{d^2}{dt^2}w_2 + \frac{d^3}{dt^3}w_2 &= 0, \\ w_1 - \frac{d}{dt}w_1 - w_2 + \frac{d}{dt}w_2 &= 0. \end{aligned} \tag{3.54}$$

(a) Determine the matrix $P(\xi) \in \mathbb{R}^{2\times 2}[\xi]$ such that (3.54) is equivalent to $P(\frac{d}{dt})w = 0$.

(b) Determine the roots of $\det P(\xi)$.

(c) Prove that every (strong) solution of (3.54) can be written as

$$w(t) = \begin{bmatrix} \alpha_1 - 3\alpha_2 \\ \alpha_1 \end{bmatrix} e^t + \begin{bmatrix} \alpha_2 \\ \alpha_2 \end{bmatrix} te^t + \begin{bmatrix} \beta \\ \beta \end{bmatrix} e^{-2t} + \begin{bmatrix} \gamma \\ \gamma \end{bmatrix} e^{-t}.$$

3.7 (a) Show that the polynomial matrix $U(\xi) \in \mathbb{R}^{2\times 2}[\xi]$ given by

$$U(\xi) := \begin{bmatrix} 1 + 3\xi + \xi^2 & -2\xi - \xi^2 \\ -2 - \xi & 1 + \xi \end{bmatrix}$$

is unimodular, and determine $(U(\xi))^{-1}$.

(b) Write $U(\xi)$ as a product of elementary unimodular matrices.

(c) Determine the behavior of $U(\frac{d}{dt})w = 0$. What general principle lies behind your answer?

3.8 Determine the behavior $\mathfrak{B}$ associated with $P(\frac{d}{dt})w = 0$, where

$$P(\xi) = \begin{bmatrix} 2 + \xi^2 & 1 \\ 2 - 2\xi - 4\xi^2 & 1 + \xi \end{bmatrix}.$$

3.9 Different polynomial matrices may have the same determinant. Let $P(\xi) \in \mathbb{R}^{2\times 2}[\xi]$ be a diagonal matrix. Given $\det P(\xi) = -2 - \xi + 2\xi^2 + \xi^3$, how many different behaviors correspond to this determinant?

3.10 The purpose of this exercise is to derive a proof of Theorem 3.2.16 that does not rely on the Smith form. Let $P(\xi)$ be given by

$$P(\xi) := \begin{bmatrix} P_{11}(\xi) & 0 \\ \\ P_{21}(\xi) & P_{22}(\xi) \end{bmatrix}$$

Consider the behavior associated with $P(\frac{d}{dt})w = 0$.

(a) Take $P_{11}(\xi) = 1 - 2\xi + \xi^2$, $P_{21}(\xi) = -3 + \xi$, and $P_{22}(\xi) = 1 + \xi$. Determine a basis of the corresponding behavior $\mathfrak{B}_{\mathrm{a}}$ and conclude that that $\mathfrak{B}_{\mathrm{a}}$ is a linear subspace of dimension three.

(b) Take $P_{11}(\xi)$ and $P_{22}(\xi)$ as in in the previous part and $P_{21}(\xi) = -3 + 2\xi - 2\xi^2 + \xi^3$. Prove that the corresponding behavior, $\mathfrak{B}_{\mathrm{b}}$, equals $\mathfrak{B}_{\mathrm{a}}$.

(c) Now let $P_{11}(\xi) \neq 0, P_{22}(\xi) \neq 0$, and $P_{22}(\xi)$ arbitrary. Prove that the corresponding behavior is a linear subspace of dimension equal to the degree of $P_{11}(\xi)P_{22}(\xi)$.

Hint: First calculate the dimension of $P_{11}(\frac{d}{dt})w_1 = 0$ by applying Theorem 3.2.5. Any solution of that equation can be plugged in as an input to the "input/output system" $P_{21}(\frac{d}{dt})w_1 + P_{22}(\frac{d}{dt})w_2 = 0$. Why don't you have to worry about the possible nonproperness of $\frac{P_{21}(\xi)}{P_{22}(\xi)}$? Now use Theorem 3.3.13 to obtain the proof.

(d) Consider the more general case

$$P(\xi) := \begin{bmatrix} P_{11}(\xi) & P_{12}(\xi) \\ P_{21}(\xi) & P_{22}(\xi) \end{bmatrix}.$$

Prove that P can be brought into lower triangular form by elementary row operations. Use this to prove that the dimension of the corresponding behavior is equal to the degree of the determinant of $P(\xi)$.

Hint: Elementary row operations do not change the determinant.

(e) Use induction on q to prove Theorem 3.2.16.

3.11 Prove Lemma 3.2.18 along the same lines as Lemma 3.2.6.

3.12 Verify that Theorem 3.2.16, specialized to the case $q = 1$, yields Theorem 3.2.5.

3.13 Consider the mechanical system shown in Figure 3.6. Assume that $q_1 = 0$

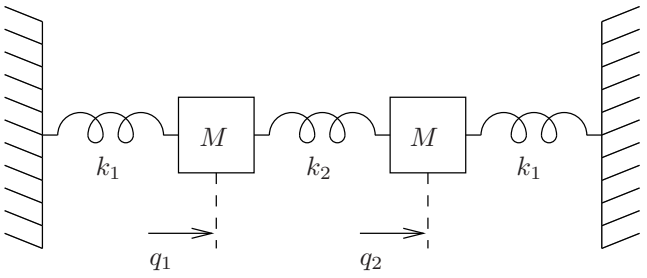


FIGURE 3.6. Mass–spring system.

corresponds to the equilibrium position of the mass on the left-hand side and that $q_2 = 0$ corresponds to that of the other mass.

(a) Determine for each of the cases below the differential equations describing

i. $\mathrm{col}(q_1, q_2)$,

ii. q_1,

iii. q_2.

(b) Use Theorem 3.2.16 to determine the behavior for the three cases above.

(c) Consider the behavior $\mathfrak{B}$ of $\mathrm{col}(q_1, q_2)$. It is of interest to see how the time behavior of q_1 relates to that of q_2. Show that the behavior $\mathfrak{B}$ may be written as $\mathfrak{B} = \mathfrak{B}_{\mathrm{s}} + \mathfrak{B}_{\mathrm{a}}$ (subscript 's' for *symmetric*, 'a' for *antisymmetric*), with $\mathfrak{B}_{\mathrm{s}}$ consisting of elements of $\mathfrak{B}$ of the form $(q_1, q_2) = (q, q)$ and $\mathfrak{B}_{\mathrm{a}}$ consisting of elements of the form $(q, -q)$. Derive differential equations describing $\mathfrak{B}_{\mathrm{s}}$ and $\mathfrak{B}_{\mathrm{a}}$.

(d) Prove that also $\mathfrak{B}_{\mathrm{s}}$ and $\mathfrak{B}_{\mathrm{a}}$ consist of pure sinusoids. Denote the respective frequencies by ω_{s} and ω_{a}. Discuss these frequencies for the cases

i. $\dfrac{k_1}{k_2} \ll 1$.

ii. $\dfrac{k_1}{k_2} \gg 1$.

These phenomena are numerically illustrated in simulation exercise A.3.

3.14 Consider the one-dimensional horizontal motion of the mechanical system depicted in Figure 3.7. Let q_1 denote the displacement of M_1 from some

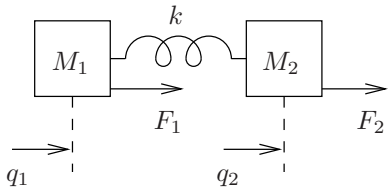


FIGURE 3.7. Mass–spring system.

reference point, and q_2 the displacement of M_2 from its equilibrium when M_1 is in the position corresponding to $q_1 = 0$. Assume that external forces F_1, F_2 act on the masses M_1 and M_2 respectively.

(a) Derive differential equations relating q_1, q_2, F_1, F_2.

(b) Derive all possible input/output partitions of q_1, q_2, F_1, F_2.

(c) Derive an integral expression relating the input $\mathrm{col}(F_1, F_2)$ to $\mathrm{col}(q_1, q_2)$.

3.15 The aim of this exercise is to prove Theorem 3.2.15. Let $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and $\det P(\xi) \neq 0$. Choose a unimodular matrix $U(\xi)$ such that $T(\xi) := U(\xi)P(\xi)$ is an upper triangular matrix (see Theorem 2.5.14):

$$T(\xi) = \begin{bmatrix} T_{11}(\xi) & \cdots & \cdots & T_{1q}(\xi) \\ 0 & T_{22}(\xi) & \cdots & T_{2q}(\xi) \\ & & \ddots & \\ 0 & \cdots & 0 & T_{qq}(\xi) \end{bmatrix}. \tag{3.55}$$

According to Theorem 2.5.4, $P(\xi)$ and $T(\xi)$ define the same behavior. Let w be a solution of $P(\frac{d}{dt})w = 0$ and hence of $T(\frac{d}{dt})w = 0$. Denote the components of w by $w_1, \ldots, w_q$ respectively.

(a) Use Theorem 3.2.4 to conclude that there exists $v_q \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ such that $w_q = v_q$ almost everywhere.

(b) Since $v_q \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$, there exists $\tilde{v}_{q-1} \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ such that

$$T_{q-1,q-1}(\frac{d}{dt})\tilde{v}_{q-1} + T_{q-1,q}(\frac{d}{dt})v_q = 0 \qquad \text{strongly,}$$

and since $w_q = v_q$ almost everywhere, we also have

$$T_{q-1,q-1}(\frac{d}{dt})w_{q-1} + T_{q-1,q}(\frac{d}{dt})v_q = 0 \qquad \text{weakly.}$$

By linearity it follows that

$$T_{q-1,q-1}(\frac{d}{dt})(w_{q-1} - \tilde{v}_{q-1}) = 0 \qquad \text{weakly.} \tag{3.56}$$

Use Theorem 3.2.4 and (3.56) to conclude that there exists $v_{q-1} \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ such that $v_{q-1} = w_{q-1}$ almost everywhere and

$$T_{q-1,q-1}(\frac{d}{dt})v_{q-1} + T_{q-1,q}(\frac{d}{dt})v_q = 0 \qquad \text{strongly.}$$

(c) Use induction to prove that there exist $v_{q-2},\ldots,v_1 \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ such that w_i and v_i are the same except on sets of measure zero ($i = q-2,\ldots,1$).

3.16 Refer to Remark 3.2.14.

(a) Let $\lambda_1,\ldots,\lambda_n \in \mathbb{C}$. In (3.18) we derived a relation in which the following Vandermonde matrix M appeared:

$$M = \begin{bmatrix} 1 & \cdots\cdots & 1 \\ \lambda_1 & \cdots\cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots\cdots & \lambda_n^{n-1} \end{bmatrix}.$$

Prove that M is nonsingular if and only if the λ_is are mutually distinct. Hint: Let $v \in \mathbb{C}^n$ such that $v^T M = 0$. Consider the entries of v as the coefficients of a polynomial. Use the Fundamental Theorem of Algebra (every complex polynomial of degree n has exactly n complex roots, counting multiplicities) to show that if the λ_is are distinct, then v must be zero.

(b) Let $\lambda_1,\ldots,\lambda_N \in \mathbb{C}$. Let $n_1,\ldots,n_N \in \mathbb{N}$, and define $n := \sum_{i=0}^{N} n_i$. If the multiplicities of the λs are allowed to be larger than one, than a relation like (3.18) still holds; see Exercise 3.3. The Vandermonde matrix is then replaced by a matrix M that is constructed as follows. The first column is $(1,\lambda_1,\ldots,\lambda_1^{n-1})^T$, the second column is the derivative with respect to λ_1 of the first column, the third column is the derivative of the second column, etc., up to the n_1th column. Repeat this for $\lambda_2,\ldots,\lambda_N$. Then M is given by

$$M = \begin{bmatrix} 1 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots \\ \lambda_1 & 1 & 0 & \cdots & \lambda_N & 1 & 0 & \cdots \\ \lambda_1^2 & 2\lambda_1 & 2 & \cdots & \lambda_N^2 & 2\lambda_N & 2 & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \\ \lambda_1^{n-1} & (n-1)\lambda_1^{n-2} & (n-1)(n-2)\lambda_1^{n-3} & \cdots & \lambda_N^{n-1} & (n-1)\lambda_N^{n-2} & & \cdots \end{bmatrix}.$$

Prove that this matrix is invertible if and only if the λ_is are mutually distinct.

3.17 Consider expression (3.8). Let

$$r_k(t) = r_k^0 + r_k^1 t + \cdots + r_k^{n_k-1} t^{n_k-1}.$$

Prove that $(\frac{d^\ell}{dt^\ell} r_k)(0) = l! r_k^l$. Deduce from this the matrix that takes the vector $\mathrm{col}(r_1^0, r_1^1, \ldots, r_1^{n_1-1}, r_2^0, r_2^1, \ldots, r_2^{n_2-1}, \ldots, r_N^0, r_N^1, \ldots, r_N^{n_N-1})$ into

$$\mathrm{col}\left(w(0), (\frac{d}{dt} w)(0), \ldots, (\frac{d^{n-1}}{dt^{n-1}} w)(0) \right)$$

with $n = n_1 + n_2 + \cdots + n_N$. Prove that this matrix is invertible and hence that there exists a bijection between the polynomials $(r_1, r_2, \ldots, r_N)$ and the initial values $\mathrm{col}(w(0), (\frac{d}{dt})w(0), \ldots, (\frac{d^{n-1}}{dt^{n-1}} w)(0))$.

3.18 Determine the partial fraction expansion of $\dfrac{1 - 6\xi + \xi^2}{-36 + 5\xi^2 + \xi^4}$.

3.19 Consider the input/output equation $2y - 3\dfrac{d}{dt}y + \dfrac{d^2}{dt^2}y = u + \dfrac{d}{dt}u$.

(a) Determine the corresponding behavior.

(b) Determine all possible ys corresponding to the input $u(t) = \sin t$.

(c) Determine y corresponding to the input $u(t) = \sin t$ and the initial condition $y(0) = \frac{d}{dt}y(0) = 0$.

3.20 Consider the i/o system defined by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u \tag{3.57}$$

with $p(\xi) = \xi - 2\xi^2 + \xi^3$ and $q(\xi) = -1 + \xi^2$.

(a) Determine the partial fraction expansion of $\dfrac{q(\xi)}{p(\xi)}$.

(b) Give an explicit characterization of the behavior $\mathfrak{B}$ of (3.57).

(c) Consider now

$$\tilde{p}(\frac{d}{dt})y = \tilde{q}(\frac{d}{dt})u \tag{3.58}$$

with $\tilde{p}(\xi) = -\xi + \xi^2$ and $\tilde{q}(\xi) = 1 + \xi$. Determine the partial fraction expansion of $\frac{\tilde{q}(\xi)}{\tilde{p}(\xi)}$. What strikes you?

(d) Give an explicit characterization of the behavior $\tilde{\mathfrak{B}}$ of (3.58).

(e) In what sense are $\mathfrak{B}$ and $\tilde{\mathfrak{B}}$ different?

(f) Give a convolution representations of $\mathfrak{B}$ and $\tilde{\mathfrak{B}}$.

3.21 Let the polynomial matrix $R(\xi)$ be given by

$$R(\xi) := \begin{bmatrix} -5\xi + \xi^2 & -5 + \xi \\ -\xi + \xi^2 & -1 + \xi \end{bmatrix}.$$

Show that $R(\frac{d}{dt})w = 0$ does *not* define an autonomous system. Write this system in input/output form. Indicate clearly which component of w is considered input and which is the output.

3.22 Consider the system of differential equations

$$\begin{aligned} 6w_1 - 5\frac{d}{dt}w_1 + \frac{d^2}{dt^2}w_1 - 3w_2 + \frac{d}{dt}w_2 = 0, \\ 2w_1 - 3\frac{d}{dt}w_1 + \frac{d^2}{dt^2}w_1 - w_2 + \frac{d}{dt}w_2 = 0. \end{aligned} \tag{3.59}$$

(a) Does this set of differential equations define an autonomous system?

(b) If the answer is no, find an input/output representation for it.

3.23 Consider the mechanical system depicted in Figure 3.8. The variables

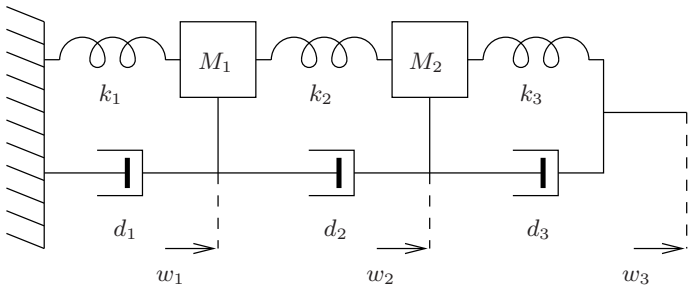


FIGURE 3.8. Mass–damper–spring system.

w_1, w_2, w_3 denote the displacements of the masses from their equilibrium positions. The damper coefficients are d_1, d_2, d_3 respectively and the spring constants are k_1, k_2, k_3. Both masses are assumed to be unity. All displacements are in the horizontal direction only; rotations and vertical movements are not possible.

(a) Show that the equations of motion are given by

$$\begin{aligned} &(k_1 + k_2)w_1 + (d_1 + d_2)\frac{d}{dt}w_1 + \frac{d^2}{dt^2}w_1 - k_2 w_2 - d_2\frac{d}{dt}w_2 = 0, \\ &-k_2 w_1 - d_2\frac{d}{dt}w_1 + (k_2 + k_3)w_2 + (d_2 + d_3)\frac{d}{dt}w_2 + \frac{d^2}{dt^2}w_2 \\ &-k_3 w_3 - d_3\frac{d}{dt}w_3 = 0. \end{aligned}$$

(b) Determine a polynomial matrix $R(\xi) \in \mathbb{R}^{2\times 3}[\xi]$ such that the behavior $\mathfrak{B}$ of the system is described by $R(\frac{d}{dt})w = 0$.

Choose $d_1 = 1$, $d_2 = 1$, $d_3 = 4$, $k_1 = 0$, $k_2 = 1$, and $k_3 = 6$.

(c) Interpret $k_1 = 0$ physically.

(d) Show that $R(\xi)$ is of full row rank and write the system in input/output form $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$; i.e., take the appropriate components of w as output and the remaining components as input.

(e) Determine $\det P(\xi)$ and its roots and the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$.

(f) Determine the behavior of $P(\frac{d}{dt})y = 0$ and the behavior of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$.

3.24 The purpose of this exercise is to prove Theorem 3.3.13 for the multivariable case. Let $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$, $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$ with $P^{-1}Q(\xi)$ proper. Assume that the partial fraction expansion of $P(\xi)^{-1}Q(\xi)$ is given by

$$P^{-1}Q(\xi) = A_0 + \sum_{k=1}^{n} A_k \frac{1}{\xi - \lambda_k}.$$

Choose $\bar{u} \in \mathfrak{L}_1^{loc}(\mathbb{R}, \mathbb{R}^m)$ and define

$$\bar{y}(t) = A_0 u(t) + \sum_{k=1}^{n} A_k \int_0^t e^{\lambda_k (t-\tau)} \bar{u}(\tau) d\tau, \quad A_k \in \mathbb{C}^{p\times m}.$$

We want to prove that with this definition of $\bar{y}$, the pair $(\bar{u}, \bar{y})$ satisfies $P(\frac{d}{dt})\bar{y} = Q(\frac{d}{dt})\bar{u}$, weakly. Choose unimodular matrices of appropriate dimensions $U(\xi)$, $V(\xi)$ such that $\tilde{P}(\xi) := U(\xi)P(\xi)V(\xi)$ is in Smith form. Define $\tilde{Q}(\xi) = U(\xi)Q(\xi)$ and $\tilde{A}_k(\xi) = V^{-1}(\xi)A_k$, $k = 0, \ldots, n$. It follows that $\tilde{P}^{-1}(\xi)\tilde{Q}(\xi)$ is given by

$$\tilde{P}^{-1}\tilde{Q}(\xi) = \tilde{A}_0(\xi) + \sum_{k=1}^{n} \tilde{A}_k(\xi) \frac{1}{\xi - \lambda_k}.$$

(a) Assume that $\bar{u} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$, and define $\tilde{y} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^p)$ by

$$\tilde{y}(t) = \tilde{A}_0(\frac{d}{dt})u(t) + \sum_{k=1}^{n} \tilde{A}_k(\frac{d}{dt}) \int_0^t e^{\lambda_k (t-\tau)} \bar{u}(\tau) d\tau, \quad A_k \in \mathbb{C}^{p\times m}.$$

Prove, in a similar way as the proof of Theorem 3.3.13 for the scalar case, that $\tilde{P}(\frac{d}{dt})\tilde{y} = \tilde{Q}(\frac{d}{dt})\bar{u}$ and conclude that indeed $P(\frac{d}{dt})\bar{y} = Q(\frac{d}{dt})\bar{u}$.

(b) Prove that $P(\frac{d}{dt})\bar{y} = Q(\frac{d}{dt})\bar{u}$ is also satisfied (weakly) if $u \in \mathfrak{L}_1^{loc}(\mathbb{R}, \mathbb{R}^m)$.

3.25 Refer to Example 3.3.18. Consider the differential equation $w_1 = \frac{d}{dt}w_2$. Take $w_2(t) = 1$, $t \geq 0$, and $w_2(t) = 0$, $t < 0$. Prove that there does not exist a $w_1 \in \mathfrak{L}_1^{\text{loc}}(\mathbb{R}, \mathbb{R})$ such that (w_2, w_1) is a weak solution of the differential equation.

3.26 Consider the behavior in Example 3.3.3. Let w_2 be as in Exercise 3.25. Prove that there does not exist a $w_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$ such that $(w_1, w_2) \in \mathfrak{B}$.

3.27 Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ be of full row rank and consider the behavior $\mathfrak{B}$ defined by $R(\frac{d}{dt})w = 0$. Assume that $w = \mathrm{col}(w_1, w_2)$ is an i/o partition; i.e., w_1 is maximally free. We want to prove that q_1, the dimension of w_1, equals $q - g$. To that end assume that $q_1 < q - g$ and argue that w_2 contains free components in that case. Alternatively, if $q_1 > q - g$, then not all components of w_1 are free.

3.28 Consider the electrical circuit of Example 2.3.1.

(a) Give an exact expression of the *short-circuit* behavior (i.e., determine all currents I compatible with $V = 0$).

(b) Give also an expression of the *open-circuit* behavior (i.e., determine all voltages V compatible with $I = 0$).

(c) Assume next that the circuit is terminated by a resistor $R > 0$; see Figure 3.9. Determine the resulting behavior of (V, I).

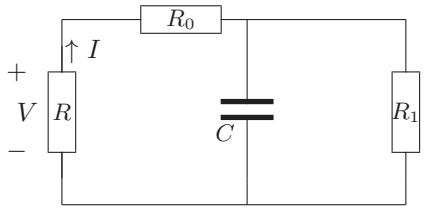


FIGURE 3.9. Electrical circuit.

3.29 Consider the electrical circuit described in Example 1.3.5. Determine for all values of $R_C > 0, R_L > 0, C > 0, L > 0$, the input/output structure, the short-circuit behavior, and the open-circuit behavior.

3.30 Prove Corollary 3.3.8. Hint: multiply both sides of (3.27) by $P(\xi)$.

3.31 Let $P(\xi), Q(\xi) \in \mathbb{R}[\xi]$, $\deg P(\xi) = n, \deg Q(\xi) = k, k \leq n$. Consider the SISO input/output system

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})u. \tag{3.60}$$

As usual, we denote by $\mathcal{C}^m(\mathbb{R}, \mathbb{R}), m \geq 0$, the functions that are m times continuously differentiable. For the sake of this exercise denote $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$ by $\mathcal{C}^{-1}(\mathbb{R}, \mathbb{R})$. Define the *relative degree* of (3.60) by $r := n - k$. Prove that if $u \in \mathcal{C}^m(\mathbb{R}, \mathbb{R}), m \geq -1$, and if (u, y) satisfies (3.60), then $y \in \mathcal{C}^{m+r}(\mathbb{R}, \mathbb{R})$. Hint: Use the integral representation (2.12) of (3.60). Use the fact that if $w \in \mathcal{C}^m(\mathbb{R}, \mathbb{R})$, then the integral of w is in $\mathcal{C}^{m+1}(\mathbb{R}, \mathbb{R})$.

3.32 Let $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ and let $u_n \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ be a sequence that converges to u in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$. Define y_n by

$$y_n(t) := \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} e^{\lambda(t-\tau)} u_n(\tau) d\tau$$

and y by

$$y(t) := \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} e^{\lambda(t-\tau)} u(\tau) d\tau.$$

Show that y_n converges to y in the sense of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$.

3.33 Consider the convolution system given by

$$y(t) = \int_{-\infty}^{\infty} h(t-\tau) u(\tau) d\tau,$$

where $h(t) = e^{-t}; t \geq 0;\ h(t) = 0,\ t < 0$.

(a) Determine an input/output system of the form

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})u$$

such that the initially-at-rest-behaviors of both systems are the same.

(b) Of course, there is a trivial nonuniqueness in the answer to the previous question: If $(p(\xi), q(\xi))$ is a possible answer, then the same is true for $(\alpha p(\xi), \alpha q(\xi))$ for every constant $\alpha \in \mathbb{R}$. Do you see a nontrivial form of nonuniqueness?

(c) Give an example of an input/output pair (u, y) that belongs to the behavior of the input/output system but that does *not* belong to the behavior of the convolution system. Conclude that although the associated initially-at-rest systems coincide, the behaviors themselves are not the same. Is there an inclusion relation?

3.34 Refer to the statement just following (3.52). Let $\mathfrak{B}$ be the behavior defined by $R(\frac{d}{dt})w = 0$. Prove that $\mathfrak{B} = \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ if and only if $R(\xi)$ is equal to the zero matrix. Hint: Write $R(\xi) = R_0 + R_1\xi + \cdots + R_L\xi^L$; then take w a nonzero constant to show that $R_0 = 0$; then take w a multiple of t; etc.

3.35 Prove Corollary 3.6.3.

3.36 Let $R_1(\xi), R_2(\xi) \in \mathbb{R}^{g\times q}[\xi]$ be of full row rank. The corresponding behaviors are denoted by $\mathfrak{B}_1$ and $\mathfrak{B}_2$ respectively.

(a) Show that $R_2(\xi)R_2^T(\xi)$ is invertible as a *rational* matrix. It suffices to prove that $\det R_2(\xi)R_2^T(\xi) \neq 0$.

(b) Show that $R_2(\xi)$ has a *right inverse*; i.e., there exists a *rational* matrix $R_2^*(\xi)$ such that $R_2(\xi)R_2^*(\xi) = I_g$.

(c) Assume that $\mathfrak{B}_1 = \mathfrak{B}_2$. Show that $R_1(\xi)R_2^*(\xi)$ is a polynomial unimodular matrix.

(d) Show by means of a simple example that the converse is not true: If $R_1(\xi)R_2^*(\xi)$ is a polynomial unimodular matrix, then we need not have that $\mathfrak{B}_1 = \mathfrak{B}_2$.

(e) Prove that $\mathfrak{B}_1 = \mathfrak{B}_2$ if and only if $R_1(\xi)R_2^*(\xi)$ is a polynomial unimodular matrix *and* $R_1(\xi) = R_1(\xi)R_2^*(\xi)R_2(\xi)$.

(f) Let $R_1(\xi)$ and $R_2(\xi)$ be given by

$$R_1(\xi) := \begin{bmatrix} 1+\xi^2 & \xi & 1+\xi \\ \xi & 0 & 1 \end{bmatrix}, \quad R_2(\xi) := \begin{bmatrix} 1 & \xi & 1 \\ \xi & 0 & 1 \end{bmatrix}.$$

Prove or disprove: $\mathfrak{B}_1 = \mathfrak{B}_2$.

4

State Space Models

4.1 Introduction

In Chapter 1 we argued that mathematical models obtained from first principles usually contain latent variables. Up to now these latent variables did not enter the mathematical development. In Chapters 5 and 6 latent variable systems will be pursued in full generality. In the present chapter we discuss a special and important class of latent variables, namely *state variables*. State variables either show up naturally in the modeling process or they can be artificially introduced. State variables have the property that they parametrize the *memory* of the system, i.e., that they "split" the past and future of the behavior. The precise meaning of this statement will be made clear in the sequel.

The chapter is structured as follows. In Section 4.2 we introduce and briefly discuss differential systems containing latent variables and formally introduce state variables. In Section 4.3 we relate state variables to differential equations that are of first order in the latent variables and of order zero in the manifest variables. Then, in Section 4.4 we consider a more structured class of state space models, namely state space models for systems in input/output form. This leads to input/state/output representations. State space transformations are treated in Section 4.6, and in Section 4.7 we study linearization of nonlinear state space models.

4.2 Differential Systems with Latent Variables

Assume that a mathematical model contains, in analogy with the discussion in Section 1.5, see (1.17), q real-valued manifest variables $w = \mathrm{col}(w_1, \ldots, w_q)$ and d real-valued latent variables $\ell = \mathrm{col}(\ell_1, \ldots, \ell_d)$. Then, assuming that the joint equations governing w and ℓ are linear constant-coefficient differential equations, we obtain the following generalization of (2.1):

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell, \tag{4.1}$$

where $w : \mathbb{R} \to \mathbb{R}^q$ denotes the manifest variable and $\ell : \mathbb{R} \to \mathbb{R}^d$ the latent variable, and $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and $M(\xi) \in \mathbb{R}^{g\times d}[\xi]$ are polynomial matrices with the same number of rows, namely g, and with q and d columns respectively. Corresponding to (4.1) we define the following behaviors:

Definition 4.2.1 The *full behavior* $\mathfrak{B}_{\mathrm{f}}$ and the *manifest behavior* $\mathfrak{B}$ represented by (4.1) are defined as

$$\mathfrak{B}_{\mathrm{f}} = \{(w, \ell) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q \times \mathbb{R}^d) \mid (w, \ell) \text{ satisfies (4.1) weakly}\}, \tag{4.2}$$

$$\mathfrak{B} = \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid \exists \ell \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^d) \text{ such that } (w, \ell) \in \mathfrak{B}_{\mathrm{f}}\}.$$

□

The idea is that (4.1) is obtained from first principles modeling, but that we are primarily interested in the manifest behavior $\mathfrak{B}$. It turns out, in fact, that $\mathfrak{B}$ can itself be described by differential equations. However, for the moment we are not concerned with the issue of how this could be proven, or how the differential equations for $\mathfrak{B}$ could be computed in a systematic way. We will come back to this in Chapter 6.

4.3 State Space Models

We now study an exceedingly important class of latent variables, *state variables*, that not only often show up naturally in applications, but that are also very useful in the analysis and synthesis of dynamical systems. We start by introducing the concept of state on an intuitive level by means of two examples.

Example 4.3.1 Consider the mass–spring system in Figure 4.1. Recall from Example 3.2.3 that the equation describing the behavior is

$$(k_1 + k_2)q + M(\frac{d}{dt})^2 q = F. \tag{4.3}$$

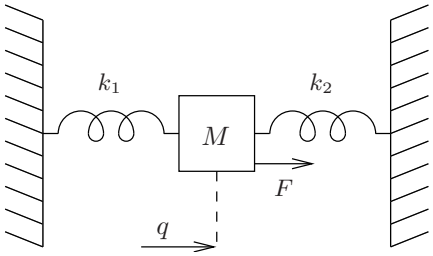


FIGURE 4.1. Mass–spring system.

We want to know to what extent the past of a trajectory determines its future. Otherwise stated, if we observe a trajectory $w = (F, q)$ up to $t = 0$ (the past), what can we then say about (F, q) after $t = 0$ (the future)? We have seen in Chapter 3 that whatever the past of (F, q), the future of F is not restricted by it. The future of q depends, on the one hand, on the future of F, and on the other hand on the position and velocity at $t = 0$. So, if $w_i = (q_i, F_i)$, $i = 1, 2$, are possible trajectories, then the trajectory $w = (q, F)$ that equals w_1 up to $t = 0$ and w_2 after $t = 0$, i.e., a trajectory that concatenates the past of w_1 and the future of w_2 at $t = 0$, is also an element of the behavior, provided that $q_1(0) = q_2(0)$ and $(\frac{d}{dt}q_1)(0) = (\frac{d}{dt}q_2)(0)$. This observation, which still needs mathematical justification, inspires us to introduce the latent variable $x := \mathrm{col}(q, \frac{d}{dt}q)$. Thus x forms what we call the *state* of this mechanical system. Notice that we can rewrite the system equation (4.3) in terms of x as

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ \dfrac{-k_1 - k_2}{M} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \dfrac{1}{M} \end{bmatrix} F, \quad q = \begin{bmatrix} 1 & 0 \end{bmatrix} x, \quad w = \mathrm{col}(q, F). \tag{4.4}$$

Using (4.4) we may reexpress the concatenation condition as follows: If (w_1, x_1) and (w_2, x_2) satisfy (4.4), then (w, x), the concatenation of (w_1, x_1) and (w_2, x_2) at $t = 0$, also satisfies (4.4) if $x_1(0) = x_2(0)$. This is the reason that we call x the *state*. Notice that (4.4) is first order in the latent variable x and order zero (static) in the manifest variables q and F. □

As a second example of a state space model, we consider an electrical circuit.

Example 4.3.2 Consider the electrical circuit consisting of a resistor, a capacitor, an inductor, and an external port shown in Figure 4.2. Suppose we want to model the relation between the voltage across and the current through the external port. Introduce the voltages across and the currents through the other elements as latent variables. Using the obvious notation,

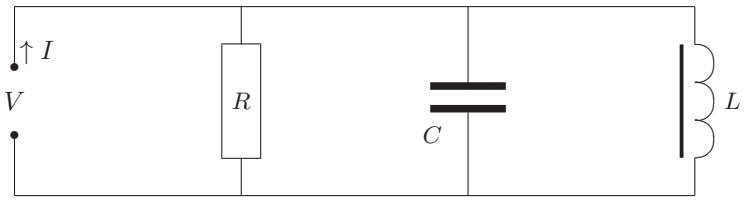


FIGURE 4.2. Electrical circuit.

the equations describing the full behavior are

$$\begin{aligned} &V = V_R = V_C = V_L, \qquad I = I_R + I_C + I_L, \qquad V_R = RI_R, \\ &I_C = C\frac{d}{dt}V_C, \qquad V_L = L\frac{d}{dt}I_L. \end{aligned} \tag{4.5}$$

This is a set of equations that implicitly determines the relation between V and I: again it contains latent variables. By eliminating V_R, I_R, I_C, and V_L in (4.5), we obtain

$$C\frac{d}{dt}V_C = -\frac{V_C}{R} - I_L + I, \qquad L\frac{d}{dt}I_L = V_C, \qquad V = V_C. \tag{4.6}$$

Now, (4.5) and (4.6) form two latent variable systems. It is not difficult to see that they define the same manifest behavior. The representation (4.6) shares some of the features with (4.4). If we define $x = \mathrm{col}(V_C, I_L)$, then (4.6) may be written as

$$\frac{d}{dt}x = \begin{bmatrix} -\frac{1}{RC} & -\frac{1}{C} \\ \frac{1}{L} & 0 \end{bmatrix} x + \begin{bmatrix} \frac{1}{C} \\ 0 \end{bmatrix} I, \qquad V = \begin{bmatrix} 1 & 0 \end{bmatrix} x. \tag{4.7}$$

Just as in (4.4), these equations are of first order in the latent variable x and of order zero in the manifest variables I and V. It turns out that also in this case, x can be seen as the state of the system. □

In many physical systems, the state has a direct interpretation in terms of physical variables, e.g., the positions and the velocities of the masses (in mechanical systems, as in Example 4.3.1) or the charges on the capacitors and the currents through the inductors (in electrical circuits, as in Example 4.3.2). Notice once more that both (4.4) and (4.7) are first order in x and order zero in the manifest variables $\mathrm{col}(q, F)$ in (4.4) and $\mathrm{col}(V, I)$ in (4.7). We will soon see that this feature is characteristic for state space systems.

In Example 4.3.1 we have made plausible that two trajectories (w_1, x_1), (w_2, x_2) may be concatenated at t_0 provided that $x_1(t_0) = x_2(t_0)$. This is what we call the *property of state*.

Definition 4.3.3 (Property of state) Consider the latent variable system defined by (4.2). Let $(w_1, \ell_1), (w_2, \ell_2) \in \mathfrak{B}_{\mathrm{f}}$ and $t_0 \in \mathbb{R}$ and suppose that ℓ_1, ℓ_2 are continuous. Define the *concatenation* of (w_1, ℓ_1) *and* (w_2, ℓ_2) *at* t_0 by (w, ℓ), with

$$w(t) = \begin{cases} w_1(t) & t < t_0, \\ w_2(t) & t \geq t_0, \end{cases} \quad \text{and} \quad \ell(t) = \begin{cases} \ell_1(t) & t < t_0, \\ \ell_2(t) & t \geq t_0. \end{cases} \tag{4.8}$$

Then $\mathfrak{B}_{\mathrm{f}}$ is said to be a *state space model*, and the latent variable ℓ is called the *state* if $\ell_1(t_0) = \ell_2(t_0)$ implies $(w, \ell) \in \mathfrak{B}_{\mathrm{f}}$. □

Remark 4.3.4 The state property expresses that ℓ splits the past and the future of w. All the information needed to decide whether or not two trajectories w_1 and w_2 can be concatenated within $\mathfrak{B}$ at time t_0 is contained in the values of the corresponding states at time t_0. This, indeed, is precisely the content of (4.8).

It can be shown that for each behavior $\mathfrak{B}$ defined by equations of the form $R(\frac{d}{dt})w = 0$, there exists a representation of the form (4.1), with ℓ having the property of state. In Chapter 6 we will demonstrate this for SISO systems.

Another useful intuitive interpretation of the state property is in terms of the *memory* of the dynamical system. Indeed, assume that a past trajectory (w^-, ℓ^-) in the behavior has been observed. What future trajectories can we expect? The state property implies that all we need to know to answer this question is $\ell^-(0)$. Any trajectory $(w^+, \ell^+) : [0, \infty) \to \mathbb{R}^q \times \mathbb{R}^n$ in the behavior can occur as a future continuation of (w^-, ℓ^-) provided that $\ell^-(0) = \ell^+(0)$ and ℓ^- and ℓ^+ are continuous. As such, $\ell^-(0)$ contains all the information about the past required to be able to understand what the future may look like. In other words, $\ell^-(0)$ is the memory of the system. □

The behavioral equations (4.4) and (4.7) are special cases of the general class of differential equations

$$E\frac{dx}{dt} + Fx + Gw = 0 \tag{4.9}$$

relating the latent variable $x \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^n)$ and the manifest variable $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. The matrices E, F, G are real matrices of appropriate sizes. Usually the state of a system is denoted by x. We follow that convention with Definition 4.3.3 as the only exception. Another convention is that x takes its values in $\mathbb{R}^n$, so that $E, F \in \mathbb{R}^{g \times n}$ and $G \in \mathbb{R}^{g \times q}$. The integer n

is the *dimension*, or the *order*, of the state space representation (4.9) and $\mathbb{R}^n$ its *state space*. Note that (4.9) is a special case of (4.1) with $R(\xi) = G$ and $M(\xi) = -F - E\xi$. We now show that (4.9) defines a state space model with the latent variable x as the state. The full behavior of (4.9) is defined as

$$\mathfrak{B}_{\mathrm{f}} = \{(w, x) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q \times \mathbb{R}^n) \mid (w, x) \text{ satisfies (4.9) weakly}\}. \quad (4.10)$$

Since $\mathfrak{B}_{\mathrm{f}}$ is governed by a set of differential equations that are *first order* in x and *order zero* in w, it has the state property.

Theorem 4.3.5 *The behavior $\mathfrak{B}_{\mathrm{f}}$ defined by (4.10) is a state space model with x as the state.*

Proof Recall that (w, x) is a weak solution of (4.9) if there exists a constant vector $c \in \mathbb{R}^g$ such that for almost all t

$$Ex(t) + F \int_0^t x(\tau)d\tau + G \int_0^t w(\tau)d\tau = c. \quad (4.11)$$

In fact, it follows from Lemma 2.3.9 that the lower limit in (4.11) is immaterial and that equivalently, (w, x) is a weak solution if and only if for all $t_0 \in \mathbb{R}$ there exists a $c_{t_0} \in \mathbb{R}^g$ such that for almost all t

$$Ex(t) + F \int_{t_0}^t x(\tau)d\tau + G \int_{t_0}^t w(\tau)d\tau = c_{t_0}. \quad (4.12)$$

We claim that x satisfies the property of state. Suppose that (w_1, x_1) and (w_2, x_2) are weak solutions of (4.9) with x_1, x_2 continuous and such that $x_1(t_0) = x_2(t_0)$. By (4.12) there exist constant vectors $c_1, c_2 \in \mathbb{R}^g$ such that for almost all t

$$Ex_i(t) + F \int_{t_0}^t x_i(\tau)d\tau + G \int_{t_0}^t w_i(\tau)d\tau = c_i, \qquad i = 1, 2. \quad (4.13)$$

Since both x_1 and x_2 are continuous, (4.13) must hold for *all* t rather than just for almost all t. To see this, suppose that, e.g., the first equation in (4.13) does *not* hold for some $\bar{t}$. Since (4.13) can fail to be true only for t in a set of measure zero, there exists a sequence t_k converging to $\bar{t}$ and such that

$$Ex_1(t_k) + F \int_{t_0}^{t_k} x_1(\tau)d\tau + G \int_{t_0}^{t_k} w_1(\tau)d\tau - c_1 = 0. \quad (4.14)$$

Since by assumption the left-hand side of (4.14) is continuous, it follows that

$$\lim_{k\to\infty} Ex_1(t_k) + F\int_{t_0}^{t_k} x_1(\tau)d\tau + G\int_{t_0}^{t_k} w_1(\tau)d\tau - c_1$$
$$= Ex_1(\bar{t}) + F\int_{t_0}^{\bar{t}} x_1(\tau)d\tau + G\int_{t_0}^{\bar{t}} w_1(\tau)d\tau - c_1 = 0.$$

In particular, (4.13) holds for $t = t_0$. By substituting $t = t_0$ in (4.13) we conclude that $c_1 = c_2$. Define (w, x) by

$$(w(t), x(t)) = \begin{cases} (w_1(t), x_1(t)) & t < t_0, \\ (w_2(t), x_2(t)) & t \geq t_0. \end{cases}$$

Now it is clear that (w, x) satisfies (4.12). For $t < t_0$ this follows from (4.13) with $i = 1$ and for $t \geq t_0$ with $i = 2$, and hence (w, x) is a weak solution of (4.9). □

Theorem 4.3.5 allows us to conclude that equations of the form (4.9) define state space representations. It can in fact be shown that the converse is also true. If the full behavior $\mathfrak{B}_{\mathrm{f}}$ of (4.1) satisfies the property of state, then the equations (4.1) are equivalent (in the sense of Definition 2.5.2) to a system of differential equations of the form (4.9). We will not need this result in the sequel, and therefore we do not prove it in this book.

Next we present a more academic example.

Example 4.3.6 Consider the (autonomous) behavior defined by

$$3w + 2\frac{d}{dt}w + \frac{d^2}{dt^2}w = 0. \tag{4.15}$$

Here $w : \mathbb{R} \to \mathbb{R}$. As we have seen in Theorem 3.2.15, we may confine ourselves to *strong* solutions of (4.15). Define $x := \mathrm{col}(w, \frac{d}{dt}w)$. Then

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -3 & -2 \end{bmatrix} x, \qquad w = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$$

From Theorem 4.3.5, it follows easily that this defines a state space representation. See also Theorem 4.4.1. □

Examples 4.3.1, 4.3.2, and 4.3.6 illustrate how in various situations state space representations occur in practice. In Example 4.3.1 the state variables $(q, \frac{d}{dt}q)$ were introduced using physical reasoning. It was shown that they form state variables. In Example 4.3.2 the latent variables (V_C, I_L) were

introduced in the modeling process. They also turned out to be state variables. Note that in both these examples the state is immediately related to the *energy* of the system: $\frac{1}{2}(k_1+k_2)q^2 + \frac{1}{2}M(\frac{d}{dt}q)^2$ in Example 4.3.1 and $\frac{1}{2}CV_C^2 + \frac{1}{2}LI_L^2$ in Example 4.3.2. In Example 4.3.6, on the other hand, our choice of the state $(w, \frac{d}{dt}w)$ was guided by the choice of the initial condition required to specify the solution uniquely. These examples show that state variables may be introduced from either physical or mathematical considerations. In Chapter 6 we will return to the question of associating to a differential system of the form $R(\frac{d}{dt})w = 0$ a state space representation of the type (4.9). Note that in Examples 4.3.1, 4.3.2, and 4.3.6 we were able to solve this representation problem in an ad hoc way.

4.4 Input/State/Output Models

In Chapter 3 we have seen that it is always possible to partition the variable w into inputs and outputs (see Corollary 3.3.23). This insightful way of viewing a system can be combined with the notion of state. We thus arrive at *input/state/output systems*, a very common way of describing linear systems. Such representations are of the form

$$\begin{array}{rcl} \frac{d}{dt}x &=& Ax + Bu, \\ y &=& Cx + Du. \end{array} \tag{4.16}$$

Here $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ is the input, $x \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^n)$ is the state, and $y \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^p)$ is the output. Consequently, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$. The matrix D is called the *feedthrough* term.

In polynomial form, (4.16) can be written as $R(\frac{d}{dt})w = M(\frac{d}{dt})x$, with $w = \mathrm{col}(u, y)$ and $R(\xi)$, $M(\xi)$ given by

$$R(\xi) := \begin{bmatrix} B & 0 \\ -D & I \end{bmatrix}, \quad M(\xi) = \begin{bmatrix} I\xi - A \\ C \end{bmatrix}. \tag{4.17}$$

Note that the "dynamics", the part of the equations that contains derivatives, of this system is completely contained in the vector (u, x) and the first-equation of (4.16), and that moreover, only first-order derivatives occur. The equation determining y from x and u is static: it does not contain derivatives. The relation between u and x is of an i/o nature in the sense of Definition 3.3.1. To see this, write $\frac{d}{dt}x = Ax + Bu$ as $(\frac{d}{dt}I - A)x = Bu$. Since $\det(I\xi - A) \neq 0$ and since $(I\xi - A)^{-1}B$ is strictly proper (see Exercise 4.19), u is a maximally free variable in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$.

The full behavior defined by (4.16) is defined by

$$\mathfrak{B}_{\mathrm{i/s/o}} := \{(u, x, y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^p) \mid (4.16) \text{ is satisfied weakly } .\}$$

The variable x is considered a *latent* variable, and hence the *manifest* behavior is given by

$$\mathfrak{B}_{\mathrm{i/o}} := \{(u,y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^p) \mid \exists x \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^n) \text{ s.t. } (u,x,y) \in \mathfrak{B}_{\mathrm{i/s/o}}\}. \tag{4.18}$$

In Chapter 6 it turns out that (4.16) is just another representation of an i/o system in the sense that each i/o system of the form $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ as studied in Section 3.3 can be expressed as the manifest behavior of a system of the form (4.16). Hence, in view of Corollary 3.3.23, every system described by $R(\frac{d}{dt})w = 0$ admits an i/s/o representation of the form (4.16). Let us now convince ourselves that (4.16) indeed defines a state space model for the behavior (4.18). We have to verify that the property of state, Definition 4.3.3, holds.

Theorem 4.4.1 *The representation (4.16) is a state space representation.*

Proof This is just a matter of writing (4.16) in the form (4.9), for which we have already proved that it satisfies the property of state. The matrices E, F, G, are readily obtained from (4.17):

$$E = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad F = \begin{bmatrix} -A \\ -C \end{bmatrix}, \quad G = \begin{bmatrix} -B & 0 \\ -D & I \end{bmatrix}.$$

□

Remark 4.4.2 The property of state is the fundamental property of x. As explained, it expresses that $x(t_0)$ splits the past and the future of the behavior. For i/s/o systems of the form (4.16) this should be understood as follows. Take two trajectories (u_1, x_1, y_1) and (u_2, x_2, y_2) in $\mathfrak{B}_{\mathrm{i/s/o}}$. Restrict the first trajectory to the interval $(-\infty, t_0)$, the *past* of (u_1, x_1, y_1), and call it (u_1^-, x_1^-, y_1^-). Analogously, denote the restriction of (u_2, x_2, y_2) to the interval $[t_0, \infty)$, the *future* of (u_2, x_2, y_2), by (u_2^+, x_2^+, y_2^+). Now, whether the past of (u_1, x_1, y_1) and the future of (u_2, x_2, y_2) can be glued together (that is, concatenated) at time t_0 to form a trajectory in the behavior is determined by whether $x_1(t_0)$ and $x_2(t_0)$ are equal. That means that the state at time t_0 contains all the information about the past that is needed to decide whether or not this gluing of trajectories is possible. Stated otherwise, given $x(t_0)$, as far as the future is concerned, we can forget everything that happened before t_0. How the system got into the state $x(t_0)$ is immaterial for its future. □

4.5 The Behavior of i/s/o Models

We now give a complete analysis of what the trajectories of the differential equation (4.16) look like. Observe that since the second equation, $y =$

$Cx + Du$, does not contain any derivatives, the difficulty lies completely in the input/state equation

$$\frac{d}{dt}x = Ax + Bu. \tag{4.19}$$

We derive an explicit expression for the behavior of (4.19) in two steps:

1. The case $u = 0$. If $u = 0$, then (4.19) reduces to the autonomous differential equation $\frac{d}{dt}x = Ax$. The solutions of this equation are characterized in terms of e^{At}, the matrix generalization of the more familiar scalar exponential function e^{at} for $a \in \mathbb{R}$.
2. The general case, $u \neq 0$.

4.5.1 The zero input case

If $u = 0$, then (4.19) reduces to

$$\frac{d}{dt}x = Ax. \tag{4.20}$$

We want to determine all solutions of (4.20). Let us first recall the scalar case, $n = 1$: $\frac{d}{dt}x = ax$, $a \in \mathbb{R}$. For this case, all solutions of (4.20) are of the form

$$x(t) = e^{at}c, \qquad c \in \mathbb{R}.$$

Recall from calculus that one way to define e^{at} is through a power series expansion:

$$e^{at} = \sum_{k=0}^{\infty} \frac{a^k t^k}{k!}. \tag{4.21}$$

From (4.21) it is easy to see that e^{at} satisfies $\frac{d}{dt}x = ax$:

$$\frac{d}{dt}\sum_{k=0}^{\infty} \frac{a^k t^k}{k!} = \sum_{k=0}^{\infty} \frac{d}{dt}\frac{a^k t^k}{k!} = \sum_{k=1}^{\infty} \frac{a^k t^{k-1}}{(k-1)!} = a\sum_{j=0}^{\infty} \frac{a^j t^j}{j!} = ae^{at}. \tag{4.22}$$

This motivates us to define the *exponential* of a matrix.

Definition 4.5.1 Let $M \in \mathbb{R}^{n\times n}$. The (matrix) exponential of M, denoted by e^M, is defined as the infinite series

$$e^M := \sum_{k=0}^{\infty} \frac{M^k}{k!}. \tag{4.23}$$

Here, M^0 is defined as the identity matrix I. In particular, if we take $M = At$, we have

$$e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}. \tag{4.24}$$

□

It is easy to prove, see Exercise 4.10, that the infinite sum in (4.23) converges absolutely, so that e^M is indeed well-defined. Mimicking (4.22), we see that the matrixvalued function of t, e^{At}, satisfies the differential equation $\frac{d}{dt}X = AX$. Indeed, since (4.24) is an absolutely convergent power series, we may interchange summation and differentiation ([51]):

$$\frac{d}{dt}e^{At} = \frac{d}{dt}\sum_{k=0}^{\infty}\frac{A^k t^k}{k!} = \sum_{k=0}^{\infty}\frac{d}{dt}\frac{A^k t^k}{k!} = \sum_{k=1}^{\infty}\frac{A^k t^{k-1}}{(k-1)!} = A\sum_{j=0}^{\infty}\frac{A^j t^j}{j!} = Ae^{At}. \tag{4.25}$$

The discussion above leads to the following characterization.

Theorem 4.5.2 *Let $A \in \mathbb{R}^{n\times n}$. All (strong) solutions of the differential equation $\frac{d}{dt}x = Ax$ are of the form*

$$x(t) = e^{At}c, \qquad c \in \mathbb{R}^n. \tag{4.26}$$

Proof From (4.25) it follows directly that every function of the form (4.26) is a solution of $\frac{d}{dt}x = Ax$.

Conversely, let $x \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n)$ be a solution of $\frac{d}{dt}x = Ax$. Define the function z as $z(t) := e^{-At}x(t)$. It follows that

$$\frac{d}{dt}z = -Ae^{-At}x(t) + e^{-At}\frac{d}{dt}x(t) = -Ae^{-At}x(t) + e^{-At}Ax(t) = 0.$$

This shows that $z(t)$ is constant, say $z(t) = c$. This implies that

$$x(t) = e^{At}c, \qquad \forall t \in \mathbb{R}.$$

□

Remark 4.5.3 An alternative proof of Theorem 4.5.2 is obtained as follows. Observe that the equation $\frac{d}{dt}x = Ax$ is a special case of the general autonomous equation $P(\frac{d}{dt})w = 0$, with $P(\xi) = I\xi - A$. From Theorem 3.2.16 it follows that the solution set is a finite-dimensional subspace of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n)$ of dimension n $(= \deg\det(I\xi - A))$. Hence it suffices to construct n independent solutions of the form (4.26). This is easy; just let c range over all n standard basis vectors: $x_i(t) = e^{At}e_i$, $i = 1, \ldots, n$, $e_i = [0 \cdots 0 \quad 1 \quad 0 \cdots 0]^T$. □

4.5.2 The nonzero input case: The variation of the constants formula

Let us now return to the input/state equation

$$\frac{d}{dt}x = Ax + Bu. \tag{4.27}$$

Define the input/state behavior:

$$\mathfrak{B}_{\mathrm{i/s}} := \{(u,x) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^n) \mid \frac{d}{dt}x = Ax + Bu, \text{ weakly}\}.$$

$\mathfrak{B}_{\mathrm{i/s}}$ is a special case of the i/o systems studied in Chapter 3. Rather than applying the general theory presented there, we explore the special (and simple) structure of (4.27) to determine the corresponding behavior.

Proposition 4.5.4 *Let $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$, and define x by*

$$x(t) := \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau. \tag{4.28}$$

Then $(u,x) \in \mathfrak{B}_{\mathrm{i/s}}$.

Proof Let x be given by (4.28) and suppose u is continuous. Then x is differentiable, and its derivative is given by

$$\begin{aligned}(\frac{d}{dt}x)(t) &= A\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau + Bu(t)\\ &= Ax(t) + Bu(t).\end{aligned}$$

This shows that (u,x) is a strong solution of (4.27), and hence $(u,x) \in \mathfrak{B}_{\mathrm{i/s}}$. For general $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$, not necessarily continuous, the proof that (4.28) defines a weak solution follows the same lines as the second part of the proof of Lemma 3.3.12. □

Corollary 4.5.5 *Every element (u,x) of $\mathfrak{B}_{\mathrm{i/s}}$ is of the form*

$$x(t) = e^{At}c + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau, \qquad c \in \mathbb{R}^n. \tag{4.29}$$

Proof Let (u,x) be of the form (4.29). It follows from Theorem 4.5.2 and Proposition 4.5.4 that $(u,x) \in \mathfrak{B}_{\mathrm{i/s}}$.
Conversely, let $(u,x) \in \mathfrak{B}_{\mathrm{i/s}}$. Define x' and x'' by

$$x'(t) = \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau \quad \text{and} \quad x'' = x - x'.$$

Then

$$\frac{d}{dt}x'' = \frac{d}{dt}x - \frac{d}{dt}x' = Ax + Bu - (Ax' + Bu) = A(x - x') = Ax''. \tag{4.30}$$

It follows from (4.30) and Theorem 4.5.2 that

$$x''(t) = e^{At}c \quad \text{for some } c \in \mathbb{R}^n.$$

This concludes the proof. □

Remark 4.5.6 The expression (4.29) is known as the *variation of the constants formula*. □

4.5.3 The input/state/output behavior

Now that we have determined $\mathfrak{B}_{\mathrm{i/s}}$, it is easy to describe $\mathfrak{B}_{\mathrm{i/s/o}}$ explicitly:

$$\mathfrak{B}_{\mathrm{i/s/o}} = \left\{ (u,x,y) \;\middle|\; \begin{array}{ll} \exists c \in \mathbb{R}^n, & x(t) = e^{At}c + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau \\ & y(t) = Cx(t) + Du(t) \end{array} \right\}.$$

From this description, x can readily be eliminated, yielding $\mathfrak{B}_{\mathrm{i/o}}$:

$$\mathfrak{B}_{\mathrm{i/o}} = \{(u,y) \mid \exists c \in \mathbb{R}^n, \quad y(t) = Ce^{At}c + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t)\}. \tag{4.31}$$

Thus each element of $\mathfrak{B}_{\mathrm{i/o}}$ in this case is completely specified by the arbitrary input $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ and the arbitrary initial state $c = x(0)$. Consequently, the elements of the behavior of the i/o system induced by (4.16) are parametrized by the input $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ and the initial state $x(0) \in \mathbb{R}^n$: once they are given, the output is determined by (4.31).

Summarizing, we have defined i/s/o systems, and we have determined the associated behaviors explicitly.

We have already shown in Theorem 4.4.1 that (4.16) has the property of state. This can also be shown by using the variation of the constants formula, (4.29), but we will not pursue this. Another useful and fundamental property is that given the state at time t_0, the state, and therefore the output from time t_0 on, is completely determined by the input after time t_0.

Property 4.5.7 *Consider the i/s/o system defined by*

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bu, \\ y &= Cx + Du. \end{aligned}$$

Then

(i) x has the state property.

(ii) The system has the property of determinism. This is defined as follows. Let $(u_1, x_1, y_1), (u_2, x_2, y_2) \in \mathfrak{B}_{\mathrm{i/s/o}}$ and suppose that for some $t_0 \in \mathbb{R} x_1(t_0) = x_2(t_0)$, and $u_1(t) = u_2(t)$ for $t \geq t_0$. Then $x_1(t) = x_2(t)$ for $t \geq t_0$, and $y_1(t) = y_2(t)$ for $t \geq t_0$.

Proof (i) That x has the state property was proved in Theorem 4.4.1.
(ii)

$$x_i(t) = e^{A(t-t_0)}x_i(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu_i(\tau)d\tau, \quad i = 1, 2. \tag{4.32}$$

If $x_1(t_0) = x_2(t_0)$ and $u_1(t) = u_2(t)$ for $t \geq t_0$, then it follows form (4.32) that $x_1(t) = x_2(t)$ for $t \geq t_0$. Then also, since $y = Cx + Du$, $y_1(t) = y_2(t)$ for $t \geq t_0$. □

Remark 4.5.8 The property of determinism expresses that the state at time t_0 and the input from time t_0 on uniquely determine the output from time t_0 on. It shows the crucial role played by the inputs as the external variables that drive the system. □

There is a strong connection between the property of state and the property of determinism. We show that determinism combined with nonanticipation (see Remark 3.3.21) implies the property of state. This is a result that holds more generally than for systems described by linear time-invariant differential equations.

Theorem 4.5.9 *Consider a behavior $\mathfrak{B} \subset \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^n)$, not necessarily linear or time-invariant, consisting of time trajectories (u, x). Assume that u is free in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$; i.e., for all $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ there exists a trajectory x such that $(u, x) \in \mathfrak{B}$. Suppose that x does not anticipate u strictly and that the property of determinism is satisfied. Then x satisfies the property of state.*

Proof Choose $(u_1, x_1), (u_2, x_2) \in \mathfrak{B}$ and suppose that $x_1(t_0) = x_2(t_0)$. We have to show that the two trajectories may be concatenated at time t_0. Define the concatenation of u_1 and u_2 at t_0 as $u(t) = u_1(t)$, $t < t_0$, and $u(t) = u_2(t)$, $t \geq t_0$. By assumption there exists x such that $(u, x) \in \mathfrak{B}$. Moreover, by strict nonanticipation, we know that x may be taken such that $x(t) = x_1(t)$, $t \leq t_0$. Finally, by the property of determinism, the future of x is uniquely determined by $x(t_0)$ and the input for $t \geq t_0$, and since $(u_2, x_2) \in \mathfrak{B}$, it follows that $x(t) = x_2(t)$ for $t \geq t_0$. □

Remark 4.5.10 The condition that the relation between u and x is *strictly* nonanticipating is essential for Theorem 4.5.9 to hold. See Exercise 4.20. □

Remark 4.5.11 We now know that (4.16) indeed defines a state space representation. The i/o system of which it is a state space representation is given by (4.31). In Chapter 6 we will see that this input/output system can also be represented in the form $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$, with $P^{-1}(\xi)Q(\xi)$ a matrix of proper rational functions. □

4.5.4 How to calculate e^{At}?

It is clear that in the characterization of the behavior of $\frac{d}{dt}x = Ax+Bu$, the matrix exponential e^{At} plays a crucial role. Definition 4.5.1 does not give a clue as to what e^{At} actually looks like, nor does it provide a constructive way to calculate it in concrete examples. Calculation of e^{At} may be achieved via several different methods. We discuss three of these methods:

1. By transforming A into Jordan normal form.
2. By applying the theory of higher-order autonomous behaviors as studied in Section 3.2.
3. Using the partial fraction expansion of $(I\xi - A)^{-1}$.

In the following proposition we have collected some useful properties of the matrix exponential.

Proposition 4.5.12
If M_1 and M_2 commute, i.e., if $M_1M_2 = M_2M_1$, then

$$e^{M_1+M_2} = e^{M_1}e^{M_2}. \tag{4.33}$$

If M_1 and M_2 are square matrices, then

$$e^{\begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix}} = \begin{bmatrix} e^{M_1} & 0 \\ 0 & e^{M_2} \end{bmatrix}.$$

If S is nonsingular, then

$$e^{S^{-1}MS} = S^{-1}e^{M}S. \tag{4.34}$$

If $\lambda_i \in \mathbb{C}$, $i = 1, \ldots, n$, then

$$e^{\mathrm{diag}(\lambda_1, \ldots, \lambda_n)} = \mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n}). \tag{4.35}$$

The matrix exponential of a matrix with ones on the upper diagonal and zeros elsewhere is given by

$$e^{\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix} t} = \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{n-1}}{(n-1)!} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \ddots & \frac{t^2}{2!} \\ & & \ddots & \ddots & t \\ & & & 0 & 1 \end{bmatrix}, \quad (4.36)$$

where n is the number of rows and columns of the matrices in (4.36).
If $\omega \in \mathbb{R}$, then

$$e^{\begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix}} = \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix}.$$

Proof The proofs are straightforward applications of Definition 4.5.1 and are left as an exercise; see Exercise 4.11. □

4.5.4.1 Calculation of e^{At} via the Jordan form

If A has a basis of eigenvectors, then e^{At} may be calculated as follows. Assume that $Av_i = \lambda_i v_i$, $i = 1, \ldots, n$, and that the vectors v_i form a basis of $\mathbb{C}^n$. Define the nonsingular matrix S and the diagonal matrix Λ by

$$S = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Then $S^{-1}AS = \Lambda$. Using (4.34) and (4.35) we conclude that

$$e^{At} = S \operatorname{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}) S^{-1}.$$

Example 4.5.13 Let A be given by

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{-3}{2} & 0 & 1 \end{bmatrix}.$$

The characteristic polynomial of A is $\det(I\xi - A) = (\xi - 1)^2(\xi + 1)$. The eigenvalues of A are $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = -1$. Corresponding eigenvectors are

$$v_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad v_3 = \begin{bmatrix} -2 \\ 1 \\ \frac{-3}{2} \end{bmatrix},$$

so that

$$e^{At} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 0 & 1 \\ 0 & 1 & \frac{-3}{2} \end{bmatrix} \begin{bmatrix} e^t & 0 & 0 \\ 0 & e^t & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 0 & 0 & -2 \\ 1 & 0 & 1 \\ 0 & 1 & \frac{-3}{2} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} e^{-t} & 0 & 0 \\ \frac{1}{2}e^t - \frac{1}{2}e^{-t} & e^t & 0 \\ \frac{3}{4}e^{-t} - \frac{3}{4}e^t & 0 & e^t \end{bmatrix}.$$

□

Not every matrix has a basis of eigenvectors. As an example of a matrix that does not have a basis of eigenvectors, consider

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

Although this matrix cannot be diagonalized by means of a similarity transformation, it is easy to compute e^{At}. Using (4.33, 4.35, 4.36) it follows that

$$e^{At} = e^{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} t + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} t} = e^{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} t} \quad e^{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} t} = \begin{bmatrix} e^t & te^t \\ 0 & e^t \end{bmatrix}. \tag{4.37}$$

The attentive reader may have recognized that A in (4.37) is in *Jordan form*. Recall that every matrix $A \in \mathbb{R}^{n \times n}$ may be transformed into Jordan form by means of a similarity transformation; i.e., there exists a nonsingular matrix S such that

$$S^{-1}AS = \begin{bmatrix} J_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & J_N \end{bmatrix}. \tag{4.38}$$

The (possibly complex) submatrices J_k in (4.38) are called the *Jordan blocks.* These are defined as follows. Let $v_1, \ldots, v_N$ be a maximal set of independent eigenvectors of A, say $Av_k = \lambda_k v_k$. To each v_k there corresponds exactly one Jordan block J_k of the form

$$J_k = \begin{bmatrix} \lambda_k & 1 & & \\ 0 & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix}.$$

The number of Jordan blocks corresponding to an eigenvalue λ of A is equal to the maximal number of independent eigenvectors of the eigenvalue and is called the *geometric multiplicity* of λ. The multiplicity of λ as a root of the characteristic polynomial of A, on the other hand, is referred to as the *algebraic multiplicity*.

Using Proposition 4.5.12, it follows that

$$e^{At} = Se^{Jt}S^{-1}, \qquad e^{Jt} = \operatorname{diag}(e^{J_1 t}, \ldots, e^{J_N t}).$$

Finally, using (4.33, 4.35, 4.36) we obtain

$$e^{J_k t} = \begin{bmatrix} e^{\lambda_k t} & te^{\lambda_k t} & \frac{t^2}{2!}e^{\lambda_k t} & \frac{t^3}{3!}e^{\lambda_k t} & \cdots & \cdots & \cdots \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & & & & \ddots & \\ & & \ddots & \ddots & \ddots & & \frac{t^3}{3!}e^{\lambda_k t} \\ & & & & & \ddots & \\ & & & \ddots & \ddots & & \frac{t^2}{2!}e^{\lambda_k t} \\ & & & & & \ddots & \\ & & & & \ddots & & te^{\lambda_k t} \\ & & & & & \ddots & \\ & & & & & & e^{\lambda_k t} \end{bmatrix}. \tag{4.39}$$

Expression (4.39) provides a clear insight as to what kind of entries the matrix e^{At} contains. Apparently, they are linear combinations of products of $e^{\lambda_k t}$s, with the λ_ks the eigenvalues of A, and polynomials in t. The maximal degree of the polynomial parts is related to the dimensions of the Jordan blocks. A Jordan block with ℓ rows and columns gives rise to polynomial parts up to and including degree $\ell - 1$.

Example 4.5.14 Let $A \in \mathbb{R}^{3\times 3}$ be given by

$$A = \begin{bmatrix} -4 & 4 & 3 \\ -12 & 11 & 8 \\ 9 & -8 & -6 \end{bmatrix}.$$

The characteristic polynomial of A is $\det(I\xi - A) = (\xi - 1)^2(\xi + 1)$. The corresponding eigenvectors are

$$v_1 = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 1 \\ 3 \\ -3 \end{bmatrix}, \quad Av_1 = v_1, \quad Av_2 = -v_2.$$

In addition to the eigenvectors v_1 and v_2, A has a *generalized eigenvector* w_1 corresponding to the eigenvalue 1:

$$w_1 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \quad Aw_1 = v_1 + w_1.$$

Define $S = [v_1 w_1 v_2]$. Then

$$S = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \\ -1 & -1 & -3 \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} 0 & 1 & 1 \\ -3 & 2 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

and

$$S^{-1}AS = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

This matrix is in Jordan form, so we can now compute e^{At} by using the fact that $e^{At} = Se^{S^{-1}AStS^{-1}}$, yielding

$$e^{At} = \begin{bmatrix} -3te^t + e^{-t} & (1+2t)e^t - e^{-t} & (1+t)e^t - e^{-t} \\ (-3-6t)e^t + 3e^{-t} & (4+4t)e^t - 3e^{-t} & (3+2t)e^t - 3e^{-t} \\ (3+3t)e^t - 3e^{-t} & (-3-2t)e^t + 3e^{-t} & (-2-t)e^t + 3e^{-t} \end{bmatrix}.$$

□

4.5.4.2 Calculation of e^{At} using the theory of autonomous behaviors

The differential equation $\frac{d}{dt}x = Ax$ is a special case of the general equations $P(\frac{d}{dt})w = 0$ studied in Section 3.2. Theorem 3.2.16 gives a complete characterization of the solutions $P(\frac{d}{dt})w = 0$, so it seems reasonable to expect that by invoking this theorem we should be able to consider e^{At} from a somewhat different perspective. Below we show how this can be done.

Definition 4.5.15 The unique function $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{n \times n}$ with the properties

- $\forall t_0, t \in \mathbb{R}: \ \dfrac{d}{dt}\Phi(t, t_0) = A\Phi(t, t_0),$
- $\forall t_0 \in \mathbb{R}: \ \Phi(t_0, t_0) = I$

is called the *state transition matrix*, or simply the *transition matrix*, of the autonomous system defined by $\frac{d}{dt}x = Ax$. □

Remark 4.5.16

- The term *transition matrix* stems from the property that to every $c \in \mathbb{R}^n$ there corresponds exactly one solution x of $\frac{d}{dt}x = Ax$ such that $x(t_0) = c$, namely $x(t) = \Phi(t, t_0)c$.
- Since the system is time-invariant, it is sufficient to restrict attention to $t_0 = 0$. By abuse of notation, $\Phi(t, 0)$ is often written as $\Phi(t)$.

□

Theorem 4.5.17 *Let $A \in \mathbb{R}^{n\times n}$ and let Φ the matrix valued function as defined in Definition 4.5.15. Then $\Phi(t, t_0) = e^{A(t-t_0)}$.*

Proof By the uniqueness property of the transition matrix, it suffices to show that $e^{A(t-t_0)}$ satisfies the requirements of Definition 4.5.15. By Theorem 4.5.2 it follows that $\frac{d}{dt}e^{A(t-t_0)} = Ae^{A(t-t_0)}$. Moreover, by definition, $e^{A(t_0-t_0)} = e^0 = I$. □

Suppose now that we have n independent solutions $x_1, \ldots, x_n$ of $\frac{d}{dt}x = Ax$. From these functions we can form the matrix $X := [x_1 \cdots x_n]$. By construction, X satisfies the *matrix differential equation*

$$\frac{d}{dt}X = AX. \tag{4.40}$$

Since the columns of X are linearly independent and since X satisfies (4.40), it follows that for every $t \in \mathbb{R}$, $X(t)$ is nonsingular; see Exercise 4.16. Define $\Phi(t, t_0) := X(t)X^{-1}(t_0)$. Then

$$\frac{d}{dt}\Phi(t, t_0) = \frac{d}{dt}X(t)X^{-1}(t_0) = A\Phi(t, t_0).$$

Moreover $\Phi(t_0, t_0) = I$. It follows that this Φ is the transition matrix, and therefore $e^{At} = X(t)X(0)^{-1}$.

The conclusion is that every n-tuple of linearly independent solutions of $\frac{d}{dt}x = Ax$ provides a means to obtain e^{At}. The question now is how to find n independent solutions. From Theorem 3.2.16 we know that every *strong* solution of (4.20) can be written as

$$x(t) = \sum_{i=1}^{N}\sum_{j=0}^{n_i-1} B_{ij}t^j e^{\lambda_i t}, \tag{4.41}$$

where the complex numbers λ_i are the *distinct* roots of the polynomial $p(\xi) := \det(I\xi - A)$ and the n_is are their respective algebraic multiplicities, and where the vectors $B_{ij} \in \mathbb{C}^n$ satisfy the linear relations

$$\sum_{j=\ell}^{n_i-1} \binom{j}{\ell} \frac{d^{j-\ell}}{ds^{j-\ell}}(sI - A)_{|s=\lambda_i} B_{ij} = 0, \quad i = 1, \ldots, N, \quad \ell = 0, \ldots, n_i - 1. \tag{4.42}$$

Since the derivatives of $sI-A$ of order larger than one are zero, the relations (4.42) reduce to

$$\begin{aligned}
(\lambda_i I - A)B_{i,n_i-1} &= 0,\\
(\lambda_i I - A)B_{i,n_i-2} + (n_i-1)B_{i,n_i-1} &= 0,\\
(\lambda_i I - A)B_{i,n_i-3} + (n_i-2)B_{i,n_i-2} &= 0,\\
&\vdots\\
(\lambda_i I - A)B_{i,1} + 2B_{i,2} &= 0,\\
(\lambda_i I - A)B_{i,0} + B_{i,1} &= 0.
\end{aligned} \tag{4.43}$$

It follows from Theorem 3.2.16 that the dimension of the autonomous behavior of $\frac{d}{dt}x = Ax$ equals the degree of $\det(I\xi - A)$; i.e., it has dimension n. As a consequence, we can find n linearly independent solutions $x_1, \ldots, x_n$ of the form (4.41).

Example 4.5.18 Take

$$A = \begin{bmatrix} 3 & -2 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix}. \tag{4.44}$$

Then $\det(I\xi - A) = -2 + 5\xi - 4\xi^2 + \xi^3 = (\xi-1)^2(\xi-2)$. The characteristic roots are $\lambda_1 = 1,\ n_1 = 2$ and $\lambda_2 = 2,\ n_2 = 1$. Every strong solution of $\frac{d}{dt}x = Ax$ is of the form

$$x(t) = B_{10}e^t + B_{11}te^t + B_{20}e^{2t}. \tag{4.45}$$

The vectors B_{ij} should satisfy the relations

$$\begin{aligned}
(I - A)B_{10} + B_{11} &= 0,\\
(I - A)B_{11} &= 0,\\
(2I - A)B_{20} &= 0.
\end{aligned} \tag{4.46}$$

Solving these equations yields

$$B_{10} = a\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad B_{11} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad B_{20} = c\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}. \tag{4.47}$$

Hence every solution x can be written as

$$x(t) = a \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} e^t + b \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^t + c \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} e^{2t}. \tag{4.48}$$

From here three linearly independent solutions are easily obtained:

$$x_1(t) = \begin{bmatrix} e^t \\ e^t \\ 0 \end{bmatrix}, \quad x_2(t) = \begin{bmatrix} 0 \\ 0 \\ e^t \end{bmatrix}, \quad x_3(t) = \begin{bmatrix} 2e^{2t} \\ e^{2t} \\ e^{2t} \end{bmatrix}. \tag{4.49}$$

The matrix X is defined as $X = [x_1\ x_2\ x_3]$. Finally

$$e^{At} = \Phi(t) := X(t)X^{-1}(0) = \begin{bmatrix} 2e^{2t} - e^t & 2e^t - 2e^{2t} & 0 \\ e^{2t} - e^t & 2e^t - e^{2t} & 0 \\ e^{2t} - e^t & e^t - e^{2t} & e^t \end{bmatrix}. \tag{4.50}$$

□

4.5.4.3 Calculation of e^{At} using the partial fraction expansion of $(I\xi - A)^{-1}$

As argued in Section 4.5.2, the behavior of $\frac{d}{dt}x = Ax + Bu$, denoted by $\mathfrak{B}_{\mathrm{i/s}}$, is in input/output form with the state as the output. From Section 3.3 we know that the pairs $(u, x) \in \mathfrak{B}_{\mathrm{i/s}}$ can be described in terms of the partial fraction expansion of $(I\xi - A)^{-1}B$. Suppose

$$(I\xi - A)^{-1}B = \sum_{i=1}^{N} \sum_{j=1}^{n_i} T_{ij} \frac{1}{(\xi - \lambda_i)^j} \ , \quad T_{ij} \in \mathbb{C}^{n \times m}.$$

Then to every $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ there corresponds $x \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^n)$ such that $(u, x) \in \mathfrak{B}_{\mathrm{i/s}}$. One such x is given by

$$x(t) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} T_{ij} \int_0^t \frac{(t-\tau)^{j-1}}{(j-1)!} e^{\lambda_i (t-\tau)} u(\tau) d\tau.$$

On the other hand, we know from (4.28) that also $(u, \tilde{x}) \in \mathfrak{B}_{\mathrm{i/s}}$, with $\tilde{x}$ given by

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

Now, since $x(0) = \tilde{x}(0)$ and since given the initial state, the solution of $\frac{d}{dt}x = Ax + Bu$ is unique, it follows that $x = \tilde{x}$. Therefore, since u was

arbitrary, we conclude that

$$e^{At}B = \sum_{i=1}^{N} \sum_{j=1}^{n_i} T_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t}. \tag{4.51}$$

In other words, $e^{At}B$ may be computed from the partial fraction expansion of $(I\xi - A)^{-1}B$. In particular, if we take $B = I$, we obtain a third method for the calculation of e^{At}.

Corollary 4.5.19 *Let* $(I\xi - A)^{-1} = \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{n_i} T_{ij} \frac{1}{(\xi - \lambda_i)^j}$, $T_{ij} \in \mathbb{C}^{n \times n}$. *Then*

$$e^{At} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} T_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t}.$$

Proof Take $B = I$ in (4.51). □

Example 4.5.20 Consider the matrix A in Example 4.5.14:

$$A = \begin{bmatrix} -4 & 4 & 3 \\ -12 & 11 & 8 \\ 9 & -8 & -6 \end{bmatrix}.$$

The partial fraction expansion of $(I\xi - A)^{-1}$ (see Remark 3.3.11) is given by

$$\begin{bmatrix} 1 & -1 & -1 \\ 3 & -3 & -3 \\ -3 & 3 & 3 \end{bmatrix} \frac{1}{\xi + 1} + \begin{bmatrix} 0 & 1 & 1 \\ -3 & 4 & 3 \\ 3 & -3 & -2 \end{bmatrix} \frac{1}{\xi - 1} + \begin{bmatrix} -3 & 2 & 1 \\ -6 & 4 & 2 \\ 3 & -2 & -1 \end{bmatrix} \frac{1}{(\xi - 1)^2}.$$

Applying Corollary 4.5.19, it follows that

$$\begin{aligned} e^{At} &= \begin{bmatrix} 1 & -1 & -1 \\ 3 & -3 & -3 \\ -3 & 3 & 3 \end{bmatrix} e^{-t} + \begin{bmatrix} 0 & 1 & 1 \\ -3 & 4 & 3 \\ 3 & -3 & -2 \end{bmatrix} e^{t} + \begin{bmatrix} -3 & 2 & 1 \\ -6 & 4 & 2 \\ 3 & -2 & -1 \end{bmatrix} te^{t} \\ &= \begin{bmatrix} -3te^t + e^{-t} & (1+2t)e^t - e^{-t} & (1+t)e^t - e^{-t} \\ (-3-6t)e^t + 3e^{-t} & (4+4t)e^t - 3e^{-t} & (3+2t)e^t - 3e^{-t} \\ (3+3t)e^t - 3e^{-t} & (-3-2t)e^t + 3e^{-t} & (-2-t)e^t + 3e^{-t} \end{bmatrix}, \end{aligned}$$

which was already derived in Example 4.5.14 by using the Jordan form of A. □

Example 4.5.21 If the matrix A has a complex eigenvalue with nonzero imaginary part, then e^{At} contains trigonometric functions. Consider for example the matrix

$$A = \begin{bmatrix} 0 & -2 \\ 1 & 2 \end{bmatrix}.$$

Its characteristic polynomial is $p(\xi) = \det(\xi - A) = 2 - 2\xi + \xi^2$. It follows that the characteristic values are given by $\lambda_1 = 1 + i$, $\lambda_2 = 1 - i$. Using either of the methods, it follows that

$$\begin{aligned} e^{At} &= \begin{bmatrix} e^t[\frac{1}{2}(e^{it} + e^{-it}) - \frac{1}{2i}(e^{it} - e^{-it})] & -2e^t\frac{1}{2i}(e^{it} - e^{-it}) \\ e^t\frac{1}{2i}(e^{it} - e^{-it}) & e^t[\frac{1}{2}(e^{it} + e^{-it}) + \frac{1}{2i}(e^{it} - e^{-it})] \end{bmatrix} \\ &= \begin{bmatrix} e^t(\cos t - \sin t) & -2e^t \sin t \\ e^t \sin t & e^t(\cos t + \sin t) \end{bmatrix}. \end{aligned}$$

□

Remark 4.5.22 We have presented three methods to calculate e^{At}. The main purpose for providing these methods is to gain insight into some of the features of e^{At}. In particular, it is clear now how the eigenvalues and their algebraic and geometric multiplicities enter the picture. In practice, we will of course use numerically reliable methods to calculate e^{At}. The three methods that we presented need not always offer reasonable numerical procedures for computing e^{At}. An overview of various other methods for computing e^{At} and their numerical properties may be found in [42]. □

4.6 State Space Transformations

In Section 2.5 we have seen that different polynomial matrices may represent the same behavior. We now study the question to what extent i/s/o representations of the same input/output behavior are nonunique. Since we are mainly interested in the input and output variables and not so much in the state variables, we use a weaker concept of equivalence.

Definition 4.6.1 Two i/s/o representations are called *input/output equivalent* if they represent the same input/output behavior. □

The i/s/o representations of a given i/o system are not unique. Indeed, consider

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bu, \\ y &= Cx + Du. \end{aligned} \tag{4.52}$$

Let $S \in \mathbb{R}^{n\times n}$ be a nonsingular matrix and let (u, x, y) satisfy (4.52). Define $\tilde{x} := Sx$. Note that this corresponds to expressing the state coordinates with respect to a new basis. The differential equation governing $\tilde{x}$ is

$$\begin{aligned} \frac{d}{dt}\tilde{x} &= SAS^{-1}\tilde{x} + SBu, \\ y &= CS^{-1}\tilde{x} + Du. \end{aligned} \tag{4.53}$$

Equations (4.52, 4.53) show that every (u, y) that belongs to the i/o behavior defined by (4.52) also belongs to the i/o behavior defined by (4.53). By applying the inverse transformation to (4.53) it follows that the converse is also true. This means that (4.52) and (4.53) represent the same i/o behavior, and therefore the representations (4.52) and (4.53) are *input/output equivalent.* We state this as a theorem.

Theorem 4.6.2 *Two state space representations of the form (4.52), parametrized by* (A_1, B_1, C_1, D_1) *and* (A_2, B_2, C_2, D_2) *respectively, are input/output equivalent if there exists a nonsingular matrix* $S \in \mathbb{R}^{n\times n}$ *such that* $SA_1S^{-1} = A_2, \quad SB_1 = B_2, \quad C_1S^{-1} = C_2, \quad D_1 = D_2.$

Correspondingly, we call two quadruples $(A_1, B_1, C_1, D_1), (A_2, B_2, C_2, D_2) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$ equivalent, or *similar*, if there exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that $SA_1S^{-1} = A_2$, $SB_1 = B_2$, $C_1S^{-1} = C_2$, and $D_1 = D_2$. The matrix S is called the corresponding state similarity transformation matrix.

Remark 4.6.3 Note that Theorem 4.6.2 shows that similarity implies the same i/o behavior. However, if two representations of the form (4.52) are i/o equivalent, then the corresponding quadruples of system matrices need not be similar. See Exercise 4.22. □

4.7 Linearization of Nonlinear i/s/o Systems

Our main interest in this book concerns linear systems. However, many systems in applications are nonlinear, particularly in areas such as mechanics and chemical reactions. However, linear systems are quite important for the analysis of nonlinear systems, since nonlinear systems can in the neighborhood of a nominal trajectory be described approximately by a linear system. This procedure of replacing the nonlinear system by a linear one

is called *linearization* and will now be explained. For simplicity, we restrict attention to nominal trajectories that are constant in time. These are called *equilibrium solutions*.

Consider the nonlinear input/state/output system described by the system of differential equations

$$\frac{dx}{dt} = f(x,u), \quad y = h(x,u). \tag{4.54}$$

Here $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is called the *state evolution function*; $f(x,u)$ shows what the derivative of the state trajectory is equal to when the system is in state x and the input value applied is u. The map $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ is called the *read-out map*; $h(x,u)$ shows what the output value is equal to when the system is in state x and the input value applied is u.

Of course, we may view (4.54) as defining a dynamical system with manifest variables $w = (u,y)$ and latent variable x. Let $\mathfrak{B}$ denote its behavior. Formally,

$$\mathfrak{B} = \{(u,y,x) : \mathbb{R} \to \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n \mid x \in \mathcal{C}^1(\mathbb{R},\mathbb{R}^n) \text{ and (4.54) is satisfied } \}.$$

With this definition of behavior, it is easy to prove that x is a state variable in the sense of Definition 4.3.3. Intuitively, it is also clear that u is an input variable (in the sense that it is free) and that y is an output variable (in the sense that y is uniquely specified by u and $x(0)$). However, in order to prove this formally, we would need to impose some smoothness conditions (of the Lipschitz continuity type) in order to ensure that the initial value problem

$$\frac{dx}{dt} = f(x,u(t)), \quad x(0) = x_0,$$

has a unique solution for all $u \in L_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^m)$ and $x_0 \in \mathbb{R}^n$. We do not enter into these considerations in this book.

Of special interest are the elements in the behavior $\mathfrak{B}$ of (4.54) that are constant in time. Let $w^* = (u^*,y^*,x^*) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^m$. It is easily seen that w defined by $w(t) = w^*$ belongs to $\mathfrak{B}$ if and only if $f(x^*,u^*) = 0$ and $y^* = h(x^*,u^*)$. An element $(u^*,y^*,x^*) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ satisfying this is called an *equilibrium point*. We will soon see how the system (4.54) can be approximated with a linear one such as (4.44) in the neighborhood of an equilibrium point. However, before we discuss this linearization, we give an example of how equations (4.54) are arrived at and how equilibria are obtained.

Example 4.7.1 Inverted pendulum Consider the mechanical system depicted in Figure 4.3. An inverted pendulum is mounted on a carriage moving on a horizontal rail. The carriage has mass M and is attached to a wall via a spring with spring constant k_2. The pendulum is mounted

on the carriage by means of a spring with spring constant k_1. The length of the pendulum is 2ℓ and its mass, assumed homogeneously distributed along the rod, is denoted by m. We can exert a force u on the carriage. The position of the center of gravity of the carriage with respect to its equilibrium position is denoted by z, and the angle of the pendulum with respect to the vertical position by θ. The input to the system is the force u, and the output is the angle θ. From the laws of mechanics it follows that

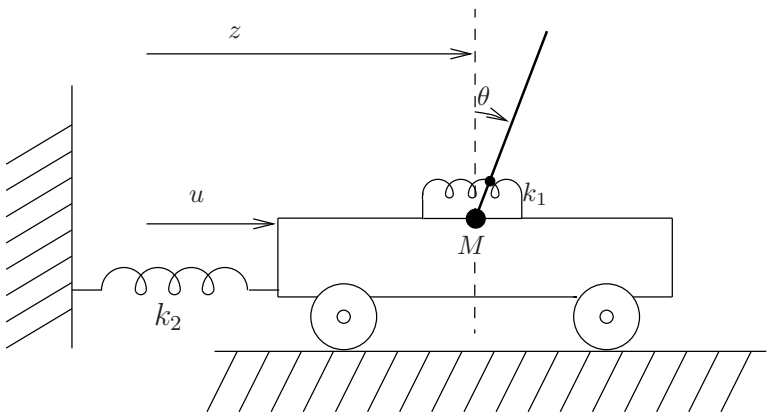


FIGURE 4.3. An inverted pendulum on a carriage.

the equations relating u, z, and θ are given by

$$\begin{aligned}(M+m)\frac{d^2}{dt^2}z + k_2 z + m\ell\cos\theta\,\frac{d^2}{dt^2}\theta &= m\ell(\frac{d}{dt}\theta)^2\sin\theta + u,\\ m\ell\cos\theta\frac{d^2}{dt^2}z + \frac{4}{3}m\ell^2\frac{d^2}{dt^2}\theta &= mg\ell\sin\theta - k_1\theta.\end{aligned} \tag{4.55}$$

Introduce the state vector $x = \mathrm{col}(z, \frac{d}{dt}z, \theta, \frac{d}{dt}\theta)$. For simplicity, take $M = 1$, $m = 1$, and $\ell = 1$. The nonlinear i/s equations can be written as $\frac{d}{dt}x = f(x,u)$, with f given by

$$\begin{bmatrix} x_2 \\ \dfrac{4k_2x_1 - 3k_1x_3\cos x_3 - 4x_4^2\sin x_3 + 3g\cos x_3\sin x_3 - 4u}{3\cos^2 x_3 - 8} \\ x_4 \\ \dfrac{3k_2x_1\cos x_3 + 6k_1x_3 - 6g\sin x_3 + 3x_4^2\cos x_3\sin x_3 + 3u\cos x_3}{3\cos^2 x_3 - 8} \end{bmatrix}. \tag{4.56}$$

The output equation is given by $y = h(x) := x_3$. The equilibria of the system when no force is acting on the carriage, i.e., $u = 0$, can be found

by solving the equation $f(x,0)=0$. It is easy to check that $x=0$, $y=0$ is an equilibrium. Physically, this corresponds to the situation where the cart is at rest in its equilibrium position and the pendulum is at rest in vertical position. In this example there are, however, two more equilibria; see Exercise 4.24. □

Since f and h are continuously differentiable, we may write, using Taylor's formula

$$\begin{aligned} f(x,u) &= f(x^*,u^*) + [\tfrac{\partial f}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial f}{\partial u}(x^*,u^*)](u-u^*) + r_f(x,u), \\ h(x,u) &= h(x^*,u^*) + [\tfrac{\partial h}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial h}{\partial u}(x^*,u^*)](u-u^*) + r_h(x,u), \end{aligned} \tag{4.57}$$

where $[\frac{\partial f}{\partial x}(x^*,u^*)]$ and $[\frac{\partial f}{\partial u}(x^*,u^*)]$ denote the matrices of partial derivatives of f with respect to x and u respectively, evaluated at the point (x^*,u^*). Similarly for h. See also (4.60, 4.61). The functions r_f and r_h satisfy

$$\lim_{(x,u)\to(x^*,u^*)} \frac{r_f(x,u)}{\|(x,u)\|} = 0 \quad \text{and} \quad \lim_{(x,u)\to(x^*,u^*)} \frac{r_h(x,u)}{\|(x,u)\|} = 0. \tag{4.58}$$

It follows from (4.57, 4.58) and the fact that $f(x^*,u^*)=0$ and $h(x^*,u^*)=y^*$ that if $\|(x-x^*,u-u^*)\|$ is small, then $f(x,u)$ and $h(x,u)$ may be approximated as

$$\begin{aligned} f(x,u) &\approx [\tfrac{\partial f}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial f}{\partial u}(x^*,u^*)](u-u^*), \\ h(x,u) &\approx y^* + [\tfrac{\partial h}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial h}{\partial u}(x^*,u^*)](u-u^*). \end{aligned}$$

As a consequence, we expect that in the neighborhood of the equilibrium, the differential equation (4.54) may be approximated by the linear differential equation

$$\begin{aligned} \tfrac{d}{dt}x &= [\tfrac{\partial f}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial f}{\partial u}(x^*,u^*)](u-u^*), \\ y-y^* &= [\tfrac{\partial h}{\partial x}(x^*,u^*)](x-x^*) + [\tfrac{\partial h}{\partial u}(x^*,u^*)](u-u^*). \end{aligned}$$

Motivated by the discussion above, we define the *linearization* of the system (4.54) about the equilibrium $(x,u)=(x^*,u^*)$ as

$$\begin{aligned} \tfrac{d}{dt}(x-x^*) &= A(x-x^*) + B(u-u^*), \\ y-y^* &= C(x-x^*) + D(u-u^*), \end{aligned} \tag{4.59}$$

where the matrices (A,B,C,D) are given by (notice that f has n components $(f_1,\ldots,f_n)$, $f_i:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ and $h=(h_1,\ldots,h_p)$, $h_j:\mathbb{R}^n\times\mathbb{R}^m\to$

$\mathbb{R}$)

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^*,u^*) \dots \frac{\partial f_1}{\partial x_n}(x^*,u^*) \\ \vdots \qquad\qquad \vdots \\ \frac{\partial f_n}{\partial x_1}(x^*,u^*) \dots \frac{\partial f_n}{\partial x_n}(x^*,u^*) \end{bmatrix}, \quad B = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(x^*,u^*) \dots \frac{\partial f_1}{\partial u_m}(x^*,u^*) \\ \vdots \qquad\qquad \vdots \\ \frac{\partial f_n}{\partial u_1}(x^*,u^*) \dots \frac{\partial f_n}{\partial u_m}(x^*,u^*) \end{bmatrix}, \tag{4.60}$$

$$C = \begin{bmatrix} \frac{\partial h_1}{\partial x_1}(x^*,u^*) \dots \frac{\partial h_1}{\partial x_n}(x^*,u^*) \\ \vdots \qquad\qquad \vdots \\ \frac{\partial h_p}{\partial x_1}(x^*,u^*) \dots \frac{\partial h_p}{\partial x_n}(x^*,u^*) \end{bmatrix}, \quad D = \begin{bmatrix} \frac{\partial h_1}{\partial u_1}(x^*,u^*) \dots \frac{\partial h_1}{\partial u_m}(x^*,u^*) \\ \vdots \qquad\qquad \vdots \\ \frac{\partial h_p}{\partial u_1}(x^*,u^*) \dots \frac{\partial h_p}{\partial u_m}(x^*,u^*) \end{bmatrix}. \tag{4.61}$$

The usefulness of the representation (4.59) lies in the fact that it provides a linear approximation of the nonlinear behavior. The closer that $x(0)$ and u are to (x^*, u^*), the better the approximation is. Thus (4.59) gives a *local* description of the nonlinear behavior.

To illustrate these formulas, we derive the linearization of the system of Example 4.7.1

Example 4.7.2 Inverted pendulum, continued Let us now determine the linearization of the system (4.56) about $(u, x) = (0, 0)$. Obviously, the D is the zero matrix. Calculation of the matrices A, B, and C according to (4.60, 4.61) yields

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{-4k_2}{5} & 0 & \frac{3k_1 - 3g}{5} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{3k_2}{5} & 0 & \frac{6g - 6k_1}{5} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \frac{4}{5} \\ 0 \\ \frac{-3}{5} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}. \tag{4.62}$$

This example is continued in Exercise 7.31. □

Linearization may also be defined for other than state space systems. Let $G : (\mathbb{R}^q)^{L+1} \to \mathbb{R}^g$ be continuously differentiable, and consider the nonlinear behavior defined by

$$G(w, \frac{d}{dt}w, (\frac{d}{dt})^2 w, \dots, (\frac{d}{dt})^L w) = 0. \tag{4.63}$$

Assume that w^* is an equilibrium solution of (4.63): $G(w^*, 0, \dots, 0) = 0$. Define matrices R_i:

$$R_i = \frac{\partial G}{\partial z_i}(w^*, 0, \dots, 0), \quad i = 0, 1, \dots, L.$$

Analogously to (4.57), $G(z)$ may be written as

$$G(z_0, z_1, \ldots, z_L) = R_0(z_0 - w^*) + R_1 z_1 + \cdots + R_L z_L + r_G(z_0, \ldots, z_L),$$

and r_G satisfies

$$\lim_{z \to (w^*, 0, \ldots, 0)} \frac{r_G(z)}{\|z\|} = 0.$$

The linearization of (4.63) about the equilibrium $(w^*, 0, \ldots, 0)$ is defined as

$$R(\frac{d}{dt})w = 0,$$

where the polynomial matrix $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ is given by

$$R(\xi) = R_0 + R_1\xi + R_2\xi^2 + \cdots + R_L\xi^L.$$

See Exercise 4.23 to apply linearization for higher-order differential equations to the equations (4.55).

4.8 Recapitulation

In this chapter we have studied a particular class of latent variable models, namely state space models. The main points were:

- Latent variable models are the result of modeling from first principles.
- A special class of latent variable models are those that are first-order in the latent variables and order zero in the manifest variables. These models have the property of state: the possible future of a trajectory is completely determined by the value of the state variable at the present time and does not depend on the past of the trajectory (Theorem 4.3.5).
- An important class of state space models is formed by the linear time-invariant input/state/output models of the form $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$. We derived a complete characterization of the trajectories in the corresponding behaviors (Section 4.5.3).
- In the analysis of the behavior of $\frac{d}{dt}x = Ax + Bu$, the exponential of a matrix, in particular e^{At}, played a central role. We discussed three methods to calculate e^{At}: via the Jordan form of A, using the theory of autonomous behaviors, and using the partial fraction expansion of $(I\xi - A)$ (Section 4.5.4).
- A change of basis in the state space of a linear i/s/o system leads to what is called a similar system. Similarity transformations do not affect the i/o behavior (Theorem 4.6.2).
- Nonlinear differential equations may be linearized about equilibrium solutions. The behavior of the resulting linear differential system approximates the nonlinear behavior in the neighborhood of the equilibrium (Section 4.7).

4.9 Notes and References

State space systems became the dominant model class used in control theory around 1960, particularly under the influence of the work of Kalman [28]; see also the preface of this book. Of course, state models have been used for a long time in mechanics and in other areas of physics. The central role of latent variables as the result of first principles modeling is emphasized and formalized in [59, 60].

4.10 Exercises

4.1 Verify the state space equations (4.7) and write them in the form (4.16); i.e., specify the matrices A, B, C, D.

4.2 Consider the i/o system described by

$$-4y + 3\frac{d}{dt}y + \frac{d^2}{dt^2}y = u.$$

Find an i/s/o representation for it.

4.3 Consider the electrical circuit shown in Figure 4.4:

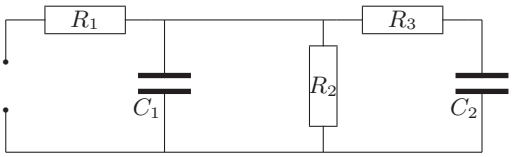


FIGURE 4.4. Electrical circuit.

The input u is the voltage across the external port, and the output y is the voltage across C_2. Take as state-vector

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} V_{C_1} \\ V_{C_2} \end{bmatrix}.$$

Assume that $R_i > 0$, $i = 1, 2, 3$, and $C_i > 0$, , $i = 1, 2$.

(a) Determine the i/s/o equations.

(b) Determine the differential equation describing the i/o behavior.

(c) Repeat the above questions for the case $R_3 = 0$.

4.4 Consider the mechanical system in Figure 4.5. Here u and y denote the horizontal displacements from the respective equilibria. Determine an i/s/o representation.

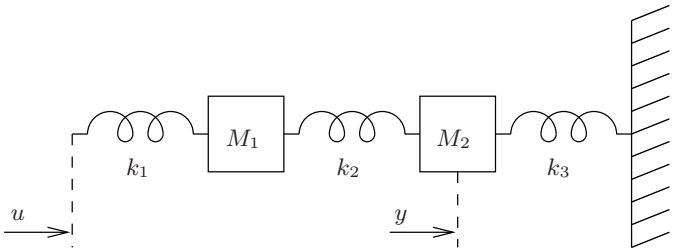


FIGURE 4.5. Mechanical system.

4.5 Determine an i/s/o representation of the electrical circuit in Figure 4.6, with u the voltage across the external port and y the current through C_2.

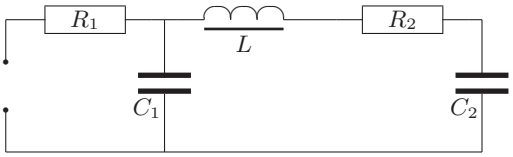


FIGURE 4.6. Electrical circuit.

4.6 Determine an i/s/o representation of the *discrete-time* system $y(t) = u(t-5),\ t \in \mathbb{Z}$. Hint: Use your intuition (the state is the memory) to choose the state.

4.7 Consider the *discrete-time* i/o system defined by $y(t) = u(t)u(t-1)u(t-2),\ t \in \mathbb{Z}$. Is the corresponding dynamical system:

(a) Linear?

(b) Time-invariant?

(c) Determine an i/s/o representation for it.

4.8 Consider the *continuous-time* i/o system defined by $y(t) = u(t-1),\ t \in \mathbb{R}$. Determine a state space model for this system. In other words, construct a latent variable system that satisfies the property of state and that has $y(t) = u(t-1)$ as the specification of the input/output behavior. Note that the state space is *not* be finite-dimensional; hence most of the theory in this chapter does not apply.

4.9 We have seen that the property of state and first-order representations of the form (4.9) are closely related. Consider, however, the autonomous behavior $\mathfrak{B}$ described by $R(\frac{d}{dt})x = 0$, where $R(\xi)$ is given by

$$R(\xi) = \begin{bmatrix} 3+3\xi & 2+5\xi+\xi^2 \\ -5+3\xi^2 & -5-4\xi+4\xi^2+\xi^3 \end{bmatrix}.$$

Prove that this system is a state space representation with x as the state. Hint: Premultiply $R(\xi)$ by a suitable unimodular matrix to obtain a polynomial matrix in which only first-order polynomials appear.

4.10 Let M be a real square matrix. Prove that the infinite series

$$\sum_{k=0}^{\infty} \frac{M^k}{k!}$$

converges absolutely.

4.11 Prove Proposition 4.5.12 by direct application of the definition of the exponential of a matrix, Definition 4.5.1, or by exploiting the fact that e^{At} is the unique solution of the matrix differential equation $\frac{d}{dt}X = AX$, $X(0) = I$ (see Theorem 4.5.17).

4.12 Compute $e^{A_i t}$ for

$$A_1 = \begin{bmatrix} \lambda & \omega \\ -\omega & \lambda \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & \omega & \epsilon & 0 \\ -\omega & 0 & 0 & \epsilon \\ 0 & 0 & 0 & \omega \\ 0 & 0 & -\omega & 0 \end{bmatrix}, \quad \epsilon = 0, 1.$$

4.13 Let $M \in \mathbb{R}^{n\times n}$. The *trace* of M, denoted by $\operatorname{Tr} M$, is the sum of its diagonal elements. Prove the following subtle property of matrix exponentials:

$$\det e^M = e^{\operatorname{Tr} M}.$$

Hint: Transform M into Jordan form and use the fact that for every nonsingular matrix S, it holds that $\operatorname{Tr} S^{-1}MS = \operatorname{Tr} M$ and $\det S^{-1}MS = \det M$.

4.14 Let $A \in \mathbb{R}^{n\times n}$. Prove that all entries of e^{At} are nonnegative for $t \geq 0$ if and only if all nondiagonal elements of A are nonnegative.

4.15 Let $A \in \mathbb{R}^{n\times n}$ be given by

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \\ \vdots & & & \ddots & \\ 0 & & & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}.$$

(a) Show that the characteristic polynomial of A equals $a_0 + a_1\xi + \cdots + a_{n-1}\xi^{n-1} + \xi^n$.

(b) Let λ be an eigenvalue of A, i.e., a root of its characteristic polynomial. Show that a corresponding eigenvector is given by

$$v_\lambda = \begin{bmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{n-1} \end{bmatrix}^T.$$

(c) Prove that for each eigenvalue λ the dimension of $\ker \lambda I - A$ is equal to one.

(d) Let A be given by

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix}.$$

Determine e^{At}.

4.16 (a) Let $x_1, \ldots, x_n : \mathbb{R} \to \mathbb{R}^n$ be strong solutions of the differential equation $\frac{d}{dt}x = Ax$. Prove that the functions $x_1, \ldots, x_n$ are linearly independent if and only if the vectors $x_1(0), \ldots, x_n(0)$ are linearly independent.

(b) Show by means of an example that the previous equivalence is no longer true for arbitrary functions $x_1, \ldots, x_n$.

4.17 Take

$$A = \begin{bmatrix} 3 & -2 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

(a) Determine e^{At}.

(b) Determine all solutions of $\frac{d}{dt}x = Ax + bu$, where u is given by $u(t) = e^{-t}$.

4.18 In Section 4.5.4 we discussed three methods to determine the exponential of a matrix. In this exercise we want to find the relation between the method based on the Jordan form and the one that uses the theory of autonomous behaviors. Let $A \in \mathbb{R}^{n\times n}$, and assume that $\lambda \in \mathbb{C}$ is an eigenvalue of A with algebraic multiplicity two and geometric multiplicity one. This means that to λ there corresponds an eigenvector $v \in \mathbb{C}^n$ and a generalized eigenvector $w \in \mathbb{C}^n$; i.e., $Av = \lambda v$ and $Aw = \lambda w + v$. From (4.41) we know that λ gives rise to solutions of $\frac{d}{dt}x = Ax$ of the form

$$x(t) = B_{10}e^{\lambda t} + B_{11}te^{\lambda t}.$$

The vectors B_{10} and B_{11} should satisfy (4.43).

(a) Show that if an eigenvector v and corresponding generalized eigenvector w are given, then B_{10} and B_{11} may be expressed as $B_{10} = \alpha v + \beta w, B_{11} = \beta v, \alpha, \beta \in \mathbb{C}$.

(b) Show that if B_{10} and B_{11} are given such that (4.43) is satisfied, then an eigenvector v and corresponding generalized eigenvector w may be expressed as $v = \gamma B_{11}, w = \gamma B_{10} + \delta B_{11}, \gamma, \delta \in \mathbb{C}$.

(c) Generalize the above formulas for the case that λ is an eigenvalue of algebraic multiplicity k and geometric multiplicity one.

4.19 Let $A \in \mathbb{R}^{n\times n}$. Prove that $(I\xi - A)^{-1}$ is a strictly proper matrix of rational functions, i.e., in each entry the degree of denominator exceeds the degree of the numerator by at least one.

4.20 Consider the behavior $\mathfrak{B}$ defined by

$$\frac{d}{dt}w = \frac{d}{dt}u.$$

(a) Show that u is free in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$.

(b) Prove that $\mathfrak{B}$ satisfies the property of determinism.

(c) Prove that w does not anticipate u.

(d) Prove that w does *not* have the property of state.

(e) Relate the above results to Theorem 4.5.9. Which of the conditions are not satisfied?

Conclusion: The property of determinism is weaker than the property of state.

4.21 Consider the set of differential equations

$$\frac{d}{dt}x = \begin{bmatrix} 0 & \omega \\ -\omega & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad y = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$$

Prove that the corresponding input/output behavior is described by $\omega^2 y + \frac{d^2}{dt^2}y = \omega u$. Hint: Use (4.31).

4.22 Consider i/s/o representations of the form $\frac{d}{dt}x = Ax + Bu, \quad y = Cx$ with

(a) $A = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$

(b) $A = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 \end{bmatrix}.$

(c) $A = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}.$

Prove that these three i/s/o representations define the same behavior. Prove that the first two are similar. Are the first and the third system also similar?

4.23 (a) Derive the linearized equations of (4.55) about $(z, u, \theta) = (0, 0, 0)$.

(b) Determine an input/state/output representation of the linearized equations with $x = \mathrm{col}(z, \frac{d}{dt}z, \theta, \frac{d}{dt}\theta)$. Does your answer coincide with (4.62)?

4.24 Consider the system of Example 4.7.1. Of course, this system always has two equilibria: one corresponding to the upright position of the rod and one corresponding to the downward position of the rod. For which values of k_1 and k_2 does the system 4.7.1 have more than two equilibria for $u = 0$? Determine these equilibria and give a physical interpretation.

5

Controllability and Observability

5.1 Introduction

In this chapter we introduce two concepts that play a central role in systems theory. The first concept is *controllability*; the second is *observability*. Loosely speaking, we call a behavior controllable if it is possible to switch from one trajectory to the other within the behavior. The advantage is that in a controllable behavior, one can, in principle, always move from an "undesirable" trajectory to a "desirable" one. Observability, on the other hand, is not a property of the behavior as such; rather it is a property related to the partition of the trajectories w into two components w_1 and w_2. We call w_2 observable from w_1 if w_1 and the laws that governing the system dynamics uniquely determine w_2. Thus observability implies that all the information about w is already contained in w_1.

In Section 5.2 we give the definition of controllability of behaviors defined by equations of the form $R(\frac{d}{dt})w = 0$. Subsequently, we study the controllability of i/s/o systems. Next, in Section 5.3, we introduce the concept of observability on a general level and, finally, specialize to the problem when the state of an i/s/o system is observable from the input/output trajectories. By combining the concepts of controllability and observability applied to i/s/o systems, we arrive at the Kalman decomposition. This is the subject of Section 5.4.

5.2 Controllability

In Section 3.2 we have studied autonomous systems. We have seen there that in autonomous systems the past of a trajectory completely determines its future. Once the system follows a particular trajectory, it stays on that trajectory forever. In this section we introduce and study behaviors that are to some extent the opposite of autonomous systems, namely *controllable* systems, in which we can always switch between between any two trajectories.

Example 5.2.1 Consider two pendula mounted on a cart; see Figure 5.1. Suppose for the time being that the masses m_1 and m_2 of the pendula

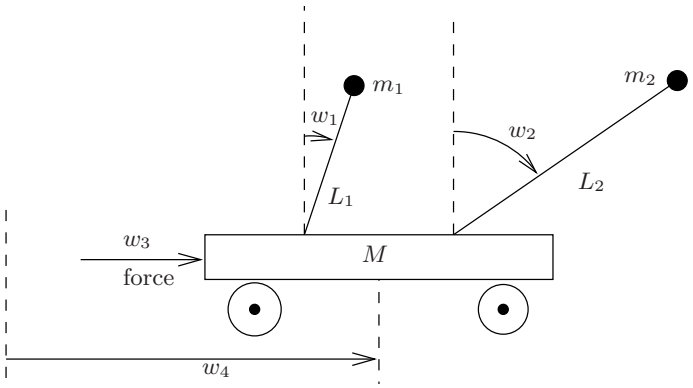


FIGURE 5.1. Two pendula mounted on a cart.

are identical. Later, in Example 5.2.12, it turns out that the values of the masses do not play any role in the present discussion. Suppose in addition that their lengths L_1 and L_2 are equal. Physical intuition indicates that the relative position of the rods is then completely determined by the initial relative position and the initial relative angular velocity, and is independent of the external force. Stated otherwise, $w_1 - w_2$ does not depend on the force w_3. That means that if the rods are in phase during a certain time interval, they remain in phase, no matter what force acts on the cart. This indicates lack of controllability. It is less intuitive what happens if $L_1 \neq L_2$. It turns out that in that case it is always possible to switch from one possible trajectory (w_1', w_2', w_3', w_4') to any other possible trajectory $(w_1'', w_2'', w_3'', w_4'')$ after a small time-delay. More precisely: If (w_1', w_2', w_3', w_4') and $(w_1'', w_2'', w_3'', w_4'')$ are possible trajectories, then there exists a third possible trajectory (w_1, w_2, w_3, w_4) and $t_1 > 0$ with the property

$$w_i(t) = \begin{cases} w_i'(t) & t \leq 0, \\ w_i''(t - t_1) & t \geq t_1, \end{cases} \qquad i = 1, 2, 3, 4. \tag{5.1}$$

Equation (5.1) implies that the system can behave according to trajectory w' until $t = 0$ and proceed according to the delayed w'' after time t_1. The time interval $(0, t_1)$ is needed for the transfer, through judicious choice of the force w_3, from w' to w''.

This example illustrates the concept of controllability. If the rods have unequal lengths, then the system is controllable. If their lengths are equal, then the system is not controllable. This will be proven rigorously in Example 5.2.12. □

We now give the formal definition of controllability.

Definition 5.2.2 Let $\mathfrak{B}$ be the behavior of a time-invariant dynamical system. This system is called *controllable* if for any two trajectories $w_1, w_2 \in \mathfrak{B}$ there exist a $t_1 \geq 0$ and a trajectory $w \in \mathfrak{B}$ with the property

$$w(t) = \begin{cases} w_1(t) & t \leq 0, \\ w_2(t - t_1) & t \geq t_1. \end{cases}$$

□

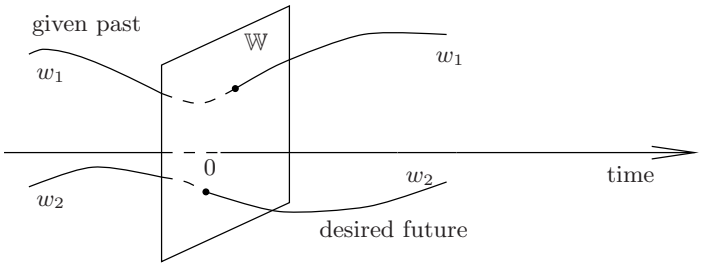
Controllability thus implies that strictly obeying the laws governing the system, i.e., within the behavior, we can switch from one trajectory to the other, provided that we allow a delay. This is in contrast to autonomous systems, where we cannot get off a trajectory once we are on it. Figure 5.2 gives a visual explanation the notion of controllability.

Controllability is a desirable property, since in principle it enables one to steer the system to a desired trajectory.

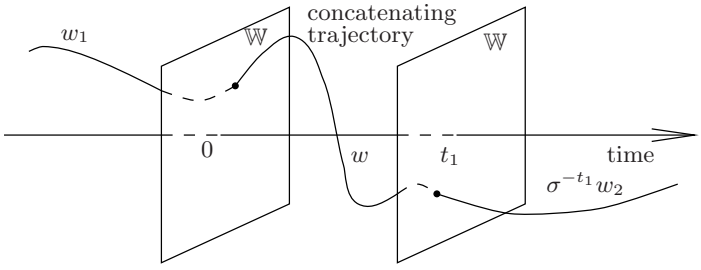
Example 5.2.3 Trivial examples of controllable systems are:

- $\mathfrak{B} := \{w : \mathbb{R} \to \mathbb{R}^q \mid w = 0\}$, corresponding to the behavioral equation $Iw = 0$.
- $\mathfrak{B} := \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)\}$, corresponding to the behavioral equation $Ow = 0$, where O denotes the zero-matrix.

As already remarked, autonomous systems are *not* controllable, with the exception of trivial systems like $\mathfrak{B} = \{0\}$. Consider, for instance, the behavior $\mathfrak{B}$ defined by $(\frac{d}{dt} - 1)w = 0$. It is given by $\{w \mid w(t) = ce^t,\ c \in \mathbb{R}\}$. To see that this system is not controllable, take two trajectories $w_1, w_2 \in \mathfrak{B}$, say $w_i(t) = c_i e^t$, $i = 1, 2$. Suppose that we could switch from w_1 to w_2. Then there would exist a trajectory $w \in \mathfrak{B}$, a constant $c \in \mathbb{R}$, and $t_1 \geq 0$ such that $w(t) = ce^t = c_1 e^t$ for $t \leq 0$ and $w(t) = c_2 e^{t - t_1}$ for $t \geq t_1$. The equality for $t \leq 0$ implies that $c = c_1$, and the equality for $t \geq t_1$ implies that $c_1 = c_2 e^{-t_1}$. If $c_1 > c_2$, then it follows that $t_1 < 0$, which is a contradiction. This shows that the system is not controllable. □



(a)



(b)

FIGURE 5.2. Controllability.

Definition 5.2.2 would not be very useful without a simple test to decide whether or not a system is controllable. We first present such a test for autonomous systems.

Lemma 5.2.4 *Let $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ with $\det(P(\xi)) \neq 0$. Then the system defined by $P(\frac{d}{dt})w = 0$ is controllable if and only if $P(\xi)$ is unimodular, i.e., if and only if $\mathfrak{B} = \{0\}$.*

Proof If $P(\xi)$ is unimodular, then $\mathfrak{B} = \{0\}$, and hence $\mathfrak{B}$ is trivially controllable. On the other hand, if $P(\xi)$ is not unimodular, then it follows from Remark 3.2.14 that the past of each trajectory uniquely determines its future. It follows from Theorem 3.2.16 that for nonunimodular $P(\xi)$, $\mathfrak{B}$ contains more than one trajectory. It follows that $\mathfrak{B}$ cannot be controllable. □

The following theorem gives a test for controllability of the $\mathcal{C}^\infty$ part of the behavior: $\mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. It is preparatory for a test for the full behavior.

Theorem 5.2.5 *Consider the system defined by $R(\frac{d}{dt})w = 0$, and denote by $\mathfrak{B}^\infty$ the $\mathcal{C}^\infty$ part of its behavior $\mathfrak{B}$; i.e., $\mathfrak{B}^\infty = \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. Then $\mathfrak{B}^\infty$ is controllable if and only if the rank of the (complex) matrix $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}$.*

Proof Choose unimodular matrices $U(\xi), V(\xi)$ such that $U(\xi)R(\xi)V(\xi) = \tilde{R}(\xi) = [D(\xi)\ 0]$ is in Smith form. Define the transformed behavior as $\tilde{\mathfrak{B}}^\infty := V^{-1}(\frac{d}{dt})\mathfrak{B}^\infty$. Let

$$D(\xi) = \begin{bmatrix} D_1(\xi) & 0 \\ 0 & 0 \end{bmatrix},$$

where $\det(D_1(\xi)) \neq 0$. Then $\tilde{\mathfrak{B}}^\infty = \{(\tilde{w}_1, \tilde{w}_2) \mid D_1(\frac{d}{dt})\tilde{w}_1 = 0\}$, where, of course, the partition of $\tilde{w}$ is made in accordance with the partitions of $\tilde{R}(\xi)$ and $D(\xi)$. Since $\tilde{w}_2$ is completely free, see Exercise 5.22, it follows that $\tilde{\mathfrak{B}}^\infty$ is controllable if and only if the behavior $\{\tilde{w}_1 \mid D_1(\frac{d}{dt})\tilde{w}_1 = 0\}$ is controllable. By Lemma 5.2.4 this is the case if and only if the square polynomial matrix $D_1(\xi)$ is unimodular. This in turn is equivalent to the condition that $\operatorname{rank} \tilde{R}(\lambda)$ is constant for $\lambda \in \mathbb{C}$.

Notice that $\operatorname{rank} R(\lambda) = \operatorname{rank} U^{-1}(\lambda)\tilde{R}(\lambda)V^{-1}(\lambda) = \operatorname{rank} \tilde{R}(\lambda)$, since $U(\xi)$ and $V(\xi)$ are unimodular. Hence $\tilde{\mathfrak{B}}^\infty$ is controllable if and only if $\operatorname{rank} R(\lambda)$ does not depend on $\lambda \in \mathbb{C}$.

The last step is to prove that $\mathfrak{B}^\infty$ is controllable if and only if $\tilde{\mathfrak{B}}^\infty$ is controllable. To that end assume that $\mathfrak{B}^\infty$ is controllable and choose $\tilde{w}', \tilde{w}'' \in \tilde{\mathfrak{B}}^\infty$. Define $w', w'' \in \mathfrak{B}^\infty$ by $w' := V(\frac{d}{dt})\tilde{w}'$ and $w'' := V(\frac{d}{dt})w''$. Since $\mathfrak{B}^\infty$ is controllable, there exist $t_1 \geq 0$ and $w \in \mathfrak{B}^\infty$ such that

$$w(t) = \begin{cases} w'(t) & t \leq 0, \\ w''(t - t_1) & t \geq t_1. \end{cases} \tag{5.2}$$

Define $\tilde{w} \in \tilde{\mathfrak{B}}^\infty$ by $\tilde{w} := V^{-1}(\frac{d}{dt})\tilde{w}$. Then it follows from (5.2) that

$$\tilde{w}(t) = \begin{cases} \tilde{w}'(t) & t \leq 0, \\ \tilde{w}''(t - t_1) & t \geq t_1. \end{cases}$$

This shows that $\tilde{\mathfrak{B}}^\infty$ is controllable. In the same way, the converse statement follows: if $\tilde{\mathfrak{B}}^\infty$ is controllable then $\mathfrak{B}^\infty$ is also controllable.

This concludes the proof. □

Corollary 5.2.6 *For $\mathfrak{B}^\infty$, the time t_1 required in Definition 5.2.2 is independent of w_1 and w_2 and can be taken to be arbitrarily small.*

Proof Let $\tilde{\mathfrak{B}}^\infty$ be as in the proof of Theorem 5.2.5. Choose $t_1 > 0$ arbitrarily. Define $\tilde{w}_i \in \tilde{\mathfrak{B}}^\infty$ by $\tilde{w}_i := V^{-1}(\frac{d}{dt})w_i$, $i = 1, 2$. Since two $\mathcal{C}^\infty$ functions can always be interpolated in a smooth way, see Exercise 5.22, there exists $\tilde{w} \in \tilde{\mathfrak{B}}$ such that

$$\tilde{w}(t) = \begin{cases} \tilde{w}_1(t) & t \leq 0, \\ \tilde{w}_2(t - t_1) & t \geq t_1. \end{cases} \tag{5.3}$$

Define w as $w := V(\frac{d}{dt})\tilde{w}$. Then $w \in \mathfrak{B}$, and by (5.3),

$$w(t) = \begin{cases} w_1(t) & t \leq 0, \\ w_2(t - t_1) & t \geq t_1. \end{cases}$$

□

Remark 5.2.7 If the rank of $R(\lambda)$ is not the same for all $\lambda \in \mathbb{C}$, then we call $\lambda \in \mathbb{C}$ for which the rank $R(\xi)$ drops a *singular value* of $R(\xi)$. Recall from Chapter 3 that in the case of a square polynomial matrix with nonzero determinant, these values were called characteristic values. □

We now extend the rank test of Theorem 5.2.5 by showing that $\mathfrak{B}$ is controllable if and only if $\mathfrak{B}^\infty$ is controllable. But first we state a lemma about polynomial matrices that do not have constant rank for all complex numbers.

Lemma 5.2.8 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Then there exist polynomial matrices $F(\xi) \in \mathbb{R}^{g\times g}[\xi]$ and $\tilde{R}(\xi) \in \mathbb{R}^{g\times q}[\xi]$, such that $R(\xi) = F(\xi)\tilde{R}(\xi)$ and* rank $\tilde{R}(\lambda)$ *is the same for all $\lambda \in \mathbb{C}$.*

Proof Choose unimodular matrices $U(\xi), V(\xi)$ such that $U(\xi)R(\xi)V(\xi) = [D(\xi)\ 0] = [\mathrm{diag}(d_1(\xi), \cdots, d_k(\xi), 0 \cdots, 0) \quad 0]$ is in Smith form with $d_j(\xi) \neq 0, j = 1, \ldots, k$. Then

$$R(\xi) = \underbrace{U^{-1}(\xi)D(\xi)}_{F(\xi)} \underbrace{\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} V^{-1}(\xi)}_{\tilde{R}(\xi)}.$$

Since $V(\xi)$ is unimodular, it is obvious that rank $\tilde{R}(\lambda)$ is the same for all $\lambda \in \mathbb{C}$. □

Theorem 5.2.9 *Let the behavior $\mathfrak{B}$ be defined by $R(\frac{d}{dt})w = 0$; i.e., $\mathfrak{B}$ is the set of weak solutions in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ of $R(\frac{d}{dt})w = 0$. Then $\mathfrak{B}^\infty := \mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ is controllable if and only if $\mathfrak{B}$ is controllable.*

Proof Suppose that $\mathfrak{B}^\infty$ is controllable. By Theorem 2.5.23 we may assume that $R(\xi)$ has full row rank. According to Corollary 3.3.23, $\mathfrak{B}$ admits an i/o representation

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})u \tag{5.4}$$

with $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$ and $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$. Let the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$ be given by

$$P^{-1}(\xi)Q(\xi) = A_0 + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{A_{ij}}{(\xi - \lambda_i)^j},$$

Define

$$H(t) := \sum_{i=1}^{N} \sum_{j=1}^{n_i} A_{ij} \frac{t^{j-1}}{(j-1)!} e^{\lambda_i t}.$$

Then by Theorem 3.3.19, every (weak) solution of (5.4) can be written as

$$y(t) := A_0 u(t) + y_{\hom}(t) + \int_0^t H(t-\tau)u(\tau)d\tau, \tag{5.5}$$

where $y_{\hom}$ satisfies $P(\frac{d}{dt})y_{\hom} = 0$.

Let $y_{\hom}$ be any solution of $P(\frac{d}{dt})y_{\hom} = 0$. By Theorem 3.2.15 we may assume that $y_{\hom} \in \mathfrak{B}^\infty$. Since $\mathfrak{B}^\infty$ is controllable and in view of Corollary 5.2.6, there exists $(u, y) \in \mathfrak{B}^\infty$ such that

$$(u(t), y(t)) = \begin{cases} (0,0) & t \le 0, \\ (0, y_{\hom}(t-1)) & t \ge 1. \end{cases} \tag{5.6}$$

From (5.5) and (5.6) it follows that for $t \le 0$,

$$0 = y(t) = y_{\hom}(t),$$

so that for $t \ge 1$,

$$\int_0^1 H(t-\tau)u(\tau)d\tau = y_{\hom}(t-1). \tag{5.7}$$

Now choose $(u_1, y_1), (u_2, y_2) \in \mathfrak{B}$ arbitrarily, say

$$y_j(t) = A_0 u_j(t) + y_{\hom,j}(t) + \int_0^t H(t-\tau)u_j(\tau)d\tau, \quad j = 1, 2.$$

From (5.7) it follows that there exists a $u_{12} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ such that for $t \ge 1$,

$$\int_0^1 H(t-\tau)u_{12}(\tau)d\tau = y_{\hom,2}(t-1) - y_{\hom,1}(t).$$

Take $\tilde{u}$ as follows:

$$\tilde{u}(t) = \begin{cases} u_1(t) & t \le 0, \\ u_{12}(t) & 0 < t < 1, \\ u_2(t-1) & t \ge 1 \end{cases}$$

and $\tilde{y}$ as

$$\tilde{y}(t) := A_0 \tilde{u}(t) + y_{\hom,1}(t) + \int_0^t H(t-\tau)\tilde{u}(\tau)d\tau.$$

Then $(\tilde{u}, \tilde{y}) \in \mathfrak{B}$. Moreover, for $t \leq 0$, we have $\tilde{y}(t) = y_1(t)$, and for $t \geq 1$,

$$\begin{aligned}
\tilde{y}(t) &= A_0\tilde{u}(t) + y_{\mathrm{hom},1}(t) + \int_0^t H(t-\tau)\tilde{u}(\tau)d\tau \\
&= A_0u_2(t-1) + y_{\mathrm{hom},1}(t) + \int_0^1 H(t-\tau)u_{12}(\tau)d\tau \\
&\quad + \int_1^t H(t-\tau)u_2(\tau-1)d\tau \\
&= A_0u_2(t-1) + y_{\mathrm{hom},1}(t) + y_{\mathrm{hom},2}(t-1) - y_{\mathrm{hom},1}(t) \\
&\quad + \int_1^t H(t-\tau)u_2(\tau-1)d\tau \\
&= A_0u_2(t-1) + y_{\mathrm{hom},2}(t-1) + \int_1^t H(t-\tau)u_2(\tau-1)d\tau \\
&= y_2(t-1).
\end{aligned}$$

This shows that controllability of $\mathfrak{B}^\infty$ implies that of $\mathfrak{B}$.

Conversely, suppose that $\mathfrak{B}^\infty$ is not controllable. Then it follows from Theorem 5.2.5 that $\operatorname{rank} R(\lambda)$ is not constant over $\mathbb{C}$. Consider the i/o form (5.4). By Lemma 5.2.8 there exist matrices $F(\xi), \tilde{P}(\xi), \tilde{Q}(\xi)$ such that

$$F(\xi)[\tilde{P}(\xi) \quad \tilde{Q}(\xi)] = [P(\xi) \quad Q(\xi)],$$

with $\operatorname{rank}[\tilde{P}(\lambda) \quad \tilde{Q}(\lambda)] = g$ for all $\lambda \in \mathbb{C}$. As a consequence, there exists at least one $\bar{\lambda} \in \mathbb{C}$ for which $\operatorname{rank} F(\bar{\lambda}) < g$. Since $\tilde{P}^{-1}(\xi)\tilde{Q}(\xi) = P^{-1}(\xi)Q(\xi)$, their respective partial fraction expansions coincide, so that the corresponding initially-at-rest behaviors, see Theorem 3.5.2, are the same. Since $\operatorname{rank} F(\bar{\lambda}) < g$ and $\operatorname{rank}[\tilde{P}(\bar{\lambda}) \quad \tilde{Q}(\bar{\lambda})] = g$, there exists a nonzero vector $v \in \mathbb{C}^p \times \mathbb{C}^m$ such that

$$[P(\bar{\lambda}) \quad -Q(\bar{\lambda})]v = 0 \quad \text{and} \quad [\tilde{P}(\bar{\lambda}) \quad -\tilde{Q}(\bar{\lambda})]v \neq 0. \tag{5.8}$$

Define the trajectory $(\bar{u}, \bar{y})$ as

$$(\bar{u}(t), \bar{y}(t)) := ve^{\bar{\lambda}t}.$$

Then by (5.8)

$$P(\frac{d}{dt})\bar{y} = Q(\frac{d}{dt})\bar{u} \quad \text{and} \quad \tilde{P}(\frac{d}{dt})\bar{y} \neq \tilde{Q}(\frac{d}{dt})\bar{u}. \tag{5.9}$$

In other words, the pair $(\bar{u}, \bar{y})$ belongs to the behavior of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$, but not to that of $\tilde{P}(\frac{d}{dt})y = \tilde{Q}(\frac{d}{dt})u$. Suppose that there exists $(u_0, y_0) \in \mathfrak{B}$ such that

$$(u_0(t), y_0(t)) = \begin{cases} (0,0) & t \leq 0, \\ (\bar{u}(t-1), \bar{y}(t-1)) & t \geq 1. \end{cases}$$

Since the initial at rest behavior of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$ equals that of $\tilde{P}(\frac{d}{dt}) = \tilde{Q}(\frac{d}{dt})u$, it follows that $\tilde{P}(\frac{d}{dt})y_0 = \tilde{Q}(\frac{d}{dt})u_0$. In particular, by time-invariance,

$$\tilde{P}(\frac{d}{dt})\bar{y} = \tilde{Q}(\frac{d}{dt})\bar{u}.$$

This contradicts the inequality in (5.9). The conclusion is that we cannot steer the system from the zero trajectory to every other trajectory within the behavior. Hence $\mathfrak{B}$ is not controllable. □

Combining Theorems 5.2.5 and 5.2.9 immediately yields the following central result:

Theorem 5.2.10 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The behavior $\mathfrak{B}$ defined by $R(\frac{d}{dt})w = 0$ is controllable if and only if the rank of the (complex) matrix $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}$.*

Corollary 5.2.11 *Let $p(\xi), q(\xi) \in \mathbb{R}[\xi]$ with $\deg p(\xi) \geq \deg q(\xi)$. The SISO system defined by $p(\frac{d}{dt})y = q(\frac{d}{dt})u$ is controllable if and only if the polynomials $p(\xi)$ and $q(\xi)$ have no common factor.*

Proof This is a direct consequence of Theorem 5.2.10. To see this, define $w = \mathrm{col}(u, y)$ and $R(\xi) = [q(\xi) \quad -p(\xi)]$. The i/o behavioral equation is now rewritten in the general form so that we can apply Theorem 5.2.10. It is clear that $R(\lambda)$ has constant rank (one) over $\mathbb{C}$ if and only if there does not exist a $\lambda \in \mathbb{C}$ such that $p(\lambda) = q(\lambda) = 0$, equivalently, if and only if $p(\xi)$ and $q(\xi)$ do not have common factors. □

To illustrate the obtained results on controllability, we apply them to the two pendula mounted on a cart of Example 5.2.1 and to the RLC-network of Example 1.3.5.

Example 5.2.12 Consider the two pendula mounted on a cart depicted in Figure 5.1. In Example 5.2.1 we claimed that this system is controllable if and only if the lengths of the rods are not identical. We want to prove this claim by considering the linearized mathematical model and using Theorem 5.2.10. Denote by w_4 the distance of the center of gravity of the cart with respect to some reference point. The (nonlinear) equations describing the relations among w_1, w_2, w_3, w_4 can be derived from the laws of mechanics and are given by

$$\begin{aligned}(M + m_1 + m_2)(\tfrac{d}{dt})^2 w_4 &= w_3 + m_1 L_1[(\sin w_1)(\tfrac{d}{dt}w_1)^2 - (\cos w_1)(\tfrac{d}{dt})^2 w_1] \\ &\quad + m_2 L_2[(\sin w_2)(\tfrac{d}{dt}w_2)^2 - (\cos w_2)(\tfrac{d}{dt})^2 w_2], \\ m_i L_i^2(\tfrac{d}{dt})^2 w_i - m_i g L_i \sin w_i + m_i L_i \cos w_i(\tfrac{d}{dt})^2 w_4 &= 0, \qquad i = 1, 2.\end{aligned} \tag{5.10}$$

For small values of the w_is the equations (5.10) can be approximated by their linearizations around the equilibrium $(0,0,0,0)$. This yields the following linear equations:

$$\begin{aligned}(M+m_1+m_2)\frac{d^2}{dt^2}w_4 &= w_3 - \frac{d^2}{dt^2}(m_1L_1w_1 + m_2L_2w_2)\\ -m_iL_igw_i + m_iL_i^2\frac{d^2}{dt^2}w_i + m_iL_i\frac{d^2}{dt^2}w_4 &= 0, \qquad i=1,2\end{aligned} \tag{5.11}$$

Let us analyze the the controllability of (5.11). First we rewrite (5.11) in the standard notation $R(\frac{d}{dt})w = 0$. Define $w = \mathrm{col}(w_1, w_2, w_3, w_4)$. The corresponding matrix $R(\xi)$ is now given by

$$R(\xi) = \begin{bmatrix} m_1L_1\xi^2 & m_2L_2\xi^2 & -1 & (M+m_1+m_2)\xi^2 \\ m_1L_1^2\xi^2 - m_1L_1g & 0 & 0 & m_1L_1\xi^2 \\ 0 & m_2L_2^2\xi^2 - m_2L_2g & 0 & m_2L_2\xi^2 \end{bmatrix}.$$

According to Theorem 5.2.10, the system is controllable if and only if the rank of the complex matrix $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}$. If $L_1 \neq L_2$, then obviously $\operatorname{rank} R(\lambda) = 3$ for all $\lambda \in \mathbb{C}$. Therefore, the system is controllable in that case. If $L_1 = L_2$, then $\operatorname{rank} R(\lambda)$ equals two for $\lambda = \pm\sqrt{\frac{g}{L}}$ and is three otherwise. This shows that the system is not controllable if the rods are of equal lengths, just as claimed on intuitive grounds in Example 5.2.1. Apparently, for controllability the masses m_1, m_2 are immaterial.

We have not been completely fair, since we tacitly identified controllability of the nonlinear system with that of its linear approximation. For this particular example it can be proved that this is justified. However, the proof is well beyond the scope of this book. □

Example 5.2.13 Consider the RLC-network of Example 1.3.5. Let us check whether the system describing the port behavior is controllable. Recall that for the case $CR_C \neq \frac{L}{R_L}$ the relation between V and I is given by

$$\left(\frac{R_C}{R_L} + (1+\frac{R_C}{R_L})CR_C\frac{d}{dt} + CR_C\frac{L}{R_L}\frac{d^2}{dt^2}\right)V = (1+CR_C\frac{d}{dt})(1+\frac{L}{R_L}\frac{d}{dt})R_CI, \tag{5.12}$$

while for the case $CR_C = \frac{L}{R_L}$ the relation is given by (see (1.12,1.13))

$$(\frac{R_C}{R_L} + CR_C\frac{d}{dt})V = (1+CR_C\frac{d}{dt})R_CI. \tag{5.13}$$

Corollary 5.2.11 reduces controllability to a common factor condition. First consider (5.12). Since the roots of the right-hand side are apparent, we should check whether $\lambda = -\frac{1}{CR_C}$ or $\lambda = -\frac{R_L}{L}$ can be roots of $\frac{R_C}{R_L} +$

$(1 + \frac{R_C}{R_L})CR_C\xi + CR_C\frac{L}{R_L}\xi^2$. It is easy to see that this is not possible, see Exercise 5.1, and hence the port behavior is controllable in the case $CR_C \neq \frac{L}{R_L}$. If $CR_C = \frac{L}{R_L}$, then the port behavior is described by (5.13). Obviously, $\frac{R_C}{R_L} + CR_C\xi$ and $1 + CR_C\xi$ have a common factor if and only if $R_C = R_L$. We conclude that the port behavior of the RLC-network is controllable unless $CR_C = \frac{L}{R_L}$ and $R_C = R_L$. □

We now prove that every behavior contains a controllable part and an autonomous part. In fact, as we show next, every behavior defined by $R(\frac{d}{dt})w = 0$ can be written as a *direct sum* of a controllable and an autonomous subbehavior.

Theorem 5.2.14 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ be of full row rank and let $\mathfrak{B}$ be the behavior defined by*

$$R(\frac{d}{dt})w = 0.$$

Then there exist subbehaviors $\mathfrak{B}_{\rm aut}$ and $\mathfrak{B}_{\rm contr}$ of $\mathfrak{B}$ such that

$$\mathfrak{B} = \mathfrak{B}_{\rm aut} \oplus \mathfrak{B}_{\rm contr},$$

where $\mathfrak{B}_{\rm contr}$ is controllable and $\mathfrak{B}_{\rm aut}$ is autonomous, and the characteristic values of $\mathfrak{B}_{\rm aut}$ are exactly those numbers $\lambda \in \mathbb{C}$ for which $\operatorname{rank} R(\lambda) < g$.

Proof As in the proof of Theorem 5.2.5, choose unimodular matrices $U(\xi), V(\xi)$ that transform $R(\xi)$ into Smith form:

$$\tilde{R}(\xi) := U(\xi)R(\xi)V(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix} \tag{5.14}$$

with $\det(D(\xi)) \neq 0$. Define the transformed behavior as $\tilde{\mathfrak{B}}^\infty := V^{-1}(\frac{d}{dt})\mathfrak{B}^\infty$. Then $\tilde{\mathfrak{B}}^\infty = \{(\tilde{w}_1, \tilde{w}_2) \mid D(\frac{d}{dt})\tilde{w}_1 = 0\}$, where, of course, the partition of $\tilde{w}$ is made in accordance with the partition of $\tilde{R}(\xi)$. If $\det D(\xi)$ is a constant, then $\mathfrak{B}^\infty$ is controllable, and we may take $\mathfrak{B}_{\rm contr} = \mathfrak{B}$ and $\mathfrak{B}_{\rm aut} = \{0\}$. Suppose $\deg\det D(\xi) \geq 1$. Then the behavior $\tilde{\mathfrak{B}}^\infty$ can easily be written as the direct sum of a controllable part and an autonomous part:

$$\tilde{\mathfrak{B}}_{\rm aut} := \{\tilde{w} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid D(\frac{d}{dt})\tilde{w}_1 = 0, \quad \tilde{w}_2 = 0\},$$

$$\tilde{\mathfrak{B}}_{\rm contr} := \{\tilde{w} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid \tilde{w}_1 = 0\}.$$

Then $\tilde{\mathfrak{B}}^\infty = \tilde{\mathfrak{B}}_{\rm aut} \oplus \tilde{\mathfrak{B}}_{\rm contr}$. Note that $\tilde{\mathfrak{B}}_{\rm aut}$ and $\tilde{\mathfrak{B}}_{\rm contr}$ are described by

$$\tilde{R}_{\rm aut}(\xi) = \begin{bmatrix} D(\xi) & 0 \\ 0 & I \end{bmatrix} \quad \text{and} \quad \tilde{R}_{\rm contr}(\xi) = \begin{bmatrix} I & 0 \end{bmatrix}$$

respectively. By transforming back we obtain the desired partition of $\mathfrak{B}^\infty$:

$$\mathfrak{B}^\infty_{\rm aut} := V(\frac{d}{dt})\tilde{\mathfrak{B}}_{\rm aut} \qquad \text{and} \qquad \mathfrak{B}^\infty_{\rm contr} := V(\frac{d}{dt})\tilde{\mathfrak{B}}_{\rm contr} \tag{5.15}$$

Of course $\mathfrak{B}^\infty = \mathfrak{B}^\infty_{\rm aut} \oplus \mathfrak{B}^\infty_{\rm contr}$ is a partition in an autonomous and a controllable part of the $\mathcal{C}^\infty$ part of $\mathfrak{B}$. The equations that describe (5.15) are obtained by transforming $\tilde{R}_{\rm aut}(\xi)$ and $\tilde{R}_{\rm contr}(\xi)$:

$$R_{\rm aut}(\xi) = \tilde{R}_{\rm aut}(\xi)V^{-1}(\xi) \qquad \text{and} \qquad R_{\rm contr}(\xi) = \tilde{R}_{\rm contr}(\xi)V^{-1}(\xi) \quad (5.16)$$

and define

$$\mathfrak{B}_{\rm aut} := \{w \in \mathfrak{L}_1^{\rm loc}(\mathbb{R}, \mathbb{R}^q) \mid R_{\rm aut}(\frac{d}{dt})w = 0, \text{ weakly}\}$$

and

$$\mathfrak{B}_{\rm contr} := \{w \in \mathfrak{L}_1^{\rm loc}(\mathbb{R}, \mathbb{R}^q) \mid R_{\rm contr}(\frac{d}{dt})w = 0, \text{ weakly}\} \qquad (5.17)$$

It follows from Theorem 2.4.13 that $\mathfrak{B} = \mathfrak{B}_{\rm aut} \oplus \mathfrak{B}_{\rm contr}$ and from Theorem 5.2.9 that $\mathfrak{B}_{\rm contr}$ is controllable. Finally, since $\det(R_{\rm aut}(\xi)) = \det D(\xi) \neq 0$, $\mathfrak{B}_{\rm aut}$ is indeed autonomous. The characteristic values of $\mathfrak{B}_{\rm aut}$ are the roots of $\det(R_{\rm aut}(\xi))$. Since $\det(R_{\rm aut}(\xi)) = \det D(\xi)$ and by (5.14), it follows that these characteristic values are indeed those complex numbers λ for which $\operatorname{rank} R(\lambda) < g$. □

Remark 5.2.15 The decomposition of a behavior into a direct sum of a controllable and an autonomous part is not unique. This can be concluded by observing that the unimodular matrices $U(\xi)$ and $V(\xi)$ that transform $R(\xi)$ into Smith form are not unique. It can be shown that the controllable part, however, is unique. The details are worked out in Exercise 5.6. □

Example 5.2.16 Consider the i/o system defined by

$$y - \frac{d}{dt}y - \frac{d^2}{dt^2}y + \frac{d^3}{dt^3}y = -2u + \frac{d}{dt}u + \frac{d^2}{dt^2}u. \qquad (5.18)$$

The corresponding polynomial matrix is given by

$$R(\xi) = \begin{bmatrix} 1 - \xi - \xi^2 + \xi^3 & 2 - \xi - \xi^2 \end{bmatrix}.$$

The unimodular matrix $V(\xi)$ that transforms $R(\xi)$ into Smith form is given by

$$V(\xi) = \begin{bmatrix} \frac{1}{3} & 2 + \xi \\ \frac{-2}{3} + \frac{1}{3}\xi & -1 + \xi^2 \end{bmatrix}.$$

Indeed,

$$\begin{bmatrix} 1 - \xi - \xi^2 + \xi^3 & 2 - \xi - \xi^2 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 2 + \xi \\ \frac{-2}{3} + \frac{1}{3}\xi & -1 + \xi^2 \end{bmatrix} = \begin{bmatrix} -1 + \xi & 0 \end{bmatrix}.$$

The inverse of $V(\xi)$ is given by

$$V^{-1}(\xi) = \begin{bmatrix} -1+\xi^2 & -2-\xi \\ \frac{2}{3}-\frac{1}{3}\xi & \frac{1}{3} \end{bmatrix}$$

The polynomials that define the controllable part and the autonomous part are, according to (5.16), given by

$$\begin{aligned} R_{\text{aut}}(\xi) &= \begin{bmatrix} -1+\xi & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1+\xi^2 & -2-\xi \\ \frac{2}{3}-\frac{1}{3}\xi & \frac{1}{3} \end{bmatrix} \\ &= \begin{bmatrix} 1-\xi-\xi^2+\xi^3 & 2-\xi-\xi^2 \\ \frac{2}{3}-\frac{1}{3}\xi & \frac{1}{3} \end{bmatrix} \end{aligned}$$

and

$$R_{\text{contr}}(\xi) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -1+\xi^2 & -2-\xi \\ \frac{2}{3}-\frac{1}{3}\xi & \frac{1}{3} \end{bmatrix} = [-1+\xi^2 \quad -2-\xi].$$

The corresponding differential equations are

$$\begin{aligned} y - \frac{d}{dt}y - \frac{d^2}{dt^2}y + \frac{d^3}{dt^3}y + 2u - \frac{d}{dt}u - \frac{d^2}{dt^2}u = 0, \\ \tfrac{2}{3}y - \tfrac{1}{3}\frac{d}{dt}y + \tfrac{1}{3}u = 0 \end{aligned}$$

for the autonomous part and

$$-y + \frac{d^2}{dt^2}y = 2u + \frac{d}{dt}u \tag{5.19}$$

for the controllable part. Notice that (5.19), the controllable part of (5.18), is obtained by just canceling the common factor $-1+\xi$ in the entries of $R(\xi)$. □

5.2.1 Controllability of input/state/output systems

5.2.1.1 Controllability of i/s systems

In Chapter 4 we have introduced i/s/o models. Part of an i/s/o model is the i/s model defined by the input-state equation $\frac{d}{dt}x = Ax + Bu$. In

this section we study the problem of to what extent the state-trajectory can be controlled by means of a proper choice of the input function. More specifically, we consider the following question: *Given two states x_1 and x_2, does there exist an input function u such that (u, x) satisfies the i/s equation and such that $x(t') = x_1$ and $x(t'') = x_2$ for some t', $t'' \in \mathbb{R}$?* If such an input function exists for all possible choices of x_1 and x_2, then we call the system *state controllable*. It will turn out that for the system given by $\frac{d}{dt}x = Ax + Bu$ that controllability is equivalent to state controllability. Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ and consider the i/s behavior defined by

$$\frac{d}{dt}x = Ax + Bu. \tag{5.20}$$

We have seen in Theorem 5.2.10 that controllability can be checked by means of a rank test of a (potentially) infinite number of matrices, namely $R(\lambda)$, $\lambda \in \mathbb{C}$. For the system (5.20) this rank test turns out to be equivalent to a rank test of a single matrix, the *controllability matrix*, directly defined in terms of the matrices A and B. The *controllability matrix* of the pair (A, B) is defined as follows:

Definition 5.2.17 Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$. Define the matrix $\mathfrak{C} \in \mathbb{R}^{n\times nm}$ as

$$\mathfrak{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}. \tag{5.21}$$

$\mathfrak{C}$ is called the *controllability matrix of the pair* (A, B). □

Theorem 5.2.18 *Consider the system $\frac{d}{dt}x = Ax + Bu$. The system is controllable if and only if its controllability matrix $\mathfrak{C}$ has rank n.*

Theorem 5.2.18 provides a simple and elegant criterion for the controllability of an i/s system. Because of the nature of the criterion, we often speak about the controllability of the *pair of matrices* (A, B) rather than of the *system defined* by (A, B) through (5.20).

For the proof of Theorem 5.2.18 we use a result whose proof uses the concept of invariant subspace.

Definition 5.2.19 (A-invariant subspace) Let $A \in \mathbb{R}^{n\times n}$ and let $\mathcal{V}$ be a linear subspace of $\mathbb{R}^n$. We call $\mathcal{V}$ an *A-invariant subspace* if for all $v \in \mathcal{V}$, $Av \in \mathcal{V}$. Notation: $A\mathcal{V} \subset \mathcal{V}$. □

See Exercise 5.24 for an example of an A-invariant space.

Lemma 5.2.20 *Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ and assume that there exists a nonzero vector $v_0 \in \mathbb{C}^n$ such that $v_0^T A^k B = 0$, $k = 0, \ldots, n-1$. Define the linear subspace $\mathcal{V}$ of $\mathbb{C}^n$ as*

$$\mathcal{V} := \operatorname*{span}_{k\geq 0}\{(A^T)^k v_0\}.$$

Then there exists a nonzero vector $v_1 \in \mathcal{V}$ *and* $\lambda_1 \in \mathbb{C}$ *such that*

$$v_1^T A = \lambda_1 v_1^T \quad \textit{and} \quad v_1^T B = 0.$$

Proof Let $p(\xi) := \det(I\xi - A) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$ be the characteristic polynomial of A. According to the theorem of Cayley–Hamilton, every matrix satisfies its own characteristic equation, and hence $p(A) = 0$. Therefore, we can express A^n in lower powers of A:

$$A^n = -(p_0 I + p_1 A + \cdots + p_{n-2}A^{n-2} + p_{n-1}A^{n-1}). \tag{5.22}$$

By induction on k it follows easily from (5.22) that for every $k \in \mathbb{N}$, A^k can be written as a linear combination of $I, A, \ldots, A^{n-1}$. By expressing A^k in terms of $I, A, \ldots, A^{n-1}$, it follows that $v_0^T A^k B = 0$ for all k, and hence for all $v \in \mathcal{V}$ we have that $v^T B = 0$. Note that $v \in \mathcal{V}$ implies that $A^T v \in \mathcal{V}$, and hence $\mathcal{V}$ is A^T-invariant. Therefore, $\mathcal{V}$ contains an eigenvector v_1 of A^T, say $A^T v_1 = \lambda_1 v_1$, so that $v_1^T A = \lambda_1 v_1^T$. Since $v_1 \in \mathcal{V}$, it follows that $v_1^T B = 0$. □

Proof of Theorem 5.2.18 Define for every $\lambda \in \mathbb{C}$ the complex matrix $H(\lambda)$ as

$$H(\lambda) := [I\lambda - A \quad B]. \tag{5.23}$$

In view of Theorem 5.2.10[1] we have to prove that $H(\lambda)$ has constant rank if and only if $\mathfrak{C}$, the controllability matrix of the pair (A, B), (5.21), has rank n. Notice that $\operatorname{rank} H(\lambda) = n$ whenever λ is *not* an eigenvalue of A. This implies that we have to prove that $\operatorname{rank} H(\lambda) = n$ for all $\lambda \in \mathbb{C}$ if and only if $\operatorname{rank} \mathfrak{C} = n$.

Assume that $\operatorname{rank} H(\lambda_1) < n$ for some $\lambda_1 \in \mathbb{C}$. This implies that there exists a nonzero vector $v_1 \in \mathbb{C}^n$ such that $v_1^T H(\lambda_1) = 0$. Therefore, $v_1^T A = \lambda_1 v_1^T$ and $v_1^T B = 0$. From that it follows that $v_1^T A^k B = 0$ for all k, and hence $v_1^T \mathfrak{C} = 0$. It follows that $\operatorname{rank} \mathfrak{C} < n$.

Now suppose $\operatorname{rank} \mathfrak{C} < n$. Then there exists $v \in \mathbb{C}^n$ such that $v^T \mathfrak{C} = 0$. Define the linear subspace $\mathcal{V}$ as

$$\mathcal{V} := \operatorname*{span}_{k \geq 0} \{(A^T)^k v\}.$$

From Lemma 5.2.20 it follows that $\mathcal{V}$ contains a left eigenvector of A contained in the left kernel of B, say $v_1^T A = \lambda_1 v_1^T$ and $v_1^T B = 0$. This implies that $v_1^T H(\lambda_1) = 0$, and hence $\operatorname{rank} H(\lambda_1) < n$. This completes the proof. □

[1] Actually, Theorem 5.2.10 refers to $R(\lambda)$, which in our case is essentially equal to $H(\lambda)$, except that B should then be replaced by $-B$. For historical reasons, and since the rank of the matrix does not depend on whether we take B or $-B$, we use $H(\lambda)$ rather than $R(\lambda)$.

Example 5.2.21 Assume that A and B are given by

$$A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \tag{5.24}$$

For what values of the λ_ks and the b_ks is the pair (A, B) controllable? It is easy to interpret this question in terms of the associated linear scalar i/s systems, which are given by

$$\frac{d}{dt}x_k = \lambda_k x_k + b_k u, \quad k = 1, 2, \ldots, n.$$

It is trivial to see that $b_k = 0$ for some k implies that this system is not controllable. Also, when $\lambda_k = \lambda_\ell$, observe that $z = b_\ell x_k - b_k x_\ell$ is then governed by

$$\frac{d}{dt}z = \lambda_k z,$$

which also shows lack of controllability. Hence (A, B) is controllable only if

$$b_k \neq 0 \quad \text{for all } k \qquad \text{and} \qquad \lambda_k \neq \lambda_\ell \quad \text{for all } k \neq \ell. \tag{5.25}$$

Thus it is easy to see that conditions (5.25) are necessary for controllability of the pair (5.24). That they are also sufficient is more difficult to see directly, but Theorem 5.2.18 provides the answer for this converse. To see this, we compute the controllability matrix $\mathfrak{C}$, see (5.21), and observe that it equals

$$\mathfrak{C} = \begin{bmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_n \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 1 & \lambda_n^1 & \lambda_n^2 & \lambda_n^{n-1} \end{bmatrix}.$$

Thus $\mathfrak{C}$ is the product of a diagonal matrix and a Vandermonde matrix. For $\mathfrak{C}$ to be nonsingular it is necessary and sufficient that both factors be nonsingular. Of course, see Exercise 3.16, this requirement is equivalent to (5.25). □

Theorem 5.2.22 *Controllability of $\frac{d}{dt}x = Ax + Bu$ implies state controllability.*

Proof Suppose that the i/s system defined by $\frac{d}{dt}x = Ax + Bu$ is controllable, and let (u_1, x_1) and (u_2, x_2) be two possible input/state trajectories. We know that there exists an input/state trajectory (u, x) and a time

instant $t_1 > 0$ such that

$$(u(t), x(t)) = \begin{cases} (x_1(t), u_1(t)) & t \le 0, \\ \\ (x_2(t - t_1), u_2(t - t_1)) & t \ge t_1. \end{cases} \tag{5.26}$$

This implies in particular that we can find an input function u that drives the system from state $x_1(0)$ at time $t = 0$ to state $x_2(0)$ at time $t = t_1$. This input function is, of course, only partially defined by (5.26), and in fact it is highly nonunique, even on the interval $[0, t_1]$. However, we can find an explicit expression for a particular input function. Define $z_1 := x_1(0)$ and $z_2 := x_2(0)$. Then, in order for (5.26) to hold, by the variation of the constants formula (4.29) we should have that

$$z_2 = e^{At_1} z_1 + \int_0^{t_1} e^{A(t_1 - \tau)} B u(\tau) d\tau.$$

Define the matrix K by

$$K := \int_0^{t_1} e^{-A\tau} B B^T e^{-A^T \tau} d\tau.$$

We claim that by the assumed controllability of the pair (A, B), K is a nonsingular matrix. To see this, assume that $Ka = 0$ for some $a \in \mathbb{R}^n$. Then also $a^T K a = 0$, and hence

$$a^T K a = \int_0^{t_1} a^T e^{-A\tau} B B^T e^{-A^T \tau} a d\tau = 0. \tag{5.27}$$

Since the integrand is nonnegative and continuous, it follows from (5.27) that for all $t \in [0, t_1]$,

$$a^T e^{-At} B B^T e^{-A^T t} a = 0. \tag{5.28}$$

Since (5.28) is equal to the square of the Euclidean norm of $a^T e^{-At} B$, it follows that for all $t \in [0, t_1]$

$$a^T e^{-At} B = 0. \tag{5.29}$$

Now differentiate (5.29) $n - 1$ times and evaluate the derivatives at $t = 0$:

$$\left.\begin{array}{rcl} a^T e^{-At} B & = & 0 \\ a^T A e^{-At} B & = & 0 \\ a^T A^2 e^{-At} B & = & 0 \\ \vdots & & \\ a^T A^{n-1} e^{-At} B & = & 0 \end{array}\right\} \Rightarrow \begin{array}{rcl} a^T B & = & 0, \\ a^T A B & = & 0, \\ a^T A^2 B & = & 0, \\ \vdots & & \\ a^T A^{n-1} B & = & 0. \end{array}$$

This implies that $a^T\mathfrak{C} = 0$, where $\mathfrak{C}$ denotes the controllability matrix of the pair (A, B). This means that a is orthogonal to the image of $\mathfrak{C}$. Since by assumption the pair (A, B) is controllable, $\mathfrak{C}$ has full rank. It follows that $a = 0$, which indeed implies that K is nonsingular.

Define u as follows:

$$u(t) := B^T e^{-A^T t} x, \tag{5.30}$$

where $x \in \mathbb{R}^n$ will be chosen later. The state at time t_1 resulting from the input (5.30) is

$$x(t_1) = e^{At_1} z_1 + \int_0^{t_1} e^{A(t_1-\tau)} B u(\tau) d\tau = e^{At_1} z_1 + \int_0^{t_1} e^{A(t_1-\tau)} B B^T e^{-A^T \tau} x d\tau$$

$$= e^{At_1} z_1 + e^{At_1} K x.$$

It is obvious how we should choose x so as to achieve that $x(t_1) = z_2$, namely

$$x = K^{-1}(e^{-At_1} z_2 - z_1). \tag{5.31}$$

This shows that the input u defined by (5.30, 5.31) drives the state from z_1 at $t = 0$ to z_2 at $t = t_1$. □

Remark 5.2.23

- We have constructed one input function that drives the system from z_1 to z_2. There are many other input functions that achieve this. In certain applications this freedom of choice can be used to optimize certain criteria. In other words, we may want to find the *best* input function that carries the state from z_1 to z_2. The problems arising from this possibility for optimization forms the subject of *optimal control* and are well beyond the scope of this book.
- If $m = 1$, the single-input case, then $\mathfrak{C}$ is a square matrix and controllability is equivalent to the condition that $\det \mathfrak{C}$ is nonzero.

□

The image of the controllability matrix has the following nice geometric interpretation.

Theorem 5.2.24 $\operatorname{im} \mathfrak{C}$ *is the smallest* A*-invariant subspace of* $\mathbb{R}^n$ *that contains* $\operatorname{im} B$*. Therefore, (5.20) is controllable if and only if* $\mathbb{R}^n$ *is the smallest* A*-invariant subspace that contains* $\operatorname{im} B$*.*

Proof Note that

$$A \operatorname{im} \mathfrak{C} = A \operatorname{im}[B \cdots A^{n-1} B] = \operatorname{im}[AB \cdots A^n B] \subset \operatorname{im}[B \ldots A^{n-1} B].$$

The last equality follows from the Cayley–Hamilton theorem, which implies that A^n can be written as a linear combination of lower powers of A. This shows that $\operatorname{im} \mathfrak{C}$ is A-invariant. That $\operatorname{im} \mathfrak{C}$ contains $\operatorname{im} B$ is obvious.

It remains to prove that $\operatorname{im} \mathfrak{C}$ is the smallest subspace with these two properties. Let V be any A-invariant subspace containing the image of B. Since V is A-invariant and since $\operatorname{im} B \subset V$, we conclude that $A \operatorname{im} B \subset AV \subset V$. This implies that $\operatorname{im} AB \subset V$. This in turn implies that $\operatorname{im} A^2B \subset \operatorname{im} AV \subset V$. Applying the same argument several times finally yields $\operatorname{im} A^{n-1}B \subset V$. From that it follows that $\operatorname{im} \mathfrak{C} \subset V$, which completes the proof. □

Using Theorem 5.2.14 yields a decomposition of the behavior of $\frac{d}{dt}x = Ax + Bu$ into a controllable and an autonomous part. Theorem 5.2.24 allows us to bring it into a form that displays this decomposition explicitly.

Corollary 5.2.25 *Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$. There exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that*

$$S^{-1}AS = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad S^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{5.32}$$

with (A_{11}, B_1) controllable.

Proof Let k be the rank of the controllability matrix $\mathfrak{C}$ corresponding to (A, B). Choose a basis $s_1, \ldots, s_k, s_{k+1}, \ldots, s_n$ of the state space $\mathbb{R}^n$ such that $s_1, \ldots, s_k$ is a basis of $\operatorname{im} \mathfrak{C}$. Define S as the matrix that has $s_1, \ldots, s_n$ as its columns. Since $\operatorname{im} \mathfrak{C}$ is A-invariant, we conclude that there exist matrices $A_{11} \in \mathbb{R}^{k\times k}$, $A_{12} \in \mathbb{R}^{k\times n-k}$, and $A_{22} \in \mathbb{R}^{n-k\times n-k}$ such that

$$AS = S \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

which proves the first equality in (5.32). Furthermore, since $\operatorname{im} B \subset \operatorname{im} \mathfrak{C}$, there exists a matrix $B_1 \in \mathbb{R}^{k\times m}$ such that

$$B = S \begin{bmatrix} B_1 \\ 0 \end{bmatrix}.$$

This proves the second equality in (5.32). Note, moreover, that

$$S^{-1}\mathfrak{C} = \begin{bmatrix} B_1 & A_{11}B_1 & \ldots & A_{11}^{n-1}B_1 \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Since the rank of $\mathfrak{C}$ is equal to k it follows that the rank of the upper part of $S^{-1}\mathfrak{C}$ is also k. It follows that the rank of $[B_1 \cdots A_{11}^{k-1}B_1]$ is equal to k; see Exercise 5.16c, which shows that (A_{11}, B_1) is indeed controllable. □

The special form (5.32) immediately yields the following result.

Corollary 5.2.26 *Consider*

$$\frac{d}{dt}x = Ax + Bu \tag{5.33}$$

with (A, B) *of the form (5.32). Let the initial state be* $x(0) = 0$. *Then for all inputs* u *and for all* t, *the resulting state-trajectory lies in the subspace* $\operatorname{im} \mathfrak{C}$. *Moreover, starting from the zero state, every state in* $\operatorname{im} \mathfrak{C}$ *can be reached. The subspace* $\operatorname{im} \mathfrak{C}$ *is therefore often called the reachable subspace of (5.33).*

Combining Theorem 5.2.18, Theorem 5.2.22, and Corollary 5.2.26, we obtain the following result.

Theorem 5.2.27 *Consider the system defined by*

$$\frac{d}{dt}x = Ax + Bu. \tag{5.34}$$

The following statements are equivalent:

1. *The system (5.34) is controllable.*
2. $\operatorname{rank}[I\lambda - A \;\; B] = n$ *for all* $\lambda \in \mathbb{C}$.
3. $\operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = n$.
4. *The system (5.34) is state controllable.*

Proof The equivalence of (1) and (2) was proven in Theorem 5.2.10, the equivalence of (1) and (3) in Theorem 5.2.18. The fact that (1) implies (4) is the content of Theorem 5.2.22 and the converse follows from Corollary 5.2.26. □

5.2.1.2 Controllability of i/s/o systems

Up to now we have analyzed the controllability of the i/s part of i/s/o systems. The question remains as to when the i/s/o representation itself is controllable. This is now a triviality: *An i/s/o representation is controllable if and only if its i/s part is controllable*. This statement follows immediately from the observation

$$\operatorname{rank}\begin{bmatrix} I\lambda - A & B & 0 \\ -C & 0 & I \end{bmatrix} = p + \operatorname{rank}\begin{bmatrix} I\lambda - A & B \end{bmatrix}, \tag{5.35}$$

where p is the number of outputs. Equation (5.35) shows that whether or not the polynomial matrix that defines the i/s/o behavior has constant rank, independent of $\lambda \in \mathbb{C}$, depends on the matrix $[I\lambda - A \;\; B]$ only.

5.2.2 Stabilizability

If a system is controllable, then it is in principle possible to steer from any trajectory in the behavior to a desired trajectory. In applications the desired trajectory is often an equilibrium point, i.e., a trajectory that does not depend on time. Controllability is not always needed to steer the system to a constant trajectory, as the following example shows.

Example 5.2.28 (**Example 5.2.1 continued.**) Consider again the two pendula mounted on a cart depicted in Figure 5.1. As argued in Example 5.2.1, this system is *uncontrollable* if the lengths of the rods are equal. In particular, the difference between the angles of the pendula, $w_1 - w_2$, behaves autonomously and is not affected by the input. In Example 5.2.1, we neglected the effect of friction in the joints of the rods. Suppose that we incorporate these in the model. Then the linearized equations (5.11) become

$$\begin{aligned}
&(M + m_1 + m_2)(\tfrac{d}{dt})^2 w_4 = w_3 - (\tfrac{d}{dt})^2 (m_1 L_1 w_1 + m_2 L_2 w_2),\\
&(k_i - m_i L_i g) w_i + d_i \tfrac{d}{dt} w_i + m_i L_i^2 \frac{d^2}{dt^2} w_i + m_i L_i \frac{d^2}{dt^2} w_4 = 0, \qquad i = 1, 2.
\end{aligned} \tag{5.36}$$

for some positive constants d_i and k_i, $i = 1, 2$. For simplicity take $M = m_1 = m_2 = 1$, $L_1 = L_2 = 1$, $d_1 = d_2 = 1$, and $k := k_1 = k_2$. Recall that we took $w = \mathrm{col}(w_1, w_2, w_3, w_4)$. The polynomial matrix representing the equations (5.36) becomes

$$R(\xi) = \begin{bmatrix} \xi^2 & \xi^2 & -1 & 3\xi^2 \\ k - g + \xi + \xi^2 & 0 & 0 & \xi^2 \\ 0 & k - g + \xi + \xi^2 & 0 & \xi^2 \end{bmatrix}.$$

It is easy to see that $\operatorname{rank} R(\lambda) = 2$ if $\lambda = \frac{1}{2}(-1 \pm \sqrt{1 + 4g - 4k})$ and $\operatorname{rank} R(\lambda) = 3$ otherwise. Therefore, the system is not controllable. As argued in Example 5.2.1, this may be explained by observing that $w_1 - w_2$ behaves autonomously. To make this more apparent, we rewrite the system equations in terms of the variables $\mathrm{col}(w_1, w_3, w_4)$ and $w_1 - w_2$:

$$\begin{aligned}
&\begin{bmatrix} \dfrac{d^2}{dt^2} & -1 & 3\dfrac{d^2}{dt^2} \\ k - g + \dfrac{d}{dt} + \dfrac{d^2}{dt^2} & 0 & \dfrac{d^2}{dt^2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_3 \\ w_4 \end{bmatrix} = 0,\\
&(k - g)(w_1 - w_2) + \frac{d}{dt}(w_1 - w_2) + \frac{d^2}{dt^2}(w_1 - w_2) = 0.
\end{aligned} \tag{5.37}$$

From (5.37) it follows that the behavior of (w_1, w_3, w_4) is controllable if $k \neq g$, whereas that of $w_1 - w_2$ is completely autonomous. Physically

speaking, this means that by applying the appropriate force, we can make the first rod reach every possible trajectory that satisfies (5.36) while the second rod is just following the first. Notice that the characteristic values of the second equation in (5.37) are the uncontrollable modes, i.e., the values for which the rank of $R(\lambda)$ drops, $\lambda_{1,2} = \frac{1}{2}(-1 \pm \sqrt{1+4g-4k})$. They give rise to trajectories of the form $w_1(t) - w_2(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$. It is interesting to notice that if $k > g$, then the λ_1 and λ_2 have real part smaller than zero. This implies that $w_1(t) - w_2(t)$ converges to zero as t tends to infinity. The conclusion is that if $k > g$, then the system may be controlled to trajectories in the behavior for which $w_1 - w_2$ goes to zero asymptotically. In particular, it is possible to bring the two rods to an upright position and keep them there. □

Example 5.2.28 indicates that a notion weaker than controllability can sometimes be useful. We call a system *stabilizable* if every trajectory in the behavior can be steered asymptotically to a desired trajectory. The formal definition is given below.

Definition 5.2.29 Let $\mathfrak{B}$ be the behavior of a time-invariant dynamical system. This system is called *stabilizable* if for every trajectory $w \in \mathfrak{B}$, there exists a trajectory $w' \in \mathfrak{B}$ with the property

$$w'(t) = w(t) \text{ for } t \leq 0 \qquad \text{and} \qquad \lim_{t\to\infty} w'(t) = 0.$$

□

An effective test for stabilizability, analogous to Theorem 5.2.10, is provided in the following theorem.

Theorem 5.2.30 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The behavior $\mathfrak{B}$ defined by $R(\frac{d}{dt})w = 0$ is stabilizable if and only if the rank of the (complex) matrix $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}_+$, where $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \operatorname{Re} s \geq 0\}$.*

Proof Without loss of generality we may assume that $R(\xi)$ has full row rank. By Theorem 5.2.14, $\mathfrak{B}$ may be written as the direct sum of an autonomous and a controllable behavior. Let $\mathfrak{B} = \mathfrak{B}_{\text{aut}} \oplus \mathfrak{B}_{\text{contr}}$ be such a decomposition. Denote the corresponding polynomial matrices by $R_{\text{aut}}(\xi)$ and $R_{\text{contr}}(\xi)$ respectively. The characteristic values of $\mathfrak{B}_{\text{aut}}$ are precisely those λs for which $\operatorname{rank} R(\lambda) < g$. Denote these by $\lambda_1, \ldots, \lambda_N$. Thus, by Theorem 3.2.16, all $w \in \mathfrak{B}_{\text{aut}}$ are linear combinations of functions of the form $B_{ij} t^i e^{\lambda_j t}$.

"If" part. By assumption, $\operatorname{Re} \lambda_i < 0$, $i = 1, \ldots N$. This implies that

$$w \in \mathfrak{B}_{\text{aut}} \quad \Rightarrow \quad \lim_{t\to\infty} w(t) = 0.$$

Choose $w \in \mathfrak{B}$. Let $w = w_1 + w_2$ with $w_1 \in \mathfrak{B}_{\rm aut}$ and $w_2 \in \mathfrak{B}_{\rm contr}$. Since $\mathfrak{B}_{\rm contr}$ is controllable, there exists $w_2' \in \mathfrak{B}_{\rm contr}$ and $t_1 \geq 0$ such that

$$w_2'(t) = 0,\ t \geq t_1 \qquad \text{and} \qquad w_2(t) = w_2'(t) \qquad \text{for} \qquad t \leq 0.$$

Since $w_1 \in \mathfrak{B}_{\rm aut}$, it vanishes asymptotically, and by defining $w' := (w_1, w_2')$, we have constructed a trajectory $w' \in \mathfrak{B}$ such that

$$\lim_{t\to\infty} w'(t) = 0 \qquad \text{and} \qquad w(t) = w'(t) \qquad \text{for} \qquad t \leq 0.$$

This shows that the system is stabilizable.

"Only if" part. Suppose that the system is stabilizable and that $\operatorname{Re}\lambda_i \geq 0$ for some $1 \leq i \leq N$. Choose $B_i \in \mathbb{C}^q$ such that $w_1(t) := B_i e^{\lambda_i t}$ belongs to $\mathfrak{B}_{\rm aut}$. Since the system is stabilizable, there exists $w_2 \in \mathfrak{B}_{\rm contr}$ such that if we define $w \in \mathfrak{B}$ as $w = w_1 + w_2$, we have that $\lim_{t\to\infty} w(t) = 0$. Since $w_2 \in \mathfrak{B}_{\rm contr}$, we conclude that $R_{\rm contr}(\frac{d}{dt})w = R_{\rm contr}(\frac{d}{dt})B_i e^{\lambda_i t} = R_{\rm contr}(\lambda_i)Be^{\lambda_i t} =: \tilde{B}e^{\lambda_i t}$. Notice that $\tilde{B} \neq 0$, since $w_1 \notin \mathfrak{B}_{\rm contr}$. Next integrate the differential equation $R_{\rm contr}(\frac{d}{dt})w = \tilde{B}e^{\lambda_i t}$ to obtain the *integral* equation

$$R^*_{\rm contr}(\textstyle\int)w + c_0 + c_1 t + \cdots + c_{L-1}t^{L-1} = R^*_{\rm contr}(\textstyle\int)\tilde{B}e^{\lambda_i t}; \qquad (5.38)$$

see Definition 2.3.7. Since $\lim_{t\to\infty} w(t) = 0$, we conclude that the left-hand side of (5.38) grows at most polynomially in t, whereas the right-hand side grows exponentially. This is absurd, and therefore the system is not stabilizable. □

An immediate consequence of Theorem 5.2.30 is the following result.

Corollary 5.2.31 *Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ be in the form (5.32). The system $\frac{d}{dt}x = Ax + Bu$, or simply the pair (A, B), is stabilizable if and only if the matrix A_{22} has all its eigenvalues in the open left half-plane.*

In Chapter 9 we will use the notion of stabilizability in the context of feedback stabilization.

5.3 Observability

In this section we introduce the notion of *observability*. This notion is intuitively explained as follows. Suppose that we have a behavior of which the variables are partitioned as (w_1, w_2). We call w_2 observable from w_1 if w_1, together with the laws of the system, determines w_2 uniquely. That means that for each w_1 there exists at most one w_2 such that (w_1, w_2) belongs to the behavior. A direct implication of w_2 being observable from w_1 is that in fact all the information of a trajectory w is already contained in its first component w_1.

Example 5.3.1 Assume that we can observe the forces acting on a mechanical system. Can we deduce its position from these observations? If the system is a simple point-mass, governed by Newton's law

$$M\frac{d^2}{dt^2}q = F, \tag{5.39}$$

then it is obvious that knowledge of F tells us only what the acceleration is, and we are unable to deduce the position (unless, of course, we know the initial position and velocity, but this is not assumed to be the case). Therefore, q is not observable from F in (5.39). This lack of observability in this simple example has important implications for inertial navigation systems. Since on-board a space vehicle we can only measure forces and accelerations, we have to integrate these twice in order to compute the position, and therefore we have to put the initial position and the initial velocity as starting conditions into the computation. Once an error is made, it is impossible to correct this using on-board measurements only, and therefore regular communication with a ground station is unavoidable in order to keep track of the position in inertial navigation.

□

Definition 5.3.2 Let $(\mathbb{R}, \mathbb{W}_1 \times \mathbb{W}_2, \mathfrak{B})$ be a time-invariant dynamical system. Trajectories in $\mathfrak{B}$ are partitioned as (w_1, w_2) with $w_i : \mathbb{R} \to \mathbb{W}_i$, $i = 1, 2$. We say that w_2 is *observable* from w_1 if for all $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$ implies $w_2 = w_2'$. □

Definition 5.3.2 formalizes the intuitive description given in the introduction to this section. Notice that if the behavior is specified by polynomial matrices as $R_(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$, then w_2 is observable from w_1, then w_2 is uniquely determined by w_1 and the polynomial matrices $R_1(\xi)$ and $R_2(\xi)$. So, given w_1, we should in principle be able to determine the corresponding w_2. Algorithms that do this are called *observers.* However, to find a means by which we actually can deduce w_2 from w_1 is in general not at all straightforward. We will treat a special case of this problem in Chapter 10.

The following rank test allows us to check observability of w_2 from w_1 in behaviors defined by $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$.

Theorem 5.3.3 *Let $R_1(\xi) \in \mathbb{R}^{g\times q_1}[\xi]$ and $R_2(\xi) \in \mathbb{R}^{g\times q_2}[\xi]$. Let $\mathfrak{B}$ be the behavior defined by $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$. Then the variable w_2 is observable from w_1 if and only if* $\operatorname{rank} R_2(\lambda) = q_2$ *for all $\lambda \in \mathbb{C}$.*

Proof Let $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$. Then by linearity of $\mathfrak{B}$, also $(0, w_2 - w_2') \in \mathfrak{B}$, and hence $R_2(\frac{d}{dt})(w_2 - w_2') = 0$. It follows that w_2 is observable from w_1 if and only if $R_2(\frac{d}{dt})w_2 = 0$ implies that $w_2 = 0$. Define $\mathfrak{B}_2 := \{w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^{q_2}) \mid R_2(\frac{d}{dt})w_2 = 0\}$. By the previous remark it suffices

to prove that $\mathfrak{B}_2 = \{0\}$ if and only if $\operatorname{rank} R_2(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$. By Theorem 2.5.23 there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that

$$U(\xi)R_2(\xi) = \begin{bmatrix} R_2'(\xi) \\ 0 \end{bmatrix} \quad \text{and } R_2'(\xi) \in \mathbb{R}^{g'\times q_2}[\xi] \text{ of full row rank.}$$

By Theorem 2.5.4 we have that $\mathfrak{B}_2 = \{w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^d) \mid R_2'(\frac{d}{dt})w_2 = 0\}$. Moreover, $\operatorname{rank} R_2(\lambda) = \operatorname{rank} R_2'(\lambda)$, so that we have to prove that $\mathfrak{B}_2 = \{0\}$ if and only if $\operatorname{rank} R_2'(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$. Since $R_2'(\xi)$ is of full row rank, $g' \leq q_2$.

Suppose that $\mathfrak{B}_2 = \{0\}$. If $g' < q_2$, then it follows from Theorem 3.3.22 and Corollary 3.3.23 that $\mathfrak{B}_2 \neq \{0\}$, which is a contradiction. Therefore, $g' = q_2$. If $\deg \det R_2'(\xi) \geq 1$, then again it follows from Theorem 3.2.16 that $\mathfrak{B}_2 \neq \{0\}$, and therefore $\det R_2'(\xi)$ must be equal to a nonzero constant; i.e., $R_2'(\xi)$ is unimodular. Equivalently, $\operatorname{rank} R_2'(\lambda) = q_2$, and therefore also $\operatorname{rank} R_2(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$.

Conversely, suppose that $\operatorname{rank} R_2'(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$. Then $g' = q_2$, and hence $\det R_2'(\xi) = c$ for some nonzero constant c. That implies that $R_2'(\xi)$ is unimodular, and therefore $\mathfrak{B}_2 = \{0\}$. □

Remark 5.3.4 Observability is particularly interesting for behaviors with latent variables described by

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell. \tag{5.40}$$

Usually, we consider w as the observed variable and ℓ as the to-be-observed variable. See Section 4.2 where the behavioral equation (5.40) was introduced and Section 6.2 where the elimination of ℓ from (5.40) is studied. The problem is then to check whether the latent variable ℓ is observable from w. By replacing w_1 by w and w_2 by ℓ, we can just apply Definition 5.3.2 and Theorem 5.3.3. If in (5.40) ℓ is observable from w in this sense, then we call the latent variable system simply *observable*. □

Example 5.3.5 Consider the electrical circuit of Example 1.3.5. The vector of latent variables is $\operatorname{col}(V_{R_C}, I_{R_C}, V_L, I_L, V_C, I_C, V_{R_L}, I_{R_L})$, while the vector of manifest variables is $w = \operatorname{col}(V, I)$. To convert the equations (1.1, 1.2, 1.3) into the standard notation $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$, we define the poly-

nomial matrices $R(\xi)$ and $M(\xi)$ as follows:

$$R(\xi) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad M(\xi) = \begin{bmatrix} 1 & -R_C & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -R_L \\ 0 & 0 & 0 & 0 & C\xi & -1 & 0 & 0 \\ 0 & 0 & -1 & L\xi & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \end{bmatrix},$$

To see whether ℓ is observable from w, we have, following Theorem 5.3.3, to determine the rank of $M(\lambda)$. To that end we bring $M(\xi)$ into a more transparent form by means of elementary row and column operations. Denote the rows of $M(\xi)$ by $r_1, \ldots, r_{11}$ and the columns by $c_1, \ldots, c_8$. Apply the following sequence of row operations:

$$\begin{aligned} &r_{11} \leftarrow r_{11} - r_9,\ r_{11} \leftarrow r_{11} + r_{10}, \quad r_5 \leftarrow r_5 - r_6, \quad r_5 \leftarrow r_5 - r_7, \\ &r_5 \leftarrow r_5 - r_8, \quad r_6 \leftarrow r_6 + r_8, \quad r_1 \leftarrow r_1 - r_9, \quad r_1 \leftarrow r_1 + R_C r_6, \\ &r_4 \leftarrow r_4 + r_{10}, \quad r_4 \leftarrow r_4 - L\xi r_7, \quad r_3 \leftarrow r_3 + C\xi r_1,\ r_9 \leftarrow r_9 + r_1, \\ &r_3 \leftarrow r_3 + r_8, \quad r_2 \leftarrow r_2 - r_4, \quad r_{10} \leftarrow r_{10} - r_4. \end{aligned}$$

Next, apply the following column operation:

$$c_8 := c_8 + R_C c_5 - L\xi c_7 - c_2 + c_4 - c_6 - R_C c_1 + L\xi c_3. \tag{5.41}$$

The result of these row and column operations is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -R_L - L\xi \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 + CR_C\xi \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} =: \tilde{M}(\xi). \tag{5.42}$$

It follows immediately from (5.42) that $\operatorname{rank} \tilde{M}(\lambda) = 8$ if and only if $\frac{L}{R_L} \neq CR_C$, and therefore also $\operatorname{rank} M(\lambda) = 8$ for all $\lambda \in \mathbb{C}$ if and only if $\frac{L}{R_L} \neq CR_C$. By Theorem 5.3.3 it follows that ℓ is observable from w if and only if $\frac{L}{R_L} \neq CR_C$. □

In the next example we derive a necessary condition for w_2 to be observable from w_1.

Example 5.3.6 Consider the system described by

$$R_1(\frac{d}{dt})w_1 + R_2(\frac{d}{dt})w_2 = 0. \tag{5.43}$$

Assume that $R(\xi) = [R_1(\xi)\ R_2(\xi)]$ is of full row rank. Let q_1 be the dimension of w_1 and q_2 the dimension of w_2. From Corollary 3.3.23 and Remark 3.3.26 we conclude that this system has $m := q_1 + q_2 - \operatorname{rank} R(\xi)$ inputs. Recall that the input variables are not constrained by the laws of the system, in this case (5.43). Therefore, for w_2 to be observable from w_1 it stands to reason that w_1 must somehow measure, directly or indirectly, all the input variables. More precisely, suppose that we want to reconstruct w_2 from w_1. In particular, we want to reconstruct the free variables, u_2 say, contained in w_2. The free part of w_1, call it u_1, contains no information about u_2. Therefore, u_2 should be constructed on the basis of the output part of w_1, namely y_1. Since u_2 is free, it seems inevitable that the dimension of y_1 should at least be equal to the dimension of u_2. Since w_1 also contains u_1, this means that the dimension of w_1 should at least be equal to the number of inputs; i.e., $q_1 \geq m$. Can we deduce this inequality rigorously from Theorem 5.3.3?

Note that it follows from Theorem 5.3.3 that w_2 is observable from w_1 if and only if $\operatorname{rank}(R_2(\lambda)) = q_2$ for all $\lambda \in \mathbb{C}$. This requirement implies that $q_2 \leq \operatorname{rank}(R(\lambda))$ must hold, and therefore $m = q_1 + q_2 - \operatorname{rank} R(\xi) \leq q_1 + q_2 - q_2 = q_1$ so that indeed $q_1 \geq m$.

Note that on the other hand, $m \leq q_1$ is not a sufficient condition for observability. For example, in the SISO system

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u,$$

y is not observable from u whenever $p(\xi)$ is of degree ≥ 1. □

5.3.1 Observability of i/s/o systems

We apply the results of the previous subsection to i/s/o systems. The state is considered to be the latent variable, and the input and output jointly form the manifest variables. Observability then means that the state trajectory can be deduced from the input and output trajectories. This is the standard situation considered in classical system theory, where observability of x from (u, y) is often referred to as *state observability*, or just *observability*. The relevance of this problem stems from the fact that in many applications the state is not directly measurable, whereas knowledge of the state is needed for purposes of control, prediction, detection, etc.

Consider the i/s/o system

$$\left\{ \begin{array}{rcl} \frac{d}{dt}x & = & Ax + Bu, \\ y & = & Cx + Du. \end{array} \right. \tag{5.44}$$

Denote the associated behavior by $\mathfrak{B}_{\mathrm{i/s/o}}$.

Applying Theorem 5.3.3 to the situation at hand immediately yields the following rank test.

Theorem 5.3.7 *The state x is observable from (u, y) if and only if the matrix*

$$\left[\begin{array}{c} I\lambda - A \\ C \end{array} \right]$$

has rank n for all $\lambda \in \mathbb{C}$.

Proof Define $R(\xi)$ and $M(\xi)$ as

$$R(\xi) := \left[\begin{array}{cc} B & 0 \\ -D & I \end{array} \right], \quad M(\xi) = \left[\begin{array}{c} I\xi - A \\ C \end{array} \right]. \tag{5.45}$$

Then (5.44) can be written as $R(\frac{d}{dt})w = M(\frac{d}{dt})x$, where $w^T = (u^T, y^T)^T$. Now apply Theorem 5.3.3. □

Notice that there is a remarkable similarity between Theorem 5.2.18 and Theorem 5.3.7. Often it is referred to as *duality*. Although much can be said about duality, we will not go more deeply into it. However, we use duality in order to obtain results about observability from their counterparts about controllability.

As was the case for controllability, observability can be checked by a simple rank criterion. It turns out that the criterion is of the same nature as that for controllability, but now in terms of the pair (A, C). Analogously to the controllability matrix of a pair (A, B), we define the observability matrix of the pair of matrices (A, C).

Definition 5.3.8 Let $(A, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$. Define the matrix $\mathfrak{O} \in \mathbb{R}^{pn\times n}$ by

$$\mathfrak{O} := \left[\begin{array}{c} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-2} \\ CA^{n-1} \end{array} \right].$$

$\mathfrak{O}$ is called the *observability matrix* of the system (5.44) or of the matrix pair (A, C). □

The counterpart of Theorem 5.2.18 is the following:

Theorem 5.3.9 *The system defined by the equations (5.44) is observable if and only if its observability matrix $\mathfrak{O}$ has rank n.*

Proof From Theorem 5.2.18 we conclude that $\mathfrak{O}^T$ has rank n if and only if the matrix $[I\lambda - A^T \quad C^T]$ has rank n for all $\lambda \in \mathbb{C}$. Since the rank of a matrix equals that of its transpose, the result now follows immediately from Theorem 5.3.7. □

Example 5.3.10 Consider the linear system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -a_0 & -a_1 \end{bmatrix} x, \quad y = [b_0 \quad b_1]x. \tag{5.46}$$

For what values of a_0, a_1, b_0, b_1 is it observable? The observability matrix equals

$$\mathfrak{O} = \begin{bmatrix} b_0 & b_1 \\ -a_0 b_1 & b_0 - a_1 b_1 \end{bmatrix}.$$

The matrix $\mathfrak{O}$ is nonsingular if and only if

$$b_0^2 - a_1 b_0 b_1 + a_0 b_1^2 \neq 0,$$

in other words, if and only if the root $-\frac{b_0}{b_1}$ of $q(\xi) = b_0 + b_1\xi$ is not a root of $p(\xi) = a_0 + a_1\xi + \xi^2$, the characteristic polynomial of the A-matrix associated with (5.46).

□

The following corollary expresses that in an observable system, knowledge of (u, y) on a given time interval of positive length determines x restricted to that time interval.

Corollary 5.3.11 *Consider the behavior defined by (5.44) and suppose x is observable from (u, y). Suppose that for some $(u_1, x_1, y_1), (u_2, x_2, y_2) \in \mathfrak{B}_{\mathrm{i/s/o}}$ and for some t_1, it holds that for all $t \in [0, t_1]$,*

$$u_1(t) = u_2(t) \qquad \text{and} \qquad y_1(t) = y_2(t). \tag{5.47}$$

Then $x_1(t) = x_2(t)$ for $t \in [0, t_1]$.

Proof Since x is observable from (u, y), it follows from Theorem 5.3.9 that $\mathfrak{O}$ has rank n. From (5.47) we conclude that for $0 \leq t \leq t_1$,

$$Ce^{At}z_1 + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t) = Ce^{At}z_2 + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t),$$

where u is any input function that is equal to u_1 and u_2 on $[0, t_1]$. Therefore, we have that for all $t \in [0, t_1]$,

$$Ce^{At}(z_1 - z_2) = 0. \tag{5.48}$$

Differentiating (5.48) $n-1$ times and evaluating the result at $t = 0$ yields

$$\left.\begin{array}{rcl} Ce^{At}(z_1 - z_2) & = & 0 \\ CAe^{At}(z_1 - z_2) & = & 0 \\ CA^2e^{At}(z_1 - z_2) & = & 0 \\ \vdots & & \\ CA^{n-1}e^{At}(z_1 - z_2) & = & 0 \end{array}\right\} \Rightarrow \begin{array}{rcl} C(z_1 - z_2) & = & 0, \\ CA(z_1 - z_2) & = & 0, \\ CA^2(z_1 - z_2) & = & 0, \\ \vdots & & \\ CA^{n-1}(z_1 - z_2) & = & 0. \end{array}$$

This implies that $z_1 - z_2 \in \ker \mathfrak{O}$. Since $\mathfrak{O}$ has rank n, it follows that $z_1 = z_2$. The state trajectories hence satisfy

$$\begin{aligned} x_1(t) &= e^{At}z_1 + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau + Du(t) \\ &= e^{At}z_2 + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau + Du(t) = x_2(t). \end{aligned}$$

This proves that $x_1(t) = x_2(t)$ for $t \in [0, t_1]$. □

Remark 5.3.12

- Observability depends only on the matrices A and C, although from the original definition it was not immediately clear that B and D play no role. Observability of x from (u, y) is often identified with the rank condition on $\mathfrak{O}$. We call the pair (A, C) observable if the associated observability matrix has full rank. Notice that (A, B) is controllable if and only if (A^T, B^T) is observable. This expresses the duality of controllability and observability.

- We call the rank tests on the matrices $H(\xi)$ in (5.23) and on $M(\xi)$ in (5.45) the *Hautus tests* for controllability and observability respectively. They are sometimes useful when the calculation of the matrices $\mathfrak{C}$ or $\mathfrak{O}$ is cumbersome.

- Note that t_1 in Corollary 5.3.11 could be an arbitrary positive number. This means that in observable systems, the initial state $x(0)$ is determined by the input and output on an arbitrarily small time interval containing 0.

- If $p = 1$, the single-output case, then $\mathfrak{O}$ is a square matrix, and hence observability is equivalent to $\det \mathfrak{O} \neq 0$.

□

Similar to the image of the controllability matrix $\mathfrak{C}$, there is an elegant geometric interpretation for the kernel of $\mathfrak{O}$.

Theorem 5.3.13 *The kernel of $\mathfrak{O}$ is the largest A-invariant subspace contained in the kernel of C. Therefore, (5.44) is observable if and only if 0 is the largest A-invariant subspace contained in* $\ker C$.

Proof Choose $x \in \ker \mathfrak{O}$. Then $Cx = CAx = \cdots = CA^{n-1}x = 0$. By the Cayley–Hamilton theorem, it follows that also $CA^n x = 0$, and hence $Ax \in \ker \mathfrak{O}$, which implies that $\ker \mathfrak{O}$ is A-invariant. Furthermore, $\mathfrak{O}x = 0$ implies that $Cx = 0$, and hence $\ker \mathfrak{O} \subset \ker C$. Therefore, $\ker \mathfrak{O}$ is an A-invariant subspace contained in $\ker C$. To show that it is the largest such subspace, assume that $\mathcal{V}$ is an A-invariant subspace contained in the kernel of C. Choose $x \in \mathcal{V}$. Then, since $\mathcal{V}$ is A-invariant, also $Ax, A^2x, \ldots, A^{n-1}x \in \mathcal{V}$, and since $\mathcal{V}$ is contained in $\ker C$, we conclude that $Cx = CAx = \cdots = CA^{n-1}x = 0$. This implies that $x \in \ker \mathfrak{O}$, and hence $\mathcal{V} \subset \ker \mathfrak{O}$, as claimed. □

From Theorem 5.3.13 we obtain a partitioning of the state space into an observable and a nonobservable part.

Corollary 5.3.14 *Let $(A, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$. There exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that*

$$S^{-1}AS = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \textit{and} \quad CS = \begin{bmatrix} 0 & C_2 \end{bmatrix}, \tag{5.49}$$

with (C_2, A_{22}) observable.

Proof Let k be the the dimension of the kernel of the observability matrix $\mathfrak{O}$ corresponding to (A, C). Choose a basis $s_1, \ldots, s_k, s_{k+1}, \ldots, s_n$ of the state space $\mathbb{R}^n$ such that $s_1, \ldots, s_k$ is a basis of $\ker \mathfrak{O}$. Define S as the matrix that has $s_1, \cdots, s_n$ as its columns. Since $\ker \mathfrak{O}$ is A-invariant, there exist matrices $A_{11} \in \mathbb{R}^{k\times k}$, $A_{12} \in \mathbb{R}^{k\times n-k}$, and $A_{22} \in \mathbb{R}^{n-k\times n-k}$ such that

$$AS = S \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{5.50}$$

which proves the first equality in (5.49). Furthermore, since $\ker \mathfrak{O} \subset \ker C$, there exists a matrix $C_2 \in \mathbb{R}^{p\times(n-k)}$ such that

$$CS = \begin{bmatrix} 0 & C_2 \end{bmatrix}.$$

This proves the second equality in (5.49). It remains to show that (A_{22}, C_2) is an observable pair. From (5.49) it follows that

$$\mathfrak{O}S = \begin{bmatrix} 0 & C_2 \\ 0 & C_2A_{22} \\ \vdots & \vdots \\ 0 & C_2A_{22}^{n-1} \end{bmatrix}. \tag{5.51}$$

Since $\dim \ker \mathfrak{O} = k$, we have that $\operatorname{rank} \mathfrak{O} = n - k$, and by (5.51) it follows that

$$\operatorname{rank} \begin{bmatrix} C_2 \\ C_2A_{22} \\ \vdots \\ C_2A_{22}^{n-1} \end{bmatrix} = n - k. \tag{5.52}$$

Using the Cayley–Hamilton theorem in (5.52) yields

$$\operatorname{rank} \begin{bmatrix} C_2 \\ C_2A_{22} \\ \vdots \\ C_2A_{22}^{n-k-1} \end{bmatrix} = n - k.$$

This shows that the pair (A_{22}, C_2) is observable. □

Combining Theorems 5.3.7 and 5.3.9 and Corollary 5.3.11, we obtain the following result.

Theorem 5.3.15 *Consider the system defined by*

$$\frac{d}{dt}x = Ax + Bu, \quad y = Cx + Du. \tag{5.53}$$

The following statements are equivalent:

1. *The system (5.53) is observable.*
2. $\operatorname{rank} \begin{bmatrix} I\lambda - A \\ C \end{bmatrix} = n$ *for all* $\lambda \in \mathbb{C}$.
3. $\operatorname{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n.$
4. *The input/output trajectory determines the state uniquely: if* $t_1 > 0$ *and* (u_1, y_1) *and* (u_2, y_2) *satisfy (5.53) and for all* $t \in [0, t_1]$, $(u_1(t), y_1(t)) = (u_2(t), y_2(t))$, *then also* $x_1(t) = x_2(t)$ *for all* $t \in [0, t_1]$.

Proof The equivalence of (1) and (2) was proven in Theorem 5.3.7, the equivalence of (1) and (3) in Theorem 5.3.9. The fact that (1) inplies (4) is the content of Corollary 5.3.11. To see that (4) implies (1), suppose that the system is not observable. It follows from Corollary 5.3.14 that the component of the state that corresponds to A_{11} does not influence u and y and is therefore not uniquely determined by $\mathrm{col}(u, y)$. □

5.3.2 Detectability

If in a behavior the variable w_2 is observable from w_1, then w_1 together with the laws of the system determine w_2 uniquely. For linear time-invariant systems this is equivalent to the property that if w_1 is the zero trajectory, then w_2 is also the zero trajectory. A slightly weaker, but still useful, property would be that if w_1 is the zero trajectory, then $w_2(t)$ converges to zero as t tends to infinity. The trajectory w_2 is then no longer uniquely determined by w_1 and the laws of the system, but we can nevertheless determine the asymptotic value(s) of the corresponding trajectory w_2. That means that to one trajectory w_1 there could correspond two trajectories, w_2 and w_2', say, but since $w_2(t) - w_2'(t)$ converges to zero as t tends to infinity, we have what we could call asymptotic uniqueness. A system in which w_1 determines w_2 asymptotically in this sense is called *detectable.* The formal definition is given below.

Definition 5.3.16 Let $(\mathbb{R}, \mathbb{W}_1 \times \mathbb{W}_2, \mathfrak{B})$ be a time-invariant dynamical system. Trajectories in $\mathfrak{B}$ are partitioned as (w_1, w_2) with $w_i : \mathbb{R} \to \mathbb{W}_i$, $i = 1, 2$. We say that w_2 is *detectable* from w_1 if $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$ implies $\lim_{t\to\infty} w_2(t) - w_2'(t) = 0$. □

Theorem 5.3.3 states that w_2 is observable from w_1 if and only if $R_2(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$. Detectability requires this rank condition only for $\lambda \in \mathbb{C}_+$, where $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \mathrm{Re}\, s \geq 0\}$.

Theorem 5.3.17 *Let $R_1(\xi) \in \mathbb{R}^{g\times q_1}[\xi]$ and $R_2(\xi) \in \mathbb{R}^{g\times q_2}[\xi]$. Let $\mathfrak{B}$ be the behavior defined by $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$. Then the variable w_2 is detectable from w_1 if and only if* $\mathrm{rank}\, R_2(\lambda) = q_2$ *for all $\lambda \in \mathbb{C}_+$, where $\mathbb{C}_+ = \{s \in \mathbb{C} \mid \mathrm{Re}\, s \geq 0\}$.*

Proof The proof follows the proof of Theorem 5.3.3 almost verbatim.

Let $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$. Then, by linearity of $\mathfrak{B}$, also $(0, w_2 - w_2') \in \mathfrak{B}$, and hence $R_2(\frac{d}{dt})(w_2 - w_2') = 0$. It follows that w_2 is detectable from w_1 if and only if $R_2(\frac{d}{dt})w_2 = 0$ implies that $\lim_{t\to\infty} w_2(t) = 0$. Define $\mathfrak{B}_2 := \{w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^{q_2}) \mid R_2(\frac{d}{dt})w_2 = 0\}$. By the previous remark it suffices to prove that all trajectories in $\mathfrak{B}_2$ asymptotically converge to zero

if and only if $\operatorname{rank} R_2(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$ with nonnegative real part. By Theorem 2.5.23 there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ such that

$$U(\xi)R_2(\xi) = \begin{bmatrix} R_2'(\xi) \\ 0 \end{bmatrix} \quad \text{and } R_2'(\xi) \in \mathbb{R}^{g'\times q_2}[\xi] \text{ of full row rank.}$$

By Theorem 2.5.4 we have that $\mathfrak{B}_2 = \{w_2 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^d) \mid R_2'(\frac{d}{dt})w_2 = 0\}$. Moreover, $\operatorname{rank} R_2(\lambda) = \operatorname{rank} R_2'(\lambda)$, so that we have to prove that $\mathfrak{B}_2$ consists of asymptotically vanishing trajectories if and only if $\operatorname{rank} R_2'(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$ with nonnegative real part.

Suppose that $\mathfrak{B}_2$ contains asymptotically vanishing trajectories only. Then $\mathfrak{B}_2$ should be autonomous, since otherwise it would contain free variables that do not necessarily converge to zero. Therefore, by Theorem 3.3.22, $R_2'(\xi)$ should have the same number of rows as the number of columns; i.e., $g' = q_2$. The roots of $\det R_2'(\xi)$ are precisely the singular values, i.e., the values for which $R_2(\xi)$ loses rank, and by Theorem 3.2.16 it follows that these roots should all have real part strictly less than zero.

Conversely, suppose that $\operatorname{rank} R_2'(\lambda) = q_2$ for all $\lambda \in \mathbb{C}$ with nonnegative real part. Then, since $R_2'(\xi)$ is of full row rank, $g' = q_2$, and hence $\det R_2'(\xi)$ can only have roots with strictly negative real part, and by Theorem 3.2.16 it follows that $\mathfrak{B}_2$ can only contain trajectories that converge to zero. □

Example 5.3.18 In Example 5.3.5 we concluded that ℓ is not observable from w if $L = CR_CR_L$, for in that case $M(\lambda)$ has a rank deficiency for $\lambda = \frac{-1}{R_C}$. By inspection of the matrix (5.42) it follows that $M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}_+$ and hence ℓ is always detectable from w, even in the case that $L = CR_CR_L$. □

An immediate consequence of Theorem 5.3.17 is the following result.

Corollary 5.3.19 *Let $(A, B, C, D) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$, with (A, C) in the form (5.49), and consider the system $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$. Then x is detectable from (u, y) if and only if the matrix A_{11} has all its eigenvalues in the open left half-plane. In that case we call the pair (A, C) a detectable pair.*

In Chapter 10 we will use the notion of detectability in the context of output feedback stabilization.

5.4 The Kalman Decomposition

Consider the i/s/o system

$$\begin{array}{rcl} \frac{d}{dt}x & = & Ax + Bu, \\ y & = & Cx + Du. \end{array} \tag{5.54}$$

In Corollary 5.2.25 we have seen how the state space of an input/state system may be decomposed in a controllable part and an autonomous part, whereas in Corollary 5.3.14 we derived a similar decomposition for state/output systems, namely into an observable and a nonobservable part. In this section we combine these two decompositions to obtain what is known as the *Kalman decomposition* of input/state/output systems.

Theorem 5.4.1 *Let $(A, B, C, D) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$, and consider the system (5.54). There exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that*

$$S^{-1}AS = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & 0 & A_{44} \end{bmatrix}, \qquad S^{-1}B = \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix}, \tag{5.55}$$

$$CS = \begin{bmatrix} 0 & C_2 & 0 & C_4 \end{bmatrix},$$

and such that

$$\left(\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \right) \tag{5.56}$$

is controllable and

$$\left(\begin{bmatrix} A_{22} & A_{24} \\ 0 & A_{44} \end{bmatrix}, \begin{bmatrix} C_2 & C_4 \end{bmatrix} \right) \tag{5.57}$$

is observable.

Proof Let $\mathfrak{C}$ be the controllability matrix of (A, B) and $\mathfrak{O}$ the observability matrix of (A, C). Denote by k_1, $k_1 + k_2$, $k_1 + k_3$ the dimensions of $\operatorname{im}\mathfrak{C} \cap \ker\mathfrak{O}$, $\operatorname{im}\mathfrak{C}$, $\ker\mathfrak{O}$ respectively, and let $k_4 = n - (k_1 + k_2 + k_3)$. Choose vectors $a_1, \ldots, a_{k_1}$, $b_1, \ldots, b_{k_2}$, $c_1, \ldots, c_{k_3}$, and $d_1, \ldots, d_{k_4}$ such that $(a_1, \ldots, a_{k_1})$ is a basis of $\operatorname{im}\mathfrak{C} \cap \ker\mathfrak{O}$, $(a_1, \ldots, a_{k_1}, b_1, \ldots, b_{k_2})$ is a basis of $\operatorname{im}\mathfrak{C}$, $(a_1, \ldots, a_{k_1}, c_1, \ldots, c_{k_3})$ is a basis of $\ker\mathfrak{O}$, and $(a_1, \ldots, a_{k_1}, b_1, \ldots, b_{k_2}, c_1, \ldots, c_{k_3}, d_1, \ldots, d_{k_4})$ is a basis of $\mathbb{R}^n$. Let S be the matrix that has $a_1, \ldots, a_{k_1}, b_1, \ldots, b_{k_2}, c_1, \ldots, c_{k_3}, d_1, \ldots, d_{k_4}$ as its columns. Since $\operatorname{im}\mathfrak{C}$ and $\ker\mathfrak{O}$ are A-invariant, so are $\operatorname{im}\mathfrak{C} \cap \ker\mathfrak{O}$ and $\operatorname{im}\mathfrak{C} + \ker\mathfrak{O}$. Hence with this definition of S, (5.55) is satisfied. Controllability of the pair (5.56) follows in the same way as in the proof of Corollary 5.2.25, and observability of the pair (5.57) in the same way as in the proof of Corollary 5.3.14. □

Theorem 5.4.1 allows a nice visualization in terms of a flow diagram. Define four subspaces of $\mathbb{R}^n$ according to the partition of $(S^{-1}AS, S^{-1}B, CS, D)$ in (5.55): $\mathcal{X}_1 = \operatorname{span}(a_1, \ldots, a_{k_1})$, $\mathcal{X}_2 = \operatorname{span}(b_1, \ldots, b_{k_2})$, $\mathcal{X}_3 = \operatorname{span}(c_1, \ldots, c_{k_3})$, and $\mathcal{X}_4 = \operatorname{span}(d_1, \ldots, d_{k_4})$. The flow diagram expressing

how the input u reaches the output y via the subspaces $\mathcal{X}_i$ and the interactions among the components of the state space is depicted in Figure 5.3. The interpretation of the four subspaces is as follows. The subspace $\mathcal{X}_1 \oplus \mathcal{X}_2$

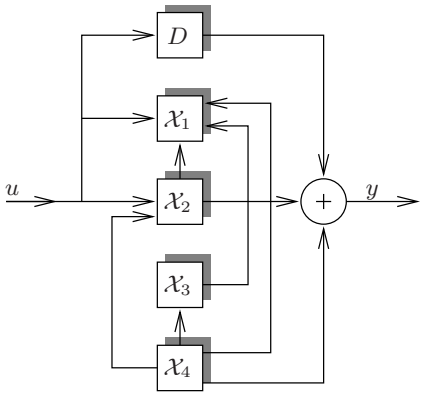


FIGURE 5.3. Flow diagram of the Kalman decomposition.

is the *controllable* part of the state space, and $\mathcal{X}_1 \oplus \mathcal{X}_3$ is the *nonobservable* part of the state space. Furthermore, $\mathcal{X}_1$ is the nonobservable part of the controllable part of the system. The three spaces $\mathcal{X}_1$, $\mathcal{X}_1 \oplus \mathcal{X}_2$, and $\mathcal{X}_2 \oplus \mathcal{X}_4$ are uniquely defined, independent of the choice of bases. The subspaces $\mathcal{X}_2$, $\mathcal{X}_3$, and $\mathcal{X}_4$ are nonunique.

The flow diagram indicates that only the feedthrough part D and the part processed by $\mathcal{X}_2$ contribute to the way the input u influences the output y. The following result formalizes this.

Corollary 5.4.2 *Consider the system (5.54) represented in the form (5.55), with $x(0) = 0$ and the controllable and observable subsystem defined by*

$$\begin{aligned} \frac{d}{dt}x_2 &= A_{22}x_2 + B_2u, \qquad x_2(0) = 0, \\ y &= C_2x_2 + Du. \end{aligned} \tag{5.58}$$

Then the responses y to any $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ in (5.54) and (5.58) are the same.

Proof From Corollary 4.5.5 it follows that for (5.54),

$$y(t) = Du(t) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau,$$

while for (5.58) we have,

$$y(t) = Du(t) + \int_0^t Ce^{A_{22}(t-\tau)}B_2u(\tau)d\tau.$$

Using Proposition 4.5.12, parts 3 and 2, we obtain

$$\begin{aligned} Ce^{At}B = & \begin{bmatrix} 0 & C_2 & 0 & C_4 \end{bmatrix} \begin{bmatrix} e^{A_{11}t} & * & * & * \\ 0 & e^{A_{22}t} & * & * \\ 0 & 0 & e^{A_{33}t} & * \\ 0 & 0 & 0 & e^{A_{44}t} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix} \\ = & C_2e^{A_{22}t}B_2, \end{aligned}$$

from which the statement immediately follows. □

Example 5.4.3 Consider the triple of matrices (A, b, c) given by

$$A = \begin{bmatrix} -2 & 3 & 4 & 1 \\ 1 & 6 & 6 & 3 \\ 5 & 6 & 6 & 4 \\ 0 & -17 & -19 & -8 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 3 & 3 & 2 & 2 \end{bmatrix}.$$

The controllability matrix $\mathfrak{C}$ of the pair (A, b) and the observability matrix $\mathfrak{O}$ of (A, c) are given by

$$\mathfrak{C} = \begin{bmatrix} 0 & -2 & -4 & -6 \\ -1 & -3 & -5 & -7 \\ 0 & -2 & -4 & -6 \\ 1 & 9 & 17 & 25 \end{bmatrix}, \quad \mathfrak{O} = \begin{bmatrix} 3 & 3 & 2 & 2 \\ 7 & 5 & 4 & 4 \\ 11 & 7 & 6 & 6 \\ 15 & 9 & 8 & 8 \end{bmatrix}.$$

It is easily checked that

$$\begin{aligned} &\operatorname{im}\mathfrak{C} \cap \ker\mathfrak{O} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \\ 1 \\ -4 \end{bmatrix}\right\}, \ \operatorname{im}\mathfrak{C} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \\ 1 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}\right\}, \\ &\ker\mathfrak{O} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \\ 1 \\ -4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ -3 \end{bmatrix}\right\}, \ \mathbb{R}^n = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \\ 1 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ -3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}\right\}. \end{aligned} \tag{5.59}$$

With respect to the basis of $\mathbb{R}^n$ in (5.59), the triple (A, b, c) takes the form

$$\tilde{A} = \begin{bmatrix} 1 & 2 & -1 & 4 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \tilde{b} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \qquad \tilde{c} = \begin{bmatrix} 0 & 1 & 0 & 2 \end{bmatrix}.$$

□

5.5 Polynomial Tests for Controllability and Observability

We now give some interesting alternative tests for controllability and observability for the single-input/single-output case. These tests will be used in Chapter 6, Theorem 6.3.1.

Theorem 5.5.1

1. *Let* $(A, c) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{1\times n}$ *and define* $p(\xi) \in \mathbb{R}[\xi]$ *and* $r(\xi) \in \mathbb{R}^{1\times n}[\xi]$ *as*

   $$p(\xi) := \det(I\xi - A), \quad r(\xi) = [r_1(\xi), \ldots, r_n(\xi)] := p(\xi)c(I\xi - A)^{-1}.$$

   Then (A, c) *is observable if and only if the* $n + 1$ *polynomials* $p(\xi)$, $r_1(\xi)$,...,$r_n(\xi)$ *have no common roots.*

2. *Let* $(A, b) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ *and define* $p(\xi) \in \mathbb{R}[\xi]$ *and* $s(\xi) \in \mathbb{R}^{n\times 1}[\xi]$ *as*

   $$p(\xi) := \det(I\xi - A), \quad s(\xi) = [s_1(\xi), \ldots, s_n(\xi)]^T := (I\xi - A)^{-1}bp(\xi).$$

   Then (A, b) *is controllable if and only if the* $n + 1$ *polynomials* $p(\xi)$, $s_1(\xi)$,...,$s_n(\xi)$ *have no common roots.*

Proof For $n = 1$ the statement is trivially true, so assume that $n \geq 2$.

Part 1. First we prove that if (A, c) is not observable, then $p(\xi), r_1(\xi), \ldots, r_n(\xi)$ should have a common factor. Consider the complex matrix

$$M(\lambda) = \begin{bmatrix} I\lambda - A \\ c \end{bmatrix}.$$

Suppose that (A, c) is not observable. By Theorem 5.3.7 there exists $\lambda_0 \in \mathbb{C}$ such that $M(\lambda_0)$ loses rank. Hence there exists an eigenvector v of A such that $cv = 0$. Consequently, there exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that

$$SAS^{-1} = \begin{bmatrix} \lambda_0 & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \text{ and } cS^{-1} = \begin{bmatrix} 0 & \tilde{c}_2 \end{bmatrix}.$$

This implies that

$$\begin{aligned}\frac{r(\xi)}{p(\xi)} &= c(I\xi - A)^{-1} = cS^{-1}(S(I\xi - A)S^{-1})^{-1}S \\ &= \begin{bmatrix} 0 & \tilde{c}_2 \end{bmatrix} \begin{bmatrix} \xi - \lambda_0 & -\tilde{A}_{12} \\ 0 & I\xi - \tilde{A}_{22} \end{bmatrix}^{-1} S \\ &= \begin{bmatrix} 0 & \tilde{c}_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\xi-\lambda_0} & \frac{1}{\xi-\lambda_0}\tilde{A}_{12}(I\xi - \tilde{A}_{22})^{-1} \\ 0 & (I\xi - \tilde{A}_{22})^{-1} \end{bmatrix} S \\ &= \begin{bmatrix} 0 & \tilde{c}_2(I\xi - \tilde{A}_{22})^{-1} \end{bmatrix} S = \frac{\tilde{r}(\xi)}{\tilde{p}(\xi)}\end{aligned} \tag{5.60}$$

for some polynomial vector $\tilde{r}(\xi)$ and $\tilde{p}(\xi) = \det(I\xi - \tilde{A}_{22})$. Since $\deg \tilde{p}(\xi)$ is obviously less than n, the degree of $p(\xi)$, (5.60) implies that the $n+1$ polynomials $p(\xi)$ and $r_1(\xi), \ldots, r_n(\xi)$ have a common factor.

Next we show that if $p(\xi), r_1(\xi), \ldots, r_n(\xi)$ have a common factor, then (A, c) cannot be observable. Suppose that $p(\xi)$ and $r_1(\xi), \ldots, r_n(\xi)$ have a common factor. Then there exists $\lambda_0 \in \mathbb{C}$ such that $p(\lambda_0) = 0$ and $r(\lambda_0) = 0$. Define the polynomial matrix $F(\xi)$ by $F(\xi) := p(\xi)(I\xi - A)^{-1}$. If A happens to be equal to $\lambda_0 I$, i.e., if $F(\lambda_0)$ is the zero matrix, then it follows immediately from Theorem 5.3.7 that (A, c) is not observable (recall that we assumed that $n \geq 2$). If $F(\lambda_0) \neq 0$, then there exists $v \in \mathbb{C}^n$ such that $F(\lambda_0)v \neq 0$. Consequently,

$$M(\lambda_0)F(\lambda_0)v = \begin{bmatrix} I\lambda_0 - A \\ c \end{bmatrix} F(\lambda_0)v = \begin{bmatrix} Ip(\lambda_0) \\ r(\lambda_0) \end{bmatrix} v = 0. \tag{5.61}$$

By Theorem 5.3.7, (5.61) implies that (A, c) is not observable.

Part 2 follows from part 1 and the observation that (A, b) is controllable if and only if (A^T, b^T) is observable; see Exercise 5.19. □

5.6 Recapitulation

In this chapter we introduced and studied the notions of *controllability* and *observability*. The main points were:

- Controllability is defined as the possibility of switching from the past of one trajectory to the future of another trajectory in the behavior of a system by allowing a time delay during which this switch takes place (Definition 5.2.2).

- The system defined by $R(\frac{d}{dt})w = 0$ is controllable if and only if the rank of $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}$ (Theorem 5.2.10).
- An i/s system $\frac{d}{dt}x = Ax + Bu$ is controllable if and only if the rank of its controllability matrix $\mathfrak{C} = [B\ AB \cdots A^{n-1}B]$ is equal to the dimension n of the state space (Theorems 5.2.18 and 5.2.27). This provides a convenient explicit test for the controllability of such systems. The same test also applies to i/s/o systems $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$.
- For i/o and i/s/o systems we introduced the notion of state controllability as the possibility of driving the state of the system from an arbitrary initial state to an arbitrary terminal state. State controllability is equivalent to controllability (Theorem 5.2.27).
- A system is called stabilizable if every trajectory in the behavior can be concatenated with a trajectory in the behavior that converges to zero as time tends to infinity (Definition 5.2.29).
- The system defined by $R(\frac{d}{dt})w = 0$ is stabilizable if and only if the rank of $R(\lambda)$ is the same for all $\lambda \in \mathbb{C}$ with nonnegative real part (Theorem 5.2.30).
- In a behavior where the variable w is partitioned as $w = (w_1, w_2)$, w_2 is called observable from w_1 if w_2 is uniquely determined by w_1 and the laws of the system (Definition 5.3.2).
- In the behavior of $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$, observability of w_2 from w_1 is equivalent to the condition that the rank of $R_2(\lambda)$ is equal to the number of columns of $R_2(\xi)$ for all $\lambda \in \mathbb{C}$ (Theorem 5.3.3).
- Observability of the state from the input and the output in an i/s/o system is usually referred to as just observability. An i/s/o system $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$ is observable if and only if the rank of its observability matrix $\mathfrak{O} = \mathrm{col}(C, CA, \ldots, CA^{n-1})$ is equal to the dimension n of the state space (Theorem 5.3.9). This provides a convenient explicit test for the observability of such systems.
- In a behavior where the variable w is partitioned as $w = (w_1, w_2)$, w_2 is called detectable from w_1 if w_2 is determined by w_1 and the laws of the system up to an asymptotically vanishing trajectory. By that we mean that if (w_1, w_2) and (w_1, w_2') belong to the behavior, then $w_2(t) - w_2'(t)$ converges to zero as t tends to infinity (Definition 5.3.16).
- In the behavior of $R_1(\frac{d}{dt})w_1 = R_2(\frac{d}{dt})w_2$, detectability of w_2 from w_1 is equivalent to the condition that the rank of $R_2(\lambda)$ is equal to the number of columns of $R_2(\xi)$ for all $\lambda \in \mathbb{C}$ with nonnegative real part (5.3.17).
- By appropriately choosing the basis in the state space, the controllability and/or observability structure of an i/s/o system may be brought into evidence. This is referred to as the Kalman decomposition (Theorem 5.4.1).

5.7 Notes and References

The notion of controllability and observability and the tests in terms of the controllability and observability matrices were introduced for input/state/output

systems by Kalman [27]. They were soon to become some of the very central notions in systems theory. They have been treated in numerous texts, for example in [15] and [25]. This last reference treats these topics in considerably more detail than we do here. The natural generalizations of these concepts to general behaviors was first presented in [59]. The Hautus test appeared in [20] and is sometimes referred to as the PBH (Popov–Belevich–Hautus) test, since it was independently derived also in [46] and [8]. More historical details about these notions may be found in [25, 29].

5.8 Exercises

As a simulation exercise illustrating the material covered in this chapter we suggest A.4.

5.1 Prove that the RLC-network of Example 5.2.13, described by (5.12), is controllable.

5.2 Consider the mechanical system depicted in Figure 5.4. Let q_1 and q_2 denote

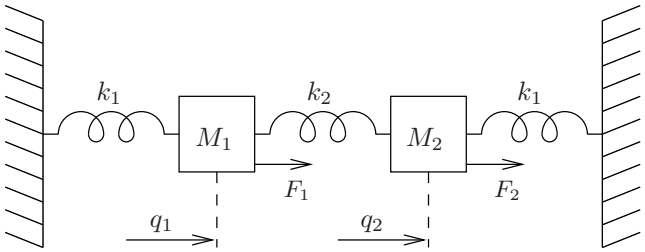


FIGURE 5.4. Mass–spring system.

the displacements of the masses M_1 and M_2 away from their equilibria. Assume that the system parameters M_1, M_2, k_1, k_2 are all positive.

(a) Derive a differential equation for the behavior of (q_1, q_2) in the absence of external forces.

(b) Assume that an external force F_1 is exerted on M_1. Obtain the behavioral differential equations for (q_1, q_2, F_1).

(c) Is this system controllable?

(d) For which values of the parameters M_1, M_2, k_1, k_2 is q_1 observable from q_2?

(e) Assume also that a force F_2 is exerted on M_2. Obtain the behavioral differential equations for (q_1, q_2, F_1, F_2).

(f) Is this system controllable?

(g) For which values of the parameters M_1, M_2, k_1, k_2 is q_1 observable from q_2?

5.3 Let the polynomials $a(\xi), b(\xi) \in \mathbb{R}[\xi]$ be given by

$$a(\xi) = a_0 + a_1\xi + \cdots + a_n\xi^n, \quad b(\xi) = b_0 + b_1\xi + \cdots + b_m\xi^m. \tag{5.62}$$

It follows immediately from Theorem 5.2.10 that the system

$$a(\frac{d}{dt})w_1 + b(\frac{d}{dt})w_2 = 0$$

in the scalar variables w_1 and w_2 is controllable if and only if the polynomials $a(\xi)$ and $b(\xi)$ are coprime. Coprimeness of polynomials occurs very naturally in the context of controllability and observability, as well as in many other problems in mathematics. In this exercise we want to derive a necessary and sufficient condition on the coefficients of $a(\xi)$ and $b(\xi)$ for the coprimeness of $a(\xi)$ and $b(\xi)$.

From Exercises 2.23 and 2.24 we know that the polynomial equation

$$a(\xi)p(\xi) + b(\xi)q(\xi) = 1 \tag{5.63}$$

has a solution $(p(\xi), q(\xi))$ with $\deg p(\xi) < m$ and $\deg q(\xi) < n$ if and only if $a(\xi)$ and $b(\xi)$ are coprime.

(a) Show that by equating the terms of equal degree on both sides of (5.63) we obtain a system of linear equations with the coefficients of $p(\xi)$ and $q(\xi)$ as unknowns:

$$\begin{bmatrix} p_0 & \cdots & p_{m-1} & q_{n-1} & \cdots & q_0 \end{bmatrix} S = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix},$$

with S given by

$$S = \begin{bmatrix} a_0 & a_1 & \cdots & a_n & & & \\ & a_0 & \cdots & a_{n-1} & a_n & & \\ & & \ddots & & & \ddots & \\ & & & a_0 & a_1 & \cdots & a_n \\ & & & b_0 & b_1 & \cdots & b_m \\ & & b_0 & b_1 & \cdots & b_m & \\ & \iddots & & & \iddots & & \\ b_0 & b_1 & \cdots & \cdots & b_m & & \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\\a\\a\end{matrix}}\right\} \text{m rows} \\ \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\\a\\a\end{matrix}}\right\} \text{n rows} \end{matrix} \tag{5.64}$$

S is called the *Sylvester resultant* of $a(\xi)$ and $b(\xi)$.

(b) Prove that the polynomials $a(\xi)$, $b(\xi)$ are coprime if and only if the Sylvester resultant S defined by (5.64) has rank $n + m$.

5.4 In Theorem 5.2.18 we have seen that the first-order system $\frac{d}{dt}x = Ax + Bu$ is controllable if and only if the controllability matrix $\mathfrak{C}$ defined by (5.21) has rank n. Similar tests can be derived for higher-order systems (see for example Exercise 5.3), but often these tend to become rather unwieldy. However, in some special cases controllability tests can be derived that are as simple as the result obtained for the first-order case.

(a) Let $r(\xi) \in \mathbb{R}[\xi]$, $w = \mathrm{col}(w_1, w_2)$, where w_1 is q_1-dimensional and w_2 is q_2-dimensional, $A \in \mathbb{R}^{q_1 \times q_1}$ and $B \in \mathbb{R}^{q_1 \times q_2}$. Assume that $r(\xi)$ is a polynomial of degree at least one. Prove that

$$r(\frac{d}{dt})w_1 + Aw_1 = Bw_2$$

is controllable if and only if $\mathrm{rank}[B \;\; AB \;\; \cdots \;\; A^{q_1-1}B] = q_1$. Hint: Mimic the proof of Theorem 5.2.18 and use (or prove) the fact that every polynomial of degree at least one defines a surjective function from $\mathbb{C}$ to $\mathbb{C}$.

(b) Mechanical systems are often described by second-order differential equations. In the absence of damping, they lead to models of the form

$$M\frac{d^2q}{dt^2} + Kq = BF$$

with q the vector of (generalized) positions, assumed n-dimensional; F the external forces; and M, K, and B matrices of suitable dimension; M is the mass matrix and K the matrix of spring constants. Assume that M is square and nonsingular. Prove that with $w = \mathrm{col}(q, F)$, this system is controllable if and only if

$$\mathrm{rank}[B \;\; KM^{-1}B \cdots (KM^{-1})^{n-1}B] = n.$$

5.5 Consider the i/o behavior $\mathfrak{B}$ defined by

$$-y + \frac{d^2}{dt^2}y = -u + \frac{d}{dt}u.$$

(a) Is this system controllable?

(b) Write $\mathfrak{B}$ as the direct sum of an autonomous part and a controllable part by applying the proof of Theorem 5.2.14 to this system.

(c) Define $\mathfrak{B}_{\mathrm{aut}} := \{(u, y) \mid -y + \frac{d}{dt}y = 0,\ u = 0\}$ and $\mathfrak{B}_{\mathrm{contr}} := \{(u, y) \mid y + \frac{d}{dt}y = u\}$. Prove that $\mathfrak{B} = \mathfrak{B}_{\mathrm{aut}} \oplus \mathfrak{B}_{\mathrm{contr}}$.

5.6 (a) Consider the behavior $\mathfrak{B}$ of $R(\frac{d}{dt})w = 0$ with $R(\xi) = [\xi^2 - 1 \;\; \xi + 1]$. Provide two *different* decompositions of $\mathfrak{B}$ as a direct sum of a controllable and an autonomous part. Hint: Carefully examine the proof of Theorem 5.2.14 to see where the nonuniqueness in the construction of $\mathfrak{B}_{\mathrm{aut}}$ occurs.

(b) Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ be of full row rank and let $\mathfrak{B}$ be the behavior of $R(\frac{d}{dt})w = 0$. Let $U(\xi) \in \mathbb{R}^{g \times g}[\xi]$ and $V(\xi) \in \mathbb{R}^{q \times q}[\xi]$ be unimodular matrices that transform $R(\xi)$ into Smith form:

$$U(\xi)R(\xi)V(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix}.$$

As we have seen in Theorem 5.2.14, a decomposition of the behavior $\mathfrak{B}$ into a controllable and an autonomous part is obtained by defining

$$R_{\mathrm{contr}}(\xi) = \begin{bmatrix} I & 0 \end{bmatrix} V^{-1}(\xi), \quad R_{\mathrm{aut}}(\xi) = \begin{bmatrix} D(\xi) & 0 \\ 0 & I \end{bmatrix} V^{-1}(\xi).$$

Let $W(\xi) \in \mathbb{R}^{q\times q}[\xi]$ be a unimodular matrix with the property that

$$\begin{bmatrix} D(\xi) & 0 \end{bmatrix} W(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix},$$

and define

$$R'_{\mathrm{aut}}(\xi) = \begin{bmatrix} D(\xi) & 0 \\ 0 & I \end{bmatrix} W^{-1}(\xi)V^{-1}(\xi).$$

Prove that $R_{\mathrm{contr}}(\xi), R'_{\mathrm{aut}}(\xi)$ also provides a decomposition of $\mathfrak{B}$ into a direct sum of a controllable and an autonomous part.

(c) In order to classify *all* possible decompositions of $\mathfrak{B}$ into a direct sum of a controllable and an autonomous part, we first classify all such decompositions of $\tilde{\mathfrak{B}}$, the behavior of $[D(\xi)\ 0]$. Let $\tilde{R}_{\mathrm{contr}}(\xi), \tilde{R}_{\mathrm{aut}}(\xi)$ define such a decomposition. Assume that both $\tilde{R}_{\mathrm{contr}}(\xi)$ and $\tilde{R}_{\mathrm{aut}}(\xi)$ are of full row rank. Prove that there exist unimodular matrices $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$, $U'(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and $W(\xi) \in \mathbb{R}^{q\times q}[\xi]$ such that

$$\tilde{R}_{\mathrm{contr}}(\xi) = U(\xi) \begin{bmatrix} I & 0 \end{bmatrix}, \quad \begin{bmatrix} D(\xi) & 0 \end{bmatrix} W(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix},$$

$$\tilde{R}_{\mathrm{aut}}(\xi) = U'(\xi) \begin{bmatrix} D(\xi) & 0 \\ 0 & I \end{bmatrix} W(\xi).$$

(d) Let $\mathfrak{B} = \mathfrak{B}_{\mathrm{contr}} \oplus \mathfrak{B}_{\mathrm{aut}}$ be a decomposition into a controllable part and an autonomous part defined by polynomial matrices $R'_{\mathrm{contr}}(\xi)$ and $R'_{\mathrm{aut}}(\xi)$. Assume that both $\tilde{R}_{\mathrm{contr}}(\xi)$ and $\tilde{R}_{\mathrm{aut}}(\xi)$ are of full row rank. Prove that there exist unimodular matrices $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$, $U'(\xi) \in \mathbb{R}^{q\times q}[\xi]$ and $W(\xi) \in \mathbb{R}^{q\times q}[\xi]$ such that

$$R'_{\mathrm{contr}}(\xi) = U(\xi)R_{\mathrm{contr}}(\xi), \quad \begin{bmatrix} D(\xi) & 0 \end{bmatrix} W(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix},$$

$$R'_{\mathrm{aut}}(\xi) = U'(\xi)R_{\mathrm{aut}}(\xi)W^{-1}(\xi).$$

(e) Characterize all unimodular matrices $W(\xi) \in \mathbb{R}^{q\times q}[\xi]$ with the property that

$$\begin{bmatrix} D(\xi) & 0 \end{bmatrix} W(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix}.$$

5.7 Consider the electrical circuit shown in Figure 5.5. Take as input $u = V$ and as output $y = I$.

(a) Choose, based on physical considerations, a state for this system.

(b) Derive the i/s/o equations.

(c) For which values of R_1, R_2, C_1, C_2 is this system controllable?

(d) For which values of R_1, R_2, C_1, C_2 is this system observable?

5.8 Consider the i/s system $\frac{d}{dt}x = Ax + Bu$, with

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 6 \end{bmatrix}.$$

(a) Is this system controllable?

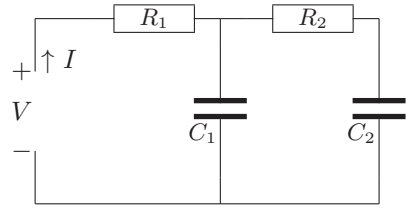


FIGURE 5.5. Electrical circuit.

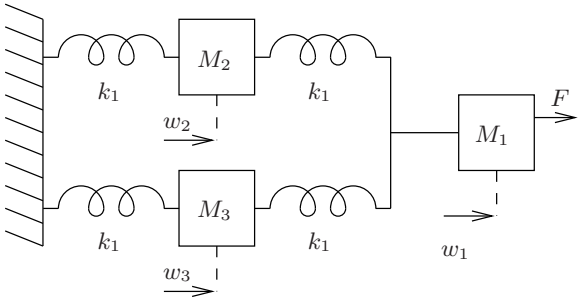


FIGURE 5.6. Mechanical system.

(b) Calculate an input function u that takes the state of the system in $\log 2$ time units from the zero state to $[1\ 0]^T$.

5.9 Consider the mechanical system in Figure 5.6. Take all spring constants to be equal to unity. Assume that an external force F acts on M_1. All displacements are in the horizontal direction only; rotations and vertical movements are excluded.

(a) Derive the equations of motion.

(b) Show that if $M_2 = M_3$, then (w_2, w_3) is not observable from (w_1, F).

(c) Which motions (w_2, w_3) are compatible with $w_1 = 0$ and $F = 0$?

(d) For which values of M_2, M_3 is w_3 observable from (w_1, w_2); that is, for which values of M_2, M_3 does $w_1 = w_2 = 0$ imply $w_3 = 0$?

5.10 Consider the mechanical system depicted in Figure 5.7. The variables w_1, w_2, and w_3 denote the displacements from the respective equilibria. All displacements are in the horizontal direction only; rotations and vertical movements are excluded. Let $M_2 = 2$, $M_3 = 1/2$, and $k_1 = 1$. Take as the state $x := [w_2, \frac{d}{dt}w_2, w_3, \frac{d}{dt}w_3]^T$ and as input $u = w_1$. The input/state equations are then given by

$$\frac{d}{dt}x = Ax + Bu,$$

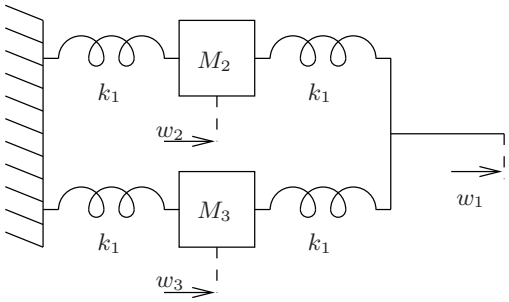


FIGURE 5.7. Mechanical system.

with

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -4 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 2 \end{bmatrix}.$$

(a) Determine e^{At}. Express the entries in simple trigonometric formulas.

(b) Assume that at time $t = 0$, the masses pass their respective equilibria in opposite directions with velocity equal to one. Determine an input function u that brings the masses to rest into their equilibrium positions at $t = 2\pi$.

(c) Check without calculation whether or not there exists an input function u that drives the system from equilibrium at $t = 0$ to state $[1, 0, -1, 0]^T$ at $t = 1$.

(d) Check without calculation whether or not there exists an input function u that drives the system from equilibrium at $t = 0$ to state $[1, 0, -1, 0]^T$ at $t = 1$ and keeps it there for $t \geq 1$.

(e) Characterize all states with the property that there exists an input function u that drives the system from the equilibrium position to that state and keeps it there.

5.11 (a) Let $n > 1$ and let $A = I\lambda \in \mathbb{R}^{n\times n}$. Prove that for all $b \in \mathbb{R}^{n\times 1}$, the pair (A, b) is not controllable.

(b) Let $B \in \mathbb{R}^{n\times m}$ and $\lambda \in \mathbb{R}$. Show that a necessary condition for controllability of $(\lambda I, B)$ is $m \geq n$.

(c) Prove that if $A \in \mathbb{R}^{n\times n}$ is such that (A, b) is controllable for all nonzero $b \in \mathbb{R}^{n\times 1}$, then $n \leq 2$. Give an example of an $A \in \mathbb{R}^{2\times 2}$ with this property.

5.12 (a) Let $A \in \mathbb{R}^{n\times n}$. Define $\tilde{A} \in \mathbb{R}^{2n\times 2n}$ by

$$\tilde{A} := \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}.$$

Prove that $(\tilde{A}, \tilde{b})$ is *not* controllable for any $\tilde{b} \in \mathbb{R}^{2n\times 1}$.

(b) Let $A_i \in \mathbb{R}^{n_i \times n_i}$, $i = 1, 2$. Define $\tilde{A} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ by

$$\tilde{A} := \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}.$$

Suppose that A_1 and A_2 have a common eigenvalue. Prove that $(\tilde{A}, \tilde{b})$ is *not* controllable for any $\tilde{b} \in \mathbb{R}^{(n_1+n_2)\times 1}$.

5.13 Consider the mechanical system shown in Figure 5.8. The masses are

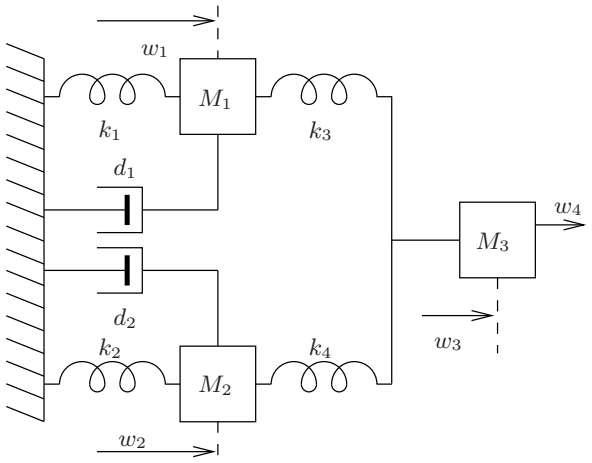


FIGURE 5.8. Mechanical system.

M_1, M_2, M_3; the spring constants are denoted by k_1, k_2, k_3, k_4; and d_1, d_2 are the damper constants. All these parameters are strictly positive. The variables w_1, w_2, w_3 denote the displacements from the respective equilibrium positions. All displacements are in the horizontal direction only; rotations and vertical movements are excluded. On the third mass we can exert a force w_4. The system equations are

$$\begin{aligned} M_1 \frac{d^2}{dt^2} w_1 &= -k_1 w_1 - d_1 \frac{d}{dt} w_1 + k_3 w_3 - k_3 w_1, \\ M_2 \frac{d^2}{dt^2} w_2 &= -k_2 w_2 - d_2 \frac{d}{dt} w_2 + k_4 w_3 - k_4 w_2, \\ M_3 \frac{d^2}{dt^2} w_3 &= w_4 + k_3 w_1 + k_4 w_2 - k_3 w_3 - k_4 w_3. \end{aligned}$$

(a) Determine a matrix $R(\xi) \in \mathbb{R}^{3\times 4}[\xi]$ such that with $w = \mathrm{col}(w_1, w_2, w_3, w_4)$, the system is described by $R(\frac{d}{dt})w = 0$.

(b) Define polynomials $r_i(\xi) = k_i + k_{i+2} + d_i\xi + M_i\xi^2$, $i = 1, 2$. Show that the system is controllable if and only if $r_1(\xi)$ and $r_2(\xi)$ are coprime. Give a physical interpretation of this coprimeness condition.

(c) According to Corollary 2.5.12, $r_1(\xi)$ and $r_2(\xi)$ are coprime if and only if the equation

$$a(\xi)r_1(\xi) + b(\xi)r_2(\xi) = 1 \tag{5.65}$$

has a solution $(a(\xi), b(\xi))$. Write $a(\xi) = a_0 + a_1\xi$ and $b(\xi) = b_0 + b_1\xi$. Rewrite (5.65) as a system of linear equations with a_0, a_1, b_0, b_1 as the unknowns and the various physical parameters as coefficients.

(d) Show that (5.65) has a solution if and only if the coefficient matrix of the linear equations that you derived in the previous question is nonsingular.

(e) Show that the values of the parameters M_i, d_i, k_i, k_{i+2}, $i = 1, 2$, for which the system is *not* controllable satisfy an algebraic equation, i.e., an equation that involves polynomial expressions in the parameters M_i, d_i, k_i, k_{i+2}, $i = 1, 2$.

(f) Assume that all parameters are equal to one. Show that (w_1, w_2) is not observable from (w_3, w_4). Give a physical interpretation explaining which motions of (w_1, w_2) are possible when w_3 and w_4 are zero.

(g) Take as input $u := w_4$ and as output $y := (w_1, w_2)$. Determine an i/s/o representation of the system with this input and output.

5.14 Let $(A, B, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n}$ be given by

$$A = \begin{bmatrix} 0 & \dots & 0 & \dots & \dots & -p_0 \\ 1 & 0 & 0 & \dots & \dots & -p_1 \\ 0 & 1 & 0 & \dots & \dots & -p_2 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & 1 & 0 & -p_{n-2} \\ 0 & 0 & 0 & 0 & 1 & -p_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ \\ \vdots \\ q_{n-2} \\ q_{n-1} \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & \dots & \dots & \dots & 0 & 1 \end{bmatrix}.$$

Define $p(\xi) := \det(I\xi - A)$ and $q(\xi) := p(\xi)C(I\xi - A)^{-1}B$.

(a) Show that $p(\xi) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$ and $q(\xi) = q_0 + q_1\xi + \cdots + q_{n-1}\xi^{n-1}$.

(b) Prove that $p(\xi)$ and $q(\xi)$ have no common factors if and only if (A, B) is controllable. Hint: Use the Hautus test for controllability.

5.15 In this exercise all matrices are assumed to be of appropriate sizes.

(a) Let S be a nonsingular matrix. Prove that (A, B) is controllable if and only if (SAS^{-1}, SB) is.

(b) Prove that (A, B) is controllable if and only if $(A + BF, B)$ is controllable.

(c) Let R be nonsingular. Prove that (A, B) is controllable if and only if (A, BR) is controllable.

(d) Let S and R be nonsingular. Prove that (A, B) is controllable if and only if $(S(A + BF)S^{-1}, SBR)$ is controllable.

5.16 (a) Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ and $\ell \in \mathbb{N}$. Prove that if

$$\operatorname{rank}\left[\begin{array}{cccc} B & AB & \cdots & A^{\ell-1}B \end{array}\right] = \operatorname{rank}\left[\begin{array}{cccc} B & AB & \cdots & A^{\ell}B \end{array}\right]$$

then

$$\operatorname{rank}\left[\begin{array}{cccc} B & AB & \cdots & A^{\ell}B \end{array}\right] = \operatorname{rank}\left[\begin{array}{cccc} B & AB & \cdots & A^{\ell+1}B \end{array}\right]$$

(b) Let $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ with $\operatorname{rank} B = r$. Prove that (A, B) is controllable if and only if

$$\operatorname{rank}\left[\begin{array}{cccc} B & AB & \cdots & A^{n-r}B \end{array}\right] = n.$$

(c) Refer to the proof of Corollary 5.2.25. To be consistent with the notation in the proof of Corollary 5.2.25, let $(A, B) \in \mathbb{R}^{k\times k} \times \mathbb{R}^{k\times m}$ and let $n \geq k$. Prove that $\operatorname{rank}[B\ AB \cdots A^{n-1}B] = k \Rightarrow \operatorname{rank}[B\ AB \cdots A^{k-1}B] = k$ (this implies that (A, B) is controllable). Hint: Use the Cayley–Hamilton theorem.

5.17 Consider the linearized equations (5.11) for the two-pendulum system of Example 5.2.12. Is w_2 observable from w_3 and $w_1 - w_3$?

5.18 Let $(A, B, C) \in \mathbb{R}^{1\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{1\times n}$ be given by

$$A = \left[\begin{array}{cccccc} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 \cdots & 0 & \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & \cdots & -p_{n-1} \end{array}\right], \quad B = \left[\begin{array}{c} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{array}\right],$$

$$C = \left[\begin{array}{cccccc} q_0 & q_1 & \cdots & \cdots & \cdots & q_{n-1} \end{array}\right].$$

Define $p(\xi) := \det(I\xi - A)$ and $q(\xi) := p(\xi)C(I\xi - A)^{-1}B$.

(a) Show that $p(\xi) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$ and $q(\xi) = q_0 + q_1\xi + \cdots + q_{n-1}\xi^{n-1}$.

(b) Prove that $p(\xi)$ and $q(\xi)$ have no common factors if and only if (A, C) is observable. Hint: Use the Hautus test for observability.

5.19 Prove that (A, B) is controllable if and only if (A^T, B^T) is observable.

5.20 Let $(A, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$. Let $\mathcal{V}$ be the linear subspace of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^p)$ defined by $\{Ce^{At}x \mid x \in \mathbb{R}^n\}$. Prove that (A, C) observable if and only if $\dim \mathcal{V} = n$. Interpret this result as linear independence in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^p)$ of the columns of Ce^{At}.

5.21 Let A, B, C be given by

$$A = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix},$$

(a) Is (A, B) controllable?

(b) Is (A, C) observable?

(c) Determine a basis of $\mathbb{R}^3$ with respect to which (A, B, C) takes the form

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix} \text{ with } (A_{11}, B_1) \text{ controllable.}$$

Determine $A_{11}, A_{12}, A_{22}, B_1, C_1$, and C_2.

(d) Determine a basis of $\mathbb{R}^3$ with respect to which (A, B, C) takes the form

$$\begin{bmatrix} A'_{11} & A'_{12} \\ 0 & A'_{22} \end{bmatrix}, \begin{bmatrix} B'_1 \\ B'_2 \end{bmatrix}, \begin{bmatrix} 0 & C'_2 \end{bmatrix} \text{ with } (A'_{22}, C'_2) \text{ observable.}$$

Determine $A'_{11}, A'_{12}, A'_{22}, B'_1, B'_2$, and C'_2.

(e) Determine a basis of $\mathbb{R}^3$ with respect to which (A, B, c) takes the Kalman form (5.55). Determine the various matrices in (5.55).

5.22 In the proof of Theorem 5.2.5 we tacitly used the fact that $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ forms a controllable behavior. In this exercise we check this for the case $q = 1$. Let $w_1, w_2 \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ and $t_1 > 0$. Prove that there exists $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ such that

$$w(t) = \begin{cases} w_1(t) & t \leq 0, \\ w_2(t - t_1) & t \geq t_1. \end{cases}$$

Hint: Prove this first with $w_2 = 0$, and use the function (2.18) given in Definition 2.4.5 as a starting point for your construction.

5.23 Consider the i/s/o system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u, \qquad y = \begin{bmatrix} 0 & 1 \end{bmatrix} x.$$

Assume that during the time interval $[0, 1]$ the input u was identically equal to 1, and the output turned out to be identically equal to $-\frac{1}{2}$. Determine the state at time $t = 0$ and $t = 1$.

5.24 Let $A \in \mathbb{R}^{n \times n}$ and $v \in \mathbb{R}^n$. Define the subspace $\mathcal{V}$ as $\mathcal{V} := \text{span}\{A^k v \mid k \geq 0\}$, ($\mathcal{V}$ is the smallest subspace containing the vectors $A^k v$, $k = 0, 1, 2, \ldots$). Prove that $\mathcal{V}$ is A-invariant. In particular, the state trajectory x_k of the discrete-time system $x_{k+1} = Ax_k$ spans an A-invariant subspace.

5.25 Consider the discrete-time i/s/o system

$$x(k + 1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k), \quad k \in \mathbb{Z}.$$

Derive tests for controllability and observability for this system.

6

Elimination of Latent Variables and State Space Representations

6.1 Introduction

In this chapter we take a closer look at dynamical systems with latent variables as introduced in Chapter 1 and briefly discussed in Chapter 4.

As we have repeatedly observed in this book, latent variables show up naturally in modeling systems from first principles. We consider two problems that occur in the context of latent variables. The first one has to do with the elimination of latent variables. The second has to do with the introduction of a convenient class of manifest variables, specifically state variables.

We have already encountered the elimination problem in Chapter 1, in the context of Examples 1.3.5 and 1.3.6. In the first of these examples, we saw that a mathematical model for a linear electrical circuit can readily be written down from the constitutive law of the electrical devices in the branches, and Kirchhoff's current and voltage laws. This leads to a model that contains, in addition to the manifest variables, the current and voltage at the external port, as well as many latent variables, notably the currents and the voltages in the external branches. For the case at hand, we were actually able to eliminate—in an ad hoc fashion—these latent variables and obtain a differential equation describing the manifest behavior that contains only the manifest variables. In Example 1.3.6, however, such an elimination could not be done. The main result of this chapter shows that elimination of latent variables in linear time-invariant differential systems is indeed always possible. We also provide a systematic algorithm for how

to do this. This leads to a general theory of eliminating latent variables in linear systems. This is treated in Section 6.2.2.

State models form an especially important class of latent variable models. The general procedure for eliminating latent variables can, of course, be applied to this case. However, for state space systems, the converse problem, the one of *introducing* variables, is also of paramount importance. Indeed, very general analysis and synthesis techniques for state models are available. We study the question of introducing state variables in the context of i/o systems. This leads to the *input/state/output representation problem* treated in Section 6.4.

Section 6.5 is devoted to equivalent and minimal state space representations. In Chapter 4 we already formulated sufficient conditions for two state space representations to be equivalent. Here we present necessary conditions. Minimality of state space representations refers to the dimension of the state space representation of a given behavior. It turns out that minimality is equivalent to observability.

The last section of this chapter is concerned with what we call *image representations*. Up to now, we have studied systems whose behavior is specified by the solution set of a system of differential equations. We call such representations *kernel representations*. Such systems need not be controllable, of course. We shall see that it is exactly the controllable systems that also admit an image representation.

6.2 Elimination of Latent Variables

6.2.1 Modeling from first principles

As argued in Chapter 1 and further elaborated in Chapter 4, models obtained from first principles invariably contain *latent* variables, in addition to the *manifest* variables, which our model aims at describing. In the context of behaviors described by differential equations as studied in Chapters 2 and 3, this leads to the following class of dynamical systems with latent variables:

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell. \tag{6.1}$$

Here $w : \mathbb{R} \to \mathbb{R}^q$ is the trajectory of the manifest variables, whereas $\ell : \mathbb{R} \to \mathbb{R}^d$ is the trajectory of the latent variables. The equating space is $\mathbb{R}^g$, and the behavioral equations are parametrized by the two polynomial matrices $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and $M(\xi) \in \mathbb{R}^{g\times d}[\xi]$.

The question that we want to consider is, *What sort of behavioral equation does (6.1) imply about the manifest variable w alone?* In particular, we wonder whether the relations imposed on the manifest variable w by the

full behavioral equations (6.1) can themselves be written in the form of a system of differential equations. In other words, we would like to know whether or not the set

$$\mathfrak{B} = \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid \exists\ \ell \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^d) \text{ s.t. } R(\frac{d}{dt})w = M(\frac{d}{dt})\ell \text{ weakly}\} \tag{6.2}$$

can be written as the (weak) solution set of a system of linear differential equations. We will see that (in a sense) it can indeed be expressed in this way. Actually, in Chapter 1, we already informally worked out an example. The RLC network of Example 1.3.5 was modeled using the constitutive equations of the components and Kirchhoff's laws. This led to the differential equations (1.1, 1.2, 1.3). We set out to model the port behavior of this circuit, and indeed, after some ad hoc manipulations we arrived at (1.12,1.13). The question is, Was the fact that the manifest behavior is also described by a differential equation a coincidence? If it is not, how can we find such a differential equation in a systematic way? Before we answer that question, we examine two more examples.

Example 6.2.1 Consider a mechanical system consisting of three masses and four springs; see Figure 6.1. Let w_1, w_2, w_3 denote the displacements of the masses from their respective equilibria. Denote the spring constants by k_1, k_2, k_3, k_4, and the masses by m_1, m_2, m_3. Suppose that we are interested

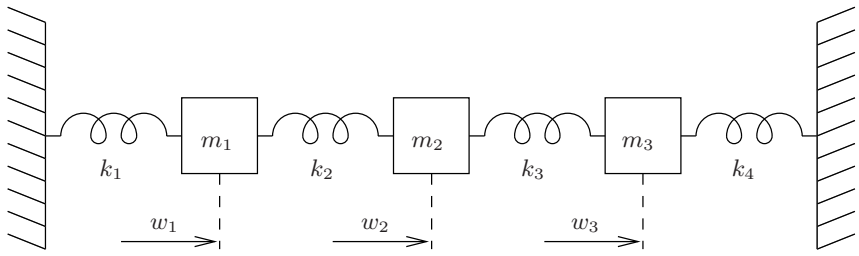


FIGURE 6.1. Mechanical system.

in a mathematical model relating the displacement w_1 of the first mass to the displacement w_3 of the third mass. The relations between the variables w_1, w_2, w_3 are given by

$$\begin{array}{rcl} m_1 \frac{d^2}{dt^2} w_1 & = & -k_1 w_1 + k_2(w_2 - w_1), \\ m_2 \frac{d^2}{dt^2} w_2 & = & k_2(w_1 - w_2) + k_3(w_3 - w_2), \\ m_3 \frac{d^2}{dt^2} w_3 & = & k_3(w_2 - w_3) - k_4 w_3. \end{array} \tag{6.3}$$

The relation that we are after is that between w_1 and w_3. The equations (6.3) determine this relation only implicitly. Implicitly, because a third variable, w_2, is also involved in (6.3). The variables w_1 and w_3 are the variables

we are interested in, whereas w_2 is just an auxiliary variable. Therefore, we call w_1 and w_3 the *manifest* variables and w_2 a *latent* variable.

If we are not satisfied with an implicit relation, in other words, if we want to obtain a differential equation in which only w_1 and w_3 appear, then we have to eliminate w_2 from (6.3). For the case at hand this can be done as follows. For simplicity, assume that all constants (masses and spring constants) are unity. From the first equation in (6.3) we obtain

$$w_2 = 2w_1 + \frac{d^2}{dt^2}w_1, \qquad \frac{d^2}{dt^2}w_2 = 2\frac{d^2}{dt^2}w_1 + \frac{d^4}{dt^4}w_1, \tag{6.4}$$

where the second expression is obtained by differentiating the first twice. Substituting this expression for w_2 and $\frac{d^2}{dt^2}w_2$ in the second and third equations of (6.3) yields

$$\begin{aligned} 3w_1 + 4\frac{d^2}{dt^2}w_1 + \frac{d^4}{dt^4}w_1 - w_3 &= 0, \\ 2w_1 + \frac{d^2}{dt^2}w_1 - 2w_3 - \frac{d^2}{dt^2}w_3 &= 0. \end{aligned} \tag{6.5}$$

Notice that (6.5) does not contain w_2. It is clear that for any triple (w_1, w_2, w_3) that satisfies (6.3), the corresponding pair (w_1, w_3) satisfies (6.5). The converse is less obvious, yet for any pair (w_1, w_3) that satisfies (6.5) there indeed exists a w_2 such that (6.3) is satisfied. Otherwise stated, we claim that the relations imposed implicitly on (w_1,w_3) by (6.4) are given explicitly by (6.5). Later in this chapter we will see how we could have arrived at these equations in a systematic way, and also it will also become clear that the relation between w_1 and w_3 is indeed determined by (6.5). □

Example 6.2.2 Consider the electrical circuit consisting of a resistor, a capacitor, an inductor, and an external port shown in Figure 6.2. Suppose

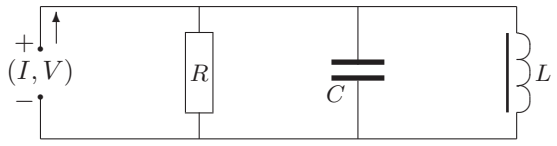


FIGURE 6.2. Electrical circuit.

that we want to model the relation between the voltage V across and the current I through the external port. Introduce the voltages across and the currents through the other elements as latent variables. The equations

describing the full behavior are, in the obvious notation (see also Example 4.3.2),

$$\begin{aligned} &V = V_R = V_C = V_L, \qquad I = I_R + I_C + I_L, \\ &V_R = I_R R, \qquad I_C = C\frac{d}{dt}V_C, \qquad V_L = L\frac{d}{dt}I_L. \end{aligned} \tag{6.6}$$

This is again a set of equations that implicitly determines the relation between the manifest variables V and I, but it contains the latent variables V_R, V_C, V_L, I_R, I_C and I_L. An explicit relation can readily be obtained. First note that $\frac{d}{dt}V$ can be expressed in terms of I and I_L by proceeding as follows:

$$C\frac{d}{dt}V = C\frac{d}{dt}V_C = I_C = I - I_R - I_L = I - \frac{V_R}{R} - I_L = I - \frac{V}{R} - I_L, \tag{6.7}$$

and hence

$$\begin{aligned} C(\frac{d}{dt})^2 V &= \frac{d}{dt}I - \frac{d}{dt}\frac{V}{R} - \frac{d}{dt}I_L = \frac{d}{dt}I - \frac{d}{dt}\frac{V}{R} - \frac{V}{L} \\ &= \frac{d}{dt}I - \frac{d}{dt}\frac{V}{R} - \frac{V}{L}. \end{aligned} \tag{6.8}$$

From (6.7) we obtain the desired equation:

$$\frac{1}{L}V + \frac{1}{R}\frac{d}{dt}V + C\frac{d^2}{dt^2}V = \frac{d}{dt}I. \tag{6.9}$$

Again, it is easy to see that (6.6) implies that that (V, I) satisfies (6.9). The converse is also true. We claim that the manifest behavior, in this case the behavior of the pair (I, V), is modeled by equation (6.9). Also, this example may be treated more systematically. See Exercise 6.2. □

Examples 6.2.1 and 6.2.2 confirm what we already argued extensively in Chapters 1 and 4, namely that in order to obtain a model of the relation between certain variables in a system of some complexity, it is natural to first model the relation among many more variables. Subsequently, the equations are then manipulated so as to eliminate the variables in which we are not interested.

Examples 6.2.1 and 6.2.2 indicate that the manifest behavior of a given full behavior described by linear differential equations with constant coefficients is described by relations of the same type. The suggestions made by the examples will be justified shortly. However, before we proceed with the linear dynamic case, we give a motivating example of the elimination problem and the difficulties that may be encountered for a *nonlinear static* mathematical model with latent variables.

Example 6.2.3 Consider in the context of Definition 1.2.9 the static model in $\mathbb{R}^2$ with $\mathfrak{B}_\mathrm{f} := \{(w,\ell) \in \mathbb{R}^2 \mid w\ell = 1\}$. Then $\mathfrak{B}$, the manifest behavior, is given by $\mathfrak{B} = \{w \in \mathbb{R} \mid \exists\ \ell \in \mathbb{R} \text{ such that } (w,\ell) \in \mathfrak{B}_\mathrm{f}\}$. The problem that we want to address is, Can $\mathfrak{B}$ be described in a similar "nice" way as $\mathfrak{B}_\mathrm{f}$? What do we mean by nice in this context? Well, $\mathfrak{B}_\mathrm{f}$ is the zero set of a polynomial equation (it is therefore called an *algebraic set*, or in this case, an *algebraic curve*). What we would like to know is whether the manifest behavior is also the zero set of an algebraic equation. It is trivial to see that in this case $\mathfrak{B} = \{w \in \mathbb{R} \mid w \neq 0\}$, which is *not* an algebraic set, since an algebraic set in $\mathbb{R}$ consists of finitely many points, or it coincides with $\mathbb{R}$. So in this example the answer is in a sense negative: the full behavior was described nicely, whereas the manifest behavior was described less nicely, namely by an *inequality* rather than by an *equation*. □

An appealing and insightful way of thinking about the problem of describing the manifest behavior $\mathfrak{B}$ is the observation that $\mathfrak{B}$ is just the *projection* of $\mathfrak{B}_\mathrm{f}$ onto $\mathfrak{L}_1^\mathrm{loc}(\mathbb{R},\mathbb{R}^q)$, the space where the manifest variables live, along $\mathfrak{L}_1^\mathrm{loc}(\mathbb{R},\mathbb{R}^d)$, the space of latent variables. The problem with Example 6.2.3 is the fact that the projection of an algebraic set in a Euclidean space onto a lower-dimensional subspace is not necessarily an algebraic set. The question that arises is the following: *Is there any chance that in the case of behaviors described by linear time-invariant differential equations, the projection of the full behavior on the signal space* $\mathfrak{L}_1^\mathrm{loc}(\mathbb{R},\mathbb{R}^q)$ *is also described by linear time-invariant differential equations?* An obvious necessary condition for this to hold is that the manifest behavior be a linear shift-invariant subspace of $(\mathbb{R}^q)^\mathbb{R}$. That this condition is indeed satisfied is easy to prove.

Theorem 6.2.4 *The manifest behavior* $\mathfrak{B}$ *defined by (6.2) is linear and shift-invariant.*

Proof See Exercise 6.10. □

Remark 6.2.5 Theorem 6.2.4 reflects only a *necessary* condition for a manifest behavior to be described by a set of linear differential equations. It is not true that every linear shift-invariant subspace of $(\mathbb{R}^q)^\mathbb{R}$ is the solution set of a system of linear time-invariant differential equations; see Exercise 6.7. So there is still work to be done in order to arrive at the result that the manifest behavior (6.2) is described by a set of linear differential equations. □

6.2.2 Elimination procedure

We now describe a general procedure for obtaining a description of the manifest behavior.

Theorem 6.2.6 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$, $M(\xi) \in \mathbb{R}^{g\times d}[\xi]$, and denote by $\mathfrak{B}_{\mathrm{f}}$ the full behavior of (6.1):*

$$\mathfrak{B}_{\mathrm{f}} = \{(w,\ell) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q \times \mathbb{R}^d) \mid \ R(\frac{d}{dt})w = M(\frac{d}{dt})\ell, \ \textit{weakly}\ \}. \quad (6.10)$$

Let the unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ be such that

$$U(\xi)M(\xi) = \begin{bmatrix} 0 \\ M''(\xi) \end{bmatrix}, \qquad U(\xi)R(\xi) = \begin{bmatrix} R'(\xi) \\ R''(\xi) \end{bmatrix}, \quad (6.11)$$

with $M''(\xi)$ of full row rank. By Theorem 2.5.23 such a unimodular matrix $U(\xi)$ exists. Then the $\mathcal{C}^\infty$ part of the manifest behavior $\mathfrak{B}$, defined by $\mathfrak{B} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ with $\mathfrak{B}$ given by (6.2), consists of the $\mathcal{C}^\infty$ solutions of

$$R'(\frac{d}{dt})w = 0.$$

Proof The partition of $U(\xi)R(\xi)$ and $U(\xi)M(\xi)$ in (6.11) provides the following equivalent description of $\mathfrak{B}_{\mathrm{f}}$:

$$R'(\frac{d}{dt})w = 0, \quad (6.12)$$

$$R''(\frac{d}{dt})w = M''(\frac{d}{dt})\ell, \quad (6.13)$$

with $R'(\xi) \in \mathbb{R}^{g'\times q}[\xi]$, $R''(\xi) \in \mathbb{R}^{g''\times q}[\xi]$, and $M''(\xi) \in \mathbb{R}^{g''\times d}$. Now, examine these equations. Equation (6.12) entails some genuine constraint on the manifest variables w. Indeed, if w is such that $(w,\ell) \in \mathfrak{B}_{\mathrm{f}}$ for some ℓ, then certainly w itself already has to satisfy (6.12). Let us now look at (6.13). We claim that this equation entails at most some smoothness constraints on w. In other words, if w needed only to satisfy (6.13) for some ℓ, then the components of w would need to be sufficiently differentiable, but no further constraints would have to be imposed. In particular, any $w \in C^\infty(\mathbb{R}, \mathbb{R}^q)$ would be permitted. In order, to see this, let $\tilde{M}''(\xi) \in \mathbb{R}^{g''\times g''}$ be a square submatrix of $M''(\xi)$ such that $\det(\tilde{M}''(\xi))$ is nonzero. Such a submatrix exists, since $M''(\xi)$ has full row rank. Assume for simplicity that $\tilde{M}''(\xi)$ consists of the first g'' columns of $M''(\xi)$. Otherwise, permute the components of ℓ so as to achieve this. Consider the matrix of rational functions

$$(\tilde{M}''(\xi))^{-1}R''(\xi).$$

This matrix of rational functions need not be proper. Let $k \in \mathbb{Z}_+$ be such that

$$\frac{1}{\xi^k}(\tilde{M}''(\xi))^{-1}R''(\xi)$$

is proper. Write the system of equations (6.13) as

$$R''(\frac{d}{dt})w = \tilde{M}''(\frac{d}{dt})\ell_1 + \tilde{\tilde{M}}''(\frac{d}{dt})\ell_2$$

and consider the related system

$$R''(\frac{d}{dt})w = \tilde{M}''(\frac{d}{dt})\frac{d^k\tilde{\ell}_1}{dt^k}. \tag{6.14}$$

Note that $(\xi^k\tilde{M}''(\xi))^{-1}R''(\xi)$ is proper. Consequently, as shown in Section 3.3, w, as constrained by (6.14), is a free variable, implying that for each $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ there exists an $\tilde{\ell}_1 \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ such that (6.14) is satisfied. Let w be such that $\frac{d^k w}{dt^k} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$. Then as we have just seen, there exists $\tilde{\ell}_1$ such that (6.14) is satisfied. From the proof of Theorem 3.3.13 it follows that since the "input" w to (6.14) is k times differentiable, so will be the "output" $\tilde{\ell}_1$. Hence for ws such that $\frac{d^k w}{dt^k} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ there exists ℓ_1 such that

$$R''(\frac{d}{dt})w = \tilde{M}''(\frac{d}{dt})\ell_1.$$

This implies that $(w_1, (\ell_1, 0)) \in \mathfrak{B}_{\mathrm{f}}$, and we conclude that for each $w : \mathbb{R} \to \mathbb{R}^q$ such that $\frac{d^k w}{dt^k} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$, there exists an ℓ such that (w, ℓ) satisfies (6.13) weakly. Consequently, for each $w : \mathbb{R} \to \mathbb{R}^q$ such that (6.12) is satisfied and such that $\frac{d^k w}{dt^k} \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$, there exists an ℓ such that $(w, \ell) \in \mathfrak{B}_{\mathrm{f}}$. □

A close examination of the above proof shows that (6.1) imposes only differentiability conditions on w in addition to the constraints imposed by the differential equation (6.12).

In the sequel we ignore these differentiability conditions and simply declare *the manifest behavior of (6.1) to be described by the differential equation (6.12). In other words, we impose that (6.12) is the result of eliminating the latent variables: it describes the laws induced on the manifest variables by the behavioral equations (6.10).*

Ignoring the smoothness constraints can be justified as follows. From a mathematical point of view it is often natural to use the closure in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ of the manifest behavior of (6.10) instead of the manifest behavior as defined in a set theoretic way. The procedure of taking the closure can be brought into connection with Example 6.2.3. By Theorem 2.4.4, the full behavior $\mathfrak{B}_{\mathrm{f}}$ is a closed subspace of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q \times \mathbb{R}^d)$. Then why is the projection of $\mathfrak{B}_{\mathrm{f}}$ not closed? This is because projection is *not* a *closed* map; i.e., projections need not map a closed set to a closed set. To see this, take the full behavior of Example 6.2.3. This is obviously a closed subset of $\mathbb{R}^2$. However, the projection onto the w space is not closed in $\mathbb{R}$. Analogously, the projection of, for instance, the full behavior of $\frac{d}{dt}w = \ell$ as a subspace

of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^2)$ on the space of w-variables $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ consists of the absolute continuous functions in $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$. This is not all of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$: it is a dense, but not a closed, subspace of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$. An example of a function $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ that does not belong the projection, is the step function: $w(t) = 0$ for t negative and $w(t) = 1$ for t positive. Indeed we have seen in Exercise 3.25 that there does not exist an ℓ such that $\frac{d}{dt}w = \ell$, weakly.

Motivated by the above discussion, we define the manifest behavior as follows.

Definition 6.2.7 Under the conditions and in the notation of Theorem 6.2.6, the manifest behavior $\mathfrak{B} \subset \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ is *defined* as the set of weak solutions of

$$R'(\frac{d}{dt})w = 0. \tag{6.15}$$

□

Note that the behavioral equation (6.15) for the $\mathcal{C}^\infty$ part of the manifest behavior was obtained by means of a *theorem*, whereas the manifest behavior viewed as a subset of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ is just *defined* to be the set of weak solutions of the same behavioral equation. As a consequence, the notation (6.2) is not completely consistent with the above definition of the manifest variable. Since the behavior defined by (6.15) is the closure of the behavior defined by (6.2) and thus only slightly larger, this mild inconsistency should not cause confusion. In the case that $M''(\frac{d}{dt})\ell = R''(\frac{d}{dt})w$ does not impose additional smoothness conditions, which means that for every $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ that satisfies $R'(\frac{d}{dt})w = 0$ weakly, there exists an ℓ such that $M''(\frac{d}{dt})\ell = R''(\frac{d}{dt})w$, the manifest behavior is closed and is therefore *exactly* described by $R'(\frac{d}{dt})w = 0$. In that case we say that *exact elimination* is possible.

Let us now come back to Example 6.2.1 and see how we can eliminate the latent variable using the general procedure discussed above.

Example 6.2.8 (Example 6.2.1 continued) Consider once more the mechanical system consisting of four springs and three masses, as depicted in Figure 6.1. Consider again w_2 as a latent variable. The problem is to eliminate w_2. We describe the solution to this problem in terms of Theorem 6.2.6.

In this example the polynomial matrices $R(\xi) \in \mathbb{R}^{3\times 2}[\xi]$ and $M(\xi) \in \mathbb{R}^{3\times 1}[\xi]$ are given by

$$R(\xi) = \begin{bmatrix} 2+\xi^2 & 0 \\ 1 & 1 \\ 0 & 2+\xi^2 \end{bmatrix}, \qquad M(\xi) = \begin{bmatrix} 1 \\ 2+\xi^2 \\ 1 \end{bmatrix}.$$

Redefine $w := \mathrm{col}(w_1, w_3)$ and $\ell := w_2$, so that (6.3) can be written as

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell.$$

In order to bring these equations into the form (6.12, 6.13), subtract the first row of $M(\xi)$ from the third row and multiplied by $\xi^2 + 2$ from the second row. Call the resulting matrix $\tilde{M}(\xi)$. Treat $R(\xi)$ analogously to obtain $\tilde{R}(\xi)$. Then

$$\tilde{R}(\xi) = \begin{bmatrix} 2+\xi^2 & 0 \\ 3+4\xi^2+\xi^4 & -1 \\ -2-\xi^2 & 2+\xi^2 \end{bmatrix}, \qquad \tilde{M}(\xi) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{6.16}$$

It is clear that the first row of $\tilde{M}(\xi)$ is of full row rank. According to Theorem 6.2.6 and Definition 6.2.7, the last two rows of $\tilde{R}(\xi)$ therefore yield the desired equations for the manifest behavior:

$$\begin{aligned} 3w_1 + 4\frac{d^2}{dt^2}w_1 + \frac{d^4}{dt^4}w_1 - w_3 &= 0, \\ -2w_1 - \frac{d^2}{dt^2}w_1 + 2w_3 + \frac{d^2}{dt^2}w_3 &= 0. \end{aligned} \tag{6.17}$$

The first row of (6.16) yields

$$2w_1 + \frac{d^2}{dt^2}w_1 = w_2. \tag{6.18}$$

As remarked before, (6.18) poses only a smoothness condition on w_1, and therefore we ignore it. However, in the case at hand, the smoothness imposed by (6.18) is already guaranteed by (6.17), so that ignoring it is completely justified. See Exercise 6.9. As a consequence, the manifest behavior is described by (6.17). This is the answer that we found in Example 6.2.1, but now we understand much better why it is indeed the correct answer. □

6.2.3 Elimination of latent variables in interconnections

Often, dynamical systems can be thought of as interconnections of "simple" subsystems. Intuitively speaking, the description of an interconnection of two systems consists of behavioral equations that describe the individual subsystems and equations relating the variables that connect the subsystems. In this section we give an example of such an interconnection, and we show how we can use the elimination procedure in order to obtain a description of the external behavior from the equations that define the subsystems and those that define the interconnection. The example that we are considering here is the *series interconnection* of two SISO systems.

Example 6.2.9 Let $p_i(\xi), q_i(\xi) \in \mathbb{R}[\xi],\ i = 1, 2$. Consider the i/o systems

$$\Sigma_1 :\ p_1(\frac{d}{dt})y_1 = q_1(\frac{d}{dt})u_1, \quad \Sigma_2 :\ p_2(\frac{d}{dt})y_2 = q_2(\frac{d}{dt})u_2. \tag{6.19}$$

The series interconnection of the associated i/o behaviors is defined by the interconnecting equation

$$y_1 = u_2. \tag{6.20}$$

The interpretation of this interconnection is that the input of the second i/o system is the output of the first. See Figure 6.3. Suppose that we are

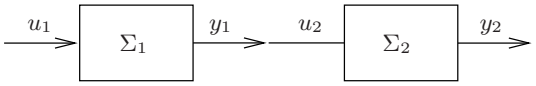


FIGURE 6.3. Series interconnection of Σ_1 and Σ_2.

interested in the relation between the "external" variables u_1 and y_2. This relation can be determined by eliminating y_1 and u_2 from (6.19) and (6.20). Define $R(\xi)$ and $M(\xi)$ as follows:

$$R(\xi) := \begin{bmatrix} q_1(\xi) & 0 \\ 0 & p_2(\xi) \\ 0 & 0 \end{bmatrix}, \quad M(\xi) := \begin{bmatrix} p_1(\xi) & 0 \\ 0 & q_2(\xi) \\ -1 & 1 \end{bmatrix}.$$

Equations (6.19), (6.20) can now be written as

$$R(\frac{d}{dt}) \begin{bmatrix} u_1 \\ y_2 \end{bmatrix} = M(\frac{d}{dt}) \begin{bmatrix} y_1 \\ u_2 \end{bmatrix}.$$

In order to find equations for the behavior of (u_1, y_2), extract the greatest common divisor from $p_1(\xi)$ and $q_2(\xi)$: Suppose that $p_1(\xi) = c(\xi)\bar{p}_1(\xi)$ and $q_2(\xi) = c(\xi)\bar{q}_2(\xi)$, where $\bar{p}_1(\xi)$ and $\bar{q}_2(\xi)$ have no further common factors. By Corollary B.1.7, Bezout, there exist polynomials $a(\xi), b(\xi)$ such that $a(\xi)\bar{p}_1(\xi) + b(\xi)\bar{q}_2(\xi) = 1$. Define unimodular matrices $U_1(\xi), U_2(\xi), U_3(\xi)$ as follows:

$$U_1(\xi) = \begin{bmatrix} 1 & 0 & p_1(\xi) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad U_2(\xi) = \begin{bmatrix} a(\xi) & b(\xi) & 0 \\ -\bar{q}_2(\xi) & \bar{p}_1(\xi) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad U_3(\xi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

One easily checks that

$$U_3(\xi)U_2(\xi)U_1(\xi)M(\xi) = \begin{bmatrix} 0 & c(\xi) \\ -1 & 1 \\ 0 & 0 \end{bmatrix}$$

and

$$U_3(\xi)U_2(\xi)U_1(\xi)R(\xi) = \begin{bmatrix} a(\xi)q_1(\xi) & b(\xi)p_2(\xi) \\ 0 & 0 \\ -q_1(\xi)\bar{q}_2(\xi) & \bar{p}_1(\xi)p_2(\xi) \end{bmatrix}.$$

It is clear that $U_3(\xi)U_2(\xi)U_1(\xi)M(\xi)$ has the required form; that is, the nonzero part has full row rank. It follows that the relation between u_1 and y_2 is described by the third row of $U_3(\xi)U_2(\xi)U_1(\xi)R(\xi)$:

$$\bar{p}_1(\frac{d}{dt})p_2(\frac{d}{dt})y_2 = q_1(\frac{d}{dt})\bar{q}_2(\frac{d}{dt})u_1.$$

□

Example 6.2.9 shows the power of the elimination procedure. It shows in a precise way how to treat common factors. It is important to observe that because of common factors, the series interconnection of Σ_1 and Σ_2 may have a different manifest behavior than the series interconnection of Σ_2 and Σ_1.

Other examples of interconnections are given in the Exercises 6.4 (feedback interconnection) and 6.5 (parallel interconnection).

6.3 Elimination of State Variables

In Chapter 4 we have introduced input/state/output models. We view the state as a special latent variable. In this section we study the problem of determining the relation between the input and output of an i/s/o model. It turns out that for SISO systems we can find a complete answer by applying the general elimination procedure presented in Section 6.2.2. We present the analysis in two steps: first for i/s/o systems of the form (4.16) with $D = 0$. Subsequently,we treat the general case.

Theorem 6.3.1 *Consider the system*

$$\begin{array}{rcl} \frac{d}{dt}x & = & Ax + bu, \\ y & = & cx, \end{array} \tag{6.21}$$

with $(A, b, c) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n}$. *Define* $\bar{p}(\xi) \in \mathbb{R}[\xi]$ *and* $\bar{r}(\xi) \in \mathbb{R}^{1\times n}[\xi]$ *by*

$$\bar{p}(\xi) := \det(I\xi - A) \quad \bar{r}(\xi) := \bar{p}(\xi)c(I\xi - A)^{-1}.$$

Let $g(\xi)$ *be the greatest common divisor of* $(\bar{p}(\xi), \bar{r}_1(\xi), \ldots, \bar{r}_n(\xi))$. *Define the polynomials* $p(\xi)$ *and* $q(\xi)$ *by*

$$p(\xi) := \frac{\bar{p}(\xi)}{g(\xi)} \text{ and } q(\xi) := \frac{\bar{r}(\xi)}{g(\xi)}b. \tag{6.22}$$

Then the i/o behavior of the i/s/o representation (6.21) is given by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u. \tag{6.23}$$

Proof In view of the discussion following Definition 4.6.1 and by Corollary 5.3.14, we may without loss of generality assume that (A, b, c) has the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad c = \begin{bmatrix} 0 & c_2 \end{bmatrix} \tag{6.24}$$

with (A_{22}, c_2) observable. Using this form, (6.21) becomes

$$\begin{array}{rcl} \frac{d}{dt}x_1 &=& A_{11}x_1 + A_{12}x_2 + b_1u, \\ \frac{d}{dt}x_2 &=& A_{22}x_2 + b_2u, \\ y &=& c_2x_2. \end{array} \tag{6.25}$$

From (6.24) it appears logical to eliminate x in two steps. First eliminate x_1, the nonobservable component of the state, and then eliminate x_2, the observable component. Elimination of x_1 from (6.25) yields straightforwardly

$$\begin{array}{rcl} \frac{d}{dt}x_2 &=& A_{22}x_2 + b_2u, \\ y &=& c_2x_2. \end{array} \tag{6.26}$$

In order to eliminate x_2, define matrices $R(\xi)$ and $M(\xi)$ by

$$R(\xi) := \begin{bmatrix} b_2 & 0 \\ 0 & 1 \end{bmatrix}, \quad M(\xi) := \begin{bmatrix} I\xi - A_{22} \\ c_2 \end{bmatrix}.$$

Then (6.26) can be written as

$$R(\frac{d}{dt}) \begin{bmatrix} u \\ y \end{bmatrix} = M(\frac{d}{dt})x_2.$$

Obviously, the row rank of $M(\xi)$ is n_2, the dimension of A_{22}, so that we should be able to create exactly one zero-row in $M(\xi)$ by means of elementary row operations. Define the polynomials $\tilde{p}(\xi), \tilde{r}(\xi)$ as follows $\tilde{p}(\xi) := \det(I\xi - A_{22})$, and $\tilde{r}(\xi) := \tilde{p}(\xi)c_2(I\xi - A_{22})^{-1}$. Since (c_2, A_{22}) is observable, it follows from Theorem 5.5.1 that $\tilde{p}(\xi)$ and $\tilde{r}(\xi)$ have no common factor. Consequently, by Theorem 2.5.10, there exist matrices $U_{11}(\xi), U_{12}(\xi)$ of appropriate dimensions such that the matrix

$$U(\xi) := \begin{bmatrix} U_{11}(\xi) & U_{12}(\xi) \\ \tilde{r}(\xi) & -\tilde{p}(\xi) \end{bmatrix}$$

is unimodular. Now,

$$U(\xi)R(\xi) = \begin{bmatrix} * & * \\ \tilde{r}(\xi)b_2 & -\tilde{p}(\xi) \end{bmatrix}, \quad U(\xi)M(\xi) = \begin{bmatrix} * \\ 0 \end{bmatrix}, \tag{6.27}$$

where as usual, the $*$s denote polynomial expressions whose exact values are immaterial. From Theorem 6.2.6 and Definition 6.2.7 it follows that the manifest behavior of (6.26), and therefore of (6.25), is given by

$$\tilde{r}(\frac{d}{dt})b_2u - \tilde{p}(\frac{d}{dt})y = 0.$$

It remains to show that $p(\xi) = \tilde{p}(\xi)$ and $\tilde{r}(\xi)b_2 = q(\xi)$, where $p(\xi)$ and $q(\xi)$ are given by (6.22). It easy to check that

$$\begin{aligned} \bar{p}(\xi) &= \det(I\xi - A_{11})\det(I\xi - A_{22}), \\ \bar{r}(\xi) &= \det(I\xi - A_{11})\det(I\xi - A_{22}) \begin{bmatrix} 0 & c_2(I\xi - A_{22})^{-1} \end{bmatrix}. \end{aligned}$$

Obviously, $\det(I\xi - A_{11})$ divides both $\bar{p}(\xi)$ and $\bar{r}(\xi)$, and since (c_2, A_{22}) is observable, it follows from Theorem 5.5.1 that g.c.d.$(\bar{p}(\xi), \bar{r}(\xi)) = \det(I\xi - A_{11})$, and hence indeed $\tilde{p}(\xi) = p(\xi)$ and $\tilde{r}(\xi) = r(\xi)$. □

The above result can straightforwardly be generalized to SISO systems in which $d \neq 0$.

Corollary 6.3.2 *Consider the system*

$$\begin{aligned} \frac{d}{dt}x &= Ax + bu, \\ y &= cx + du, \end{aligned} \tag{6.28}$$

where $(A, b, c, d) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n} \times \mathbb{R}$. *Define* $\bar{p}(\xi) \in \mathbb{R}[\xi]$ *and* $\bar{r}(\xi) \in \mathbb{R}^{1\times n}[\xi]$ *by*

$$\bar{p}(\xi) := \det(I\xi - A), \quad \bar{r}(\xi) := \bar{p}(\xi)c(I\xi - A)^{-1}. \tag{6.29}$$

Let $g(\xi)$ *be the greatest common divisor of* $(\bar{p}(\xi), \bar{r}(\xi))$. *Define*

$$p(\xi) := \frac{\bar{p}(\xi)}{g(\xi)} \text{ and } q(\xi) := \frac{\bar{r}(\xi)}{g(\xi)}b + dp(\xi). \tag{6.30}$$

Then the i/o behavior of the i/s/o representation (6.28) is given by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u.$$

Proof Define $\tilde{y} := y - du$. Then $\tilde{y} = cx$. Define $\tilde{q}(\xi) := q(\xi) - dp(\xi)$. According to Theorem 6.3.1 the relation between u and $\tilde{y}$ is given by $p(\frac{d}{dt})\tilde{y} = \tilde{q}(\frac{d}{dt})u$. This implies that

$$p(\frac{d}{dt})y = p(\frac{d}{dt})\tilde{y} + dp(\frac{d}{dt})u = (\tilde{q}(\frac{d}{dt}) + dp(\frac{d}{dt}))u = q(\frac{d}{dt})u.$$

□

For observable systems the i/o behavior is particularly easy to describe. This case is treated in the next corollary.

Corollary 6.3.3 *Consider the i/s/o system (6.28). Assume that* (A, c) *is an observable pair. Define* $p(\xi) \in \mathbb{R}[\xi]$ *and* $q(\xi) \in \mathbb{R}[\xi]$ *by*

$$p(\xi) := \det(I\xi - A) \quad q(\xi) := p(\xi)c(I\xi - A)^{-1}b + dp(\xi).$$

Then the i/o behavior of the i/s/o representation (6.28) is given by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u.$$

Proof Since (A, c) is observable, it follows from Theorem 5.5.1 that $p(\xi)$ and $r(\xi)$ have no common factors. The statement now follows from Theorem 6.3.1. □

Remark 6.3.4 As remarked in Section 6.2.2, the equation defined by the first row of $U(\xi)R(\xi)$ and $U(\xi)M(\xi)$ in (6.27) could impose a smoothness condition on w. It can be proved that for this particular problem, the elimination of the state, this is not the case. That means that the i/o behavior defined by (6.21) is *exactly* equal to the i/o behavior defined by (6.23). In other words, exact elimination is possible; see Exercise 6.25. □

Remark 6.3.5 The common factor $g(\xi)$ of $\bar{p}(\xi)$ and $\bar{r}(\xi)$ corresponds to the nonobservable part of the state space and is canceled in the elimination procedure. In the final i/o description (6.23) it could very well be the case that $p(\xi)$ and $q(\xi)$ still have a common factor. This factor corresponds to a noncontrollable part of the state space and should *not* be canceled. See also Exercise 6.20. □

Example 6.3.6 Consider the i/s/o system

$$\begin{aligned} \frac{d}{dt}x &= Ax + bu, \\ y &= cx. \end{aligned} \tag{6.31}$$

with

$$A = \begin{bmatrix} 2 & 4 & -5 \\ -1 & -3 & 15 \\ 0 & 0 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 0 & 1 & -2 \end{bmatrix}.$$

Note that this system is neither controllable nor observable. Straightforward calculations yield

$$\det(I\xi - A) = 6 - 5\xi - 2\xi^2 + \xi^3,$$

$$[\det(I\xi - A)]\,(I\xi - A)^{-1} = \begin{bmatrix} -9 + \xi^2 & -12 + 4\xi & 45 - 5\xi \\ 3 - \xi & 6 - 5 + \xi^2 & -25 + 15\xi \\ 0 & 0 & -2 + \xi + \xi^2 \end{bmatrix},$$

so that $\bar{p}(\xi) = 6 - 5\xi - 2\xi^2 + \xi^3 = (\xi - 3)(\xi + 2)(\xi - 1)$, and $\bar{r}(\xi) = [\,3 - \xi \;\; 6 - 5\xi + \xi^2 \;\; -21 + 13\xi - 2\xi^2\,] = (\xi - 3)\,[\,-1 \;\; -2 + \xi \;\; 7 - 2\xi\,]$. It follows that the greatest common divisor of $\bar{p}(\xi)$ and $\bar{r}(\xi)$ is given by $g(\xi) = \xi - 3$. Define $p(\xi)$ and $q(\xi)$ by

$$p(\xi) := \frac{\bar{p}(\xi)}{g(\xi)} = -2 + \xi + \xi^2, \qquad q(\xi) := \frac{\bar{r}(\xi)}{g(\xi)} b = 1 - \xi.$$

According to Theorem 6.3.1, the i/o behavior of the system (6.31) is therefore described by

$$-2y + \frac{d}{dt}y + \frac{d^2}{dt^2}y = u - \frac{d}{dt}u. \tag{6.32}$$

Notice that the polynomials $p(\xi)$ and $q(\xi)$ still have a factor $\xi - 1$ in common. This factor corresponds to a noncontrollable but observable part of the behavior, and therefore this factor should not be canceled. The common factor $\xi - 3$ in $\bar{p}(\xi)$ and $\bar{r}(\xi)$ corresponds to the nonobservable part of the system, and this factor is canceled by the elimination procedure.

□

6.4 From i/o to i/s/o Model

In the previous section we have seen how to obtain the i/o behavior of an i/s/o model. The last question to be answered is that of finding an i/s/o representation for a given i/o behavior. Otherwise stated, rather than eliminating a latent variable, we want to introduce a latent variable, but a special one: the state. We treat only the SISO case.

Let the polynomials $p(\xi)$ and $q(\xi)$ be given by

$$\begin{aligned} p(\xi) &= p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n, \\ q(\xi) &= q_0 + q_1\xi + \cdots + q_{n-1}\xi^{n-1} + q_n\xi^n, \end{aligned} \tag{6.33}$$

and consider the i/o system described by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u. \tag{6.34}$$

The problem under consideration is the following: *Given an i/o system of the form (6.34), does there exist an i/s/o representation of it, and if the answer is affirmative, how can we obtain this i/s/o representation?*

The i/o behavior is defined by

$$\mathfrak{B}_{\mathrm{i/o}} := \{(u,y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}\times\mathbb{R}) \mid p(\frac{d}{dt})y = q(\frac{d}{dt})u, \text{ weakly } \}. \tag{6.35}$$

In mathematical terms, the state representation problem is this:

Definition 6.4.1 State Representation Problem Given $\mathfrak{B}_{\mathrm{i/o}}$, defined by (6.35), find $n' \in \mathbb{N}$ and four matrices $A, b, c, d \in \mathbb{R}^{n'\times n'} \times \mathbb{R}^{n'\times 1} \times \mathbb{R}^{1\times n'} \times \mathbb{R}^{1\times 1}$ such that the i/o behavior of

$$\begin{array}{rcl} \frac{d}{dt}x &=& Ax + bu, \\ y &=& cx + du \end{array} \tag{6.36}$$

is exactly $\mathfrak{B}_{\mathrm{i/o}}$. □

We present two methods for obtaining such an i/s/o representation. Both are based on Corollary 6.3.2. There it was shown what i/o equations correspond to a given quadruple (A, b, c, d). The state representation problem can therefore be rephrased as, *Given polynomials $p(\xi)$ and $q(\xi)$, find matrices (A, b, c, d) of appropriate dimensions such that the correspondence between (A, b, c, d) and $(p(\xi), q(\xi))$ is given by (6.29, 6.30).*

6.4.1 The observer canonical form

The first solution to the state representation problem yields what is called *the observer canonical form.*

Let $p(\xi), q(\xi)$ be given by (6.33). Let $d \in \mathbb{R}$ and $q(\xi) \in \mathbb{R}[\xi]$ be such that $q(\xi) = dp(\xi) + \tilde{q}(\xi)$ and $\deg \tilde{q}(\xi) < \deg p(\xi)$. Note that $\tilde{q}(\xi)$ is given by $\tilde{q}(\xi) = q(\xi) - q_n p(\xi)$. Denote the coefficients of the polynomial $\tilde{q}(\xi)$ by

$\tilde{q}_0, \ldots, \tilde{q}_{n-1}$ and define $(A, b, c, d) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n}$ by

$$A := \begin{bmatrix} 0 & \dots & \dots & \dots & 0 & -p_0 \\ 1 & 0 & \dots & \dots & \dots & -p_1 \\ 0 & 1 & 0 & \dots & \dots & -p_2 \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & 0 & -p_{n-2} \\ 0 & \dots & \dots & 0 & 1 & -p_{n-1} \end{bmatrix}, \quad b := \begin{bmatrix} \tilde{q}_0 \\ \tilde{q}_1 \\ \vdots \\ \vdots \\ \tilde{q}_{n-2} \\ \tilde{q}_{n-1} \end{bmatrix}, \tag{6.37}$$

$$c := \begin{bmatrix} 0 & \dots & \dots & \dots & 0 & 1 \end{bmatrix}, \qquad d := q_n.$$

Theorem 6.4.2 *Let (A, b, c, d) be defined by (6.37). Then*

$$\begin{aligned} \frac{d}{dt}x &= Ax + bu, \\ y &= cx + du \end{aligned} \tag{6.38}$$

is an i/s/o representation of (6.34).

Proof Using the notation of Corollary 6.3.2, define $\bar{p}(\xi) := \det(I\xi - A)$ and $\bar{r}(\xi) := \bar{p}(\xi)c(I\xi - A)^{-1}$. In order to determine an explicit expression for $\bar{p}(\xi)$, we apply the following sequence of elementary row operations to $(I\xi - A)$. Multiply the last row by ξ and add it to the last but one row. Then, multiply in the resulting matrix the $(n-1)$th row by ξ and add to the $(n-2)$th row. Etc., etc. Finally, multiply the second row by ξ and add it to the first row. The resulting matrix is

$$\begin{bmatrix} 0 & \cdots & 0 & \cdots & \cdots & p_0 + p_1(\xi) + \cdots + p_{n-1}\xi^{n-1} + \xi^n \\ -1 & 0 & 0 & \cdots & \cdots & p_1 + p_2\xi + \cdots + p_{n-1}\xi^{n-2} + \xi^{n-1} \\ 0 & -1 & 0 & \cdots & \cdots & p_2 + p_3\xi + \cdots + p_{n-1}\xi^{n-3} + \xi^{n-2} \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & -1 & 0 & p_{n-2} + p_{n-1}\xi + \xi^2 \\ 0 & 0 & 0 & 0 & -1 & p_{n-1} + \xi \end{bmatrix}. \tag{6.39}$$

From (6.39) we conclude that

$$\bar{p}(\xi) = p(\xi). \tag{6.40}$$

By direct calculation it is easily seen that

$$\bar{r}(\xi) = \begin{bmatrix} 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \end{bmatrix}. \tag{6.41}$$

Obviously, $\bar{p}(\xi)$ and $\bar{r}(\xi)$ have no common factors. It follows from Corollary 6.3.2 that the i/o behavior of (6.38) is described by $p(\frac{d}{dt})y = q(\frac{d}{dt})u$, where $q(\xi) := \bar{r}(\xi)b + dp(\xi)$. This proves the theorem. □

The following theorem explains why the i/s/o representation (6.38) is called the *observer* canonical form. The adjective "canonical" will be explained in Section 6.5.

Theorem 6.4.3 *The representation (6.38) is observable. It is also controllable if and only if $p(\xi)$ and $q(\xi)$ have no common factors. In other words, the state space model (6.38) is controllable if and only if the i/o model (6.34) is controllable.*

Proof According to Theorem 5.3.9, (6.38) is observable if and only if the rank of the associated observability matrix $\mathfrak{O}$ is n. It is easily verified that in this case $\mathfrak{O}$ is of the form

$$\mathfrak{O} = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & & 0 & 1 & * \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 0 & 1 & \iddots & & \vdots \\ 1 & * & \cdots & \cdots & * \end{bmatrix}.$$

Obviously, $\mathfrak{O}$ has rank n. Of course, the observability could also have been determined from the fact that $\bar{r}(\xi)$ and $\bar{p}(\xi)$ have no common factors and from Theorem 5.5.1, Part 1.

The second part of the statement is left to the reader as Exercise 5.14. □

Remark 6.4.4 The observer canonical form admits a nice visualization in terms of a signal flow diagram, as depicted in Figure 6.4. □

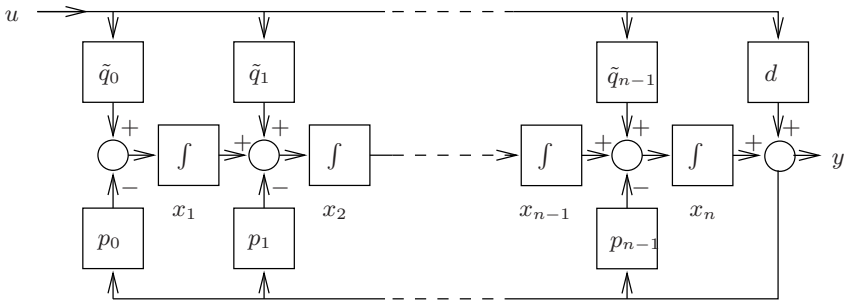


FIGURE 6.4. Signal flow diagram of the observer canonical form (6.37).

Remark 6.4.5 It is of interest to relate the state vector constructed in (6.38) directly in terms of u and y and their derivatives. If we confine ourselves to the $\mathcal{C}^\infty$ trajectories, this can be done as follows.

Denote the components of the state vector x by $x_1, \ldots, x_n$. From (6.37, 6.38) it follows that

$$\begin{aligned}
x_n &= y - du, \\
x_{n-1} &= \tfrac{d}{dt}(y - du) + p_{n-1}(y - du) - \tilde{q}_{n-1}u, \\
x_{n-2} &= \tfrac{d^2}{dt^2}(y - du) + \tfrac{d}{dt}(p_{n-1}(y - du) - \tilde{q}_{n-1}u) + p_{n-2}(y - du) - \tilde{q}_{n-2}u, \\
&\vdots \\
x_2 &= \tfrac{d^{n-2}}{dt^{n-2}}(y - du) + \tfrac{d^{n-3}}{dt^{n-3}}(p_{n-1}(y - du) - \tilde{q}_{n-1}u) + \cdots + p_2(y - du), \\
&\quad -\tilde{q}_2 u \\
x_1 &= \tfrac{d^{n-1}}{dt^{n-1}}(y - du) + \tfrac{d^{n-2}}{dt^{n-2}}(p_{n-1}(y - du) - \tilde{q}_{n-1}u) + \cdots + p_1(y - du). \\
&\quad -\tilde{q}_1 u
\end{aligned} \tag{6.42}$$

These expressions show how state variables can be created from suitable combinations of (u, y) and their derivatives up to order $n - 1$. The meticulous reader may wonder how to interpret (6.42) in the case that u is not sufficiently differentiable. Of course, if u (and y) are not sufficiently differentiable, then (6.42) has no interpretation in the classical sense. In that case, (6.42) should interpreted in the sense of *weak* solutions as discussed in Chapter 2. □

Example 6.4.6 (Example 6.3.6 continued) In Example 6.3.6 we have derived the i/o representation (6.32) of the i/s/o system (6.31). We could again represent (6.32) in state space form. The observer canonical representation of (6.32) is given by

$$\begin{aligned}
\frac{d}{dt}\tilde{x} &= \tilde{A}\tilde{x} + \tilde{b}u, \\
y &= \tilde{c}\tilde{x},
\end{aligned}$$

with

$$\tilde{A} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \tilde{c} = \begin{bmatrix} 0 & 1 \end{bmatrix}. \tag{6.43}$$

The interesting feature of the representation (6.43) is that its order, the dimension of its state space, is only two, whereas the original dimension of (6.31) was three. Also, it is obvious that $(\tilde{c}, \tilde{A})$ in (6.43) is an observable pair. It appears therefore that we have removed the nonobservable part. In fact, that is exactly what has happened, whence the reduction of the dimension of the state space. Notice that we could as well have done this directly, without first deriving an i/o representation, by transforming the i/s/o system into the form (5.49).

A natural question to ask is whether or not we could find an even lower-dimensional i/s/o representation. We will prove in Section 6.5 that there does not exist a lower-dimensional i/s/o representation of a given i/s/o behavior if and only if it is observable; see Theorem 6.5.11. □

6.4.2 The controller canonical form

Our second solution to the state representation problem yields what is called *controller canonical form*. This representation exists only if the i/o system is controllable.

Let $p(\xi), q(\xi)$ be given by (6.33). Consider the i/o system

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u. \tag{6.44}$$

Let $d \in \mathbb{R}$ and $\tilde{q}(\xi) \in \mathbb{R}[\xi]$ be such that

$$q(\xi) = dp(\xi) + \tilde{q}(\xi), \quad \text{with } \deg \tilde{q}(\xi) < \deg p(\xi). \tag{6.45}$$

Note that $\tilde{q}(\xi)$ is given by $\tilde{q}(\xi) = q(\xi) - q_n p(\xi)$. Denote the coefficients of $\tilde{q}(\xi)$ by $\tilde{q}_0, \ldots, \tilde{q}_{n-1}$ and define (A, b, c, d) by

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 \cdots & 0 & \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ -p_0 & -p_1 & \cdots & \cdots & \cdots & -p_{n-1} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \\ c &= \begin{bmatrix} \tilde{q}_0 & \tilde{q}_1 & \cdots & \cdots & \cdots & \tilde{q}_{n-1} \end{bmatrix}, \qquad d = q_n. \end{aligned} \tag{6.46}$$

Theorem 6.4.7 *Let (A, b, c, d) be defined by (6.45, 6.46). Consider the i/s/o system defined by*

$$\begin{aligned} \frac{d}{dt}x &= Ax + bu, \\ y &= cx + du. \end{aligned} \tag{6.47}$$

Assume that $p(\xi)$ and $q(\xi)$ have no common factors, i.e., that (6.44) is controllable. Then the i/o behavior of (6.47) is described by $p(\frac{d}{dt})y = q(\frac{d}{dt})u$.

Proof Using the notation of Corollary 6.3.2, define

$$\bar{p}(\xi) := \det(I\xi - A), \quad \bar{r}(\xi) := \bar{p}(\xi)c(I\xi - A)^{-1}, \quad \bar{s}(\xi) := \bar{p}(\xi)(I\xi - A)^{-1}b. \tag{6.48}$$

It is easy to see, see also (6.40, 6.41), that

$$\bar{p}(\xi) = p(\xi), \qquad \bar{s}(\xi) = [1 \ \xi \ \xi^2 \cdots \xi^{n-1}]^T. \tag{6.49}$$

From (6.49) it follows that $\tilde{q}(\xi) = c\bar{s}(\xi)$, and hence, by (6.48), that $\tilde{q}(\xi) = \bar{r}(\xi)b$. Since by assumption $p(\xi)$ and $q(\xi)$ have no common factors, neither

do $p(\xi)$ and $\tilde{q}(\xi)$, and as a consequence, $\bar{p}(\xi)$ and $\bar{r}(\xi)$ have no common factors. It follows from Corollary 6.3.2 that the i/o behavior of (6.47) is described by $p(\frac{d}{dt})y = q(\frac{d}{dt})u$, where $q(\xi) := \bar{r}(\xi)b + d$. This proves the theorem. □

Remark 6.4.8 Notice that in Theorem 6.4.7 we assumed that $p(\xi)$ and $q(\xi)$ have no common factors. This is in contrast to the situation in Theorem 6.4.2, where this requirement was not needed. In Exercise 6.18 it is shown that the result of Theorem 6.4.7 does not hold if $p(\xi)$ and $q(\xi)$ have a common factor. □

Remark 6.4.9 The i/s/o representation (6.47) is called the *controller canonical form*. The reason why it is called *controller* canonical form is now explained. The adjective "canonical" is explained in Section 6.5. □

Theorem 6.4.10 *Assume that $p(\xi)$ and $q(\xi)$ have no common factor. Then the system defined by the controller canonical form (6.47) is both controllable and observable.*

Proof According to Theorem 5.2.18, (6.47) is controllable if and only if the rank of the associated controllability matrix $\mathfrak{C}$ is n. It is easily verified that in this case $\mathfrak{C}$ is of the form

$$\mathfrak{C} = \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & & 0 & 1 & * \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 0 & 1 & \iddots & & \vdots \\ 1 & * & \cdots & \cdots & * \end{bmatrix}.$$

Obviously, $\mathfrak{C}$ has rank n.

The proof of observability is left to the reader as Exercise 5.18. □

Remark 6.4.11 As for the observer canonical form, we can express in the representation (6.47) the state x in terms of u and y and their derivatives, provided that we restrict our attention again to the $\mathcal{C}^\infty$ trajectories.

Consider the i/o system

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u, \tag{6.50}$$

Where $p(\xi), q(\xi) \in \mathbb{R}[\xi]$ are of the form

$$\begin{aligned} p(\xi) &= p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n, \\ q(\xi) &= q_0 + q_1\xi + \cdots + q_{n-1}\xi^{n-1} + q_n\xi^n \end{aligned}$$

and have no common factors. Let (A, b, c, d) be given by (6.46). By Theorem 6.4.7 we know that $\frac{d}{dt}x = Ax + bu$, $y = cx + du$ is a state representation of (6.50). We want to express the state x in u, y, and their derivatives.

A latent variable representation of (6.50) is given by

$$y = q(\frac{d}{dt})\ell, \qquad u = p(\frac{d}{dt})\ell. \tag{6.51}$$

This is easily proved by applying the elimination procedure, see Exercise 6.12. For our purposes, however, it is more convenient to prove the input/output equivalence of (6.50) and (6.51) directly. Choose (u, y, ℓ) such that (6.51) is satisfied. Then, $p(\frac{d}{dt})y = p(\frac{d}{dt})q(\frac{d}{dt})\ell = q(\frac{d}{dt})p(\frac{d}{dt})\ell = q(\frac{d}{dt})u$. This shows that (u, y) satisfies (6.50).

To prove the converse, choose (u, y) such that (6.50) is satisfied. Since $p(\xi)$ and $q(\xi)$ have no common factors, there exist, by Corollary B.1.7 (Bezout), polynomials $a(\xi)$ and $b(\xi)$ such that

$$a(\xi)p(\xi) + b(\xi)q(\xi) = 1. \tag{6.52}$$

Define

$$\ell := b(\frac{d}{dt})y + a(\frac{d}{dt})u. \tag{6.53}$$

Then

$$q(\frac{d}{dt})\ell = q(\frac{d}{dt})b(\frac{d}{dt})y + q(\frac{d}{dt})a(\frac{d}{dt})u = (q(\frac{d}{dt})b(\frac{d}{dt}) + p(\frac{d}{dt})a(\frac{d}{dt}))y = y. \tag{6.54}$$

In the same way one proves that

$$p(\frac{d}{dt})\ell = u. \tag{6.55}$$

From (6.54) and (6.55) we conclude that (u, y, ℓ) satisfies (6.51).

Define the vector-valued function z as

$$z := \begin{bmatrix} \ell \\ \frac{d}{dt}\ell \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}}\ell \end{bmatrix}. \tag{6.56}$$

It follows from (6.54, 6.55) and the definition of z that (u, y, z) also satisfies

$$\begin{aligned} \frac{d}{dt}z &= Az + bu, \\ y &= cz + du, \end{aligned} \tag{6.57}$$

with (A, b, c) defined by (6.46). Since (u, y, x) also satisfies (6.57) and since by Theorem 6.4.10, (A, c) is an observable pair, it follows that $x = z$. Combining (6.53) and (6.56) we conclude that the state is given by

$$x = \begin{bmatrix} b(\frac{d}{dt})y + a(\frac{d}{dt})u \\ \frac{d}{dt}[b(\frac{d}{dt})y + a(\frac{d}{dt})u] \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}}[b(\frac{d}{dt})y + a(\frac{d}{dt})u] \end{bmatrix}. \tag{6.58}$$

It should be noted that the right-hand side in (6.58) is independent of the choice of the polynomials $a(\xi)$ and $b(\xi)$ satisfying (6.52). See Exercise 6.6 for a proof of this statement. □

Thus in both the observer and controller canonical forms, the state can be expressed in terms of u and y and their derivatives. In the observer canonical form these expressions are readily obtained from the coefficients of the polynomials $p(\xi)$ and $q(\xi)$, see (6.42), whereas in the controller canonical form we first have to solve the Bezout equation (6.52).

Remark 6.4.12 Also the controller canonical form admits a nice visualization in terms of a signal flow diagram, as depicted in Figure 6.5. □

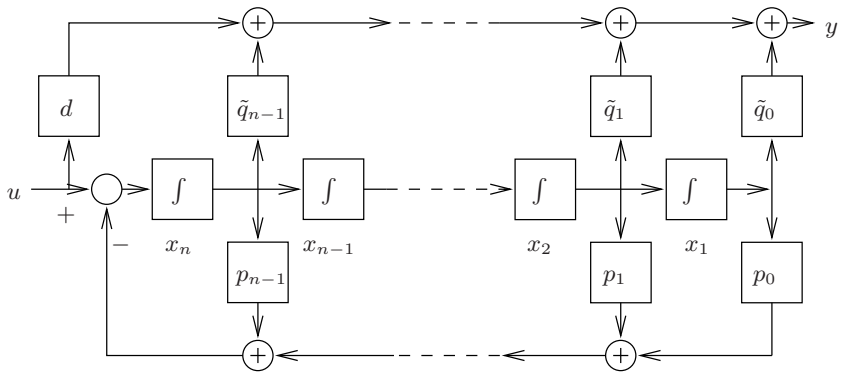


FIGURE 6.5. Signal flow diagram of the controller canonical form (6.46).

Example 6.4.13 Consider the i/o system defined by

$$y + \frac{d}{dt}y + \frac{d^2}{dt^2}y = 2u + \frac{d}{dt}u. \tag{6.59}$$

The corresponding polynomials are $p(\xi) = 1+\xi+\xi^2$, $q(\xi) = 2+\xi$. Obviously, $p(\xi)$ and $q(\xi)$ have no common factor so that we can form the controller

canonical form. This yields

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 2 & 1 \end{bmatrix}.$$

According to Theorem 6.4.7 the controller canonical representation of (6.59) is now given by $\frac{d}{dt}x = Ax + bu$, $y = cx$. In order to express the state in terms of u and y, we need polynomials $a(\xi)$ and $b(\xi)$ such that $a(\xi)p(\xi) + b(\xi)q(\xi) = 1$. It follows by inspection that we may take $a(\xi) = \frac{1}{3}$ and $b(\xi) = \frac{1}{3}(1 - \xi)$. By (6.58) it follows that x is given by:

$$x = \begin{bmatrix} \frac{1}{3}y - \frac{1}{3}\frac{d}{dt}y + \frac{1}{3}u \\ \\ \frac{1}{3}\frac{d}{dt}y - \frac{1}{3}\frac{d^2}{dt^2}y + \frac{1}{3}\frac{d}{dt}u \end{bmatrix}.$$

Using (6.59) we can eliminate $\frac{d^2}{dt^2}y$ from this expression and obtain

$$x = \begin{bmatrix} \frac{1}{3}y - \frac{1}{3}\frac{d}{dt}y + \frac{1}{3}u \\ \\ \frac{1}{3}y + \frac{2}{3}\frac{d}{dt}y - \frac{2}{3}u \end{bmatrix}. \tag{6.60}$$

See Exercise 6.24 for an alternative calculation. □

Note that both the controller and observer canonical forms are of dimension n, the degree of the polynomial $p(\xi)$.

6.5 Canonical Forms and Minimal State Space Representations

In Section 4.6 we have shown that the i/o behavior of an i/s/o representation is invariant under state space transformations. This observation gives rise to several questions. The first question that we study in this section is, *Given the set of all i/s/o representations of the same i/o system, do there exist some natural representatives?* It turns out that for general i/o systems the observer canonical form provides such a representative, and so does the controller canonical form for controllable i/o systems. The way to approach this mathematically is through the notions of equivalence relations and canonical forms. The second question is, *When are two i/s/o representations of the same i/o system equivalent in the sense that they can be transformed into one another by means of a state space transformation?* We show that this is the case when both i/s/o representations are observable. The third

result that we prove in this section is that among all possible state space representations of a given i/o behavior, the observable ones require the smallest dimension of the state space.

6.5.1 Canonical forms

Let us first discuss the notions of *canonical forms* and *trim canonical forms.* Let $\mathfrak{A}$ be a nonempty set. A binary relation $\sim$ on $\mathfrak{A}$ is simply a subset of $\mathfrak{A} \times \mathfrak{A}$. It is called an *equivalence relation* if (i) for all $a \in \mathfrak{A}$: $a \sim a$, (ii) for all $a, b \in \mathfrak{A}$: $a \sim b$ implies that $b \sim a$, and (iii) for all $a, b, c \in \mathfrak{A}$: $a \sim b$ and $b \sim c$ implies that $a \sim c$. If $\sim$ is an equivalence relation on $\mathfrak{A}$, then we define for each $a \in \mathfrak{A}$ the *equivalence class* of a as the set of all $b \in \mathfrak{A}$ such that $a \sim b$. We denote this set by $\bar{a}$. The equivalence relation $\sim$ partitions $\mathfrak{A}$ into equivalence classes. The set of equivalence classes is denoted by $\mathfrak{A}/\sim$. A subset $\mathfrak{K} \subset \mathfrak{A}$ is called a *canonical form* for $\sim$ if $\mathfrak{K}$ contains *at least* one element from each equivalence class. It is called a *trim canonical form* if $\mathfrak{K}$ contains *precisely one* element from each equivalence class.

Following the terminology in Section 4.6, we call two systems of the type (4.52), or equivalently two quadruples (A_1, B_1, C_1, D_1), (A_2, B_2, C_2, D_2), *similar* if there exists a nonsingular matrix S such that $(S^{-1}A_1S, S^{-1}B_1, C_1S, D_1) = (A_2, B_2, C_2, D_2)$. Similarity defines an equivalence relation on the set of quadruples (A, B, C, D). We now prove that for SISO systems the observer canonical form provides a trim canonical form on the set of (A, B, C, D)s for which (A, C) is observable.

We start with the following preliminary result.

Theorem 6.5.1 *Let $(A_1, c_1), (A_2, c_2) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{1\times n}$ be observable pairs, and assume that A_1 and A_2 have the same characteristic polynomial:* $\det(I\xi - A_1) = \det(\xi I - A_2) =: p_0 + \cdots + p_{n-1}\xi^{n-1} + \xi^n$. *Then there exists exactly one nonsingular matrix S such that*

$$(S^{-1}A_1S, c_1S) = (A_2, c_2).$$

Proof (Existence) Denote by $\mathfrak{O}_i$ the observability matrices of (A_i, c_i), $i = 1, 2$:

$$\mathfrak{O}_i = \begin{bmatrix} c_i \\ c_iA_i \\ \vdots \\ c_iA_i^{n-1} \end{bmatrix}.$$

Define the matrix S as $S := \mathfrak{O}_1^{-1}\mathfrak{O}_2$. Then $c_1A_1^kS = c_2A_2^k$, for $k = 0, \ldots, n-1$. Since A_1 and A_2 have the same characteristic polynomial, it follows from the Cayley–Hamilton theorem that also $c_1A_1^nS = c_2A_2^n$. We conclude that $\mathfrak{O}_1A_1S = \mathfrak{O}_2A_2$, and therefore$S^{-1}A_1S = A_2$.

(Uniqueness) Suppose that $(S^{-1}A_1S, c_1S) = (A_2, c_2)$. Then it follows by direct calculation that $\mathfrak{O}_1 S = \mathfrak{O}_2$, and hence that $S = \mathfrak{O}_1^{-1}\mathfrak{O}_2$. □

A direct consequence of Theorem 6.5.1 is that every observable pair (A, c) may be transformed into observer canonical form, as we show next.

Theorem 6.5.2 *Let $(A, c) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{1\times n}$ be an observable pair and let $p(\xi) = p_0 + \cdots + p_{n-1}\xi^{n-1} + \xi^n$ be the characteristic polynomial of A. Then there exists a unique nonsingular matrix S such that $(S^{-1}AS, cS) = (\tilde{A}, \tilde{c})$, with*

$$\tilde{A} = \begin{bmatrix} 0 & \cdots & 0 & \cdots & \cdots & -p_0 \\ 1 & 0 & 0 & \cdots & \cdots & -p_1 \\ 0 & 1 & 0 & \cdots & \cdots & -p_2 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & 0 & 1 & 0 & -p_{n-2} \\ 0 & 0 & 0 & 0 & 1 & -p_{n-1} \end{bmatrix}, \quad \tilde{c} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \tag{6.61}$$

Proof This is direct consequence of Theorem 6.5.1, using the observability of the pair $(\tilde{A}, \tilde{c})$ and the fact that $\det(I\xi - A) = \det(I\xi - \tilde{A})$. □

An immediate consequence of Theorem 6.5.2 is the following result.

Corollary 6.5.3 *The observer canonical form is a trim canonical form for the observable SISO systems.*

Remark 6.5.4 As we have shown in Section 6.4.1, every SISO system may be represented by an observable state space representation. Corollary 6.5.3 formalizes that the observer canonical representation is in some sense a natural choice. □

By duality, the controller canonical form yields a trim canonical form for controllable i/s/o systems.

Corollary 6.5.5 *The controller canonical form is a trim canonical form for the controllable SISO systems.*

Proof Let (A, b, c, d) represent a controllable SISO i/s/o system. Then (A^T, b^T) is an observable pair, and according to Theorem 6.5.2 there exists a unique nonsingular matrix S that transforms (A^T, b^T) into the form (6.61). It follows that $(S^T)^{-1}$ transforms (A, b, c, d) into the controller canonical form (6.46). □

Remark 6.5.6 As we have shown in Section 6.2, every controllable SISO system may be represented by a controllable state space representation. Corollary 6.5.5 formalizes that the controller canonical representation is in some sense a natural choice. □

6.5.2 Equivalent state representations

Using Theorem 6.5.1 we can now prove the converse of Theorem 4.6.2, namely that two observable i/s/o representations of the same i/o behavior are similar. In the proof we use a small technical lemma.

Lemma 6.5.7 *Let (A, c) be an observable pair and let $p(\xi) = \det(I\xi - A)$. Define the n-dimensional polynomial row vector $r(\xi) := p(\xi)c(I\xi - A)^{-1}$. Then there exist $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ such that the n vectors $r(\lambda_i)$, $i = 1, \ldots, n$, are linearly independent.*

Proof By Theorem 6.5.2 there exists a nonsingular matrix S such that $(\tilde{A}, \tilde{c}) = (S^{-1}AS, cS)$ is as in (6.61). Define $\tilde{r}(\xi) := p(\xi)\tilde{c}(I\xi - \tilde{A})^{-1}$. It follows from (6.41) that $\tilde{r}(\xi) = [1\ \xi\ \xi^2 \cdots\ \xi^{n-1}]$. Choose n distinct complex numbers λ_i, $i = 1, \ldots, n$. Then the n vectors $\tilde{r}(\lambda_i)$, $i = 1, \ldots, n$, form a nonsingular Vandermonde matrix and are hence independent. Since $r(\xi) = \tilde{r}(\xi)S$, the vectors $r(\lambda_i)$, $i = 1, \ldots, n$, are also linearly independent. □

Theorem 6.5.8 *Let $(A_k, b_k, c_k, d_k) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n} \times \mathbb{R}$, $k = 1, 2$, be such that (A_k, c_k) are observable pairs, $k = 1, 2$. Let $\mathfrak{B}_{\mathrm{i/s/o},k}$ be defined by*

$$\begin{array}{rcll} \dfrac{d}{dt}x_k & = & A_k x_k + b_k u, & \\ y_k & = & c_k x_k + d_k u, & \end{array} \qquad k = 1, 2.$$

Then the state representations $\mathfrak{B}_{\mathrm{i/s/o},1}$ and $\mathfrak{B}_{\mathrm{i/s/o},2}$ define the same input/output behavior if and only if there exists a nonsingular matrix S such that

$$(S^{-1}A_1S, S^{-1}b_1, c_1S, d_1) = (A_2, b_2, c_2, d_2).$$

Proof The "if" part is just Theorem 4.6.2.

"Only if" part. Since (A_1, c_1) and (A_2, c_2) are observable, we may apply Corollary 6.3.3: the i/o behavior defined by $\mathfrak{B}_{\mathrm{i/s/o},k}$ is given by

$$p_k(\frac{d}{dt})y = q_k(\frac{d}{dt})u, \qquad k = 1, 2,$$

with

$$p_k(\xi) := \det(I\xi - A_k), \quad q_k(\xi) := p_k(\xi)c_k(I\xi - A_k)^{-1}b_k + d_k p_k(\xi), \tag{6.62}$$

Since $\mathfrak{B}_{\mathrm{i/s/o},1}$ and $\mathfrak{B}_{\mathrm{i/s/o},2}$ define the same i/o behavior, we conclude that

$$p_1(\xi) = p_2(\xi) =: p(\xi), \qquad q_1(\xi) = q_2(\xi) =: q(\xi), \tag{6.63}$$

and therefore $d_1 = d_2$. From (6.63) it follows that A_1 and A_2 have the same characteristic polynomial, and by Theorem 6.5.1 it follows that there exists

a nonsingular matrix S such that $(S^{-1}A_1S, c_1S) = (A_2, c_2)$. Combining this with (6.62, 6.63), we conclude that $p(\xi)c_1(I\xi - A_1)^{-1}S^{-1}b_1 = p(\xi)c_1(I\xi - A_1)^{-1}b_2$. Using Lemma 6.5.7, we obtain that $b_2 = S^{-1}b_1$. This concludes the proof. □

A direct consequence of Theorem 6.5.8 is:

Theorem 6.5.9 *Consider the i/o behavior $\mathfrak{B}$ defined by $p(\frac{d}{dt})y = q(\frac{d}{dt})u$, with $p(\xi), q(\xi) \in \mathbb{R}[\xi]$, $p(\xi)$ monic of degree n, and $\deg q(\xi) \leq n$. Assume that $p(\xi)$ and $q(\xi)$ have no common factors. Then the controller and observer canonical i/s/o representations of $\mathfrak{B}$ are equivalent.*

6.5.3 Minimal state space representations

Let us now come back to the issue of state space representations of minimal dimension. As claimed in the introduction to this section, minimality of the dimension of the state space turns out to be equivalent to observability. This statement follows immediately from the following theorem.

Theorem 6.5.10 *Assume that $\frac{d}{dt}x_i = A_ix_i + b_iu$, $y = c_ix_i + d_iu$, $(A_i, b_i, c_i, d_i) \in \mathbb{R}^{n_i \times n_i} \times \mathbb{R}^{n_i \times 1} \times \mathbb{R}^{1 \times n_i} \times \mathbb{R}$, define the same i/o behavior and that (A_1, c_1) is observable. Then $n_1 \leq n_2$.*

Proof According to Corollary 6.3.3 the i/o behavior corresponding to the quadruple (A_1, b_1, c_1, d_1) is of the form

$$p_1(\frac{d}{dt})y = q_1(\frac{d}{dt})u, \tag{6.64}$$

with $p_1(\xi) = \det(I\xi - A_1)$. By Corollary 6.3.2, the i/o behavior defined by (A_2, b_2, c_2, d_2) is given by

$$p_2(\frac{d}{dt})y = q_2(\frac{d}{dt})u, \tag{6.65}$$

where $p_2(\xi)$ divides $\det(I\xi - A_2)$. Since the i/o behaviors defined by (6.64) and (6.65) are the same, we conclude from Theorem 3.6.2 that $p_1(\xi) = p_2(\xi)$. This implies that $\det(I\xi - A_1)$ divides $\det(I\xi - A_2)$ and hence that $n_1 \leq n_2$. □

As an immediate consequence of Theorem 6.5.10 we obtain the following result.

Theorem 6.5.11 *Consider the SISO system (6.34). There exists an observable i/s/o representation, namely the observer canonical form (6.37). All observable state space representations are of the same dimension. Moreover, this dimension is minimal among all possible state space representations.*

The above result also holds for multivariable i/o systems, as discussed in Section 3.3. We do not provide the details. We call an i/s/o representation of a given i/o behavior *minimal* if among all possible i/s/o representations its state space has minimal dimension. It follows from Theorem 6.5.11 that for systems of the form (6.36) minimality is equivalent to observability of the pair (A, c). Note that a minimal i/s/o representation need not be controllable (see Exercise 6.18). In fact, a minimal i/s/o representation is controllable if and only if the i/o system that it represents is also controllable; see Exercise 6.26.

6.6 Image Representations

Thus far we have studied several representations of linear time-invariant differential systems:

1. $R(\frac{d}{dt})w = 0$. This is the type of representation that is at the core of this book. For obvious reasons we could call this a *kernel representation.*

2. $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$. Such equations with latent variables are usually obtained as a result of modeling from first principles. We have shown in Section 6.2 that by eliminating the latent variable ℓ we obtain a kernel representation for the manifest behavior.

3. $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$, with $P^{-1}(\xi)Q(\xi)$ proper. Every kernel representation can be brought into such an input/output form by an appropriate partition of w in u and y.

4. $E\dfrac{d}{dt}x+Fx+Gw = 0$. These state space representations form a special class of latent variable models. The latent variable x has the property of state.

5. $\frac{d}{dt}x = Ax + Bu$, $y = Cx + Du$. These input/state/output models are state space models of a special structure. They are compatible with both the input/output structure and the state space structure of the behavior. All systems defined by a kernel representation can be brought into this form, although we only proved this for SISO systems.

In this section we want to take a brief look at latent variable models described by $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$. A special case is obtained when $R(\xi)$ is just the identity matrix:

$$w = M(\frac{d}{dt})\ell, \qquad M(\xi) \in \mathbb{R}^{q\times m}[\xi]. \tag{6.66}$$

Representations of the form (6.66) are, for obvious reasons, called *image representations.* In (6.66), the manifest behavior is the image of $\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ under the differential operator $M(\frac{d}{dt})$. The question that we want to address is under what conditions a system defined by a kernel representation is equivalent to one defined by an image representation. More precisely, let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$ and denote the corresponding behavior by $\mathfrak{B}_{\mathrm{ker}}$. Denote the behavior of w induced by (6.66) by $\mathfrak{B}_{\mathrm{im}} := \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid \exists \ell \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m) \text{ such that } w = M(\frac{d}{dt})\ell\}$. We want to find conditions on $R(\xi)$ under which there exists $M(\xi)$ such that $\mathfrak{B}_{\mathrm{ker}} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) = \mathfrak{B}_{\mathrm{im}} \cap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$.

Because the matrix $[I \quad M(\lambda)]$ has rank q for all $\lambda \in \mathbb{C}$, image representations are always controllable. Hence a necessary condition on $R(\xi)$ is that it represents a controllable system, see Exercise 6.26. The somewhat surprising result is that controllability of $\mathfrak{B}_{\mathrm{ker}}$ is also a sufficient condition for the existence of an image representation.

Theorem 6.6.1 *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Then there exists an integer m and a matrix $M(\xi) \in \mathbb{R}^{q\times m}[\xi]$ such that $\mathfrak{B}_{\mathrm{ker}} = \mathfrak{B}_{\mathrm{im}}$ if and only if $\mathfrak{B}_{\mathrm{ker}}$ is controllable.*

Proof The "only if" part is shown in Exercise 6.26. We therefore consider only the "if" part. In view of Theorem 2.5.23, we may assume that $R(\xi)$ is of full row rank. Since $R(\xi)$ represents a controllable behavior, we know that for all $\lambda \in \mathbb{C}$, $\operatorname{rank} R(\lambda) = g$. We prove that there exists $R'(\xi) \in \mathbb{R}^{(q-g)\times q}[\xi]$ such that $\operatorname{col}(R(\xi), R'(\xi))$ is unimodular. In order to see this, choose unimodular matrices $U(\xi)$ and $V(\xi)$ such that $U(\xi)R(\xi)V(\xi)$ is in Smith form: $U(\xi)R(\xi)V(\xi) = \begin{bmatrix} D(\xi) & 0 \end{bmatrix}$. Now $R(\lambda)$ is of full rank for all $\lambda \in \mathbb{C}$, and therefore the diagonal matrix $D(\xi)$ can be taken to be the identity matrix. This implies that

$$R(\xi)\underbrace{V(\xi)\begin{bmatrix} U(\xi) & 0 \\ 0 & I \end{bmatrix}}_{W(\xi)} = \begin{bmatrix} U^{-1}(\xi) & 0 \end{bmatrix}\begin{bmatrix} U(\xi) & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \end{bmatrix}. \tag{6.67}$$

It follows from (6.67) that there exists a unimodular matrix $W(\xi)$ such that $R(\xi) = \begin{bmatrix} I & 0 \end{bmatrix} W(\xi)^{-1}$. This means that $R(\xi)$ forms the first g rows of a unimodular matrix (for $g = 1$ this is Theorem 2.5.10, except that there it forms the last row). Form $R'(\xi)$ by the remaining $q - g$ rows of $W(\xi)^{-1}$ and define the latent variable system

$$\begin{bmatrix} R(\frac{d}{dt}) \\ R'(\frac{d}{dt}) \end{bmatrix} w = \begin{bmatrix} 0 \\ I \end{bmatrix} \ell. \tag{6.68}$$

According to Definition 6.2.7, (6.68) is a latent variable representation of $\mathfrak{B}_{\ker}$. Finally, premultiplication of both sides of (6.68) by $W(\frac{d}{dt})$ yields

$$w = W(\frac{d}{dt}) \begin{bmatrix} 0 \\ I \end{bmatrix} \ell =: M(\frac{d}{dt})\ell.$$

Notice that $M(\xi) \in \mathbb{R}^{q \times (q-g)}[\xi]$, so that $m = q - g$. □

In the subsection on the controller canonical form we have already used an image representation of controllable SISO systems. Indeed, if $p(\frac{d}{dt})y = q(\frac{d}{dt})u$ is a controllable SISO system, then an image representation of this SISO system is

$$\begin{aligned} y &= q(\tfrac{d}{dt})\ell, \\ u &= p(\tfrac{d}{dt})\ell. \end{aligned}$$

See Remark 6.4.11 for details.

6.7 Recapitulation

In this chapter we discussed two related topics. First, the elimination of latent variables and second the relation between i/o models and i/s/o representations. The main points were:

- Disregarding smoothness issues, the manifest behavior of a behavior with latent variables described by differential equations of the form $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$ can be described by $R'(\frac{d}{dt})w = 0$ for a suitable polynomial matrix $R'(\xi)$. An algorithm was derived to calculate the polynomial matrix $R'(\xi)$ from $M(\xi)$ and $R(\xi)$ (Theorem 6.2.6).
- The elimination algorithm was applied to obtain the i/o behavioral equation $p(\frac{d}{dt})y = q(\frac{d}{dt})u$ from the i/s/o equations $\frac{d}{dt}x = Ax + bu$, $y = cx + du$ (Theorem 6.3.1).
- Two canonical i/s/o representation of a given i/o system were derived: the observer canonical form and the controller canonical form. The latter representation applies to controllable systems only (Theorems 6.4.2 and 6.4.7).
- We derived a complete characterization of equivalent observable i/s/o representations of a given i/o system. The main result states that all observable i/s/o representations of a given i/o system can be transformed into each other by means of state space transformation (Theorem 6.5.8). Moreover, the dimension of the state space of an observable i/s/o representation is minimal among all possible i/s/o representations (Theorem 6.5.11).
- In the last section, we studied representations of the form $w = M(\frac{d}{dt})\ell$. These are referred to as image representations. A behavior in kernel representation with $R(\frac{d}{dt})w = 0$ admits an image representation if and only if it is controllable (Theorem 6.6.1).

6.8 Notes and References

The importance of latent variables and the relevance of the elimination theorem in the context of differential systems originated in [59, 60]. However, not very surprisingly in view of their natural occurrence in first principles modeling, there were earlier attempts to incorporate latent variables in the description of dynamical systems. In this context it is worth mentioning Rosenbrock's notion of partial state [48] and, of course, the state space theory of dynamical systems. Elimination of latent variables was treated in depth, including the exact elimination question and the related smoothness issue, in [44]. The construction of state space representations originates in the work of Kalman [28, 31], where this problem area was called realization theory. It is one of the basic problems in systems theory, with many interesting and practical aspects, in particular the theory of approximate realizations [18]. However, these ramifications fall far beyond the scope of this book. The use of the controller and observer canonical forms belong to the early state space theory. See [25] for a number of other canonical state space representations of SISO and multivariable systems. The treatment of these canonical forms using the elimination theorem appears here for the first time. That controllable systems allow an image representation was first proven in [59]. It is also shown there that every controllable linear time-invariant differential systems allow an *observable* image representation. In the literature on nonlinear systems, systems that allow an observable image representation are called *flat systems*, [17].

6.9 Exercises

6.1 Consider the electrical circuit of Example 1.2.7. We want to derive the relation between V and I. To that end take V_i, I_i, $i = 1, \ldots, 5$ as latent variables and I and V as manifest variables. Write the equations describing the relations among the manifest and latent variables in the form $Rw = M\ell$ (all equations are static, therefore R and M are just real matrices). Apply the elimination procedure to deduce the relation between I and V. In case the calculations turn out to be too cumbersome, you can alternatively use the Gauss elimination procedure in Maple applied to the matrix $\begin{bmatrix} R & -M \end{bmatrix}$.

6.2 Consider the electrical circuit of Example 6.2.2. Determine the relation between V and I by applying the general elimination procedure.

6.3 Let $R(\xi), M(\xi) \in \mathbb{R}^{2\times 1}[\xi]$ and consider

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell,$$

with $R(\xi) = [R_1(\xi) \quad R_2(\xi)]^T$ and $M(\xi) = [M_1(\xi) \quad M_2(\xi)]^T$. We want to eliminate ℓ.

(a) Assume that $M_1(\xi)$ and $M_2(\xi)$ have no common factor. Use Theorem 2.5.10 to prove that the manifest behavior is described by

$$(M_2(\frac{d}{dt})R_1(\frac{d}{dt}) - M_1(\frac{d}{dt})R_2(\frac{d}{dt}))w = 0.$$

(b) Determine the differential equation for the manifest behavior when $M_1(\xi)$ and $M_2(\xi)$ may have a common factor.

6.4 Consider the SISO systems

$$\Sigma_1 : \ p_1(\frac{d}{dt})y_1 = q_1(\frac{d}{dt})u_1, \quad \Sigma_2 : \ p_2(\frac{d}{dt})y_2 = q_2(\frac{d}{dt})u_2. \tag{6.69}$$

Define the *feedback interconnection* of Σ_1 and Σ_2 by (6.69) and the interconnection equations $u_2 = y_1$, $u_1 = u + y_2$, and $y = y_1$. Here u is the external input and y is the external output; see Figure 6.6.

We are interested in the relation between u and y. To that end we have to eliminate u_1, u_2, y_1, and y_2. Elimination of u_2 and y_1 is straightforward, since $u_2 = y_1 = y$. In order to eliminate u_1 and y_2, define ℓ and w as

$$\ell := \begin{bmatrix} u_1 \\ y_2 \end{bmatrix}, \quad w := \begin{bmatrix} u \\ y \end{bmatrix}.$$

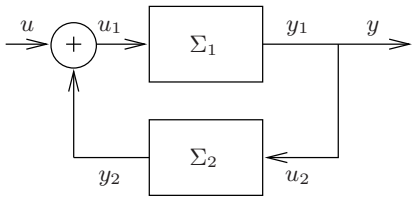


FIGURE 6.6. Feedback interconnection of Σ_1 and Σ_2.

(a) Determine matrices $R(\xi), M(\xi)$ of appropriate dimensions such that the behavior with these latent variables is described by $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$.

(b) Eliminate ℓ from $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$. Conclude that the relation between u and y is given by

$$(p_1(\frac{d}{dt})\bar{p}_2(\frac{d}{dt}) - \bar{q}_1(\frac{d}{dt})q_2(\frac{d}{dt}))y = \bar{p}_2(\frac{d}{dt})q_1(\frac{d}{dt})u,$$

with $p_2(\xi) = c(\xi)\bar{p}_2(\xi)$ and $q_1(\xi) = c(\xi)\bar{q}_1(\xi)$, such that $\bar{p}_2(\xi)$ and $\bar{q}_1(\xi)$ have no common factors.

6.5 Repeat 6.4 for the *parallel interconnection* $p_1(\frac{d}{dt})y_1 = q_1(\frac{d}{dt})u$, $p_2(\frac{d}{dt})y_2 = q_2(\frac{d}{dt})u$, $y = y_1 + y_2$. See Figure 6.7. The answer in this case is $(\bar{p}_2(\frac{d}{dt})q_1(\frac{d}{dt}) + \bar{p}_1(\frac{d}{dt})q_2(\frac{d}{dt}))u = \bar{p}_1(\frac{d}{dt})p_2(\frac{d}{dt})y$, where $p_1(\xi) = c(\xi)\bar{p}_1(\xi)$ and $p_2(\xi) = c(\xi)\bar{p}_2(\xi)$, such that $\bar{p}_1(\xi)$ and $\bar{p}_2(\xi)$ have no common factors.

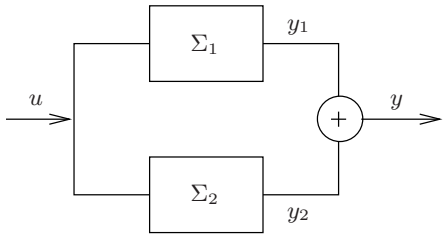


FIGURE 6.7. Parallel interconnection of Σ_1 and Σ_2.

6.6 Refer to Remark 6.4.11. For given polynomials $p(\xi)$ and $q(\xi)$, the polynomials $a(\xi)$ and $b(\xi)$ satisfying (6.52) are not unique. In fact, since (6.52) is a *linear* equation, it follows that every pair $(a(\xi), b(\xi))$ satisfying (6.52) can be written as $(a(\xi), b(\xi)) = (a_p(\xi), b_p(\xi)) + (a_h(\xi), b_h(\xi))$, with $(a_p(\xi), b_p(\xi))$ a particular solution of (6.52) and $(a_h(\xi), b_h(\xi))$ an arbitrary pair of polynomials that satisfies $a_h(\xi)p(\xi) + b_h(\xi)q(\xi) = 0$.

Prove that the expression for the state (6.58) is independent of the choice of the pair $(a(\xi), b(\xi))$ satisfying (6.52). Hint: define $v := a_h(\frac{d}{dt})u + b_h(\frac{d}{dt})y$ and show that $p(\frac{d}{dt})v = 0$ and $q(\frac{d}{dt})v = 0$. Conclude that since $p(\xi)$ and $q(\xi)$ are coprime, v must be zero.

6.7 Refer to Remark 6.2.5.

(a) Let $\mathfrak{B}$ be given by

$$\mathfrak{B} = \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \,|\, \exists n \in \mathbb{N}, c_k, \lambda_k \in \mathbb{R}, k = 1, \ldots, n : \; w(t) = \sum_{k=0}^{n} c_k e^{\lambda_k t}\}.$$

i. Prove that $\mathfrak{B}$ is linear and time-invariant.

ii. Prove that $\mathfrak{B}$ is an infinite-dimensional space. Hint: Assume that $\dim \mathfrak{B} = N$ then there should exist $\lambda_1, \ldots \lambda_N$ such that the functions $w_i(t) = e^{\lambda_i t}$ form a basis of $\mathfrak{B}$. Choose $\lambda \neq \lambda_i$, $i = 1, \ldots, N$. By assumption there exist $\alpha_i \in \mathbb{R}$ such that $w = \sum_{k=1}^{n} \alpha_i w_i$. Apply the differential operator $\prod_{k=1}^{n} (\frac{d}{dt} - \lambda_k)$ to w to arrive at a contradiction. Alternatively, apply the appropriate results in Chapter 3.

iii. Conclude, by invoking results from Chapter 3, that $\mathfrak{B}$ is *not* described by linear differential equations with constant coefficients.

(b) As a second example of a linear time-invariant behavior that is not described by linear differential equations with constant coefficients, consider

$$\mathfrak{B} = \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \mid w(t) = w(t-1) \;\; \forall t \in \mathbb{R}\}.$$

i. Prove that $\mathfrak{B}$ is linear time-invariant, and autonomous (see Definition 3.2.1).

ii. Prove that $\mathfrak{B}$ is infinite-dimensional.

iii. Use an appropriate result from Chapter 3 to conclude that since $\mathfrak{B}$ is autonomous and infinite-dimensional, it cannot be the solution set of a set of linear differential equations with constant coefficients.

6.8 Consider the static latent variable model with $\mathbb{U} = \mathbb{R}^2$ and $\mathbb{U}_\ell = \mathbb{R}$ defined as

$$\mathfrak{B}_{\mathrm{f}} := \{(w_1, w_2, \ell) \in \mathbb{R}^3 \mid \ell^2 = w_1^2 - 1 = w_2^2 - 1, \}$$
$$\mathfrak{B} := \{(w_1, w_2) \in \mathbb{R}^2 \mid \exists \ell \in \mathbb{R} \text{ such that } (w_1, w_2, \ell) \in \mathbb{B}_{\mathrm{f}}\}.$$

(a) Is $\mathfrak{B}_{\mathrm{f}}$ an algebraic subset of $\mathbb{R}^3$?

(b) Determine an explicit characterization of the manifest behavior $\mathfrak{B}$ as a subset of $\mathbb{R}^2$.

(c) Is $\mathfrak{B}$ an algebraic subset of $\mathbb{R}^2$?

(d) Is the closure (in the sense of the Euclidean metric) of $\mathfrak{B}$ in $\mathbb{R}^2$ an algebraic set?

(e) Is $\mathfrak{B}$ a *semi-algebraic* set?

Remark. A subset of $\mathbb{R}^n$ that is the solution set of a finite number of polynomial equations is called an *algebraic* set. If it is the solution set of a finite number of polynomial equations and polynomial inequalities then it is a *semi-algebraic* set.

6.9 Consider equations (6.17) and (6.18) in Example 6.2.8. In principle, (6.18) could impose a smoothness restriction on the solutions of (6.17). In this specific example the smoothness of w_1 is already guaranteed by (6.17). Prove this and conclude that the manifest behavior is *exactly* described by (6.17).

6.10 Prove Theorem 6.2.4

6.11 Assume that the scalar variables w_1, w_2 are governed by

$$\begin{bmatrix} r_{11}(\frac{d}{dt}) & r_{12}(\frac{d}{dt}) \\ r_{21}(\frac{d}{dt}) & r_{22}(\frac{d}{dt}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0,$$

with $r_{ij}(\xi) \in \mathbb{R}[\xi]$.

(a) Assume that $r_{12}(\xi)$ and $r_{22}(\xi)$ are coprime. Prove, by eliminating w_2, that the dynamics of w_1 are governed by

$$[r_{11}(\frac{d}{dt})r_{22}(\frac{d}{dt}) - r_{12}(\frac{d}{dt})r_{21}(\frac{d}{dt})]w_1 = 0.$$

(b) What are the dynamics of w_1 if $r_{12}(\xi)$ and $r_{22}(\xi)$ are *not* coprime?

(c) Prove that if $r_{11}(\xi)r_{22}(\xi) - r_{12}(\xi)r_{21}(\xi) \neq 0$, then the smoothness conditions as discussed just after the proof of Theorem 6.2.6 that could arise in the elimination procedure are automatically fulfilled; i.e., show that exact elimination is possible in this case.

6.12 Prove that (6.51) is a latent variable representation of (6.50) by eliminating ℓ using the general elimination procedure.

6.13 Refer to Example 1.3.5. Consider I as input and V as output.

(a) Prove that (1.6, 1.7, 1.8) define an i/s/o representation.

(b) Derive the observer canonical form.

(c) Derive the controller canonical form for the case that the system is controllable.

(d) Prove that in the case $CR_C \neq \dfrac{L}{R_L}$ these three i/s/o representations are equivalent.

Repeat the above questions with V considered as the input and I as the output.

6.14 For each of the following cases, determine an i/o representation of the i/s/o representation by eliminating the state from the i/s/o equations

$$\frac{d}{dt}x = Ax + bu, \quad y = cx.$$

(a) $A = \begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 0 \end{bmatrix}.$

(b) $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 0 \end{bmatrix}.$

(c) $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 1 \end{bmatrix}.$

(d) $A = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 2 \end{bmatrix}.$

(e) $A = \begin{bmatrix} 8 & -3 & -3 \\ -9 & 2 & 5 \\ 23 & -7 & -10 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \quad c = \begin{bmatrix} 4 & 1 & -2 \end{bmatrix}.$

6.15 Consider the i/o system

$$-12y - 11\frac{d}{dt}y + 2\frac{d^2}{dt^2}y + \frac{d^3}{dt^3}y = -8u + 2\frac{d}{dt}u + \frac{d^2}{dt^2}u.$$

(a) Determine the observer canonical representation.

(b) Is this representation controllable?

6.16 Consider the i/s/o system of Exercise 5.21. Eliminate the state so as to obtain an i/o representation.

6.17 Consider the i/s/o representations

$$\Sigma_i \quad \begin{aligned} \frac{d}{dt}x_i &= A_i x_i + b_i u_i, \\ y_i &= c_i x_i, \end{aligned} \qquad i = 1, 2. \tag{6.70}$$

Just as in Example 6.2.9, the series interconnection of Σ_1 and Σ_2 is defined by (6.70) and the equations $u_2 = y_1, u = u_1, y = y_2$. Assume that both systems are observable. An i/s/o representation of the series interconnection is given by

$$\begin{array}{rcl} \frac{d}{dt}x & = & Ax + bu, \\ y & = & cx, \end{array}$$

with

$$A = \begin{bmatrix} A_1 & 0 \\ b_2c_1 & A_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 0 & c_2 \end{bmatrix}. \tag{6.71}$$

Define the following polynomial vectors:

$$\begin{array}{l} p_i(\xi) := \det(I\xi - A_i), \quad r_i(\xi) := p_i(\xi)c_i(I\xi - A_i)^{-1}, \quad q_i(\xi) := r_i(\xi)b_i, \quad i = 1, 2, \\ p(\xi) := \det(I\xi - A), \quad r(\xi) := p(\xi)c(I\xi - A)^{-1}, \quad q(\xi) := r(\xi)b. \end{array}$$

(a) Show that $r(\xi) = \begin{bmatrix} q_2(\xi)r_1(\xi) & p_1(\xi)r_2(\xi) \end{bmatrix}$.

(b) Assume that $p_1(\xi)$ and $q_2(\xi)$ have no common factors. Prove that (c, A), as defined by (6.71), is an observable pair.

6.18 Consider the the controller canonical form (6.46). Assume that $p(\xi)$ and $q(\xi)$ have a common factor. Where does the proof of Theorem 6.4.7 break down? Prove that in this case (6.44) is *not* the manifest behavior of (6.47) by observing that (6.47) is controllable, whereas (6.44) is not controllable.

6.19 In Theorem 6.5.2 we have seen that every observable pair (A, c) may be transformed into observer canonical form. Equivalently, there exists a basis of the state space with respect to which (A, c) takes the observer canonical form. The dual statement is, of course, that for every *controllable* pair (A, b) there exists a basis with respect to which (A, b) is in *controller canonical form*. We want to construct this basis. Let $(A, b) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ be a controllable pair. Define vectors $d_1, \ldots, d_n \in \mathbb{R}^n$ as follows:

$$\begin{array}{rcl} d_n & := & b, \\ d_{n-1} & := & Ab + p_{n-1}b, \\ d_{n-2} & := & A^2b + p_{n-1}Ab + p_{n-2}b, \\ & \vdots & \\ d_{n-k} & := & A^kb + p_{n-1}A^{k-1}b + \cdots + p_{n-k}b, \\ & \vdots & \\ d_2 & := & A^{n-2}b + p_{n-1}A^{n-3}b + \cdots + p_2b, \\ d_1 & := & A^{n-1}b + p_{n-1}A^{n-2}b + \cdots + p_1b. \end{array}$$

Here $p_0, \ldots, p_{n-1}$ are the coefficients of the characteristic polynomial of A: $\det(I\xi - A) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$.

(a) Prove that $d_1, \ldots, d_n$ are linearly independent and hence that they form a basis of $\mathbb{R}^n$.

(b) Prove that $d_{n-k} = Ad_{n-k+1} + p_{n-k}d_n$, $k = 1, \ldots, n-1$.

(c) Express Ad_1 in terms of $d_1, \ldots, d_n$. (Hint: use Cayley–Hamilton).

(d) Represent the matrix A in terms of the basis $d_1, \ldots, d_n$.

(e) Prove that in the basis $d_1, \ldots, d_n$, the vector b takes the form: $[0 \cdots 0\ 1]^T$.

(f) Take

$$A = \begin{bmatrix} 1 & 2 & 1 \\ -2 & 3 & 8 \\ 1 & 0 & 5 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Determine the basis d_1, d_2, d_3 for this case.

6.20 Let α and β be real numbers. Consider the system

$$\frac{d}{dt}x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ \alpha \end{bmatrix} u, \qquad y = \begin{bmatrix} 1 & \beta \end{bmatrix} x.$$

(a) Eliminate the state from the above equations. Distinguish the cases $\beta = 0$ and $\beta \neq 0$.

(b) Show that for $\alpha = 0$ and $\beta \neq 0$ the i/o behavior is described by

$$(\frac{d}{dt} - 1)(\frac{d}{dt} + 1)y = (\frac{d}{dt} + 1)u$$

and hence is not controllable. Explain why β does not enter this equation.

6.21 Consider

$$-2\frac{d}{dt}y + \frac{d^2}{dt^2}y + \frac{d^3}{dt^3}y = 2u + 3\frac{d}{dt}u + \frac{d^2}{dt^2}u.$$

(a) Determine the observer canonical i/s/o representation.

(b) Is this representation controllable?

(c) Does there exist a controllable i/s/o representation of the given i/o system?

6.22 Let $(A, b, c) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1} \times \mathbb{R}^{1\times n}$. Define $p(\xi) := \det(I\xi - A)$ and $q(\xi) := p(\xi)c(I\xi - A)^{-1}b$. Prove that $p(\xi)$ and $q(\xi)$ have no common factors if and only if (A, b) is controllable and (c, A) is observable. Hint: Use Theorem 5.5.1 for the proof "from left to right" and use Exercises 6.19, 5.18 for the opposite direction.

6.23 Consider the matrix A in (6.37). Prove that $\det(I\xi - A) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$.

6.24 In Remarks 6.4.5 and 6.4.11 we showed how the state of the observer and controller canonical form, respectively can be expressed in u and y. Consider an *observable* SISO system $\frac{d}{dt}x = Ax + bu$, $y = cx + du$.

(a) Prove that

$$\frac{d^k}{dt^k}y = c(A^k x + A^{k-1}bu + A^{k-2}b\frac{d}{dt}u + \cdots + Ab\frac{d^{k-2}}{dt^{k-2}}u + b\frac{d^{k-1}}{dt^{k-1}}u) + d\frac{d^k}{dt^k}u. \tag{6.72}$$

(b) Use the observability of (A, c) to solve x from (6.72), and use the result to derive (6.60).

6.25 Prove the claim made in Remark 6.3.4. Hint: An alternative way to eliminate the state variables is given in Chapter 4, (4.31). From this expression it follows easily that the i/o behavior is closed.

6.26 (a) Consider the full behavior $\mathfrak{B}_{\mathrm{f}}$ with latent variables defined by

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell.$$

Assume that the manifest behavior $\mathfrak{B}$ is defined by

$$R'(\frac{d}{dt})w = 0.$$

Prove, using the elimination theorem, that if the full behavior is controllable, see Definition 5.2.2, then the manifest behavior is also controllable.

(b) Consider the system

$$\frac{d}{dt}x = Ax + Bu, \qquad y = Cx + Du.$$

Assume that (A, C) is observable. Prove that the manifest behavior, i.e., the corresponding i/o behavior, is controllable if and only if (A, B) is controllable.

6.27 Consider the SISO system defined by

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u, \tag{6.73}$$

with $p(\xi), q(\xi) \in \mathbb{R}[\xi]$ coprime polynomials and $\deg p(\xi) > \deg q(\xi)$. Assume that $p(\xi)$ has only simple real roots $\lambda_1, \ldots, \lambda_n$. Let the partial fraction expansion of $\frac{q(\xi)}{p(\xi)}$ be given by

$$\frac{q(\xi)}{p(\xi)} = \sum_{k=1}^{n} \frac{\gamma_k}{\xi - \lambda_k}.$$

(a) Prove that $\frac{d}{dt}x = Ax + bu$, $y = cx$ with

$$A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \quad b = \mathrm{col}(\gamma_1, \ldots, \gamma_n) \quad c = [1 \cdots 1]$$

defines a state space representation for (6.73).

(b) Repeat the previous question for

$$A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \quad b = \mathrm{col}[1, \ldots, 1] \quad c = (\gamma_1, \ldots, \gamma_n).$$

(c) In addition to the controller and observer canonical forms we have now obtained two more state space representations. Prove that these four state space representations are similar. What are the similarity transformations that connect them? In principle we are asking for twelve nonsingular matrices. However, if, e.g., S_1 connects the first two representations and S_2 the second and the third, then the transformation that connects the first and the third representations is easily derived from S_1 and S_2.

6.28 Consider the system described by

$$Kw + M(\frac{d}{dt})^2 w = 0, \tag{6.74}$$

with $K = K^T$ and $M = M^T > 0$. Such second-order models occur frequently in mechanics.

(a) Prove that the system (6.74) is autonomous (see Section 3.2).

(b) Give a state space representation of (6.74) with state

$$x = \begin{bmatrix} w \\ \frac{d}{dt} w \end{bmatrix}.$$

(c) Define the *momentum* by $p = M\frac{d}{dt}w$. Give a state space representation with state

$$x = \begin{bmatrix} w \\ p \end{bmatrix}. \tag{6.75}$$

(d) Provide a similarity transformation relating the two state space representations.

(e) Define the function L as $L(w, v) = \frac{1}{2} w^T K w - \frac{1}{2} v^T M v$. Show that (6.74) can be written as

$$\frac{d}{dt}\frac{\partial L}{\partial v}(w, \frac{d}{dt}w) - \frac{\partial L}{\partial w}(w, \frac{d}{dt}w) = 0. \tag{6.76}$$

Define $H(w, p) = \frac{1}{2} w^T K w + \frac{1}{2} p^T M^{-1} p$. Show that the state equations (6.75) can be written as

$$\frac{d}{dt}w = \frac{\partial H}{\partial p}(w, p), \quad \frac{d}{dt}p = -\frac{\partial H}{\partial w}(w, p). \tag{6.77}$$

Interpret L and H in terms of the potential and kinetic energy respectively, and (6.76) and (6.77) in terms of Lagrangian and Hamiltonian mechanics.

6.29 Consider the latent variable system defined by

$$R(\frac{d}{dt})w = M(\frac{d}{dt})\ell. \tag{6.78}$$

Assume that the full behavior, i.e, the behavior of (w, ℓ), is controllable. According to Theorem 6.6.1, the full behavior admits an image representation, say

$$w = M'(\frac{d}{dt})\ell' \quad \ell = M''(\frac{d}{dt})\ell',$$

Prove that $w = M'(\frac{d}{dt})\ell'$ is an image representation of the manifest behavior, i.e., the behavior of w, of (6.78). Use this result to obtain an alternative solution to Exercise 6.26a.

6.30 This exercise is concerned with nonlinear systems. To avoid difficulties with existence of solutions and smoothness we assume that all maps and all trajectories are infinitely differentiable. Let $f : (\mathbb{R}^d)^{L+1} \to \mathbb{R}^q$ and consider the latent variable representation

$$w = f(\ell, \frac{d}{dt}\ell, \ldots, (\frac{d}{dt})^L\ell).$$

Prove that the manifest behavior is controllable (in the sense of Definition 5.2.2). This shows that the existence of an image representation is a sufficient condition for controllability. For the linear case this is also necessary (Theorem 6.6.1). For the nonlinear case this equivalence does not hold in general.

7
Stability Theory

7.1 Introduction

In this chapter we study the stability of dynamical systems. Stability is a very common issue in many areas of applied mathematics. Intuitively, stability implies that small causes produce small effects. There are several types of stability. In *structural stability*, one wants small parameter changes to have a similar small influence on the behavior of a system. In *dynamic stability*, which is the topic of this chapter, it is the effect of disturbances in the form of initial conditions on the solution of the dynamical equations that matters. Intuitively, an equilibrium point is said to be *stable* if trajectories that start close to it remain close to it. Dynamic stability is thus not in the first instance a property of a system, but of an equilibrium point. However, for linear systems we can, and will, view stability as a property of the system itself. In *input/output stability* small input disturbances should produce small output disturbances. Some of these concepts are intuitively illustrated by means of the following example.

Example 7.1.1 In order to illustrate the stability concept, consider the motion of a pendulum; see Figure 7.1. The differential equation describing the angle θ is

$$\frac{d^2}{dt^2}\theta + \frac{g}{L}\sin\theta = 0. \tag{7.1}$$

L denotes the length of the pendulum, and g the gravitational constant. The system (7.1) has $\theta^* = 0$ and $\theta^* = \pi$ as equilibria: if the pendulum

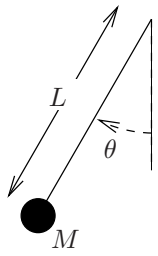


FIGURE 7.1. A pendulum.

starts with zero velocity in the initial position $\theta(0) = 0$ or $\theta(0) = \pi$, then it remains in this initial state for all $t \geq 0$. However, if we disturb $\theta(0)$ slightly and keep $\frac{d}{dt}\theta(0) = 0$, then two situations occur: if we disturb the equilibrium $\theta^* = 0$, then the pendulum oscillates around $\theta = 0$, and the distance of $\theta(t)$ from $\theta^* = 0$ remains small. This property is called *stability* of the equilibrium point $\theta^* = 0$. If, however, we disturb the equilibrium $\theta^* = \pi$, then a small perturbation causes the pendulum to fall, leading to a solution such that the distance from the equilibrium $\theta^* = \pi$ becomes large. The equilibrium $\theta^* = \pi$ is therefore called *unstable*.

Equation (7.1) assumes that the pendulum moves without friction. If, however, there is friction (e.g., the unavoidable air friction, or friction in the joint where the pendulum is suspended), then the equation for the motion of the pendulum becomes

$$\frac{d^2}{dt^2}\theta + D\frac{d}{dt}\theta + \frac{g}{L}\sin\theta = 0, \tag{7.2}$$

where D is the friction coefficient. The solutions of this differential equation show a different behavior for $D \neq 0$ than is the case when $D = 0$. If $D > 0$, for example, it can be seen that small initial disturbances from $\theta^* = 0$ are damped out, and the solution now approaches the equilibrium $\theta^* = 0$. This is called *asymptotic stability*. Of course, for small initial disturbances from $\theta^* = \pi$, the pendulum again falls, resulting in instability. Next, think of the case of negative damping $D < 0$. This comes down to assuming that the pendulum is accelerated by a term proportional to its velocity. It is not easy to think up a simple physical mechanism that produces such an effect, but one could think of an external force that is being applied and that pushes the pendulum proportional to its velocity $\frac{d}{dt}\theta$. Small disturbances of the initial position away from $\theta^* = 0$ then lead to instability. This sensitive dependence of the stability properties of an equilibrium on the system parameters is an example of lack of *structural stability*. We do not discuss this concept in depth in this book. Observe that the solutions in the neighborhood of the equilibrium $\theta^* = 0$ have a completely different behavior when $D > 0$ and $D < 0$. Thus the system (7.2) is not structurally

stable around $D = 0$, and the parameter value $D = 0$ is called a *bifurcation point*.

The trajectories of (7.2) for $D = 0, D > 0$, and $D < 0$ are shown in Figure 7.2. □

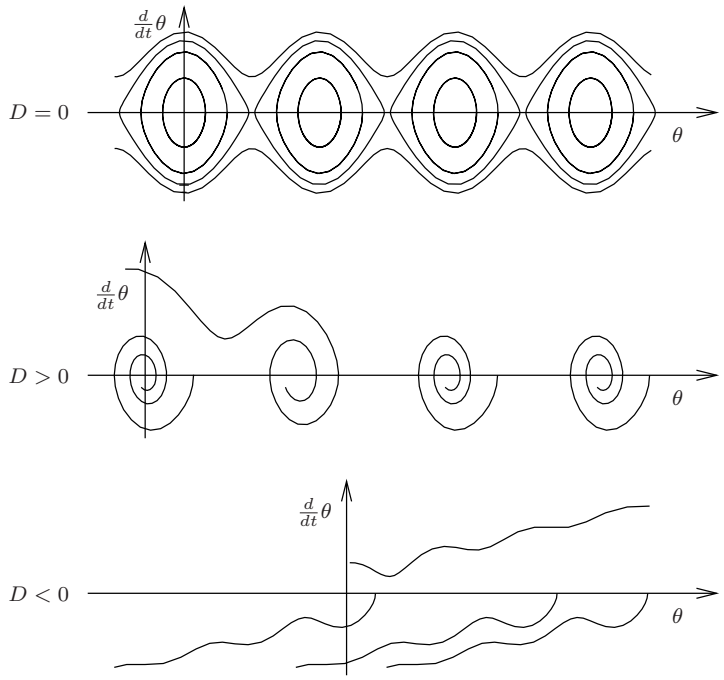


FIGURE 7.2. Phase portraits of the motion of a pendulum.

We first look at the stability properties of the linear autonomous system

$$P\left(\frac{d}{dt}\right)w = 0 \tag{7.3}$$

studied in Section 3.2. This system is called stable if all the trajectories in its behavior are bounded on the half line $[0, \infty)$, and asymptotically stable if all its trajectories approach 0 as $t \to \infty$. Our purpose is to derive conditions for (asymptotic) stability in terms of the polynomial matrix $P(\xi)$. We subsequently apply these results to the special case of autonomous state models, i.e., systems of the form

$$\frac{d}{dt}x = Ax. \tag{7.4}$$

A common theme in stability theory is the construction of energy-like functions (called the *Lyapunov functions*) that are nonincreasing along solutions and from which stability may be deduced. For systems as (7.4) such functions can be constructed using a linear matrix equation, called the *Lyapunov equation*. Lyapunov functions can also be used very effectively for the determination of the stability of an equilibrium point of the autonomous nonlinear system

$$\frac{d}{dt}x = f(x), \tag{7.5}$$

where f is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$.

A final issue on our agenda in this chapter is the *input/output stability* of the system

$$P\left(\frac{d}{dt}\right)y = Q\left(\frac{d}{dt}\right)u.$$

7.2 Stability of Autonomous Systems

Let $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$, with $\det P(\xi) \neq 0$; i.e., the polynomial $\det P(\xi)$ is assumed not to be the zero polynomial. Consider the system of differential equations (7.3). This defines, as explained in detail in Chapter 3, the dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$ with $\mathfrak{B} = \{w : \mathbb{R} \to \mathbb{R}^q \mid w \text{ satisfies } P\left(\frac{d}{dt}\right)w = 0 \text{ weakly}\}$. Since we assume $\det P(\xi) \neq 0$, the resulting system (7.3) is an autonomous one, which implies that $\mathfrak{B}$ is finite-dimensional and that each weak solution of (7.3) is a strong one. Moreover, all solutions are infinitely differentiable in this case, and the general form of the solution to (7.3) has been given in Theorem 3.2.16. In fact, we showed that $w \in \mathfrak{B}$ if and only if w is of the following form: $w = w_1 + w_2 + \cdots + w_N$, where each of the w_ks is associated with one of the distinct roots $\lambda_1, \lambda_2, \ldots, \lambda_N$ of $\det P(\xi)$. This w_k is given by

$$w_k(t) = \left(\sum_{\ell=0}^{n_k-1} B_{k\ell}t^\ell\right)e^{\lambda_k t}, \tag{7.6}$$

where n_k is the multiplicity of the root λ_k of $\det P(\xi)$ and the $B_{k\ell}$s are suitable constant complex vectors. How these vectors are obtained is not important at this point. It has been explained in Theorem 3.2.16. What is important, however, is the fact that the set of admissible polynomials

$$\sum_{\ell=0}^{n_k-1} B_{k\ell}t^\ell$$

obtained this way forms an n_k dimensional linear space.

In (7.6) we have assumed that we are considering complex solutions. The real solutions are simply obtained by taking the real part in (7.6). Since the distinction between the real and the complex case is not relevant in stability considerations, we continue by silently assuming that we are considering complex as well as real solutions.

As can be seen from Example 7.1.1, in nonlinear systems some equilibria may be stable, others may be unstable. Thus stability is not a property of a dynamical system, but of a trajectory, more specifically, of an equilibrium of a dynamical system. However, for linear systems it can be shown (see Exercise 7.26) that all equilibria have the same stability properties. For simplicity of exposition, this fact has been incorporated in the definition that follows.

Definition 7.2.1 The linear dynamical system Σ described by (7.3) is said to be *stable* if all elements of its behavior $\mathfrak{B}$ are bounded on the half-line $[0, \infty)$, precisely, if $(w \in \mathfrak{B}) \Rightarrow$ (there exists $M \in \mathbb{R}$ such that $\|w(t)\| \leq M$ for $t \geq 0$). Of course, this bound M depends on the particular solution $w \in \mathfrak{B}$. It is said to be *unstable* if it is not stable; it is said to be *asymptotically stable* if all elements of $\mathfrak{B}$ approach zero for $t \to \infty$ (precisely, if $(w \in \mathfrak{B}) \Rightarrow (w(t) \to 0$ as $t \to \infty)$). □

In order to state stability conditions in terms of the polynomial $P(\xi)$, we need to introduce the notion of a semisimple root of the square polynomial matrix $P(\xi)$. The *roots* (or *singularities*) of $P(\xi)$ are defined to be those of the scalar polynomial $\det P(\xi)$. Hence $\lambda \in \mathbb{C}$ is a root of $P(\xi)$ if and only if the complex matrix $P(\lambda) \in \mathbb{C}^{q \times q}$ has rank less than q. The root λ is called *simple* if it is a root of $\det P(\xi)$ of multiplicity one, and *semisimple* if the rank deficiency of $P(\lambda)$ equals the multiplicity of λ as a root of $P(\xi)$ (equivalently, if the dimension of $\ker P(\lambda)$ is equal to the multiplicity of λ as a root of $\det P(\xi)$). Clearly, for $q = 1$ roots are semisimple if and only if they are simple, but for $q > 1$ the situation is more complicated. For example, 0 is a double root of both the polynomial matrices

$$\begin{bmatrix} \xi & 0 \\ 0 & \xi \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \xi & 1 \\ 0 & \xi \end{bmatrix}. \tag{7.7}$$

This root is semisimple in the first case, but not in the second.

With Theorem 3.2.16 at hand, it is very easy to decide on the stability of (7.3).

Theorem 7.2.2 *The system defined by (7.3) is:*

(i) asymptotically stable *if and only if all the roots of* $\det P(\xi)$ *have negative real part;*

(ii) stable *if and only if for each* $\lambda \in \mathbb{C}$ *that is a root of* $\det P(\xi)$*, there must hold either (i)* $\operatorname{Re}\lambda < 0$*, or (ii)* $\operatorname{Re}\lambda = 0$ *and* λ *is a semisimple root of* $P(\xi)$*.*

(iii) unstable *if and only if* $\det P(\xi)$ *has a root with positive real part and/or a nonsemisimple root with zero real part.*

Proof (i) Let $w \in \mathfrak{B}$. Then w is given by an expression as given in Theorem 3.2.16, (3.19). Since ($\operatorname{Re}\lambda < 0$ and $\ell \in \mathbb{Z}_+$) $\Rightarrow$ ($\lim_{t\to\infty} t^\ell e^{\lambda t} = 0$), the "if" part of part 1 of the theorem follows. To prove the "only if" statement, observe that if $\lambda \in \mathbb{C}$ is a root of $\det P(\xi)$, then there exists a $0 \neq B \in \mathbb{C}^q$ such that $P(\lambda)B = 0$, and hence the function $t \mapsto Be^{\lambda t}$ belongs to $\mathfrak{B}$. Now, if $Be^{\lambda t}$ is to go to zero as $t \to \infty$, $\operatorname{Re}\lambda$ must be < 0. This yields the "only if" part.

(ii) To prove stability, the argument of part 1 needs to be refined in one point only: by considering more closely what happens to the roots of $P(\xi)$ with zero real part. In particular, if λ_k is such a root, and if its multiplicity n_k is larger than one, then (7.6) shows that there may be elements $w \in \mathfrak{B}$ of the form

$$w(t) = \left(\sum_{\ell=0}^{n_k-1} B_{k\ell} t^\ell \right) e^{\lambda_k t}. \tag{7.8}$$

We must show that $B_{k1} = B_{k2} = \cdots = B_{k(n_k-1)} = 0$ if and only if the root λ_k is semisimple. We know on the one hand that the set of polynomials

$$\sum_{\ell=0}^{n_k-1} B_{k\ell} t^\ell.$$

such that (7.8) yields a $w \in \mathfrak{B}$ form an n_k-dimensional space (see 3.2.16). On the other hand, the function $t \mapsto B_k e^{\lambda_k t}$ belongs to $\mathfrak{B}$ if and only if $P(\lambda_k)B_k = 0$. These B_ks form an n_k dimensional vector space if and only if λ_k is a semisimple root of $P(\xi)$. This shows that there are ws in $\mathfrak{B}$ of the form (7.8) that are unbounded if and only if λ_k is semisimple. This yields part 2.

(iii) Follows from part 2. □

Example 7.2.3

1. Consider the scalar first-order system $aw + \frac{d}{dt}w = 0$. The associated polynomial is $P(\xi) = a + \xi$. Its root is $-a$. Hence this system is asymptotically stable if $a > 0$, stable if $a = 0$, and unstable if $a < 0$. This, of course, is easily verified, since the solution set consists of the exponentials Ae^{-at}.

2. Consider the scalar second-order system $aw + \frac{d^2}{dt^2}w = 0$. The associated polynomial is $P(\xi) = a + \xi^2$. Its roots are $\lambda_{1,2} = \pm\sqrt{-a}$ for $a < 0$, $\lambda_{1,2} = \pm i\sqrt{a}$ for $a > 0$, and $\lambda = 0$ is a double, not semisimple root when $a = 0$. Thus, according to Theorem 7.2.2, we have ($a < 0 \Rightarrow$ instability), ($a > 0 \Rightarrow$ stability), and ($a = 0 \Rightarrow$ instability). These conclusions correspond to the results obtained in Theorem 3.2.5 and Corollary 3.2.13. Indeed, for $a < 0$ the solution set is given by the time trajectories of the form $Ae^{\sqrt{-a}t} + Be^{-\sqrt{-a}t}$, and hence ($A = 1, B = 0$) there are unbounded solutions; for $a > 0$ by $A\cos\sqrt{a}t + B\sin\sqrt{a}tA$, and hence all solutions are bounded; for $a = 0$ by $A + Bt$, and hence ($B \neq 0$) there are unbounded solutions.

□

We now apply Theorem 7.2.2 to derive conditions on the stability of state equations. Let $A \in \mathbb{R}^{n\times n}$, and consider the autonomous state system (7.4). The roots of the polynomial $\det(I\xi - A)$ are the eigenvalues of A. Accordingly, we call an eigenvalue λ of A *semisimple* if λ is a semisimple root of $\det(I\xi - A)$, in other words, if the dimension of $\ker(\lambda I - A)$ is equal to the multiplicity of λ as a root of the characteristic polynomial $\det(I\xi - A)$ of A. In Exercise 7.7 an equivalent condition for semisimplicity of an eigenvalue of A is derived.

Corollary 7.2.4 *The system defined by* $\frac{d}{dt}x = Ax$ *is:*

1. asymptotically stable *if and only if the eigenvalues of A have negative real part;*

2. stable *if and only if for each* $\lambda \in \mathbb{C}$ *that is an eigenvalue of A, either (i)* $\mathrm{Re}\lambda < 0$*, or (ii)* $\mathrm{Re}\ \lambda = 0$ *and* λ *is a semisimple eigenvalue of A;*

3. unstable *if and only if A has either an eigenvalue with positive real part or a nonsemisimple one with zero real part.*

Proof Apply Theorem 7.2.2 with $P(\xi) = I\xi - A$. □

Example 7.2.5 The free motion in $\mathbb{R}^3$ of a particle with mass m is described by $m\frac{d^2}{dt^2}w = 0$. Hence $P(\xi) = mI\xi^2$, with I the (3×3) identity matrix. The determinant of $P(\xi)$ is $m^3\xi^6$, and $P(0) = 0$. Hence the root 0 is not semisimple, since $\dim\ker P(0) = 3$ but 0 is a root of $\det P(\xi)$ of multiplicity 6. In fact, the behavior of this system consists of all functions of the form $at + b$, with $a, b \in \mathbb{R}^3$. Whenever the initial velocity $a \neq 0$, this leads to a function that is unbounded on $\mathbb{R}_+$, showing instability. □

7.3 The Routh–Hurwitz Conditions

Let $p(\xi) \in \mathbb{R}[\xi]$, written out in terms of its coefficients

$$p(\xi) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + p_n\xi^n, \tag{7.9}$$

with $p_0, p_1, \ldots, p_n \in \mathbb{R}$. Now consider the problem of finding conditions on the coefficients $p_0, p_1, \ldots, p_n$ such that all the roots of $p(\xi)$ have negative real part. The question arises, *Is it necessary to compute the roots of $p(\xi)$ in order to decide whether their real part is negative, or do there exist relatively simple tests on the coefficients of $p(\xi)$ for the roots to have negative real part?* Note that this question, when applied to $\det P(\xi)$ or to $\det(I\xi - A)$, arises very naturally as a result of the asymptotic stability condition of systems (7.3) and (7.4) as established in Theorem 7.2.2 and Corollary 7.2.4.

This question, nowadays called the *Routh–Hurwitz problem*, has a history going back more than a century. Maxwell (indeed, he of the basic equations describing electromagnetic fields) was the first scientist who ran into this problem, but he was unable to give a satisfactory answer. The question itself may sound a bit quaint in the age of computers, where roots of high-order polynomials can be evaluated to great accuracy in a matter of seconds. Nevertheless, verifying that the roots of a high-order polynomial have negative real part by actually computing them all explicitly certainly feels like overkill. However, until a few decades ago, scientists did not suffer the comfort of computers, and it was natural that the Routh–Hurwitz question became a very belabored one in view of the importance of the dynamic stability question.

In order to appreciate the difficulty of the Routh–Hurwitz problem, consider the cases $n = 1, 2$. Assume for simplicity that $p_n = 1$. Clearly, $p_0 + \xi$ has its root $-p_0$ in the left half of the complex plane if and only if $p_0 > 0$. The roots of $p_0 + p_1\xi + \xi^2$ are given by $-\frac{p_1}{2} \pm \sqrt{(\frac{p_1}{2})^2 - p_0}$ and it is easy to sort out that these have negative real part if and only if $p_0 > 0$, and $p_1 > 0$. We invite the reader to try to do the case $n = 3$ without reading the sequel to this chapter. The question is, can we come up with a test for any n? We say that $p(\xi)$ is a *Hurwitz* polynomial if all its roots have negative real part. If all the eigenvalues of $A \in \mathbb{R}^{n\times n}$ have negative real part, then A is called a *Hurwitz matrix*. If $P(\xi) \in \mathbb{R}^{q\times q}[\xi]$ has $\det P(\xi) \neq 0$ and if $\det P(\xi)$ is a Hurwitz polynomial, then we call $P(\xi)$ a *Hurwitz polynomial matrix*.

Assume that the degree of $p(\xi)$ is n; hence $p_n \neq 0$. We are looking for conditions on $p_0, p_1, \ldots, p_{n-1}, p_n$ for $p(\xi)$ to be Hurwitz. Note that we may as well assume that $p_n > 0$; otherwise, consider the polynomial $-p(\xi)$.

We now state two equivalent conditions on the coefficients $p_0, p_1, \ldots, p_{n-1}, p_n$ for $p(\xi)$ to be Hurwitz. The first condition is due to Routh. It is difficult to

state, but straightforward to apply. The second condition is due to Hurwitz. It is easier to state, but it requires the evaluation of large determinants.

7.3.1 The Routh test

In order to state the *Routh test*, consider first the following procedure for forming from two sequences of real numbers a third one:

$$\begin{array}{llllll} \text{sequence } 1: & a_1 & a_2 & a_3 & \cdots, \\ \text{sequence } 2: & b_1 & b_2 & b_3 & \cdots, \\ \text{sequence } 3: & c_1 & c_2 & c_3 & \cdots, \end{array} \tag{7.10}$$

with $c_k = b_1 a_{k+1} - a_1 b_{k+1}$. Note that it easy to compute the c_ks in a systematic way, since c_k is simply minus the determinant of $\left[\begin{smallmatrix} a_1 & a_{k+1} \\ b_1 & b_{k+1} \end{smallmatrix}\right]$. Now form the sequences derived from, respectively, the coefficients of the even and odd parts of the polynomial $p(\xi)$:

$$\begin{array}{llllll} \text{row } 1: & p_0 & p_2 & p_4 & \cdots, \\ \text{row } 2: & p_1 & p_3 & p_5 & \cdots. \end{array} \tag{7.11}$$

When, while setting up these sequences, one meets coefficients beyond ξ^n, take them to be equal to zero. Now compute a third sequence

$$\text{row } 3: \quad c_1 \quad c_2 \quad c_3 \quad \cdots$$

from the rows (7.11), using the procedure explained above. Next, compute the fourth row

$$\text{row } 4: \quad d_1 \quad d_2 \quad d_3 \quad \cdots$$

from row 2 and row 3, also in the manner indicated above. Proceeding this way yields the *Routh table*:

$$\begin{array}{llll} p_0 & p_2 & p_4 & \cdots \\ p_1 & p_3 & p_5 & \cdots \\ c_1 & c_2 & c_3 & \cdots \\ d_1 & d_2 & d_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots\vdots \end{array} \tag{7.12}$$

Simple bookkeeping shows that only the first $(n+1)$ rows of this table contain elements that are not zero; see Exercise 7.12. Let $(r_0, r_1, r_2, \ldots, r_n)$ denote the elements in the *first column* of the Routh table. This array is called the *Routh array*. Of course, $r_0 = p_0, r_1 = p_1, r_2 = p_1 p_2 - p_0 p_3$, etc., but $r_3, r_4, \ldots$ are increasingly complicated functions of $p_0, p_1, p_2, \ldots$. However, using the recursive procedure explained for the Routh table, they are straightforward to compute.

Theorem 7.3.1 (Routh test) *Assume that $p_n > 0$. Then all the roots of $p(\xi)$ have negative real part if and only if $r_0 > 0, r_1 > 0, \ldots, r_n > 0$, i.e., the elements of the Routh array are all positive.*

Proof We do not give the proof of this result. However, Exercise 7.15 leads the interested reader through an elementary inductive proof. □

The test of Theorem 7.3.1 is called the Routh test. It is an amazing result. The following examples illustrate the construction of the Routh table and the Routh array, and hence the determination of the asymptotic stability of the associated differential equation.

Example 7.3.2 1. Determine whether $p(\xi) = 1 + 2\xi + 3\xi^2 + \xi^3$ is Hurwitz. Computation of the Routh table (blank elements are zero) yields

$$\begin{array}{ll} 1 & 3 \\ 2 & 1 \\ 5 & \\ 5 & \end{array}$$

The Routh array equals $(1, 2, 5, 5)$, and hence the polynomial is Hurwitz.

2. Determine for what $\alpha \in \mathbb{R}$ the polynomial $p(\xi) = \alpha + \xi + 2\xi^2 + 3\xi^3 + 2\xi^4 + \xi^5$ is Hurwitz. The Routh table becomes

$$\begin{array}{ll} \alpha & 2 \quad 2 \\ 1 & 3 \quad 1 \\ 2 - 3\alpha & 2 - \alpha \\ 4 - 8\alpha & 2 - 3\alpha \\ 4(2-\alpha)^2 - (2-3\alpha)^2 & \\ (4(2-\alpha)^2 - (2-3\alpha)^2)(2-3\alpha) & \end{array}$$

The Routh array equals $(\alpha, 1, (2 - 3\alpha), 4 - 8\alpha, 4(2 - \alpha)^2 - (2 - 3\alpha)^2, (4(2-\alpha)^2-(2-3\alpha)^2)(2-3\alpha))$. Hence this polynomial is Hurwitz if and only if $0 < \alpha < 2/3$.

3. Determine for what p_0, p_1, p_2, p_3 the following polynomials are Hurwitz:

$$\begin{array}{lrcl} \text{(i)} & p(\xi) & = & p_0 + \xi. \\ \text{(ii)} & p(\xi) & = & p_0 + p_1\xi + \xi^2. \\ \text{(iii)} & p(\xi) & = & p_0 + p_1\xi + p_2\xi^2 + \xi^3. \\ \text{(iv)} & p(\xi) & = & p_0 + p_1\xi + p_2\xi^2 + p_3\xi^3 + \xi^4. \end{array}$$

Using the Routh test, the following conditions are readily derived:

case (i): $p_0 > 0$.
case (ii): $p_0 > 0, p_1 > 0$.
case (iii): $p_0 > 0, p_1 > 0, p_1p_2 > p_0$.
case (iv): $p_0 > 0, p_1 > 0, p_1p_2 - p_0p_3 > 0, p_1p_2p_3 - p_0p_3^2 - p_1^2 > 0$.

Note that the conditions of case (iv) are equivalent to $p_0 > 0, p_1 > 0, p_2 > 0, p_3 > 0$, and $p_1 p_2 p_3 - p_0 p_3^2 - p_1^2 > 0$.

□

7.3.2 The Hurwitz test

We now state another necessary and sufficient condition for the roots of a polynomial to have negative real part. Form the following $n \times n$ matrix $H \in \mathbb{R}^{n \times n}$ from the coefficients of $p(\xi)$:

$$H = \begin{bmatrix} p_1 & p_0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ p_3 & p_2 & p_1 & p_0 & 0 & 0 & \cdots & 0 & 0 \\ p_5 & p_4 & p_3 & p_2 & p_1 & p_0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & p_{n-3} & p_{n-4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & p_{n-1} & p_{n-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & p_n \end{bmatrix}.$$

Now, let $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}, \Delta_n$ denote the leading principal minors of H. Recall that a *minor* of a matrix is the determinant of a square submatrix. A *principal* minor is one obtained from a submatrix that is formed by taking rows and columns with the same indices. The kth *leading* principal minor is obtained from a submatrix that is formed the first k rows and the first k columns. The determinants $\Delta_1, \Delta_2, \ldots, \Delta_n$ are called the *Hurwitz determinants* associated with p. The following result is due to Hurwitz.

Theorem 7.3.3 (Hurwitz test) *Assume that $p_n > 0$. Then all the roots of $p(\xi)$ have negative real part if and only if $\Delta_1 > 0, \Delta_2 > 0, \ldots, \Delta_n > 0$, i.e., its Hurwitz determinants are all positive.*

Proof We do not give the proof of this result. However, Exercise 7.16 leads the reader through a proof of the Hurwitz test. □

A necessary condition for $p(\xi)$ to be Hurwitz is that all its coefficients have the same sign. Of course, this must be a consequence of the Routh–Hurwitz conditions, but it is much more easily derived by factoring $p(\xi)$ in terms of its real and complex roots, yielding

$$p(\xi) = p_n \Big(\prod_k (\xi - \lambda_k)\Big)\Big(\prod_{k'} ((\xi - \lambda_{k'})^2 + \omega_{k'}^2)\Big).$$

The first product in the above expression runs over the real roots λ_k of $p(\xi)$, the second over the complex roots $\lambda_{k'} \pm i\omega_{k'}$. Since for a Hurwitz polynomial all the real numbers appearing in these factors are individually positive, we obtain the following result.

Theorem 7.3.4 *If $p(\xi) \in \mathbb{R}[\xi]$ is a Hurwitz polynomial of degree n, then its coefficients $p_0, p_1, \ldots, p_n$ all have the same sign (and in particular, none of these coefficients can be zero).*

Example 7.3.2, part 3, demonstrates that this sign condition is also sufficient when $n \leq 2$, but not when $n \geq 3$.

As already mentioned, the Routh–Hurwitz conditions have to some extent lost their appeal as a test for asymptotic stability because nowadays it is easy to calculate the roots of a polynomial on a computer. However, there are other useful results that can be derived very nicely from the Routh–Hurwitz conditions. As an illustration of this we prove that asymptotic stability is a *robustness* property. Consider the system (7.3). Write out the polynomial matrix $P(\xi)$ in terms of its coefficient matrices: $P(\xi) = P_0 + P_1\xi + \ldots + P_L\xi^L$. In applications these coefficient matrices are usually functions of the physical parameters of the system modeled. We have seen earlier in this book examples of mechanical systems where the values of the masses, spring constants, and damping coefficients define these coefficient matrices, and of electrical circuits where the values of the resistors, capacitors, and inductors define these coefficient matrices. Hence in many applications it is natural to view the matrices $P_0, P_1, \ldots, P_L$ as functions of a physical parameter vector $\alpha \in \mathbb{R}^N$, yielding mappings $P_k : \mathbb{R}^N \to \mathbb{R}^{q\times q}, k = 0, 1, \ldots, L$, that take α into the coefficient matrices $P_0(\alpha), P_1(\alpha), \ldots, P_L(\alpha)$. This yields the system described by the differential equations

$$P_0(\alpha)w + P_1(\alpha)\frac{d}{dt}w + \cdots + P_L(\alpha)\frac{d^L}{dt^L}w = 0. \tag{7.13}$$

Assume that for $\alpha = \alpha_0$, the resulting system (7.13) is asymptotically stable. Then the question arises whether (7.13) remains asymptotically stable for all αs close to this α_0. If this is the case, then we call (7.13) *robustly asymptotically stable* at α_0. There holds:

Theorem 7.3.5 *Assume that the maps $P_k : \mathbb{R}^N \to \mathbb{R}^{q\times q}$ are continuous in a neighborhood of $\alpha_0 \in \mathbb{R}^N$. Let $P_\alpha(\xi) \in \mathbb{R}^{q\times q}[\xi]$ be defined by*

$$P_\alpha(\xi) := P_0(\alpha) + P_1(\alpha)\xi + \cdots + P_L(\alpha)\xi^L$$

and assume that $\det P_{\alpha_0}(\xi)$ is Hurwitz and that $P_L(\alpha_0) \neq 0$, i.e., that its degree is constant for α in a neighborhood of α_0. Then (7.13) is robustly asymptotically stable at α_0.

Proof Of course, readers who are aware of the (somewhat tricky) result that the roots of a polynomial are continuous functions of its coefficients will immediately believe this theorem too. However, for these readers as well it should be apparent that the following proof circumvents this continuity argument.

Consider the Hurwitz determinants associated with $\det P_\alpha(\xi)$. Note that since the P_ks are continuous functions of α, so are the Hurwitz determinants. Now, in the neighborhood of α_0, the number of such determinants is equal to the degree of $\det P_{\alpha_0}(\xi)$. Furthermore, at $\alpha = \alpha_0$ all the Hurwitz determinants are positive. By continuity, they remain positive in a neighborhood of α_0. This proves the theorem. □

Now consider the system (7.4) in which the matrix A is assumed to be a function of a parameter vector $\alpha \in \mathbb{R}^N$, yielding

$$\frac{d}{dt}x = A(\alpha)x. \tag{7.14}$$

Corollary 7.3.6 *Assume that the map $A : \mathbb{R}^N \to \mathbb{R}^{n\times n}$ in (7.14) is continuous in a neighborhood of $\alpha_0 \in \mathbb{R}^N$. If $A(\alpha_0)$ is Hurwitz, then (7.14) is robustly asymptotically stable at $\alpha_0 \in \mathbb{R}^N$.*

Proof Observe that the degree of $\det(I\xi - A(\alpha))$ is n for all α and apply Theorem 7.3.5. □

The constant degree condition on $\det P_\alpha$ in Theorem 7.3.5 is important. In Corollary 7.3.6 it was automatically satisfied. Examination of the asymptotic stability of the differential equation

$$w + \frac{d}{dt}w + \alpha\frac{d^2}{dt^2}w = 0$$

around $\alpha = 0$ shows that this degree condition is not superfluous.

7.4 The Lyapunov Equation

In this section we discuss Lyapunov functions. We first introduce the intuitive idea in the context of the system (7.5) and subsequently work out the details in the case (7.4). For notational convenience, we assume that the equilibrium of (7.5) for which we examine the stability is $x^* = 0$.

Consider again (7.5)

$$\frac{d}{dt}x = f(x), \tag{7.15}$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$ and $f(0) = 0$. Note that this implies that 0 is an equilibrium point of (7.15). We would like to find conditions that ensure that every solution $x : \mathbb{R} \to \mathbb{R}^n$ (or every solution starting close to 0) of (7.15) goes to zero as $t \to \infty$, without having to solve this differential equation explicitly.

Example 7.4.1 As an example to illustrate the idea of a Lyapunov function, consider the scalar (possibly nonlinear) system described by

$$w + D(w)\frac{d}{dt}w + \frac{d^2}{dt^2}w = 0. \tag{7.16}$$

Think of (7.16) as describing the displacement from equilibrium of a mass that is dragged by springs over a rough surface (see Figure 7.3). The dependence on w of the friction coefficient $D(w)$ signifies that the friction exerted by the surface may depend on the place along the surface. Writing

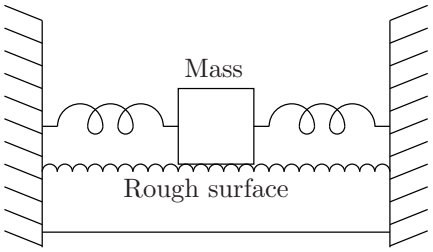


FIGURE 7.3. Mass–spring system with friction.

(7.16) in state form with $x = \text{col}(x_1, x_2)$, with $x_1 = w$ and $x_2 = \frac{d}{dt}w$, yields

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -1 & -D(x_1) \end{bmatrix} x. \tag{7.17}$$

The energy stored in this mechanical system is $V(x) = \frac{1}{2}||x||^2$. When the state is x then the rate of change of the stored energy equals the rate of energy dissipated by the friction, i.e., $-D(x_1)x_2^2$. This may be verified by computing the derivative of $V(x(\cdot))$ along a solution $x(\cdot)$ of (7.17). This is given by $\frac{d}{dt}V(x(t)) = \frac{d}{dt}\frac{1}{2}(x_1^2(t) + x_2^2(t)) = x_1(t)\frac{d}{dt}x_1(t) + x_2(t)\frac{d}{dt}x_2(t) = x_1(t)x_2(t) - x_1(t)x_2(t) - D(x_1(t))x_2^2(t) = -D(x_1(t))x_2^2(t)$. The important thing is that we can determine what this derivative is *without* knowing the solution $x(\cdot)$ explicitly. Note that when $D(x_1) \geq 0$ for all x_1, then the system dissipates energy, and hence the energy is nonincreasing in time. Hence $V(x(t)) \leq V(x(0))$ for $t \geq 0$ along solutions. This implies $||x(t)||^2 \leq ||x(0)||^2$ and demonstrates the stability of the equilibrium point. Hence by examining the sign of the derivative of V along solutions, we are able to infer stability without having computed the solution. This is the idea behind a Lyapunov function.

□

We now generalize this idea to (7.15). Suppose for a moment that we come up with a function $V : \mathbb{R}^n \to \mathbb{R}$ with $V(0) = 0$ and $V(x) > 0$ for $x \neq 0$,

and having the property that its derivative along every solution of (7.15) is nonpositive. Then $V(x(t))$ is nonincreasing, and it is reasonable to expect that, perhaps under some additional requirements, $V(x(t)) \to 0$ for $t \to \infty$. If, on the other hand, $V(x) < 0$ for some $x \in \mathbb{R}^n$ and $V(0) = 0$, then it simply cannot happen that $x(t) \to 0$ as $t \to \infty$, establishing lack of asymptotic stability.

This reasoning can be justified intuitively by drawing the level sets of V. Let us illustrate this graphically in the case $n = 2$. Draw the level sets of V in $\mathbb{R}^2$, i.e., the contours where $V(x)$ is constant. We may have elliptic-looking level sets or hyperbolic-looking level sets. If $V(x(t))$ is nonincreasing, then trajectories of (7.15) tend towards the origin in the case of elliptic contours, but not in the case of hyperbolic contours (see Figure 7.4).

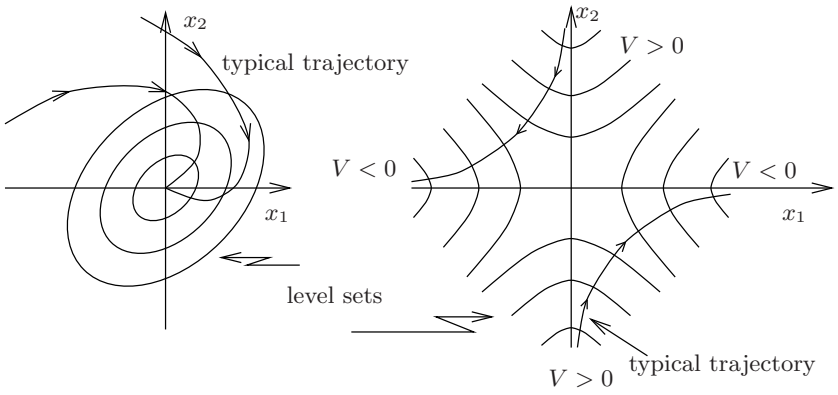


FIGURE 7.4. Phase portraits and level sets.

A nonnegative function V that has the property that its derivative along solutions of (7.15) is negative is called a *Lyapunov function* for (7.15). Two questions arise: *How do we establish that $V(x(t))$ is nonincreasing along solutions?* and *How do we find such a Lyapunov function V?*

The first question is easy to answer. Indeed, by the chain rule of differentiation the derivative of $V(x(t))$ at t is given by $(\text{grad } V)(x(t)) \cdot f(x(t))$, where $\text{grad } V = (\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}, \ldots, \frac{\partial V}{\partial x_n})$. Hence if the function $\dot{V} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\dot{V} := \text{grad } V \cdot f \tag{7.18}$$

is nonpositive, then $V(x(\cdot))$ is nonincreasing along solutions.

Definition 7.4.2 A differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ is said to be a *Lyapunov function* for the system (7.15) on the set $S \subset \mathbb{R}^n$ for (7.15) if $\dot{V}$, defined by (7.18), satisfies $\dot{V}(x) \leq 0$ for all $x \in S$. □

The second question is more difficult. In physical systems that dissipate energy, the stored energy is a good candidate for a Lyapunov function. In thermodynamic systems the negative of the entropy is a good candidate. For nonlinear systems the choice of a good Lyapunov function remains much a matter of experience, luck, and trial and error. For linear systems we shall see that the construction of (quadratic) Lyapunov functions can be carried out quite explicitly.

Now return to the linear system (7.4), and examine this technique in detail for quadratic Lyapunov functions.

Definition 7.4.3 Let $M \in \mathbb{R}^{n\times n}$ be symmetric: $M = M^T$. Then the function from $\mathbb{R}^n$ to $\mathbb{R}$ defined by $x \mapsto x^T Mx$ is called the *quadratic form defined by* M. The symmetric matrix M is said to be:

- *nonnegative definite* (denoted $M \geq 0$) if $x^T Mx \geq 0$ for all $x \in \mathbb{R}^n$;
- *positive definite* (denoted $M > 0$) if in addition, $x^T Mx = 0$ implies $x = 0$;
- *nonpositive definite* (denoted $M \leq 0$) if $-M$ is nonnegative definite;
- *negative definite* (denoted $M < 0$) if $-M$ is positive definite.

□

Let us now carry out the computation of $\dot{V}$ in the case that V a quadratic form and the system is given by (7.4). Consider the quadratic form $V(x) = x^T Px$ with $P = P^T$, and examine how it evolves along solutions of (7.4). Let $x : \mathbb{R} \to \mathbb{R}^n$ be a solution of (7.4). Then x satisfies $\frac{d}{dt}x = Ax$ and it is easily seen, using the chain rule of differentiation, that $\frac{d}{dt}V(x(t))$ is given by

$$\frac{d}{dt}V(x(t)) = (\text{grad } V)(x(t)) \cdot Ax(t) = x^T(t)(A^T P + PA)x(t).$$

In other words, if at time t, $x(t) = a$, then the derivative of $V(x(t))$ at time t is $a^T(A^T P + PA)a$. We denote this function by $\dot{V}$. Explicitly, $\dot{V} : \mathbb{R}^n \to \mathbb{R}$ is thus given by

$$\dot{V}(x) = x^T(\underbrace{A^T P + PA}_{Q})x.$$

Note that $\dot{V}$ is also a quadratic form, $\dot{V}(x) = x^T Qx$, with Q the matrix

$$Q = A^T P + PA. \tag{7.19}$$

Equation (7.19) shows the relation between the system matrix $A \in \mathbb{R}^{n\times n}$, the symmetric matrix $P = P^T$ defining the quadratic Lyapunov function

$V(x) = x^T Px$, and the symmetric matrix $Q = Q^T$ defining its derivative along solutions of (7.3), the quadratic function $\dot{V}(x) = x^T Qx$. Equation (7.19) viewed as a relation between A, P, and Q is called the *Lyapunov equation*. Sometimes, one wants to find Q for a given A and P in order to see if $\dot{V} \leq 0$. Sometimes, however, one wants to find P for a given A and Q. Note that if $Q \leq 0$, then $\dot{V} \leq 0$, and so $V : x \mapsto x^T Px$ is then a Lyapunov function for (7.4). Thus solving (7.19) for P corresponds to constructing a Lyapunov function. As mentioned in the introduction, this should allow us to draw conclusions regarding the stability of (7.4). However, before we can make this statement precise, we need one more thing. Recall that the pair of matrices (A, Q) is said to be *observable* if the associated state system $\frac{d}{dt}x = Ax$, $w = Qx$ is observable, equivalently, if and only if the rank of $\mathrm{col}(Q, QA, \ldots, QA^{n-1})$ is n (see Theorem 5.3.9), or if and only if the only A-invariant subspace contained in $\ker Q$ is $\{0\}$ (see Theorem 5.3.13). Note, in particular, that if Q is nonpositive or nonnegative definite, then (A, Q) is observable if and only if $\left(\frac{d}{dt}x = Ax,\ x^T Qx = 0\right) \Rightarrow (x = 0)$. This follows easily from the implication that then $(x^T Qx = 0) \Leftrightarrow (Qx = 0)$.

Theorem 7.4.4 *Consider (7.4). Assume that $A, P = P^T$, and $Q = Q^T$ satisfy the Lyapunov equation (7.19). Then*

1. *$(P > 0,\ Q \leq 0) \Rightarrow$ ((7.4) is stable).*
2. *$(P > 0,\ Q \leq 0$, and (A, Q) observable$) \Rightarrow$ ((7.4) is asymptotically stable).*
3. *$(P$ not $\geq 0,\ Q \leq 0$, and (A, Q) observable$) \Rightarrow$ ((7.4) is unstable).*

Proof We use the following fact from linear algebra. Consider the symmetric matrix $P = P^T$. If $P > 0$, then there exist $\epsilon, M \in \mathbb{R}$, with $0 < \epsilon \leq M$, such that for all $x \in \mathbb{R}^n$ there holds $\epsilon x^T x \leq x^T Px \leq Mx^T x$.

1. Let $x : \mathbb{R} \to \mathbb{R}^n$ be a solution of (7.4). Consider as Lyapunov function V the quadratic form defined by P. Then $\frac{d}{dt}V(x)(t) = x^T(t)Qx(t)$ is nonpositive because $Q \leq 0$. Hence, for $t \geq 0$,

$$V(x(t)) - V(x(0)) = \int_0^t \frac{d}{dt}V(x(\tau))d\tau \leq 0.$$

Consequently, $x^T(0)Px(0) \geq x^T(t)Px(t)$. Hence $Mx^T(0)x(0) \geq x^T(0)P(x(0) \geq x^T(t)Px(t) \geq \epsilon x^T(t)x(t)$, which shows that for $t \geq 0$, $\|x(t)\|^2 \leq \frac{M}{\epsilon}\|x(0)\|^2$ for $t \geq 0$. Boundedness of x on $[0, \infty)$ and hence stability of (7.4) follow.

2. From (1), we know that (7.4) is stable. If it were not asymptotically stable, then by Corollary 7.2.4, A must have an eigenvalue on the imaginary axis. Therefore, (7.4) would have a nonzero periodic solution. Let $\tilde{x}$ be this

periodic solution, and assume that it has period $T > 0$. Define the subspace $\mathfrak{L}$ of $\mathbb{R}^n$ by $\mathfrak{L} = \text{span}\{\tilde{x}(t), t \in [0,T]\}$. Now verify that $\mathfrak{L}$ is A-invariant ($Ax = \lim_{t\to 0} \frac{e^{At}x_0 - x_0}{t}$ then belongs to $\mathfrak{L}$ if x_0 does). Furthermore, since

$$0 = V(\tilde{x}(T)) - V(\tilde{x}(0)) = \int_0^T \frac{d}{dt} V(\tilde{x}(\tau))d\tau = \int_0^T \tilde{x}^T(\tau)Q\tilde{x}(\tau)d\tau,$$

$\tilde{x}^T(t)Q\tilde{x}(t)$ must be zero for $t \in [0,T]$. Since $Q \leq 0$, this implies that $Q\tilde{x}(t) = 0$ for $t \in [0,T]$. Hence $\mathfrak{L}$ is an A-invariant subspace contained in $\ker Q$, implying, by the observability of (A,Q), that $\mathfrak{L} = \{0\}$. Hence $\tilde{x} = 0$, which establishes by contradiction that (7.4) is indeed asymptotically stable.

3. In order to prove (3), first use the same argument as in (2) to prove that (7.4) cannot have nonzero periodic solutions. Hence A has no eigenvalues with zero real part. Therefore it suffices to prove that (7.4) is not asymptotically stable. Since P is not ≥ 0, there is an $a \in \mathbb{R}^n$ such that $a^TPa < 0$. Now consider the solution $x : \mathbb{R} \to \mathbb{R}^n$ of (7.4) with $x(0) = a$. By the Lyapunov argument used in (1), it follows that for $t \geq 0$, $x^T(t)Px(t) = V(x(t)) \leq V(x(0)) = V(a) < 0$. Therefore, $x^T(t)Px(t) \leq V(a) < 0$ for all $t \geq 0$. By continuity, this shows that for this solution x, $\lim_{t\to\infty} x(t)$ cannot be zero. Hence (7.4) is unstable in this case. □

Example 7.4.5 Assume that $A + A^T \leq 0$. Then Theorem 7.4.4 with $P = I$ shows that (7.4) is stable. If $A + A^T < 0$, then it is asymptotically stable. More generally, if $A + A^T \leq 0$, then (7.4) is asymptotically stable if $(A, A + A^T)$ is an observable pair of matrices. □

Example 7.4.6 Consider the system described by the scalar second-order differential equation

$$bw + a\frac{d}{dt}w + \frac{d^2}{dt^2}w = 0. \tag{7.20}$$

Recall from Example 3.2.2 that we can think of (7.20) as describing the motion of a unit mass in a mass–damper–spring combination, with a the damping coefficient and b the spring constant. By the results in Section 7.3, we know that this system is

(asymptotically stable) $\Leftrightarrow$ ($a > 0$ and $b > 0$);

(stable) $\Leftrightarrow$ (($a \geq 0$ and $b > 0$) or ($a > 0$ and $b \geq 0$));

(unstable) $\Leftrightarrow$ (($a < 0$) or ($b < 0$) or ($a = b = 0$)).

Let us see whether we can deduce this also from Theorem 7.4.4. Introduce the state variables $x_1 = w$ and $x_2 = \frac{d}{dt}w$. This leads to

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix} x. \tag{7.21}$$

Consider the following quadratic Lyapunov function V.

- For $a \geq 0 : V(x_1, x_2) = bx_1^2 + x_2^2$. Its derivative along solutions of (7.21) is $\dot{V}(x_1, x_2) = -2ax_2^2$. In terms of the notation of theorem 7.4.4, we have $P = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix}$ and $Q = \begin{bmatrix} 0 & 0 \\ 0 & -2a \end{bmatrix}$. Note that (A, Q) is observable if and only if $a > 0$ and $b \neq 0$. Furthermore, $(P > 0) \Leftrightarrow (b > 0)$, $(P \geq 0) \Leftrightarrow (b \geq 0)$, and $(P \text{ not } \geq 0) \Leftrightarrow (b < 0)$. Theorem 7.4.4 therefore allows us to conclude that (7.21) is
  - asymptotically stable if $a > 0$ and $b > 0$.
  - stable if $a \geq 0$ and $b > 0$.
  - unstable if $a > 0$ and $b < 0$.
- For $a \leq 0 : V(x_1, x_2) = -bx_1^2 - x_2^2$. Its derivative is $\dot{V}(x_1, x_2) = 2ax_2^2$. Following the above argument, we conclude that
  - (7.21) is unstable if $a < 0$ and $b \neq 0$.
- For $a = 0 : V(x_1, x_2) = x_1 x_2$. Its derivative is $\dot{V}(x_1, x_2) = -bx_1^2 + x_2^2$. Applying theorem 7.4.4, we conclude, after some calculations, that
  - (7.21) is unstable if $a = 0$ and $b < 0$.
- For $b = 0 : V(x_1, x_2) = (ax_1 + x_2)^2 + x_2^2$. Its derivative is $\dot{V}(x_1, x_2) = -2ax_2^2$. Conclusion:
  - (7.21) is stable if $a > 0$ and $b = 0$.

We have now covered all cases, except when $b = 0$ and $a \leq 0$, for which (7.21) is unstable. In fact, Theorem 7.4.4 cannot be used to prove instability in this case, since it is easy to verify that then there exist no $P = P^T$ and $Q = Q^T \leq 0$, with (A, Q) observable, satisfying the Lyapunov equation (7.19). □

Example (7.4.6) shows at the same time the power and some of the pitfalls of Theorem 7.4.4. The choice of V as a Lyapunov function in the first two cases is rather natural if one identifies the system (7.20) with a mass–damper–spring combination, the Lyapunov function with the stored energy, and its derivative with the rate of dissipation of energy. However, the Lyapunov function for the other two cases has no such simple physical interpretation. Finally, we also saw that Theorem 7.4.4 was unable to provide a complete analysis even in this simple example. Indeed, on the basis of Theorem 7.4.4 we could not conclude instability in the case $b = 0$ and $a \leq 0$.

Next, we establish the converse of part 1 of Theorem 7.4.4. In other words, we show that for asymptotically stable systems it is always possible to find a suitable quadratic Lyapunov function.

Theorem 7.4.7 *Assume that A is a Hurwitz matrix.*

1. Then for any $Q = Q^T$ there exists a unique $P = P^T$ such that (7.19) is satisfied.

2. Moreover, $(Q \leq 0) \Rightarrow (P \geq 0)$.

3. Finally, if $Q \leq 0$, then $(P > 0) \Leftrightarrow ((A, Q)$ is observable).

Proof 1. Let A be Hurwitz and let $Q = Q^T$ be given. Consider the symmetric $(n \times n)$ matrix

$$-\int_0^\infty e^{A^T t} Q e^{At} dt. \tag{7.22}$$

Note that since A is Hurwitz, this matrix is well-defined; in other words, the infinite integral converges. Furthermore,

$$\begin{aligned} A^T(-\int_0^\infty e^{A^T t} Q e^{At} dt) + (-\int_0^\infty e^{A^T t} Q e^{At} dt) A \\ = -\int_0^\infty \frac{d}{dt}(e^{A^T t} Q e^{At}) dt = -e^{A^T t} Q e^{At} \,]_0^\infty = Q. \end{aligned}$$

Hence the matrix defined by (7.22) indeed satisfies the Lyapunov equation (7.19).

Next, we prove that it is the *unique* solution to this Lyapunov equation with P viewed as the unknown. We give two proofs of uniqueness. The reason for giving two proofs is purely pedagogical. The proofs differ widely in mathematical character, and both are useful for generalization to more general situations. The first proof is based on a property of linear transformations on finite-dimensional vector spaces. Consider the map $L : X \mapsto A^T X + XA$. Clearly, L maps the set of symmetric $(n \times n)$ matrices into itself. Thus L is a linear mapping from a real $\frac{n(n+1)}{2}$ dimensional vector space into itself. We have just proved that it is surjective, since for any $Q = Q^T$, (7.22) provides us with a solution to $L(X) = Q$. Since L is a surjective linear map from a finite-dimensional vector space onto itself, it is also injective. (Indeed, a square matrix has full column rank if and only if it has full row rank.) Hence (7.22) is the only $P = P^T$ that satisfies (7.19).

For the second proof of uniqueness, assume that both P_1 and P_2 satisfy (7.19). Then $\Delta := P_1 - P_2$ satisfies $A^T\Delta + \Delta A = 0$. Consider $M(t) := e^{A^T t}\Delta e^{At}$ and note that $\frac{d}{dt}M(t) = e^{A^T t}(A^T\Delta + \Delta A)e^{At} = 0$. Hence $M(t)$ is a constant matrix as a function of t. Obviously, $M(0) = \Delta$, and $M(t) \to 0$ as $t \to \infty$. Therefore, $\Delta = 0$, i.e., $P_1 = P_2$, which establishes uniqueness.

2. The expression (7.22) shows that for $a \in \mathbb{R}^n$, $a^T P a = \int_0^\infty (e^{At}a)^T(-Q)(e^{At}a)dt \geq 0$ when $Q \leq 0$. Hence $(Q \leq 0) \Rightarrow (P \geq 0)$.

3. From (7.22), we obtain that for all $a \in \mathbb{R}^n$,

$$a^T P a = a^T \Big(\int_0^\infty e^{A^T t}(-Q)e^{At} dt \Big) a. \tag{7.23}$$

Now, observability of (A, Q) implies that the right-hand side of (7.23) is zero only if $a = 0$. Since we already know from part 2 that $P \geq 0$, this yields $P > 0$. This establishes the implication ($\Leftarrow$) of part 2.

To show the converse implication, observe that $P > 0$ implies that the left-hand side of (7.23) is zero only if $a = 0$. Therefore, $Qe^{At}a = 0$ for $t \geq 0$ only if $a = 0$, which establishes the observability of (A, Q). □

Summarizing Theorems 7.4.4 and 7.4.7 for asymptotically stable systems shows that if (A, P, Q) satisfy the Lyapunov equation (7.19), then $P = P^T > 0, Q = Q^T \leq 0$, and (A, Q) observable imply that A is Hurwitz. Conversely, if A is Hurwitz and if we pick any $Q = Q^T \leq 0$ with (A, Q) observable, then there is a unique solution P to (7.19), and it satisfies $P = P^T > 0$.

Example 7.4.6 (continued):

Theorem 7.4.7 allows us to conclude that for $a > 0$ and $b > 0$, there must exist a $V(x_1, x_2)$ such that $\dot{V}(x_1, x_2) = -x_1^2 - x_2^2$. Let us compute it. The relevant Lyapunov equation is

$$\begin{bmatrix} 0 & -b \\ 1 & -a \end{bmatrix} \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} + \begin{bmatrix} p_1 & p_2 \\ p_2 & p_3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -b & -a \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Solving for p_1, p_2, p_3 yields

$$p_1 = \frac{a}{2b} + \frac{b+1}{2a}\ ,\ p_2 = \frac{1}{2b}\ ,\ p_3 = \frac{b+1}{2ab}.$$

Asymptotic stability is easy whenever $V(x)$ is a positive definite quadratic form and $\dot{V}(x)$ is a negative definite one. Unfortunately, while such Lyapunov functions exist (see (7.4.7)), they are not easy to obtain. Thus if we interpret this example as a mass–damper–spring combination, we see that using the stored energy $\frac{1}{2}(bx_1^2 + x_2^2)$ yields a positive definite Lyapunov function V, but with derivative $\dot{V}$ that is nonpositive definite but not negative definite (since the dissipation depends only on the velocity). From a physical point of view, this is a very natural Lyapunov function. However, in order to allow us to conclude asymptotic stability, we need to invoke observability. On the other hand, there always exist positive definite quadratic Lyapunov functions with a negative definite derivative. We have just seen that

$$V(x_1, x_2) = \left(\frac{a}{2b} + \frac{b+1}{2a} \right) x_1^2 + \frac{1}{b} x_1 x_2 + \frac{b+1}{2ab} x_2^2$$

indeed yields $\overset{\bullet}{V}(x_1, x_2) = -x_1^2 - x_2^2$.

7.5 Stability by Linearization

Let us now consider again the nonlinear system (7.15)

$$\frac{d}{dt}x = f(x), \tag{7.24}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ may be nonlinear and is now assumed to be differentiable. Let $\Sigma = (\mathbb{R}, \mathbb{R}^n, \mathfrak{B})$ be the dynamical system defined by (7.24). Since in stability considerations we agreed to consider only smooth solutions, the behavior is defined by $\mathfrak{B} := \{x : \mathbb{R} \to \mathbb{R}^n \mid x \text{ is differentiable and } \frac{d}{dt}x(t) = f(x(t)) \text{ for all } t\}$. Furthermore, if, for example, $f' : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, the Jacobi matrix of derivatives of f, is bounded on $\mathbb{R}^n$, then for each $a \in \mathbb{R}^n$, there exists precisely one element $x \in \mathfrak{B}$ such that $x(0) = a$. These existence and uniqueness properties are standard results from the theory of differential equations. However, they are of no real concern to us in the sequel, and we mention them merely for completeness.

Let $x^* \in \mathbb{R}^n$ be an *equilibrium point* of (7.24). This means that $f(x^*) = 0$ and hence that the constant trajectory $x : \mathbb{R} \to \mathbb{R}^n$ with $x(t) = x^*$ for all $t \in \mathbb{R}$ belongs to $\mathfrak{B}$. It is the stability of this equilibrium point that matters to us in this section.

Definition 7.5.1 The equilibrium point $x^* \in \mathbb{R}^n$ of (7.24) is said to be *stable* if for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$(x \in \mathfrak{B}, \|x(0) - x^*\| \leq \delta) \Rightarrow (\|x(t) - x^*\| \leq \epsilon \text{ for all } t \geq 0).$$

It is said to be an *attractor* if there exists an $\epsilon > 0$ such that

$$(x \in \mathfrak{B}, \|x(0) - x^*\| \leq \epsilon) \Rightarrow (\lim_{t\to\infty} x(t) = x^*).$$

It is said to be *asymptotically stable* if it is a stable attractor, and *unstable* if it is not stable. □

These definitions are illustrated in Figure 7.5. See Exercises 7.25 and 7.26 for the relations between Definitions 7.2.1 and 7.5.1 for linear systems.

It turns out that the stability properties of the equilibrium point x^* can to a large extent be decided by the linearization of the system (7.24) at the equilibrium point x^*. Linearization has been discussed extensively in Section 4.7. Recall that the linear system

$$\frac{d}{dt}\Delta = f'(x^*)\Delta \tag{7.25}$$

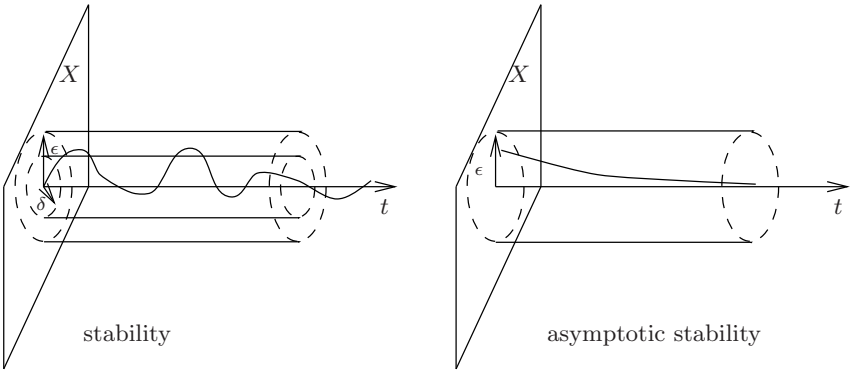


FIGURE 7.5. Stability and asymptotic stability.

is the *linearization* of (7.24) at the equilibrium point x^*. Here f' denotes the Jacobi matrix of f, i.e., the matrix of first-order partial derivatives

$$f' = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots\!\vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \tag{7.26}$$

where f is given by

$$f(x_1, x_2, \dots, x_n) = \begin{bmatrix} f_1(x_1, x_2, \dots, x_n) \\ f_2(x_1, x_2, \dots, x_n) \\ \vdots \\ f_n(x_1, x_2, \dots, x_n) \end{bmatrix}.$$

Now, $f'(x^*)$ is the matrix obtained by evaluating f' at the equilibrium point x^*. It is important to realize that $f'(x^*)$ is a constant $(n \times n)$ matrix and hence (7.25) becomes a system of first-order differential equations like (7.4) with $A = f'(x^*)$. Also, remember from Section 4.7 that if $x \in \mathfrak{B}$, then $x(t)$ is equal to $x^* + e^{At}(x(0) - x^*)$ up to terms of order $\|x(0) - x^*\|^2$. From this it stands to reason that there is a close relation between the stability of x^* as an equilibrium point of (7.24) and the stability of (7.25).

Theorem 7.5.2 *Consider (7.24) and assume that $f(x^*) = 0$.*

1. Assume that all the eigenvalues of the matrix $f'(x^*)$ have negative real parts. Then x^* is an asymptotically stable equilibrium of (7.24).

2. Assume that at least one eigenvalue of $f'(x^*)$ has positive real part. Then x^* is an unstable equilibrium of (7.24).

Proof 1. Consider the Lyapunov equation

$$(f'(x^*))^T P + P(f'(x^*)) = -I.$$

It follows from Theorem 7.4.7 that this equation has a unique solution $P = P^T > 0$. Consider the rate of change of the function $V : \mathbb{R}^n \to \mathbb{R}$ defined by $V(x) = (x-x^*)^T P(x-x^*)$, along solutions of (7.24). Let $x \in \mathfrak{B}$ be a solution of (7.24) and compute $\frac{d}{dt}(V(x))$. Obviously,

$$\frac{d}{dt}V(x)(t) = 2(x(t) - x^*)^T P f(x(t)).$$

Since $f(x) = f(x^*) + f'(x^*)(x-x^*)$ + terms of order higher than $\|x-x^*\|^2$ and since $f(x^*) = 0$, it follows that

$$\frac{d}{dt}V(x)(t) = -\|x(t) - x^*\|^2 + \text{ terms of order higher than } \|x(t) - x^*\|^2.$$

This implies that there exists an $\epsilon > 0$ such that

$$\left(\|x(t) - x^*\|^2 \leq \epsilon\right) \Rightarrow \left(\frac{d}{dt}V(x)(t) \leq -\frac{1}{2}\|x(t) - x^*\|^2\right). \tag{7.27}$$

On the other hand, since $P = P^T > 0$, there exists $\delta > 0$ such that

$$\left((x(t) - x^*)^T P(x(t) - x^*) \leq \delta\right) \Rightarrow \left(\|x(t) - x^*\| \leq \epsilon\right). \tag{7.28}$$

Furthermore, since $P = P^T > 0$, there exists $\alpha > 0$ such that

$$\|x(t) - x^*\|^2 \geq \alpha(x(t) - x^*)^T P(x(t) - x^*). \tag{7.29}$$

Using (7.27), (7.28), and (7.29), we obtain

$$\begin{aligned}&((x(t) - x^*)^T P(x(t) - x^*) \leq \delta) \Rightarrow \\ &\left(\frac{d}{dt}(x(t) - x^*)^T P(x(t) - x^*) \leq -\frac{\alpha}{2}(x(t) - x^*)^T P(x(t) - x^*)\right).\end{aligned} \tag{7.30}$$

From (7.30) we conclude that

$$\begin{aligned}&\left((x(0) - x^*)^T P(x(0) - x^*) \leq \delta\right) \Rightarrow \\ &\left(x(t) - x^*\right)^T P(x(t) - x^*) \leq e^{-\frac{\alpha}{2}t}\left((x(0) - x^*)^T P(x(0) - x^*)\right),\end{aligned}$$

which yields asymptotic stability.

2. The proof of the second part of the theorem is omitted. See Exercise 7.24. □

Example 7.5.3 The motion of a damped pendulum (see Example 7.1.1) is governed by the behavioral differential equation (7.2):

$$\frac{d^2}{dt^2}\phi + D\frac{d}{dt}\phi + \frac{g}{L}\sin\phi = 0,$$

where $L > 0$ denotes the length, $g > 0$ the constant of gravity, and D the friction coefficient. Take $x_1 = \phi$ and $x_2 = \frac{d}{dt}\phi$. The state space equations become

$$\begin{aligned} \frac{d}{dt}x_1 &= x_2, \\ \frac{d}{dt}x_2 &= -\frac{g}{L}\sin x_1 - Dx_2. \end{aligned}$$

The equilibria are
1. $x_1^* = 0, \quad x_2^* = 0$ (the pendulum is hanging down),
2. $x_1^* = \pi, \quad x_2^* = 0$ (the pendulum is standing up).

Linearization around these equilibria leads to

$$\frac{d}{dt}\Delta = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L} & -D \end{bmatrix}\Delta$$

for the first equilibrium, and

$$\frac{d}{dt}\Delta = \begin{bmatrix} 0 & 1 \\ \frac{g}{L} & -D \end{bmatrix}\Delta$$

for the second equilibrium. Application of Theorem 7.5.2 shows that when $D > 0$, the first equilibrium point is asymptotically stable, and unstable when $D < 0$. The second equilibrium point is unstable for both $D \geq 0$ and $D \leq 0$. It can be shown that in fact, the first equilibrium is also stable but not asymptotically stable when $D = 0$ (see Exercise 7.30), but that is a result that does not follow from Theorem 7.5.2. It requires analysis of the nonlinear system, instead of the linearized one (see Exercise 7.30). □

7.6 Input/Output Stability

In this section we examine the stability of the i/o system

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})u, \tag{7.31}$$

where $P(\xi) \in \mathbb{R}^{p\times p}[\xi]$, $\det P(\xi) \neq 0$, $Q(\xi) \in \mathbb{R}^{p\times m}[\xi]$, and $P^{-1}(\xi)Q(\xi) \in \mathbb{R}^{p\times m}(\xi)$ is a matrix of proper rational functions. We have seen in Section 3.3 that the behavior of (7.31) is given by

$$y(t) = H_0u(t) + \int_0^t H_1(t-\tau)u(\tau)d\tau + y_a(t), \tag{7.32}$$

where $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ and $H_0 \in \mathbb{R}^{p\times m}$, $H_1 : \mathbb{R} \to \mathbb{R}^{p\times m}$ are defined through the partial fraction expansion of $P^{-1}(\xi)Q(\xi)$ (see Theorem 3.3.13), and where y_a ranges over the set of solutions of the autonomous system

$$P(\frac{d}{dt})y = 0.$$

In i/o stability considerations, we are interested in the solutions on the half-line $[0, \infty)$. More specifically, we examine whether small inputs generate small outputs.

Definition 7.6.1 Let p be a real number, $1 \leq p < \infty$. The system (7.31) is said to be $\mathfrak{L}_p$-*i/o-stable* if

$$((u, y) \in \mathfrak{B} \text{ and } \int_0^\infty \|u(t)\|^p dt < \infty) \Rightarrow (\int_0^\infty \|y(t)\|^p dt < \infty).$$

It is said to be $\mathfrak{L}_\infty$-*i/o-stable* if

$$((u, y) \in \mathfrak{B} \text{ and } \sup_{t \geq 0} \|u(t)\| < \infty) \Rightarrow (\sup_{t \geq 0} \|y(t)\| < \infty).$$

Especially important in applications are $\mathfrak{L}_1$-, $\mathfrak{L}_2$-, and $\mathfrak{L}_\infty$-i/o stability. The third type of stability is often referred to as *BIBO (bounded input-bounded output)-stability.* □

Theorem 7.6.2 *1. Let $1 \leq p < \infty$. System (7.31) is $\mathfrak{L}_p$-i/o-stable if and only if all the roots of* $\det P(\xi)$ *have negative real parts.*

2. System (7.31) is $\mathfrak{L}_\infty$-i/o-stable if and only if each root of $\det P(\xi)$ *satisfies one of the following conditions:*

1. *its real part is negative;*
2. *its real part is zero, it is a semisimple singularity of $P(\xi)$, and it is* not *a pole of the transfer function $P^{-1}(\xi)Q(\xi)$. In the scalar case, this means that the roots of $P(\xi)$ on the imaginary axis must also be roots of $Q(\xi)$.*

Remark 7.6.3 The second condition of part 2 of the above theorem can be interpreted in terms of the uncontrollable modes of (7.31). Indeed, it states that the controllable part (see Section 5.2, in particular Theorem 5.2.14) of (7.31) cannot have poles on the imaginary axis. However, the uncontrollable part can have poles on the imaginary axis, provided that they are semisimple. □

For the proof of Theorem 7.6.2 we need the following lemma, which is of interest in its own right.

Lemma 7.6.4 *Let $p(\xi), q(\xi) \in \mathbb{R}[\xi]$, $p(\xi) \neq 0$, $p^{-1}(\xi)q(\xi)$ be proper, and assume that $p^{-1}(\xi)q(\xi)$ has a pole on the imaginary axis. Then the dynamical system represented by*

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u \tag{7.33}$$

is not $\mathfrak{L}_\infty$-i/o-stable.

Proof Let $i\omega_0 \in \mathbb{C}$, $\omega_0 \in \mathbb{R}$, be a pole of $p^{-1}(\xi)q(\xi)$. We show that bounded inputs of the form $u_{\omega_0} : t \mapsto \alpha e^{i\omega_0 t}$, $0 \neq \alpha \in \mathbb{C}$, generate unbounded outputs. Note that the solutions corresponding to this input $(u_{\omega_0}, y_{\omega_0})$ satisfy the set of differential equations

$$\begin{aligned} p(\frac{d}{dt})y_{\omega_0} &= q(\frac{d}{dt})u_{\omega_0}, \\ (\frac{d}{dt} - i\omega_0)u_{\omega_0} &= 0. \end{aligned} \tag{7.34}$$

The second equation guarantees that u_{ω_0} has the desired form $t \mapsto \alpha e^{i\omega_0 t}$, while the first one guarantees that $(u_{\omega_0}, y_{\omega_0})$ satisfies (7.33).

From Theorem 7.2.2 it follows that there are unbounded solutions $(u_{\omega_0}, y_{\omega_0})$ to (7.34) if and only if $i\omega_0$ is not a semisimple singularity of

$$\begin{bmatrix} p(\xi) & -q(\xi) \\ 0 & \xi - i\omega_0 \end{bmatrix}. \tag{7.35}$$

We now show that this is the case. Note that since $i\omega_0$ is a pole of $p^{-1}(\xi)q(\xi)$, $i\omega_0$ is certainly a root of $p(\xi)$, and if it happens also to be a root of $q(\xi)$, its multiplicity as a root of $q(\xi)$ must be less than its multiplicity as a root of $p(\xi)$. Now the dimension of the kernel of

$$\begin{bmatrix} p(i\omega_0) & q(i\omega_0) \\ 0 & 0 \end{bmatrix}$$

is 1 if $q(i\omega_0) \neq 0$ and 2 if $q(i\omega_0) = 0$, in which case $i\omega_0$ is a root of $p(\xi)$ of multiplicity at least two. The multiplicity of $i\omega_0$ as a root of

$$\det \begin{bmatrix} p(\xi) & -q(\xi) \\ 0 & \xi - i\omega_0 \end{bmatrix} = p(\xi)(\xi - i\omega_0)$$

equals one plus the multiplicity of $i\omega_o$ as a root $p(\xi)$. This shows that $i\omega_0$ is not a semisimple singularity of (7.35).

Hence, from Theorem 7.2.2, it follows that (7.34) has a solution of the form $t \mapsto (\alpha e^{i\omega_0 t}, (\beta + \gamma t)e^{i\omega_0 t})$ with $\gamma \neq 0$. This proves the lemma. □

Proof of Theorem 7.6.2 We prove only part 2, the BIBO case, $p = \infty$. The case $p = 1$ is also easily proven, while the other cases $1 < p < \infty$ are more tricky.

In order to prove the "if" part, observe that the assumptions imply that all the poles of the matrix of rational functions $P^{-1}(\xi)Q(\xi)$ have negative real parts. It then follows from (3.31) that

$$\int_0^\infty \|H_1(t)\|dt < \infty.$$

Let $(u, y) \in \mathfrak{B}$ and assume that $\sup_{t\geq 0} \|u(t)\| \leq \infty$. Of course, u and y are related by (7.32). Clearly, $\sup_{t\geq 0} \|H_0 u(t)\| < \infty$. Also, since the roots of $\det P(\xi)$ have negative real parts, or zero real part and are semisimple as singularities of $P(\xi)$, it follows from Theorem 7.2.2 that $\sup_{t\geq 0} \|y_a(t)\| < \infty$. Furthermore, for $t \geq 0$,

$$\begin{aligned} \|\int_0^t H_1(t-\tau)u(\tau)d\tau\| &\leq \int_0^t \|H_1(t-\tau)\| \, \|u(\tau)\|d\tau \\ &\leq (\int_0^t \|H_1(t-\tau)\|d\tau)(\sup_{t\geq 0}\|u(t)\|) \\ &\leq (\int_0^\infty \|H_1(t)\|dt)(\sup_{t\geq 0}\|u(t)\|). \end{aligned}$$

Therefore, y as given by (7.32) is the sum of three functions that are bounded on $\mathbb{R}_+$. It follows that $\sup_{t\geq 0} \|y(t)\| < \infty$, as claimed.

In order to prove the "only if" part, observe first that $\mathfrak{L}_\infty$-i/o-stability requires that all solutions of $P(\frac{d}{dt})y = 0$ be bounded on $\mathbb{R}_+$. Theorem 7.2.2 then shows that the roots of $\det P(\xi)$ must either have negative real part or have zero real part and be semisimple singularities of $P(\xi)$. It remains to be shown that if $P(\xi)^{-1}Q(\xi)$ has a pole on the imaginary axis, then (7.31) is not $\mathfrak{L}_\infty$-i/o-stable. In Lemma 7.6.4 we have proven this for the single-input/single-output case.

We now show in the multivariable case that if the transfer function $P^{-1}(\xi)Q(\xi)$ has a pole on the imaginary axis, then (7.31) is not $\mathfrak{L}_\infty$-i/o-stable. Let $V_1(\xi) \in \mathbb{R}^{p\times p}[\xi]$ and $V_2(\xi) \in \mathbb{R}^{p\times p}[\xi]$ be unimodular polynomial matrices such that $V_1(\xi)P(\xi)V_2(\xi)$ is in diagonal form. Let $P'(\xi) := V_1(\xi)P(\xi)V_2(\xi)$ and $Q'(\xi) := V_1(\xi)Q(\xi)$. Then $(P')^{-1}(\xi)Q'(\xi)$ has also a pole on the imaginary axis. Hence one of its entries, say the (k, ℓ)th entry, has a pole on the imaginary axis. Let $i\omega_0$ be such a pole. Now consider for the system described by

$$P'(\frac{d}{dt})y' = Q'(\frac{d}{dt})u' \tag{7.36}$$

the input $u' = \mathrm{col}(u'_1, \ldots, u'_m)$ with the ℓth entry given by $u'_\ell : t \mapsto \alpha e^{i\omega_0 t}$ with $\alpha \neq 0$, and the other elements zero. Then the kth element of y' in (7.36) is governed by

$$p'_{kk}(\frac{d}{dt})y'_k = q'_{k\ell}(\frac{d}{dt})u'_\ell, \tag{7.37}$$

where p_{kk} denotes the kth element on the diagonal of the (diagonal) polynomial matrix $P'(\xi)$ and $q'_{k\ell}$ denotes the (k, ℓ)th element of $Q'(\xi)$. By Lemma 7.6.4 it follows that (7.37) admits an unbounded solution of the form $(\beta + \gamma t)e^{i\omega_0 t}$ with $\gamma \neq 0$. It follows from this that the set of solutions

to

$$\begin{aligned} P'(\frac{d}{dt})y' &= Q'(\frac{d}{dt})u', \\ (\frac{d}{dt} - i\omega_0)u' &= 0 \end{aligned} \tag{7.38}$$

contains unbounded solutions. Now consider the solution set of

$$\begin{aligned} P(\frac{d}{dt})y &= Q(\frac{d}{dt})u, \\ (\frac{d}{dt} - i\omega_0)u &= 0. \end{aligned} \tag{7.39}$$

The definition of $P'(\xi)$ and $Q'(\xi)$ shows that (u', y') is a solution to (7.38) if and only if $(u', V_2^{-1}(\frac{d}{dt})y')$ is a solution to (7.39). Since $V_2(\xi)$ is unimodular, this shows that also (7.39) has unbounded solutions (see Exercise 7.2). This ends the proof of Theorem 7.6.2. □

The proof of Theorem 7.6.2 shows that when $P(\xi)$ has a singularity at $i\omega_0$, then the system (7.31) with input $u = 0$ has solutions of the form $e^{i\omega_0 t}\alpha$, $\alpha \neq 0$. If $i\omega_0$ is a pole of $P^{-1}(\xi)Q(\xi)$, then (7.31) has unbounded solutions of the form $u : t \mapsto \alpha e^{i\omega_0 t}$, $\quad y : t \mapsto (\beta + \gamma t)e^{i\omega_0 t}$, with $\gamma \neq 0$. Note that this unbounded solution is generated by a bounded input applied to a system that is zero–input stable (but not asymptotically stable). This phenomenon is called *resonance* and $\frac{\omega_0}{2\pi}$ is called a *resonant frequency*. It implies that periodic inputs such as $u : t \mapsto A\cos\omega_0 t$ yield unbounded outputs of the form $y : t \mapsto (B + Ct)\cos(\omega_0 t + \varphi)$. Periodic inputs at resonant frequencies are hence "pumped up" to generate unbounded outputs.

Example 7.6.5 Consider the motion of the position q of the mass of a mass–spring–damper combination under influence of an external force F. See Examples 3.2.2 and 3.2.3. This system is governed by the scalar differential equation

$$Kq + D\frac{d}{dt}q + M\frac{d^2}{dt^2}q = F$$

with $M > 0$ the mass, $K > 0$ the spring constant, and $D \geq 0$ the damping. If $D > 0$, then Theorem 7.6.2 allows us to infer $\mathfrak{L}_\infty$-i/o-stability. However, if $D = 0$, the transfer function

$$\frac{1}{K + M\xi^2}$$

has a pole at $\pm i\sqrt{\frac{K}{M}}$, showing that the system is not $\mathfrak{L}_\infty$-i/o-stable.

This can be illustrated by computing the solution q resulting from applying the input force $F = \sin\sqrt{\frac{K}{M}}t$ to the undamped system starting at rest with $q(0) = 0$ and $\frac{d}{dt}q(0) = 0$. The resulting response q is given by

$$q(t) = \frac{1}{2K}\sin\sqrt{\frac{K}{M}}t - \frac{1}{2\sqrt{KM}}t\cos\sqrt{\frac{K}{M}}t.$$

The second term is the resonance term and shows that the undamped system is not $\mathfrak{L}_\infty$-i/o-stable. □

Resonance is an important phenomenon in applications. Undamped or lightly damped systems generate very large responses when subject to small inputs containing a periodic component with period equal to the natural frequency of the system. It is in order to avoid this resonance response that in older times marching soldiers had to fall out of step when crossing a bridge. Resonance is also responsible for the singing of glasses and vases that sometimes occurs when playing high toned opera music (such as *the Queen of the Night*) loudly in a room.

7.7 Recapitulation

The topic of this chapter is stability of dynamical systems. Stability is one of the important concepts in systems theory. It is often the most central issue in the synthesis of control systems. The main points of this chapter are:

- The mathematical definition of stability. For linear autonomous systems, stability concepts refer to boundedness and convergence to zero of solutions (Definition 7.2.1). For nonlinear autonomous systems, stability is a property of an equilibrium solution and refers to the behavior of solutions in the neighborhood of the equilibrium (Definition 7.5.1). For input/output systems, stability means that bounded inputs should produce bounded outputs (Definition 7.6.1).
- Stability of autonomous linear systems can be determined explicitly by the location of the roots of the determinant of the polynomial matrix specifying the kernel representation or the eigenvalues of the system matrix specifying the state representation. In particular, the system is asymptotically stable if and only if the roots of the characteristic polynomial or the eigenvalues of the system matrix have negative real part (Theorem 7.2.2 and Corollary 7.2.4).
- There are explicit tests that allow one to deduce that the roots of a polynomial have negative real part, in particular the Routh test and the Hurwitz test (Theorems 7.3.1 and 7.3.3).
- An effective way of examining stability of a dynamical system is by means of a Lyapunov function, an energy-like function whose rate of change can be evaluated without computing the solutions (Definition 7.4.2). For linear systems, quadratic Lyapunov functions can be explicitly constructed through a linear matrix equation, called the Lyapunov equation (Theorems 7.4.4 and 7.4.7).
- The asymptotic stability and instability of an equilibrium point of a nonlinear system are closely related to the analogous property of the linearized system (Theorem 7.5.2).

- Input/output stability can be decided in terms of the roots of the determinant of a polynomial matrix specifying the zero-input behavior (Theorem 7.6.2). An interesting phenomenon occurring in the context of bounded-input/bounded-output stability is that of resonance.

7.8 Notes and References

Many textbooks on control and on the theory of differential equations treat stability problems, for example [62] and [23]. The Routh–Hurwitz problem originated in the paper [39] by J.C. Maxwell in 1868. This paper can be considered to be the first mathematical paper on control. Maxwell posed the problem of finding conditions on the coefficients of a polynomial for the real part of its roots to be negative as a public problem for the Adams prize. The prize was awarded to Routh for his work leading to the Routh test [49]. The Hurwitz test appeared in [24]. There have been uncountable papers devoted to variations on the Routh–Hurwitz problem. A book that treats many aspects of this problem is [7]. The proof of the Routh test outlined in Exercise 7.15 is inspired by [41]. The nice stability results for interval polynomials, Theorem 7.9.2 of Exercise 7.17, appeared in [32]. An elementary treatment of Lyapunov methods can be found in [62]. Input/output stability is a more recent development: see e.g. [54], where earlier references can be found.

7.9 Exercises

7.1 Determine for all parameters $\alpha \in \mathbb{R}$ whether the systems described by the following differential equations represent stable, asymptotically stable, or unstable systems. Do not use Routh–Hurwitz, but determine the roots of the corresponding polynomial (matrix) explicitly.

(a) $\alpha w + \frac{d^2}{dt^2} w = 0$.

(b) $\alpha^2 w + 2\alpha \frac{d^2}{dt^2} w + \frac{d^4}{dt^4} w = 0$.

(c) $\begin{bmatrix} \frac{d}{dt}(\frac{d}{dt}+1) & \alpha \\ 0 & \frac{d}{dt}(\frac{d}{dt}+1) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0.$

7.2 Consider the autonomous systems $R_1(\frac{d}{dt})w = 0$ and $R_2(\frac{d}{dt})w = 0$, with $R_1(\xi)$ and $R_2(\xi) \in \mathbb{R}^{q\times q}[\xi]$, $\det R_1(\xi) \neq 0$ and $\det R_2(\xi) \neq 0$. Assume that there exist unimodular polynomial matrices $U(\xi), V(\xi) \in \mathbb{R}^{q\times q}[\xi]$ such that $R_2(\xi) = U(\xi)R_1(\xi)V(\xi)$. Prove that the first system is asymptotically stable, stable, or unstable if and only if the second one is.

7.3 Let $P(\xi) \in \mathbb{R}[\xi]$, $P(\xi) \neq 0$. Consider the dynamical system represented by $P(\frac{d^2}{dt^2})w = 0$. Prove that it is asymptotically stable if and only if P is

of degree 0. Prove that it is stable if and only if all the roots of $P(\xi)$ are strictly negative and simple. Prove that this system is hence stable if and only if all solutions are bounded on all of $\mathbb{R}$.

7.4 Determine for what $\omega_1, \omega_2, \alpha \in \mathbb{R}$ the following system represents a stable, asymptotically stable, or unstable system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & \omega_1 & \alpha & 0 \\ -\omega_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega_2 \\ 0 & 0 & -\omega_2 & 0 \end{bmatrix} x.$$

7.5 Prove that if $A \in \mathbb{R}^{n\times n}$ is Hurwitz, then $\mathrm{Tr}(A) < 0$ and $(-1)^n \det(A) > 0$. Prove that these conditions imply that A is Hurwitz in the case $n = 2$.

7.6 Prove that $p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + p_n\xi^n$ is a Hurwitz polynomial if and only if $p_n + p_{n-1}\xi + \cdots + p_1\xi^{n-1} + p_0\xi^n$ is. Prove that $A \in \mathbb{R}^{n\times n}$ is a Hurwitz matrix if and only if A^{-1} is, and if and only if A^T is.

7.7 Let $A \in \mathbb{R}^{n\times n}$. Its *characteristic polynomial* $\chi_A(\xi)$ is $\det(I\xi - A)$. A polynomial $p(\xi) \in \mathbb{R}[\xi]$ is said to *annihilate* A if $p(A)$ is the zero matrix. In other words, if $p_0 I + p_1 A + \cdots + p_d A^d$ is the zero matrix, where $p(\xi) = p_0 + p_1\xi + \cdots + p_d\xi^d$. The Cayley-Hamilton theorem states that $\chi_A(\xi)$ annihilates A. The monic polynomial of minimal degree that annihilates A is called the *minimal polynomial* of A and is denoted by $\mu_A(\xi)$. It is easy to prove that $\chi_A(\xi)$ and $\mu_A(\xi)$ have the same roots but that the multiplicities of the roots of $\mu_A(\xi)$ may be less than those of A. Prove that $\lambda \in \mathbb{C}$ is a semisimple eigenvalue of A if and only if it is a simple root of $\mu_A(\xi)$.

7.8 Consider the discrete-time system ($\mathbb{T} = \mathbb{Z}$) with behavioral difference equation $P(\sigma)w = 0$, with σ the shift operator ($(\sigma w)(t) = w(t+1)$). The associated difference equation is thus $P_0 w(t) + P_1 w(t+1) + \cdots + P_2 w(t+L) = 0$. Assume that $P(\xi)$ is square and that $\det P(\xi) \neq 0$. Define stability, instability, and asymptotic stability fully analogously as in the continuous-time case. Prove that this system is

(a) *asymptotically stable* if and only if all the roots of $\det P(\xi)$ are inside the unit disc $|\lambda| < 1$. A polynomial matrix having this property is called Schur (see Exercise 7.19);

(b) *stable* if and only if for each $\lambda \in \mathbb{C}$ that is a root of $P(\xi)$ there must hold either (i) $|\lambda| < 1$, or (ii) $|\lambda| = 1$ and λ is a semisimple root of $P(\xi)$;

(c) *unstable* if and only if $P(\xi)$ has either a root with $|\lambda| > 1$ and/or a nonsemisimple root with $|\lambda| = 1$.

7.9 Consider the discrete-time analogue of (7.4), $x(t+1) = Ax(t)$. Define stability, asymptotic stability, and instability fully analogously as in continuous-time case. Prove the analogue of Corrolary 7.2.4.

7.10 Which of the following polynomials are Hurwitz?

(a) $1 + \xi + \xi^2 + \xi^3 + \xi^4 + \xi^5$.

(b) $1 + 5\xi + 10\xi^2 + 10\xi^3 + 5\xi^4 + \xi^5$.

7.11 Determine necessary and sufficient conditions on the coefficients $a, b, c, d \in \mathbb{R}$ for the polynomial $d + c\xi + b\xi^2 + a\xi^3 + \xi^4$ to be Hurwitz.

7.12 Prove that the $(n+2)$th row of the Routh table of an nth order polynomial is zero.

7.13 Call $p(\xi) \in \mathbb{R}[\xi]$ *anti-Hurwitz* if all its roots have positive real part. Give a Routh-type test in terms of the coefficients of $p(\xi)$ for it to be anti-Hurwitz.

7.14 Prove the following refinements of Theorems 7.3.3 and 7.3.4.

(a) Assume that $p_n > 0$. Prove that if all the roots of $p(\xi)$ have nonpositive real part, then $\Delta_1 \geq 0, \Delta_2 \geq 0, \ldots, \Delta_n \geq 0$. Provide a counterexample for the converse.

(b) Prove that if all the roots of $p(\xi) \in \mathbb{R}[\xi]$ have nonpositive real part, then no two coefficients of $p(\xi)$ can have opposite sign, but some can be zero.

7.15 The purpose of this exercise is to lead the reader through a step-by-step proof of the Routh test. Let $p(\xi) \in \mathbb{R}[\xi]$ be given by (7.9), and assume that it has degree n. Write $p(\xi)$ in terms of its even and odd parts as

$$p(\xi) = E_0(\xi^2) + \xi E_1(\xi^2).$$

Note that $E_0(\xi), E_1(\xi) \in \mathbb{R}[\xi]$ are given by

$$E_0(\xi) = p_0 + p_2\xi + p_4\xi^2 + \cdots, \quad E_1(\xi) = p_1 + p_3\xi + p_5\xi^2 + \cdots.$$

The coefficients of the polynomials $E_0(\xi)$ and $E_1(\xi)$ form the first and second rows of the Routh table. The third row of the Routh table consists of the coefficients of the polynomial

$$E_2(\xi) = \xi^{-1}(E_1(0)E_0(\xi) - E_0(0)E_1(\xi)).$$

Prove that with the obvious notation, the $(k+1)$th row of the Routh table consists of the coefficients of the polynomial

$$E_k(\xi) = \xi^{-1}(E_{k-1}(0)E_{k-2}(\xi) - E_{k-2}(0)E_{k-1}(\xi)).$$

The Routh test thus states that if $p_n > 0$, then $p(\xi)$ is Hurwitz if and only if the constant term coefficients of the $E_k(\xi)$s, $E_1(0), E_2(0), \ldots, E_n(0)$ are all positive.

Define $q(\xi) \in \mathbb{R}[\xi]$ by $q(\xi) = E_1(\xi^2) + \xi E_2(\xi^2)$. Prove that $q(\xi)$ has degree less than n. The key to the Routh test is provided by the following lemma.

Lemma 7.9.1 *Denote the leading coefficients of $p(\xi)$ and $q(\xi)$ by p_n and q_{n-1}, respectively. The following statements are equivalent:*

(i) $p(\xi)$ is Hurwitz and $p_n > 0$;

(ii) $q(\xi)$ *is Hurwitz,* $q_{n-1} > 0$*, and* $p(0) > 0$*.*

Organize your proof of this lemma as follows:

- Consider the convex combination of $p(\xi)$ and $q(\xi)$, $q_\alpha(\xi) = (1-\alpha)p(\xi) + \alpha q(\xi)$ for $\alpha \in [0,1]$. Write $q_\alpha(\xi)$ in terms of $E_0(\xi^2)$ and $E_1(\xi^2)$, and prove that if $p(0) > 0$ and $q(0) > 0$ then all the polynomials $q_\alpha(\xi)$ have the same imaginary axis roots for $\alpha \in [0,1]$.
- Prove (i) $\Rightarrow$ (ii). Hint: Use the fact that no roots of $q_\alpha(\xi)$ cross the imaginary axis to show $q_\alpha(\xi)$ is Hurwitz for all $\alpha \in [0,1]$.
- Prove (ii) $\Rightarrow$ (i). Hint: Use the fact that no roots of $q_\alpha(\xi)$ cross the imaginary axis to show that $q_\alpha(\xi)$ has at least $n-1$ roots in the open left half of the complex plane for $\alpha \in [0,1]$. Prove that $p(0)p_n > 0$ implies that the nth root of $p(\xi) = q_0(\xi)$ lies also in the open left half of the complex plane.

Finally, prove the Routh test by induction on n, using this lemma.

7.16 The purpose of this exercise is to lead the reader through a proof of the Hurwitz test. We use the notation introduced in Exercise 7.15. Let H_p be the Hurwitz matrix associated with $p(\xi)$, and H_q the one associated with $q(\xi)$.

(i) Prove that

$$H_p P = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ p_3 & & & \\ p_5 & & H_q & \\ \vdots & & & \end{bmatrix},$$

where

$$P = \begin{bmatrix} 1 & -p_0 & 0 & 0 & \cdots \\ 0 & p_1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -p_0 & \cdots \\ 0 & 0 & 0 & p_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

(ii) Let $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}, \Delta_n$ denote the Hurwitz determinants associated with $p(\xi)$, and $\Delta'_1, \Delta'_2, \ldots, \Delta'_{n-1}$ those associated with $q(\xi)$. Prove from the above relation between H_p and H_q that

$$\begin{aligned} \Delta_1 &= p_1 \\ p_1^{\lfloor \frac{k}{2} \rfloor - 1} \Delta_{k-1} &= \Delta'_k \text{ for } k = 1, 2, \ldots, n-1; \end{aligned}$$

where $\lfloor \alpha \rfloor$ denotes the largest integer $\leq \alpha$.

(iii) Use Lemma 7.9.1 to prove the Hurwitz test by induction on n.

7.17 In this exercise we study a Hurwitz type test for the asymptotic stability of differential equations when only bounds on the coefficients are known.

(i) Assume that the two polynomials $p_0 + p_1\xi + \cdots + p_{k-1}\xi^{k-1} + p'_k\xi^k + p_{k+1}\xi^{k+1} + \cdots + p_n\xi^n$ and $p_0 + p_1\xi + \cdots + p_{k-1}\xi^{k-1} + p''_k\xi^k + p_{k+1}\xi^{k+1} + \cdots + p_n\xi^n$ are both Hurwitz. Use the induction lemma used in the proof of the Routh test to prove that $p_0 + p_1\xi + p_{k-1}\xi^{k-1} + \cdots + p_k\xi^k + p_{k+1}\xi^{k+1} + \cdots + p_n\xi^n$ is also Hurwitz for all $p'_k \leq p_k \leq p''_k$.

(ii) Let $[a_k, A_k], k = 0, 1, \ldots, n$, be $n+1$ intervals in $\mathbb{R}$. Consider the *interval family* of polynomials consisting of all polynomials $p_0 + p_1\xi + \ldots + p_n\xi^n$ with $a_k \leq p_k \leq A_k$ for $k = 0, 1, \ldots, n$. Its *extreme points* consist of the 2^{n+1} polynomials with $p_k \in \{a_k, A_k\}$ for $k = 0, 1, \cdots, n$. Use (i) to prove that all the polynomials in this interval family are Hurwitz if and only if its extreme points are.

(iii) The result of (ii) concerning an interval family of polynomials can be dramatically simplified. Define the four polynomials $k_1(\xi), k_2(\xi), k_3(\xi), k_4(\xi)$ as follows:

$$\begin{aligned}
k_1(\xi) &= a_0 + a_1\xi + A_2\xi^2 + A_3\xi^3 + a_4\xi^4 + a_5\xi^5 + A_6\xi^6 + \cdots, \\
k_2(\xi) &= a_0 + A_1\xi + A_2\xi^2 + a_3\xi^3 + a_4\xi^4 + a_5\xi^5 + A_6\xi^6 + \cdots, \\
k_3(\xi) &= A_0 + A_1\xi + a_2\xi^2 + a_3\xi^3 + A_4\xi^4 + A_5\xi^5 + a_6\xi^6 + \cdots, \\
k_4(\xi) &= A_0 + a_1\xi + a_2\xi^2 + A_3\xi^3 + A_4\xi^4 + a_5\xi^5 + a_6\xi^6 + \cdots.
\end{aligned}$$

Note that these polynomials follow the pattern

$$\ldots, \max, \min, \min, \max, \max, \min, \min, \max, \max,$$

(the *Kharitonov melody*).

The purpose of this exercise is to prove the following result:

Theorem 7.9.2 (Kharitonov test) *All polynomials in the interval family are Hurwitz if and only if the four polynomials $k_1(\xi), k_2(\xi), k_3(\xi)$, and $k_4(\xi)$ are Hurwitz.*

Prove this result as follows

1. First prove that if $k_1(\xi), k_2(\xi), k_3(\xi)$, and $k_4(\xi)$ are Hurwitz, then any convex combination of these polynomials is also Hurwitz. In order to see this, write these four polynomials as

$$\begin{aligned}
&E'_0(\xi^2) + \xi E'_1(\xi^2), E'_0(\xi^2) + \xi E''_1(\xi^2), \\
&E''_0(\xi^2) + \xi E'_1(\xi^2), E''_0(\xi^2) + \xi E''_1(\xi^2),
\end{aligned}$$

   and use the induction used in the proof of the Routh test.

2. Next, prove that if $p(\xi)$ is any element of the interval family of polynomials, then

$$\begin{aligned}
&\mathrm{Re}(k_1(i\omega)) = \mathrm{Re}(k_2(i\omega)) \leq \mathrm{Re}(p(i\omega)) \leq \mathrm{Re}(k_3(i\omega)) = \mathrm{Re}(k_4(i\omega)), \\
&\mathrm{Im}(k_1(i\omega)) = \mathrm{Im}(k_4(i\omega)) \leq \mathrm{Im}(p(i\omega)) \leq \mathrm{Im}(k_2(i\omega)) = \mathrm{Im}(k_3(i\omega)).
\end{aligned}$$

   for all $\omega \in \mathbb{R}, 0 \leq \omega < \infty$.

3. Combine 2 and 3 to prove that $p(\xi)$ cannot have roots on the imaginary axis.
4. Finally, prove Theorem 7.9.2.

7.18 It is well known that by choosing $\alpha_1, \beta, \gamma, \delta \in \mathbb{R}$ suitably, the map $s \in \mathbb{C} \mapsto \frac{\alpha s+\beta}{\gamma s+\delta} \in \mathbb{C}$ maps the imaginary axis onto any line parallel to the imaginary axis, or a circle with center on the real axis. This construction leads to Routh–Hurwitz type tests for the roots of a polynomial to lie strictly to the left or to the right of any line parallel to the imaginary axis, or inside or outside any circle centered on the real axis, by considering the polynomial

$$(\gamma\xi + \delta)^n p(\frac{\alpha\xi + \beta}{\gamma\xi + \delta}).$$

Use this idea to find conditions on p_0, p_1, p_2 for $p_0 + p_1\xi + p_2\xi^2 + \xi^3$ to have its roots strictly inside the unit disc.

7.19 As shown in Exercise 7.8, the scalar difference equation $p(\sigma)w = 0$ is asymptotically stable if and only if the polynomial $p(\xi) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + p_n\xi^n$ has all its roots inside the unit disc $\{z \in \mathbb{C} \mid |z| < 1\}$. A polynomial which has all its roots inside the unit disc is called *Schur*. The purpose of this exercise is to derive a Routh-type test for $p(\xi)$ to be Schur. Define $p^*(\xi) := p_n + p_{n-1}\xi + \cdots + p_1\xi^{n-1} + p_0\xi^n$. Denote $q_1(\xi) := p(\xi)$ and define

$$q_{k+1}(\xi) := \frac{q_k^*(0)q_k(\xi) - q_k(0)q_k^*(\xi)}{\xi}.$$

(i) Prove that $p(\xi)$ is Schur if and only if $|q_k(0)| < |q_k^*(0)|$ for $k = 1, 2, \ldots, n$. Hint: Use the ideas of Exercise 7.15: Prove that $p(\xi)$ is Schur if and only if $|p(0)| < |p^*(0)|$ and $q_2(\xi)$ is Schur. Proceed by induction on n.

(ii) Determine the resulting conditions on the coefficients of $p(\xi)$ for $n = 1, 2, 3, 4$.

7.20 Consider Theorem 7.3.5. Discuss whether stability or instability of (7.3) is a robust property, assuming that the degree of $\det P_\alpha(\xi)$ is constant in a neighborhood of α_0. Is instability a robust property if $\det P_{\alpha_0}(\xi)$ has a root with positive real part? Use Exercise 7.7 to formulate a result concerning robust asymptotic stability, stability, and instability properties of the dynamical system represented by $P(\frac{d^2}{dt^2})w = 0$.

7.21 Construct for the following asymptotically stable system a positive definite Lyapunov function with a negative definite derivative

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -1 \end{bmatrix} x.$$

7.22 The purpose of this exercise is to prove Corollary 7.3.6 using Theorem 7.4.7. Use the notation of Corollary 7.3.6. Assume that $A(\alpha_0)$ is Hurwitz. Let P be the solution of $A^T(\alpha_0)P + PA(\alpha_0) = -I$. Now prove that $A^T(\alpha)P +$

$PA(\alpha) < 0$ for all α sufficiently close to α_0. Conclude that $A(\alpha)$ is Hurwitz for α sufficiently close to α_0.

7.23 Assume that $P = P^T$ and $Q = Q^T$ satisfy the Lyapunov equation $A^T P + PA = Q$. In Theorem 7.4.4 it has been shown that the conditions (P not ≥ 0, $Q \leq 0$, and (A, Q) observable) imply that (7.4) is unstable. Theorem 7.4.7, however, did not claim the converse. In other words, it was *not* claimed that if (7.4) is unstable, then there exists a $P = P^T$ not ≥ 0 and a $Q = Q^T \leq 0$ with (A, Q) observable. So the question arises, *Are there unstable systems for which there do not exist such P and Q*? Prove that this is indeed the case. Specifically, prove that if $A, P = P^T$, and $Q = Q^T \leq 0$ satisfy the Lyapunov equation, then (A, Q) cannot be observable whenever A has eigenvalues with zero real part. Prove that if A has no eigenvalues with zero real part, then (7.4) is unstable if and only if there exist such P and Q.

7.24 Assume that $A \in \mathbb{R}^{n\times n}$ has at least one eigenvalue with real part positive. Prove then that there exists $Q = Q^T \leq 0$ and $P = P^T \leq 0, P \neq 0$, satisfying the Lyapunov equation (7.19). Now consider system (7.24) with $f(x^*) = 0$ and assume that $f'(x^*)$ has at least one eigenvalue with real part positive. Use the Lyapunov function $(x - x^*)^T P(x - x^*)$ to prove that x^* is an unstable equilibrium, as claimed in part 2 of Theorem 7.5.2.

7.25 Consider $\frac{d}{dt}x = Ax$. This is a special case of both (7.3) and (7.24). Of course, $x^* = 0$ is an equilibrium point of this system. What is the relation between the notions of stability introduced in Definitions 7.2.1 and 7.5.1 respectively? This exercise shows that these notions correspond.

(a) Prove that all solutions of $\frac{d}{dt}x = Ax$ are bounded on $[0, \infty)$ if and only if 0 is a stable equilibrium.

(b) Prove that all solutions of $\frac{d}{dt}x = Ax$ converge to zero as $t \to \infty$ if and only if 0 is an asymptotically stable equilibrium.

7.26 Consider the system (7.4). Of course, $x^* = 0$ is always an equilibrium, but there may be more. Obviously, $x^* = 0$ is the only equilibrium if and only if A is nonsingular.

(i) Prove that (7.4) is asymptotically stable if and only if $x^* = 0$ is an asymptotically stable equilibrium, in which case it is the only equilibrium solution.

(ii) Prove that the following are equivalent:
1. (7.4) is stable.
2. $x^* = 0$ is a stable equilibrium.
3. Let $a \in \mathbb{R}^n$ be another equilibrium. Then it is stable.

(iii) Repeat (ii) with "stable" replaced by "unstable."

7.27 Let us take a look at the Lyapunov equation (7.19) *ipso suo*. Let $A \in \mathbb{R}^{n\times n}$, and define the linear map $L : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ by $L(X) := A^T X + XA$. Assume that the matrix A^T has a basis of real eigenvectors $v_1, \ldots, v_n$, say $A^T v_k = \lambda_k v_k; k = 1, \ldots, n$.

(a) Prove that $v_k v_\ell^T$ is an eigenvector (eigenmatrix if you like) of L.

(b) Show that the vectors (matrices) $v_k v_\ell^T; k, \ell = 1, 2, \ldots n$ are linearly independent.

(c) Conclude that the eigenvalues of L are given by $\lambda_k + \lambda_\ell; k, \ell = 1, \ldots, n$. Note that these numbers are never distinct, even when $\lambda_1, \lambda_2, \ldots, \lambda_n$ are distinct.

(d) State necessary and sufficient conditions in terms of A for L to be a bijective map.

(e) Denote the linear subspace of real symmetric $n \times n$ matrices by S. Prove that S is L-invariant and determine the dimension of S.

(f) From part (c), we conclude that L_S, the restriction of L to S, is well-defined as a linear map from S to S. Prove that $v_k v_\ell^T + v_\ell v_k^T; k, \ell = 1, 2, \ldots, n; k \geq \ell$, forms an independent system of eigenvectors (eigenmatrices) of L_S. Conclude that the eigenvalues of L_S are given by $\lambda_k + \lambda_\ell; k, \ell = 1, \ldots, n,\ k \geq \ell$.

(g) Argue that the above results are valid for general matrices $A \in \mathbb{R}^{n \times n}$ (without the assumption of distinct eigenvalues or the existence of a basis of eigenvectors).

(h) Generalize all this to the case that the λ_ks and v_ks could be complex.

7.28 Linear mechanical systems can often be described by systems of second–order differential equations of the form

$$Kw + D\frac{d}{dt}w + M\frac{d^2}{dt^2}w = 0,$$

with M, D, and $K \in \mathbb{R}^{q \times q}$; M represents the masses (and $M\frac{d^2}{dt^2}w$ the inertial forces), D represents the damping (and $D\frac{d}{dt}w$ friction forces), and K represents the springs (and Kw the restoring forces). Assume that $M = M^T > 0$ and $K = K^T > 0$. Prove that this system is stable if $D + D^T \geq 0$, asymptotically stable if $D + D^T > 0$, and unstable if $D + D^T$ is not ≥ 0.

Hint: The total energy of the system is a good candidate Lyapunov function. The idea is that if the system dissipates energy through the dampers, then we have asymptotic stability. Introduce $\mathrm{col}(x_1, x_2)$ as the state, with $x_1 = w$ and $x_2 = \frac{d}{dt}w$, and consider as Lyapunov function the total energy $\frac{1}{2}x_1^T K x_1 + \frac{1}{2}x_2^T M x_2$. Show that $\dot{V}(x_1, x_2) = -x_2^T(D + D^T)x_2$.

7.29 Consider the system of differential equations

$$\begin{array}{rcl} \frac{d}{dt}x_1 & = & x_2, \\ \frac{d}{dt}x_2 & = & -x_1 - (\alpha + x_1^2)x_2. \end{array}$$

Use Theorem 7.5.2 to classify the stability properties of the equilibrium $x^* = 0$ for all $\alpha \neq 0$. For $\alpha = 0$, Theorem 7.5.2 does not allow us to reach a conclusion. Use a direct argument to show that the equilibrium is asymptotically stable when $\alpha = 0$.

7.30 Consider the equilibria of the undamped pendulum of Example 7.5.3 (with $D = 0$). Prove that

$$-\frac{g}{L}\cos x_1 + \frac{1}{2}x_2^2$$

is invariant along solutions. Examine the level sets of this function around $x_1^* = 0$, $x_2^* = 0$, and around $x_1^* = \pi$, $x_2^* = 0$. Use this to prove that the first equilibrium is stable, but the second is unstable.

7.31 Consider the linearized system (4.62) of Examples 4.7.1 and 4.7.2, the inverted pendulum on a carriage. Investigate the stability.

7.32 The dynamical equations of the rotation of a spinning body are given by

$$\begin{aligned} I_1\frac{d\omega_1}{dt} &= (I_2 - I_3)\omega_2\omega_3, \\ I_2\frac{d\omega_2}{dt} &= (I_3 - I_1)\omega_3\omega_1, \\ I_3\frac{d\omega_3}{dt} &= (I_1 - I_2)\omega_1\omega_2. \end{aligned} \tag{7.40}$$

These equations are called the *Euler equations*. Here $\omega_1, \omega_2, \omega_3$ denote the rotation rates of the body around its principal axes, and I_1, I_2, I_3 denote the moments of inertia of the body with respect to these principal axes. The Euler equations describe only the spinning of the body. The complete equations of motion that describe both the motion of the center of gravity and the attitude of the body are more complicated and are not discussed here. Assume for simplicity that $0 < I_1 < I_2 < I_3$ (implying a certain lack of symmetry).

(a) Describe the equilibria of (7.40).

(b) Linearize around the equilibria.

(c) Use Theorem 7.5.2 to prove that steady spinning around the second principal axis is unstable.

(d) Does Theorem 7.5.2 allow you to conclude something about the stability of steady spinning around the other axes?

(e) Prove that the quadratic forms $I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2$ and $I_1^2\omega_1^2 + I_2^2\omega_2^2 + I_3^2\omega_3^2$ are both invariants of motion (in other words, prove that they are constant along solutions of (7.40)). Sketch on a surface where the first quadratic form is constant (assume $I_1 = \frac{1}{2}, I_2 = 1, I_3 = 2$) the curves on which the second is also constant. Use this to prove that steady spinning around the first and third principal axes is stable, but not asymptotically stable.

Conclusion: A spinning body spins stably around the principal axis with the smallest and the largest moment of inertia, but not around the principal axis with the intermediate moments of inertia. This can be demonstrated by (carefully) tossing *this book* into the air. You will see that you can get it to spin nicely around its largest and smallest axes, but if you try to spin it around the middle axis, the motion will be very wobbly, suggesting instability. See [1] for an in–depth analysis of this problem.

7.33 For which $1 \leq p \leq \infty$ are the following systems $\mathfrak{L}_p$-i/o-stable?

(a) $(1 - \frac{d^2}{dt^2})y = u$.

(b) $(1 + \frac{d^2}{dt^2})y = u$.

(c) $(1 + \frac{d}{dt} + \frac{d^2}{dt^2})y = (1 - \frac{d}{dt})u$.

(d) $(1 - \frac{d^2}{dt^2})y = (1 - \frac{d}{dt})u$.

7.34 Consider the system described by $(\omega_0^2 + \frac{d^2}{dt})y = u$. Let $\omega_0 \in \mathbb{R}$, and consider the input $u = \cos \omega_0 t$. Compute the solution corresponding to the initial condition $y(0) = 0$, $\frac{d}{dt}y(0) = 0$. Conclude that this system is not $\mathfrak{L}_\infty$-i/o-stable. Verify that this agrees with Theorem 7.6.2.

7.35 Consider the nonlinear input/output system

$$\frac{1}{4}y - \frac{4}{10 - y} + \frac{d}{dt}y + \frac{d^2}{dt^2}y = 0 \tag{7.41}$$

(a) Define $x := [y \ \frac{d}{dt}y]^T$, and determine $f : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\frac{d}{dt}x = f(x)$, $y = x_1$ is a state space representation of (7.41).

(b) Determine the equilibrium points of $\frac{d}{dt}x = f(x)$.

(c) Linearize the system about each of the equilibrium points.

(d) Investigate the local (in)stability of the equilibrium points.

8

Time- and Frequency-Domain Characteristics of Linear Time-Invariant Systems

8.1 Introduction

The purpose of this chapter is twofold. First, we explain how a linear time-invariant system acts in the frequency domain. An important feature of such systems is that (in an input/output setting) they transform sinusoidal (and, more generally, exponential) inputs into sinusoidal (exponential) outputs. This leads to the transfer function and the frequency response as a convenient way of describing such systems. The second purpose of this chapter is to study properties of the time- and frequency-domain response. Thus we describe important characteristics of a system that can be deduced from its step-response, or from its Bode and Nyquist plots.

In Chapters 2 and 3, we studied two related classes of linear time-invariant dynamical systems. The first class consists of the systems described by differential equations

$$R(\frac{d}{dt})w = 0 \tag{8.1}$$

defined by the polynomial matrix $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. The second class consists of the systems described by the convolution

$$y(t) = \int_{-\infty}^{+\infty} H(t-t')u(t')dt', \tag{8.2}$$

defined in terms of the kernel $H \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^{p\times m})$. Often, the limits of the integral in (8.2) are, or can be taken to be, $\int_{-\infty}^{t}$ or $\int_{0}^{t}$, but for the time being, we need not be concerned with that. The function H is called the *impulse response matrix* of the system; see Section 3.4 for an explanation of this terminology. The system of equations (8.1) defines the dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B}_R)$ with behavior $\mathfrak{B}_R$ defined by

$$\mathfrak{B}_R = \{w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q) \mid R(\frac{d}{dt})w = 0 \ \text{ weakly}\}$$

while (8.2) defines the dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^p, \mathfrak{B}_H)$ with behavior $\mathfrak{B}_H$ defined by

$$\mathfrak{B}_H = \{(u, y) \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m \times \mathbb{R}^p) \mid \ (8.2) \text{ is satisfied.}\}$$

We have also seen that (8.1) can always be reduced to a system of equations of the form

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})u, \tag{8.3}$$

with $P(\xi) \in \mathbb{R}^{p\times p}[\xi], Q(\xi) \in \mathbb{R}^{p\times m}[\xi], \det P(\xi) \neq 0$, and $P^{-1}(\xi)Q(\xi) \in \mathbb{R}^{p\times m}(\xi)$ a matrix of proper rational functions. In this case, (8.3) defines an *input/output* dynamical system: its behavior allows for any function $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ as input, and the output y is completely specified by the input u and by the appropriate initial conditions. We have also seen that the system descriptions (8.3) and (8.2) are very closely related whenever H is a Bohl function (see also Section 3.5). Indeed, if $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ and assuming that system is initially at rest, then the response y to (8.3) is given by (8.2) with the impulse response matrix specified by (3.45).

In the remainder of this chapter, we occasionally silently assume that we are considering complex-valued inputs and outputs.

8.2 The Transfer Function and the Frequency Response

In this section we study systems from what is called the *frequency-domain* point of view. In this context, we basically expand the time functions in their frequency components and study how the individual frequency (or exponential) components are constrained or transformed by the dynamical system. In order to understand the system behavior it suffices then to "add" the behavior for the individual frequencies. This feature, of being able to view the response to a system as a sum of exponential terms, each of which individually satisfies the system equations, is characteristic for linear time-invariant systems. The mathematics that underlies this is the *Fourier and Laplace transforms*. A brief introduction to these is given in Appendix B.

We treat convolution systems and differential systems separately.

8.2.1 Convolution systems

Consider the system described by (8.2). Assume that the impulse response matrix has a two-sided Laplace transform

$$G(s) = \int_{-\infty}^{+\infty} H(t)e^{-st}dt.$$

Obviously, $G : \mathbb{C} \to \mathbb{C}^{p\times m}$ is a matrix of complex functions. Its domain of definition consists of the domain of convergence of the Laplace transform G, i.e., of all $s \in \mathbb{C}$ such that $H \exp_s \in \mathfrak{L}_1(\mathbb{R}, \mathbb{C}^{p\times m})$, where the exponential function with exponent s, $\exp_s : \mathbb{C} \to \mathbb{C}$, is defined by $\exp_s(t) := e^{st}$. The function G is called the *transfer function* of the system (8.2).

Consider the input exponential $u : \mathbb{R} \to \mathbb{C}^m$ defined by $u = u_s \exp_s$, with $u_s \in \mathbb{C}^m$. If s belongs to the domain of convergence of G, then

$$\begin{aligned} y(t) &= \int_{-\infty}^{+\infty} H(t-t')u_s e^{st'}dt' \\ &= (\int_{-\infty}^{+\infty} H(t-t')e^{-s(t-t')}dt')e^{st}u_s \\ &= G(s)u_s e^{st}. \end{aligned} \tag{8.4}$$

This shows that the output corresponding to an exponential input is also an exponential. The significance of the transfer function therefore is that it shows how exponential inputs are transformed into exponential outputs, namely, $u_s \exp_s$ is transformed into $y_s \exp_s$, with $y_s = G(s)u_s$. The vector u_s of the exponential input is thus multiplied by the transfer function matrix $G(s)$ in order to produce the corresponding output vector y_s.

The special case that $H \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}^{p\times m})$, i.e., that $\int_{-\infty}^{+\infty} \|H(t)\|dt < \infty$, is of particular importance. In this case H is Fourier transformable. Its Fourier transform

$$G(i\omega) = \int_{-\infty}^{+\infty} H(t)e^{-i\omega t}dt$$

is called the *frequency response* of (8.2). In this case, the output corresponding to the sinusoidal input $e^{i\omega t}u_\omega$ equals the sinusoidal output $e^{i\omega t}y_\omega$, with $y_\omega = G(i\omega)u_\omega$.

The effect of the transfer function and the frequency response on more general inputs is described in the following theorem.

Theorem 8.2.1 *(i) Consider the system (8.1) and assume that H is (2-sided) Laplace transformable, with its Laplace transform, called the transfer function, denoted by G. Let $u : \mathbb{R} \to \mathbb{C}^m$ be also (2-sided) Laplace transformable and denote its Laplace transform by $\hat{u}$. Assume that the intersection of the domains of convergence of G and $\hat{u}$ is nonempty. Then the output of (8.2) is also (2-sided) Laplace transformable. Denote its Laplace transform by $\hat{y}$. Then $\hat{y}(s) = G(s)\hat{u}(s)$, and the domain of convergence of $\hat{y}$ contains the intersection of those of G and $\hat{u}$.*

(ii) Assume that $H \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}^{p\times m})$ and $u \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}^m)$. Then y, defined by (8.2), belongs to $\mathfrak{L}_1(\mathbb{R}, \mathbb{R}^p)$, and the $\mathfrak{L}_1$-Fourier transforms of u and y satisfy $\hat{y}(i\omega) = H(i\omega)\hat{u}(i\omega)$.

(ii)′ Assume that $H \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}^{p\times m})$ and $u \in \mathfrak{L}_2(\mathbb{R}, \mathbb{R}^m)$. Then y, defined by (8.2), belongs to $\mathfrak{L}_2(\mathbb{R}, \mathbb{R}^p)$ and the $\mathfrak{L}_2$-Fourier transforms of u and y satisfy $\hat{y}(i\omega) = H(i\omega)\hat{u}(i\omega)$.

Proof We sketch only the proof of part (i); the other cases can be proven analogously (see Exercise B.3).

The input and output u and y are related by (8.2). Hence

$$\begin{aligned}
\hat{y}(s) &= \int_{-\infty}^{\infty} y(t)e^{-st}dt \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t-t')u(t')dt'e^{-st}dt \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t-t')e^{-s(t-t')}u(t')e^{-st'}dt\ dt' \\
&= \int_{-\infty}^{\infty} H(t)e^{-st}dt \int_{-\infty}^{\infty} u(t')e^{-st'}dt' \\
&= G(s)\hat{u}(s).
\end{aligned}$$

□

Example 8.2.2 Consider the system

$$y(t) = \frac{1}{2\Delta}\int_{t-\Delta}^{t+\Delta} u(t')dt'. \tag{8.5}$$

This is an example of a (simple) *smoother*, in which the output computes a windowed average of the input. It is a special case of (8.2) with

$$H(t) = \begin{cases} 1 & \text{for } |t| \leq \Delta, \\ 0 & \text{for } |t| > \Delta. \end{cases}$$

The frequency response of this system is given by

$$G(i\omega) = \int_{-\Delta}^{\Delta} e^{-i\omega t} dt = 2\frac{\sin \omega\Delta}{\omega}.$$

The frequency-response function G is shown in Figure 8.1. This figure

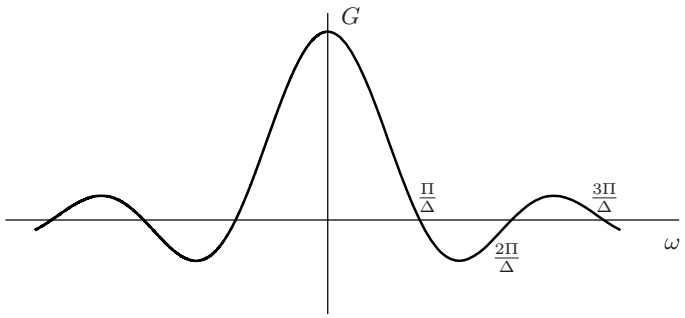


FIGURE 8.1. Frequency response of Example 8.2.2.

shows that, as can be intuitively expected from the fact that (8.5) averages the input, a high-frequency input is transformed into an output that is practically zero. It is also worthwhile to observe that there are certain frequencies that are completely cut off by this system, and hence the input $u_\omega e^{i\omega t}$, with $\omega = k\frac{\pi}{\Delta}$ for $k = 1, 2, \ldots$, results in a zero output. □

8.2.2 Differential systems

We now study the transfer function and the frequency response for differential systems (8.1). A differential system can be represented in input/output form (8.3), and as we have seen in Section 3.5, this leads to a convolution system (8.2) with impulse response matrix (3.45). In fact, the initially-at-rest response of (8.3) is exactly given by this convolution. As such, we can in principle apply the theory of Section 8.2.1 to this class of systems. Thus the transfer function is given as the Laplace transform of (3.45), and it is well known (see Exercise 8.2) that this transfer function is given by

$$G(s) = P^{-1}(s)Q(s). \tag{8.6}$$

The domain of convergence of (8.6) viewed as the Laplace transform of (3.45) includes the open half plane to the right of the root of $P(s)$ with largest real part.

This domain of convergence consideration is an annoying element of the interpretation of (8.6) as a transfer function. In particular, it implies that we cannot simply view $G(i\omega)$ as the frequency response of (8.3) unless $P(\xi)$ is a Hurwitz polynomial matrix, more precisely, unless the impulse response (3.45) belongs to $\mathfrak{L}_1(\mathbb{R}, \mathbb{R}^{p\times m})$. Note, however, that there are only a finite number of elements $s \in \mathbb{C}$, the roots of $\det P(\xi)$, where the expression $P^{-1}(s)Q(s)$ is not defined, and as such, the domain of definition of $P^{-1}(s)Q(s)$ equals not just a half plane, but all of $\mathbb{C}$ minus this finite set of points. So it seems reasonable that one should be able to interpret $P^{-1}(s)Q(s)$ as the transfer function without reference to domains of convergence. We shall see that the behavioral interpretation of (8.1) and (8.3) indeed allows us to do that. An important advantage of this is that the frequency response for differential systems is thus always a well-defined complex matrix, except at most at a finite number of points.

Let $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathfrak{B})$ be the linear time-invariant system represented by (8.1). We assume, for the reasons already explained, that we are considering complex systems obtained, for example, by complexifying a real system.

Definition 8.2.3 The *exponential behavior* of Σ is defined as the elements of $\mathfrak{B}$ of the special form $w = b\ \exp_s$ with $b \in \mathbb{C}^q$ and $s \in \mathbb{C}$. □

Thus the exponential behavior of Σ induces the mapping $\mathfrak{E}$ from $\mathbb{C}$ to the subset of $\mathbb{C}^q$ such that

$$\mathfrak{E}(s) = \{b \in \mathbb{C}^q \mid b \exp_s \in \mathfrak{B}\}.$$

Let us now consider the exponential behavior of (8.1). Since $R(\frac{d}{dt})(b \exp_s) = R(s)b \exp_s$ (see Exercise 8.4), we immediately obtain the following lemma.

Lemma 8.2.4 *The exponential behavior of (8.1) is characterized by*

$$\mathfrak{E}(s) = \ker R(s). \tag{8.7}$$

Thus, in particular, $\mathfrak{E}(s)$ is a linear subspace of $\mathbb{C}^q$.

Consider now the exponential behavior of (8.3). Obviously, for $s \in \mathbb{C}$ such that $\det P(s) \neq 0$, there holds

$$\mathfrak{E}(s) = \{(u_s, y_s) \in \mathbb{C}^m \times \mathbb{C}^p \mid y_s = P^{-1}(s)Q(s)u_s\}.$$

Therefore, in analogy to what has been obtained in Section 8.2.1, we call $G(s) = P^{-1}(s)Q(s)$ the *transfer function* of (8.3).

We emphasize that we view the transfer function as the mapping that produces from the exponential input $u_s \exp_s$, the corresponding exponential output $y_s \exp_s$ (easily seen to be unique when $\det P(s) \neq 0$), the relation being given by premultiplication of u_s by the *transfer function* in order to obtain y_s.

The following theorem shows the significance of the exponential behavior and of the transfer function for differential systems. A signal $w \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$ that is Laplace transformable is in the behavior of (8.1) if and only if its Laplace transform is in the exponential behavior for all s where the Laplace transform is defined.

Theorem 8.2.5 *Consider the system defined by (8.1). Let $w : \mathbb{R} \to \mathbb{C}^q$ be given. If w is (2-sided) Laplace transformable with Laplace transform $\hat{w}$, then $w \in \mathfrak{B}$ if and only if $\hat{w}(s) \in \mathfrak{E}(s)$ for all s in the domain of convergence for $\hat{w}$.*

Proof We give the proof only under the additional assumption that also $R(\frac{d}{dt})w$ is Laplace transformable.

(if): $R(\frac{d}{dt})w$ has (2-sided) Laplace transform $R(s)\hat{w}(s)$. Since $\hat{w}(s) \in \mathfrak{E}(s) = \ker R(s)$ for all s in the domain of convergence of $\hat{w}$, this implies that the Laplace transform of $R(\frac{d}{dt})w$ is zero. Hence $R(\frac{d}{dt})w = 0$, as desired.

(only if): Assume that $R(\frac{d}{dt})w = 0$. Then also its Laplace transform is zero, and hence $R(s)\hat{w}(s) = 0$ for all s in the domain of convergence of $\hat{w}$. Hence $\hat{w}(s) \in \mathfrak{E}(s)$, as claimed. □

It follows immediately from the above theorem that if $w \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}^q)$, then it belongs to the behavior of (8.1) if and only if its Fourier transform $\hat{w}(i\omega)$ satisfies $\hat{w}(i\omega) \in \mathfrak{E}(i\omega)$ for all $\omega \in \mathbb{R}$. The same holds for signals $w \in \mathfrak{L}_2(\mathbb{R}, \mathbb{R}^q)$, with $\hat{w}$ the $\mathfrak{L}_2$-Fourier transform.

In particular, for input/output systems, we obtain that if (u, y) is Laplace transformable, then (u, y) belongs to the behavior of (8.3) if and only if $\hat{y}(s) = G(s)\hat{u}(s)$ for all s in the domain of convergence of $(\hat{u}, \hat{y})$. Applied to ($\mathfrak{L}_1$- or $\mathfrak{L}_2$-) Fourier transformable pairs (u, y), this leads to $\hat{y}(i\omega) = G(i\omega)\hat{u}(i\omega)$. This has the following important consequence. Let $u \in \mathfrak{L}_2(\mathbb{R}, \mathbb{R}^q)$ and assume that $G(i\omega)\hat{u}(i\omega)$, viewed as a mapping from $\omega \in \mathbb{R}$ to $\mathbb{C}^q$, belongs to $\mathfrak{L}_2(\mathbb{R}, \mathbb{C}^q)$. This is automatically the case if $\det P(\xi)$ has no roots on the imaginary axis, since in that case $G(i\omega)$, viewed as a mapping from $\omega \in \mathbb{R}$ to $\mathbb{C}^{p\times m}$, is bounded. However, $G(i\omega)\hat{u}(i\omega)$ can obviously be in $\mathfrak{L}_2(\mathbb{R}, \mathbb{C}^q)$, even when $\det P(\xi)$ does have roots on the imaginary axis. Let $y \in \mathfrak{L}_2(\mathbb{R}, \mathbb{R}^q)$ be the inverse $\mathfrak{L}_2$-Fourier transform of $G(i\omega)\hat{u}(i\omega)$. Then (u, y) belongs to the behavior of (8.3). This shows that for differential systems (8.3) the frequency response $G(i\omega) = P^{-1}(i\omega)Q(i\omega)$ always has a clear significance in terms of the behavior, and $P(\xi)$ need not be Hurwitz for this interpretation to hold.

Example 8.2.6 Let us illustrate this by means of an example. Consider the mechanical system with behavioral differential equation relating the displacement q and the force F given by

$$Kq + M\frac{d^2}{dt^2}q = F,$$

where M is the mass and K the spring constant. Writing this in our standard form yields

$$\left[(K + M\frac{d^2}{dt^2}) \quad -1\right] \left[\begin{array}{c} q \\ F \end{array}\right] = 0.$$

Note that F is the input, and that q is the output. The corresponding convolution (8.2) is in this case

$$q(t) = \int_{-\infty}^{t} \sin\sqrt{\frac{K}{M}}(t-t') \cdot \frac{F(t')}{\sqrt{MK}}dt'.$$

- The exponential behavior (8.7) becomes

$$\mathfrak{E}(s) = \mathrm{Im} \left[\begin{array}{c} 1 \\ Ms^2 + K \end{array}\right].$$

- The transfer function equals

$$\frac{1}{Ms^2 + K}.$$

- If we apply a periodic input force F with Fourier series

$$F(t) = \sum_{k=-\infty}^{+\infty} \hat{F}_k e^{ik\frac{2\pi}{T}},$$

then there is a periodic q corresponding to this force if and only if

$$\hat{F}_k = 0 \text{ for } k = \pm\frac{\sqrt{\frac{K}{M}}}{2\pi/T},$$

in which case this output is

$$q(t) = \sum_{k=-\infty}^{+\infty} \frac{\hat{F}_k}{K - M\left(\frac{k2\pi}{T}\right)^2} e^{ik\frac{2\pi}{T}t}.$$

- If we apply an $\mathfrak{L}_2$-input force F with $\mathfrak{L}_2$-Fourier transform $\hat{F}(i\omega)$, then there is a corresponding $\mathfrak{L}_2$-output q if and only if

$$\omega \mapsto \frac{1}{K - M\omega^2}\hat{F}(i\omega)$$

is in $\mathfrak{L}_2(\mathbb{R};\mathbb{C})$, in which case q has $\frac{\hat{F}(i\omega)}{K-M\omega^2}$ as its $\mathfrak{L}_2$-Fourier transform.

- If we apply a Laplace transformable input F with Laplace transform $\hat{F}(s)$ that converges in a strip S such that $S \cap \{s \mid \mathrm{Re}(s) = 0\} \neq \varnothing$, then there is a corresponding Laplace transformable q with

$$\hat{q}(s) = \frac{1}{Ms^2 + K}\hat{F}(s)$$

as its Laplace transform.

□

8.2.3 The transfer function represents the controllable part of the behavior

From the definition of the transfer function it is clear that the behavior defines the transfer function uniquely. The converse, however, is not true. In fact, we now prove that two systems of the form (8.3) have the same transfer function if and only if the controllable parts of their behaviors coincide. Hence *the transfer function determines only the controllable part of a behavior* and is therefore a useful concept mainly in the context of controllable systems.

Theorem 8.2.7 *The controllable parts of the behaviors of two systems of the form (8.3) are the same if and only if their transfer functions are the same.*

Proof We give the proof only for single-input/single-output systems. Let $\mathfrak{B}_i$ be represented by $p_i(\frac{d}{dt})y = q_i(\frac{d}{dt})u$ $(i = 1, 2)$. By Theorem 5.2.14, (5.17), the controllable parts are represented by canceling the common factors of $p_i(\xi)$ and $q_i(\xi)$ yielding $\bar{p}_i(\frac{d}{dt})y = \bar{q}_i(\frac{d}{dt})u$. Since the transfer functions $G_1(s)$ and $G_2(s)$ are the same, we have $\frac{\bar{q}_1}{\bar{p}_1}(s) = \frac{\bar{q}_2}{\bar{p}_2}(s)$. Since the common factors have been canceled, this implies that $\bar{q}_1(s) = \alpha\bar{q}_2(s)$ and $\bar{p}_1(s) = \alpha\bar{p}_2(s)$, for some $\alpha \neq 0$. This implies that the controllable parts of the behaviors coincide, since then $\bar{p}_i(\frac{d}{dt})y = \bar{q}_i(\frac{d}{dt})u$ and $\bar{p}_2(\frac{d}{dt})y = \bar{q}_2(\frac{d}{dt})u$ represent the same system.

To show the converse, observe that $(u_s, G_i(s)u_s)\exp_s$ belongs to the controllable part of $\mathfrak{B}_i$. Therefore, if for $s' \in \mathbb{C}, G_1(s') \neq G_2(s')$, then there is an exponential response in the controllable part of $\mathfrak{B}_1$ but not in that of $\mathfrak{B}_2$. □

8.2.4 The transfer function of interconnected systems

One of the main applications of transfer functions as system representations is that it becomes very simple to calculate the transfer function of a system

that is specified as an interconnection of component subsystems through a signal flow diagram. We illustrate this for series, parallel, and feedback connection.

Let the transfer functions of the i/o systems Σ_1 and Σ_2 be given by $G_1(s)$ and $G_2(s)$ respectively. Assume that the input and output signal spaces have dimension such that the required interconnections are well defined.

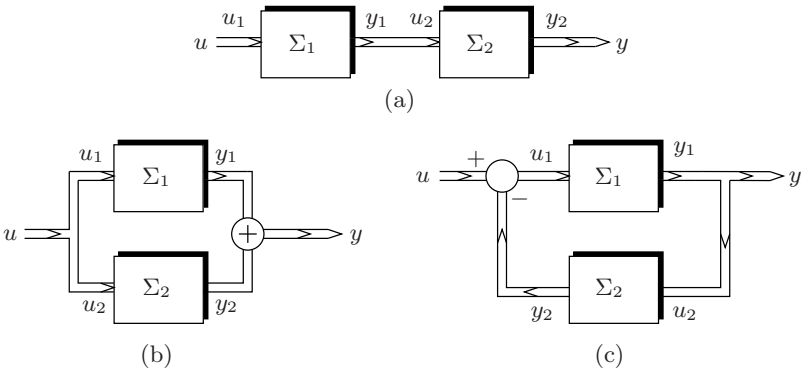


FIGURE 8.2. (a) Series, (b) parallel, and (c) feedback interconnection.

1. *Series interconnection* (see Figure 8.2). The series interconnection of Σ_1 and Σ_2 is defined by $u_2 = y_1$, $u = u_1$, and $y = y_2$. The transfer function of the series interconnection is given by $G(s) = G_2(s)G_1(s)$.

   Proof $\hat{y}(s) = \hat{y}_2(s) = G_2(s)\hat{u}_2(s) = G_2(s)\hat{y}_1(s) = G_2(s)G_1(s)\hat{u}_1(s) = G_2(s)G_1(s)\hat{u}(s)$. □

2. *Parallel interconnection* (see Figure 8.2). The parallel interconnection of Σ_1 and Σ_2 is defined by $u_1 = u_2 = u$ and $y = y_1 + y_2$. The transfer function of the parallel interconnection is given by $G(s) = G_1(s) + G_2(s)$.

   Proof $\hat{y}(s) = \hat{y}_1(s) + \hat{y}_2(s) = G_1(s)\hat{u}(s) + G_2(s)\hat{u}(s) = (G_1(s) + G_2(s))\hat{u}(s)$. □

3. *Feedback interconnection* (see Figure 8.2). The feedback interconnection of Σ_1 and Σ_2 is defined by $u = u_1 - y_2$, $u_2 = y_1$, $y = y_1$. The transfer function of the feedback interconnection is given by $G(s) = (I - G_1(s)G_2(s))^{-1}G_1(s)$.

   Proof $\hat{y}(s) = G_1(s)\hat{u}_1(s) = G_1(s)(\hat{u}(s) + \hat{y}_2(s)) = G_1(s)(\hat{u}(s) + G_2(s)\hat{y}(s))$. This implies $(I - G_1(s)G_2(s))\hat{y}(s) = G_1(s)\hat{u}(s)$, which in turn yields $\hat{y}(s) = (I - G_1(s)G_2(s))^{-1}G_1(s)\hat{u}(s)$. □

8.3 Time-Domain Characteristics

We now study some characteristic features of the time response of dynamical systems described by (8.2) or (8.3). We consider the single-input/single-output (SISO) case only. Multi-input/multi-output (MIMO) systems are usually analyzed by considering the transfer characteristics input channel by output channel. Some aspects of this section were already mentioned in Section 3.4.

The system under consideration is a special case of (8.2):

$$y(t) = \int_{-\infty}^{t} H(t-t')u(t')dt'. \tag{8.8}$$

Thus we in effect assumed that $H(t) = 0$ for $t < 0$: the system is assumed to be *nonanticipating*. Two important characteristics of (8.8) are the *impulse response* and the *step response*. The *impulse response* is the response as $\varepsilon \to 0$ to the input

$$u(t) = \begin{cases} 0 & \text{for } t < 0, \\ 1/\varepsilon & \text{for } 0 \le t \le \varepsilon, \\ 0 & \text{for } t > \varepsilon. \end{cases} \tag{8.9}$$

The corresponding output (assuming that H is continuous) is given, in the limit as $\varepsilon \to 0$, by

$$y(t) = \begin{cases} 0 & \text{for } t < 0, \\ H(t) & \text{for } t \ge 0. \end{cases}$$

A mathematically more sophisticated way of approaching this is to consider the Dirac delta distribution as the input. From this point of view, (8.9) defines a family of inputs that approach the delta distribution as $\epsilon \to 0$. In the engineering literature the Dirac delta distribution is called an impulse, whence the name impulse response.

The *step response* is the response to the *unit step* (sometimes called the *Heaviside step function*)

$$u(t) = \begin{cases} 0 & \text{for } t < 0, \\ 1 & \text{for } t \ge 0. \end{cases}$$

The corresponding output, denoted by s, for step response, is of course given by

$$s(t) = \begin{cases} 0 & \text{for } t < 0, \\ \int_0^t H(t')dt' & \text{for } t \ge 0. \end{cases}$$

A typical step response is shown in Figure 8.3. This figure shows a number of characteristic features of a step response. These are now formally defined.

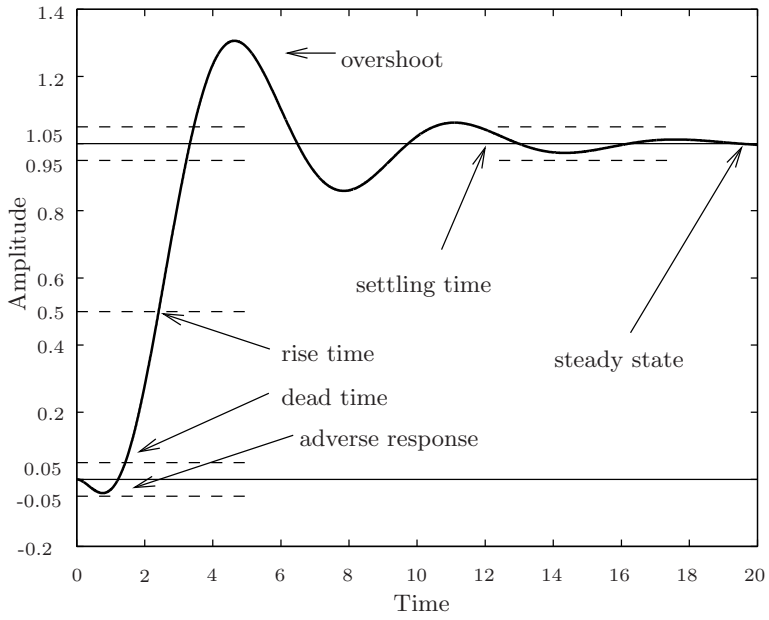


FIGURE 8.3. Step response of $(1-0.5s)/(1+s)(1+0.5s+s^2)$.

- The *steady-state*, or *static gain* is defined as

$$s_\infty = \lim_{t\to\infty} s(t).$$

  For asymptotically stable systems this limit s_∞ exists. For systems defined by the convolution (8.8) it equals $\int_0^\infty H(t)dt$; for systems described by (8.3) it equals $P^{-1}(0)Q(0)$. The steady-state gain is a measure of the amplification of constant inputs into constant outputs.

- The % *overshoot* is given by

$$\max_{t\geq 0} \frac{s(t)-s_\infty}{s_\infty} 100.$$

  The overshoot is a measure of the extent to which the output exceeds its steady-state value before settling down.

- The 5% *settling time* is given by

$$\min\{t \geq 0 \mid \ \mid s(t') - s_\infty \mid \leq 0.05 s_\infty \text{ for } t' \geq t\}.$$

  The settling time is a measure for the time at which the output settles down to its steady-state value.

- The *rise time* is given by

$$\min\{t \geq 0 \mid s(t) = 0.5 s_\infty\}.$$

  The rise time is a measure of the speed at which the output responds when it is driven by an input to a new steady-state value. Sometimes the value $0.9 s_\infty$ (instead of $0.5 s_\infty$) is used for the comparison level. The rise time is a measure of the global time delay that is present in a system.

- The 5% *deadtime* is given by

$$\max\{t \geq 0 \mid\mid s(t') \mid \leq 0.05 s_\infty \text{ for } \ 0 \leq t' < t\}.$$

  The deadtime is a measure of the hard time delay in the system, the time required to have the system react at all.

- The notion of the *timeconstant* of a system is very close in spirit to the rise time, and, to a lesser extent, to the deadtime and the settling time. However, the notion of *timeconstants* usually refers to the behavior of the autonomous system $\frac{d}{dt}x = Ax$, or more generally of $P(\frac{d}{dt})y = 0$, obtained by setting $u = 0$ in (8.3). In (3.16) and Section 4.5.4 we have seen that the solution of these differential equations consists of terms involving sums of exponentials $e^{\lambda_k' t}$, with the λ_k's the real parts of the roots of $\det P(s)$ or the eigenvalues of A. Assuming asymptotic stability, i.e., that these real parts are negative, then the times $T_k = -1/\lambda_k$ such that $e^{\lambda_k T_k} = e^{-1}$ are called the *timeconstants* of the system. Often, the largest of these is called *the timeconstant*. From the timeconstants one gets an idea of how fast the system reacts and how long it takes before transients die out.

- We say that a system has an *adverse response* if $s(t) \cdot s_\infty < 0$ for some t. Usually, the adverse response occurs in an interval containing the initial time 0. The % *(adverse) undershoot* is given by

$$\max_{t \geq 0}\{-\frac{s(t)}{s_\infty} 100\}.$$

  Not all systems have an adverse response. A system with an adverse response reacts (initially) in a direction that is opposite to its steady state response. A system with an adverse response is often difficult to understand intuitively and to control, since its initial response points in the wrong direction as compared to the direction of its ultimate response.

We warn the reader that there is no uniformity in the terminology used above. These concepts should be used with common sense (some of them

are meaningless, for example, if $s_\infty = 0$). The question arises as to what are good characteristics for a system. This, of course, depends on the type of application. In some applications the system is designed such that the output "tracks" the input. One can think of a radar whose direction is required to point towards an airplane that it is following. Or one can think of a servo in which the output is required to adjust to an imposed input path (robot motion, systems in which a desired temperature profile is to be followed, or a sensor, such as a thermometer, in which the sensor output should be a mere scaled version of the input). "Good" tracking systems require

- small overshoot,
- small settling time,
- small deadtime,
- small rise time,
- no adverse response.

Of course, these requirements are often incompatible. In other applications the system is designed so that the input disturbance is suppressed, and the output is insensitive to the input. For example, the suspension of an automobile should absorb the forces due to road irregularities; and the control system of a heating system should suppress the disturbances due to changes in the ambient temperature. For a "good" disturbance rejection, one should have

- small steady state gain,
- small overshoot,
- small adverse response.

In Section 8.5 we study the step response of second-order systems in detail. The reader is referred to that section for examples.

8.4 Frequency-Domain Response Characteristics

In Section 8.2 we have seen that a system of the type (8.2) or (8.3) transforms sinusoidal inputs into sinusoidal outputs. Thus the input $u(t) = Ae^{i\omega t}$ is transformed into the output $y(t) = G(i\omega)Ae^{i\omega t}$, where the frequency response $G(i\omega)$ equals $\int_{-\infty}^{+\infty} H(t)e^{-i\omega t}dt$ in the case (8.2), (assuming $\int_{-\infty}^{+\infty} \mid H(t) \mid dt < \infty$), and $G(i\omega) = P^{-1}(i\omega)Q(i\omega)$ in the case (8.3). From the response to elementary complex exponentials (= trigonometric functions) we

can then derive the response to arbitrary inputs. The important observation in this is that there is no interference between the different frequencies: if the input is a sum of trigonometric functions, then the output is the sum of the corresponding outputs at the different frequencies.

If we concentrate on real-valued trajectories, we see that the sinusoidal input with frequency $\frac{\omega}{2\pi}$, amplitude $A \geq 0$, and phase $\varphi \in \mathbb{R} : A\sin(\omega t + \varphi)$ is transformed into the output $|G(i\omega)|A\sin(\omega t + \varphi + \ \mathrm{Arg}G(i\omega))$, where $G(i\omega) = |G(i\omega)|e^{i\mathrm{Arg}G(i\omega)}$. This output is a sinusoid with the same frequency $\frac{\omega}{2\pi}$ as the input, but with amplitude $|G(i\omega)|A$ and phase $\varphi + \mathrm{Arg}G(i\omega)$: the amplitude is multiplied by the modulus of $G(i\omega)$, and the phase is shifted by an amount that equals the argument of $G(i\omega)$. Thus we see the importance of the modulus and the phase of the frequency-response function. There is, however, a small problem of definition here. In general, the argument of a complex number is assumed to lie in $[0, 2\pi)$ or, better, in $\mathbb{R}(\mathrm{mod}\ 2\pi)$. However, in the case of the frequency response it often makes good sense to talk about negative phase shifts and let the phase shift continue past 2π or -2π. We therefore use the following definition of gain and phase of a transfer function.

Definition 8.4.1 Assume that the transfer function $G : \mathbb{R} \to \mathbb{C}$ is continuous with $G(i\omega) = \bar{G}(-i\omega)$, where $^-$ denotes complex conjugate, and assume that $G(i\omega) \neq 0$ for all $\omega \in \mathbb{R}$. Then the *gain* is defined as $A : \mathbb{R} \to \mathbb{R}_+$ with $A(\omega) = |G(i\omega)|$ and the *phase* is defined as the *continuous* function $\phi : \mathbb{R} \to \mathbb{R}$ such that $G(i\omega) = A(\omega)e^{i\phi(\omega)}$ with

$$\phi(0) = \begin{cases} 0 & if \quad G(0) > 0, \\ -\pi & if \quad G(0) < 0. \end{cases}$$

If $G : \mathbb{R} \to \mathbb{C}$ has a zero on the imaginary axis, say $G(i\omega_0) = 0,\ \omega_0 > 0$, then define $A(\omega_0) = 0$ and $\phi(\omega_0^+) = \phi(\omega_0^-) + k\pi$, where k is the multiplicity of the zero. If it has a pole at $\omega_0 > 0$, then define $A(\omega_0) = +\infty$ and $\phi(\omega_0^-) = \phi(\omega_0^-) - k\pi$ where k is the multiplicity of the pole. If $G(s)$ has a zero at $s = 0$, factor out the zero, $G(s) = s^k G_1(s)$, such that G_1 has no zeros at $s = 0$, and take $A(0) = 0$ and $\phi_G(0) = \phi_{G_1}(0) + k\pi$. If $G(s)$ has a pole at $s = 0$, factor out the pole $G(s) = \frac{1}{s^k}G_1(s)$ with G_1 such that it has no poles at $s = 0$, take $A(0) = \infty$ and $\phi_G(0) = \phi_{G_1}(0) - k\pi$. □

Example 8.4.2 The phase of $\frac{1}{s^2+\omega_0^2}$ is given by $\phi(\omega) = -\pi$ for $|\omega| < \omega_0$ and $\phi(\omega) = 0$ for $|\omega| > \omega_0$. The input $\sin\omega t$ with $\omega \neq \omega_0$ is transformed into $\frac{1}{\omega^2-\omega_0^2}\sin\omega t$. The fact that for $|\omega| < \omega_0$ the sinusoid at the output has opposite sign of the input sinusoid is reflected by the phase shift $-\pi$ for $|\omega| < \omega_0$. □

8.4.1 The Bode plot

There are several ways of graphically representing the frequency response of a system or, more generally, the Fourier transform of a signal. In the *Bode plot* the gain A and the phase ϕ are plotted for $\omega > 0$ in two graphs. For the gain axis a logarithmic scale is used; for the phase axis, a linear scale is used; and for the frequency[1] axis a logarithmic scale. As units for the magnitude A, $20 \log$ is used (a unit is called a *decibel*, abbreviated dB); for the phase ϕ, degrees are used; and for the frequency ω the unit used is a 10-fold (a unit is called a *decade*). Sometimes for ω the unit used is a 2-fold (in which case one calls it an *octave*: don't be surprised—there are 8 full tones between a frequency and its double). As the origin of the ω-axis, a characteristic frequency ω_0, for example the resonant frequency, is taken. A typical Bode plot is shown in Figure 8.4. Note on the dB scale that 6dB "up" means "doubling," 6dB "down" means "halving." Decibel is used for $20 \log$ (while "deci" should remind one of 10, not 20), since $10 \log A^2 = 20 \log A$, and for signals, $|\hat{f}(i\omega)|^2$ often signifies the energy in a signal at frequency ω. So in the future, when you want to ask your friends to keep the noise down, ask them, *"6dB down, please,"* so that they will know that you have become a nerd.

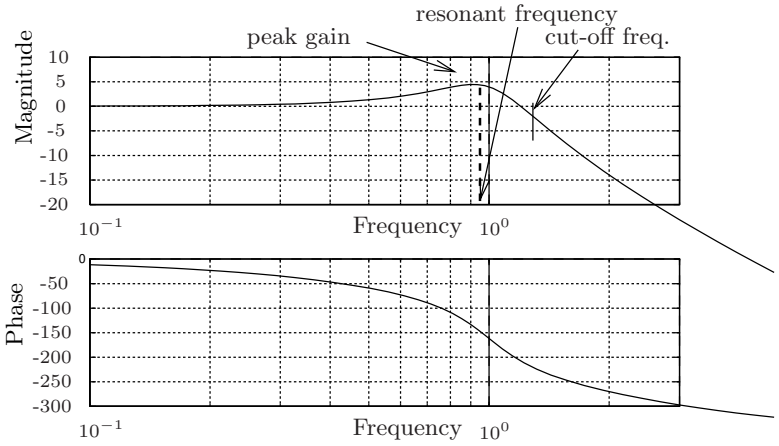


FIGURE 8.4. Bode plot of $(1 - 0.5s)/(1 + s)(1 + 0.5s + s^2)$.

We now describe a number of characteristic features of the frequency response that can be easily deduced from the Bode plot.

---

[1] We do not distinguish the variable ω (radians/unit time) and $f = \frac{\omega}{2\pi}$ (cycles/unit time). Thus, for example, the resonance frequency ω_0 is taken as being expressed in radians per second. In order to get it in cycles/second, divide by 2π.

- The *peak gain* is defined as

$$\max_{\omega} A(\omega) =: A_{\max}.$$

  The peak relative gain is defined as $\frac{A_{\max}}{A(0)}$. The peak gain is a measure of the degree to which signals are transmitted through a system, as opposed to being cut off by it.

- The *peak*, or *resonant, frequency* ω_r is defined as the frequency such that $A(\omega_r) = A_{\max}$. The peak frequency is the frequency at which a sinusoidal signal passes most easily through the system.

- The (6 dB) *pass-band* is defined as the set of ωs such that $A(\omega) \geq \frac{1}{2}A_{\max}$. Often, this set is an interval, $[\omega_{\min}, \omega_{\max}]$. The pass-band is the interval of frequencies that, relatively speaking, dominate the output. Sometimes it is a family of intervals. In that case, there are many pass-bands. If $\omega_{\min} = 0$, we speak about a *low-frequency* signal or a *low-pass* filter; if $\omega_{\max} = \infty$, of a *high frequency* signal or a *high-pass* filter; if $\omega_{\min} > 0$ and $\omega_{\max} < \infty$, of a *band-pass* signal or filter. The interval $\omega_{\max} - \omega_{\min}$ is called the *bandwidth*.

- The frequencies such that

$$\frac{A(\omega)}{A_{\max}} = \frac{1}{2}$$

  are called the *cut-off frequencies* (for example, $\omega_{\min}$ and $\omega_{\max}$).

- The rate at which $A(\omega) \to 0$ as $\omega \to \infty$ can be measured by

$$\lim_{\omega \to \infty} 20 \log \frac{A(10\omega)}{A(\omega)} = \alpha.$$

  We call α the *high-frequency roll-off* and say that A rolls off at α dB per decade. In particular, it is easily verified that thus $\frac{1}{s^n}$ rolls off at the rate of $n \times 20$ dB/decade or $n \times 6$ dB/octave.

8.4.2 The Nyquist plot

A second way of representing graphically the frequency response is by means of the *Nyquist plot*. This is a graph in the complex plane consisting of the set $\{G(i\omega) | \omega \in [0, \infty)\}$. Usually, this graph also displays the parametrization by ω. A typical Nyquist plot is shown in Figure 8.5.

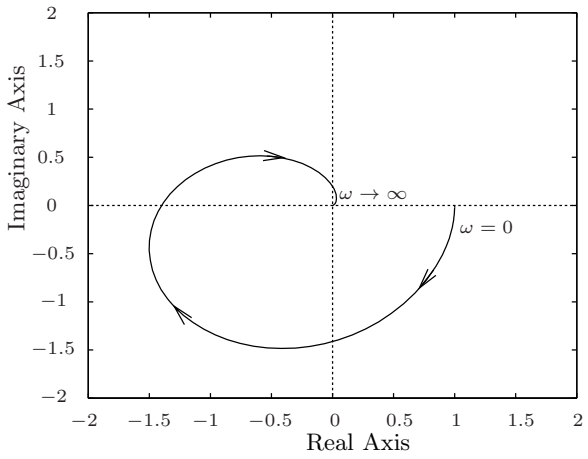


FIGURE 8.5. Nyquist plot of $(1 - 0.5s)/(1 + s)(1 + 0.5s + s^2)$.

8.5 First- and Second-Order Systems

8.5.1 First-order systems

In the previous sections we have introduced a number of characteristic features related to the time- and frequency-domain description of systems. In this section, we study first- and second-order systems, and relate these characteristics to the system parameters.

Consider the system described by

$$\alpha y + \frac{d}{dt} y = \beta u.$$

The timeconstant of this system is $\frac{1}{\alpha}$, and the steady-state gain equals $\frac{\beta}{\alpha}$ (assuming $\alpha > 0$). Its impulse response is $\beta e^{-\alpha t}$, its step response is $\frac{\beta}{\alpha}(1 - e^{-\alpha t})$, and its transfer function is $\frac{\beta}{s+\alpha}$. This system is asymptotically stable if and only if $\alpha > 0$, stable if and only if $\alpha \geq 0$, and unstable if $\alpha < 0$. The impulse response, the step response, and the Bode and Nyquist plots of a first-order system are shown in Figure 8.6. In these plots, we used the normalized variables $y' = \alpha/\beta y, t' = t/\alpha, \omega' = \omega/\alpha$.

8.5.2 Second-order systems

Consider the second-order system

$$p_0 y + p_1 \frac{d}{dt} y + p_2 \frac{d^2}{dt^2} y = q_0 u. \tag{8.10}$$

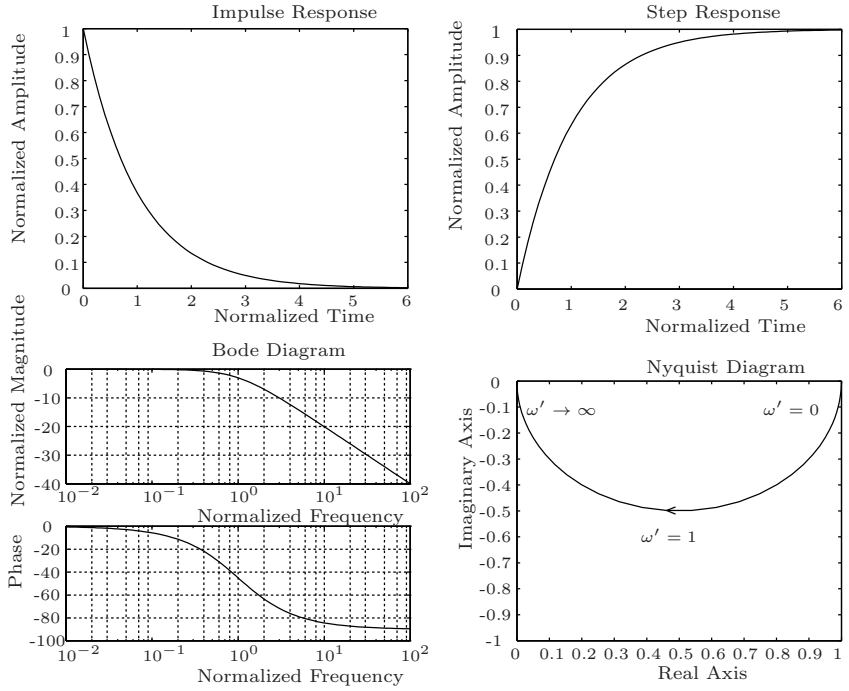


FIGURE 8.6. The response of a first-order system.

Assume $p_0, p_2 > 0$, $p_1 \geq 0$, and $q_0 \neq 0$. Note that the system is thus assumed to be stable, and asymptotically stable if $p_1 > 0$. The steady-state gain of this system is given by $s_\infty = \frac{q_0}{p_0}$. In order to analyze (8.10), we will reduce the number of parameters. This can be achieved by choosing convenient units for the output and for the time axis. Thus we renormalize y and t by using appropriate scales. For the y scale, choose $s_\infty = \frac{q_0}{p_0}$ as the unit, and for the time scale, choose $\sqrt{\frac{p_2}{p_0}}$ as the unit. The system equation in terms of the normalized variables

$$y' = \frac{y}{q_0/p_0} \text{ and } t' = \frac{t}{\sqrt{\frac{p_2}{p_0}}}$$

becomes

$$y' + 2\zeta \frac{d}{dt'} y' + \frac{d^2}{dt'^2} y' = u, \tag{8.11}$$

where $\zeta := \frac{1}{2} \frac{p_1}{\sqrt{p_0 p_2}}$. The time $\sqrt{\frac{p_2}{p_0}}$ is called the *characteristic time*, $\omega_0 = \sqrt{\frac{p_0}{p_2}}$ the *characteristic frequency*, and the coefficient ζ the *damping coefficient* of the system.

Note that the choice of the scales has reduced the number of parameters in (8.11 to one, the damping coefficient ζ. For $0 < \zeta < 1$, the roots of $p(\xi) = 1 + 2\zeta\xi + \xi^2$ are complex, and for $\zeta \geq 1$ both roots are real. The undriven system ($u = 0$) has thus a (damped) oscillatory response for $0 < \zeta < 1$; the system is called *underdamped*. For $\zeta > 1$ the undriven response consists of the sum of two real exponentials; the system is said to be *overdamped*. When $\zeta = 1$ the system is called *critically damped*; the zero input response is of the form $(a+bt)e^{-t}$ in that case. For $\zeta = 0$ the response consists of a periodic response. The system has no damping in this case.

Figure 8.7 shows the response of (8.11) as a function of ζ. Here, we have used the normalized variables t', y' and the normalized frequency $\omega' = \omega/\omega_0$. Note that as ζ becomes large (which corresponds to high damping), the system becomes more and more sluggish, while as $\zeta \to 0$ (which corresponds to low damping), the system exhibits large overshoot and oscillatory behavior. The response of this second-order system, viewed as a tracking servo, i.e., when we desire y to follow u, is nice for ζ between 0.7 and 1, whence good tracking requires a reasonable, but not an excessive, amount of damping. Thus a damper should be properly tuned in order to function well. Systems with too much damping respond very slowly to external inputs, which implies that correction or command inputs will be slow to have effect, and that disturbance inputs are noticed (when it is too) late. These are clearly undesirable features of systems. On the other hand, systems with too little damping will overreact to commands and will amplify disturbances. This is obviously also an undesirable situation. This fact, that systems should be finely tuned in order to work well, is one of the reasons that control is an important subject.

The required tuning and trade-off between too much and too little damping can be observed already in very common low-tech devices. For example, in dampers of automobiles, or in door-closing mechanisms. In these devices, too much and too little damping are both undesirable. A door closing mechanism with too much damping causes the door to bang, too little damping causes it to close too slowly. A car with too little damping will cause uncomfortable overshoots at bumps, etc.

The Bode and Nyquist plots of the system (8.11) are also shown in Figure 8.7. The system has a low-pass characteristic for $\zeta > 1$, but it obtains the character of a band-pass system as $\zeta \to 0, 0 < \zeta < 1$. For ζ close to zero, the system resonates, and the output will be dominated by the frequency content of the input signal around $\omega = \omega_0$. In particular, if a lightly damped system will be excited by an input that contains energy in the frequency-band around ω_0, then the output will be strongly amplified. In many situations, this is undesirable and often dangerous.

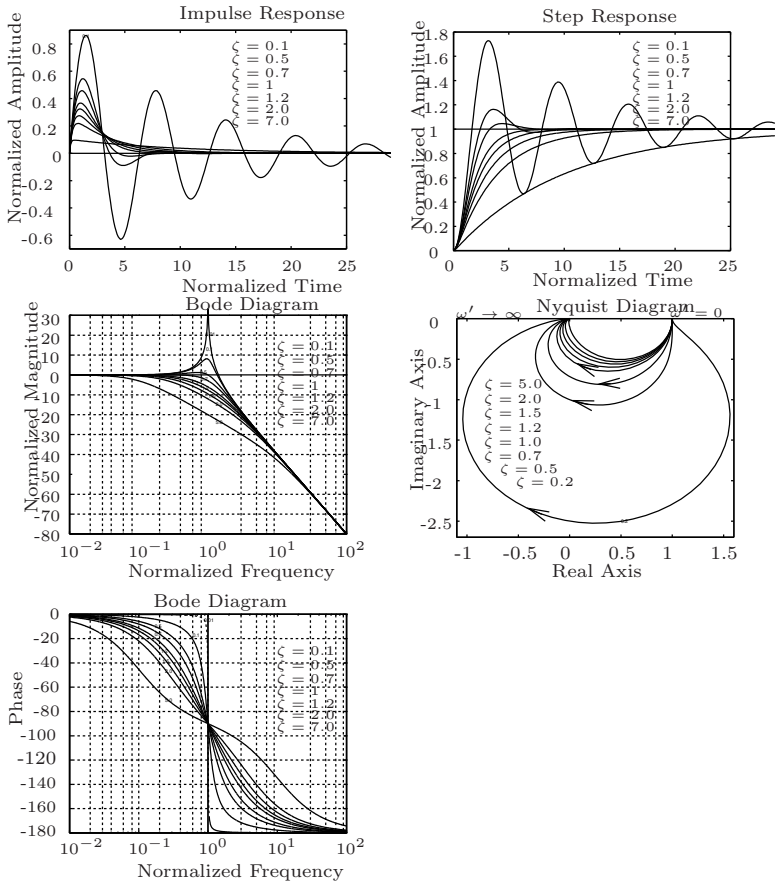


FIGURE 8.7. Responses of second-order systems.

8.6 Rational Transfer Functions

Let $G(s)$ be a transfer function. If it is of the form $G(s) = P^{-1}(s)Q(s)$ with $P(\xi), Q(\xi) \in \mathbb{R}[\xi]$, then it is called a *rational* transfer function. Such transfer functions occur very frequently in applications. As we have seen, transfer function obtained from systems of differential equations (8.1) via the input/output representation (8.3) are rational. In Chapter 6, we have seen that systems described by state space equations lead to differential systems for the their external behaviors. Conversely, we have also seen in Section 6.4 that i/o systems can be represented in state space form. Thus we expect that state space systems also lead to rational transfer functions. We show that this is indeed the case. For simplicity, we consider only the single-input/single-output case.

In this section we study the characteristics of rational transfer functions. We shall see that the Bode plot for such systems can be readily obtained by combining Bode plots of first- and second-order systems.

8.6.1 Pole/zero diagram

Definition 8.6.1 Consider the system (8.3) with $m = p = 1$. Then the roots of $P(\xi)$ are called the *poles* of this system, and the roots of $Q(\xi)$ are called the *zeros*. Thus poles and zeros are complex numbers with multiplicities. Let $G(\xi) = P^{-1}(\xi)Q(\xi)$ be the associated transfer function. If a pole coincides with a zero, then we say that there is a *cancellation*. □

From the theory developed in Chapter 2, it follows that the poles determine the dynamics when the input is zero,

$$P(\frac{d}{dt})y = 0,$$

while the zeros determine the dynamics when the output is zero,

$$Q(\frac{d}{dt})u = 0.$$

In the *pole/zero diagram* the poles and zeros are marked in the complex plane. A pole is marked as a cross, $\times$, a zero as a circle, $\circ$. Multiple poles and zeros are marked as $\divideontimes$ and $\circledcirc$, etc. Note that no zeros and poles coincide if and only if the system is controllable. Thus (8.3) is controllable if it has no poles and zeros in common.

Example 8.6.2 Consider the system

$$y + 2\frac{d}{dt}y + 2\frac{d^2}{dt^2}y + \frac{d^3}{dt^3}y = 4u - 3\frac{d^2}{dt^2}u - \frac{d^3}{dt^3}u.$$

The corresponding polynomials are $p(\xi) = (1+\xi)(1+\xi+\xi^2)$ and $q(\xi) = -(-1+\xi)(2+\xi)^2$. The pole/zero diagram is shown in Figure 8.8. □

8.6.2 The transfer function of i/s/o representations

In this subsection we derive the expression of the transfer function of a system in i/s/o form. Consider the system studied extensively in Chapter 4,

$$\begin{aligned} \tfrac{d}{dt}x &= Ax + Bu, \\ y &= Cx + Du. \end{aligned}$$

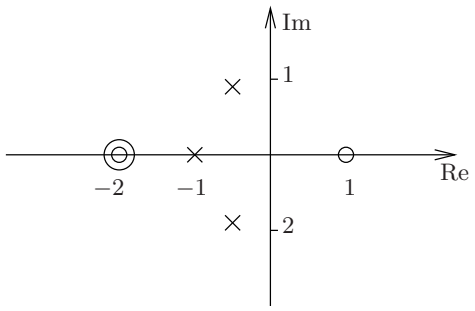


FIGURE 8.8. Pole/zero diagram of Example 8.6.2.

Assume that $s \in \mathbb{C}$ is not an eigenvalue of A. Let $u(t) = e^{st}u_s$ be an exponential input. Then obviously, $x_s = (Is - A)^{-1}Bu_s e^{st}$ is a corresponding exponential state response, and $y_s = (D + C(Is - A)^{-1}B)u_s$ is the corresponding exponential output. Hence, the transfer function is given by

$$G(s) = D + C(sI - A)^{-1}B.$$

This expression for the transfer function can also be obtained from the convolution description (4.31). Take $u(t) = e^{st}u_s$. Then assuming that $s \in \mathbb{C}$ is to the right of every eigenvalue of A, there holds

$$y(t) = \int_{-\infty}^{t} Ce^{A(t-\tau)}Be^{s\tau}u_s d\tau + De^{st}u_s = (C(sI - A)^{-1}B + D)u_s e^{st}.$$

As a consequence,

$$G(s) = D + C(sI - A)^{-1}B.$$

The transfer function of the i/s/o system is thus $D + C(Is - A)^{-1}B$. In particular, it is a matrix of rational functions. Combining this result with the theory of Section 6.4 shows that the following families of systems are equivalent in the sense that to each system in one family there corresponds one (but, in general, more than one) in the other class:

- Systems described by linear constant-coefficient differential equations,
- Finite-dimensional i/s/o systems,
- Systems with rational transfer functions.

8.6.3 The Bode plot of rational transfer functions

Let $G(\xi) = \frac{q(\xi)}{p(\xi)}$ be a rational function. Now write these numerator and denominator polynomials in their elementary factors to obtain

$$G(s) = Ks^r \frac{\prod_i (1 + \frac{s}{z'_i}) \prod_j (1 + 2\zeta'_j(\frac{s}{\omega'_j}) + (\frac{s}{\omega'_j})^2)}{\prod_k (1 + \frac{s}{z''_k}) \prod_\ell (1 + 2\zeta''_\ell(\frac{s}{\omega''_\ell}) + (\frac{s}{\omega''_\ell})^2)}, \tag{8.12}$$

with $r \in \mathbb{Z}, \mid z'_k \mid \neq 0, \mid z''_k \mid \neq 0, \omega'_j > 0, \omega''_\ell > 0, \mid \zeta'_j \mid < 1$, and $\mid \zeta''_\ell \mid < 1$. The factor s^r in this expression corresponds to the poles or zeros at the origin, the $-z'_i$s to the real zeros, and the $-z''_k$s to the real poles. The complex zeros are given by $\omega'_j(-\zeta'_j \pm i\sqrt{1 - (\zeta'_j)^2})$, and the complex poles by $\omega''_\ell(-\zeta''_\ell \pm i\sqrt{1 - (\zeta''_\ell)^2})$. The factor K in front of the expression (8.12) equals the steady-state gain (in the case of asymptotic stability). The expression 8.12 shows that the rational function $G(s)$ is the product of a finite number of elementary first-, zero-th, and second-order factors of the form

$$K, s^{\pm 1}, (1 + \frac{s}{z})^{\pm 1}, (1 + 2\zeta\frac{s}{\omega} + (\frac{s}{\omega})^2)^{\pm 1}.$$

Because of the scales chosen, the Bode plot of the product $G_1(i\omega)G_2(i\omega)$ can be obtained by adding both the magnitude and the phase graphs of the Bode plots of $G_1(i\omega)$ and $G_2(i\omega)$. Similarly, the Bode plot of $G^{-1}(i\omega)$ can be obtained from the Bode plot of $G(i\omega)$ by taking the negative of both the magnitude and the phase graphs of the Bode plot of $G(i\omega)$. Consequently, the Bode plot of $G(s)$ can be obtained by adding and subtracting the magnitude and phase graphs of elementary factors such as

$$K, s, 1 + \frac{s}{z}, 1 + 2\zeta\frac{s}{\omega} + (\frac{s}{\omega})^2.$$

Moreover, the Bode plot of the elementary factor $1 + \frac{s}{z}$ can be obtained by scaling the frequency axis from the Bode plot of $1 + s$ or $1 - s$, depending on whether $z > 0$ or $z < 0$. However, $1 + s$ and $1 - s$ have the same magnitude plot but the opposite phase. Similarly, the Bode plot of $1 + 2\zeta\frac{s}{\omega} + (\frac{s}{\omega})^2$ can be obtained from the Bode plot of $1 + 2\zeta s + s^2$ by scaling the frequency axis. In addition, $1 + 2\zeta s + s^2$ and $1 - 2\zeta s + s^2$ have the same magnitude plot, but opposite phase. It follows from all this that the Bode plot of $G(s)$ can be obtained by adding, subtracting, and the frequency scaling of the Bode plots of the simple zero-th, first-, and second-order systems $K, s, 1 + s, 1 + 2\zeta s + s^2$, with $0 \leq \zeta \leq 1$.

We can use these ideas in order to obtain very quickly a rough idea of the magnitude part of the Bode plot of a system with rational transfer function $G(s)$. Note, at least when $\mid \zeta \mid$ is not too small, that the magnitude parts of $1 + \frac{s}{z}$ and $1 + 2\zeta\frac{s}{\omega} + (\frac{s}{\omega})^2$ are reasonably well approximated by the straight line curves shown in Figure 8.9.

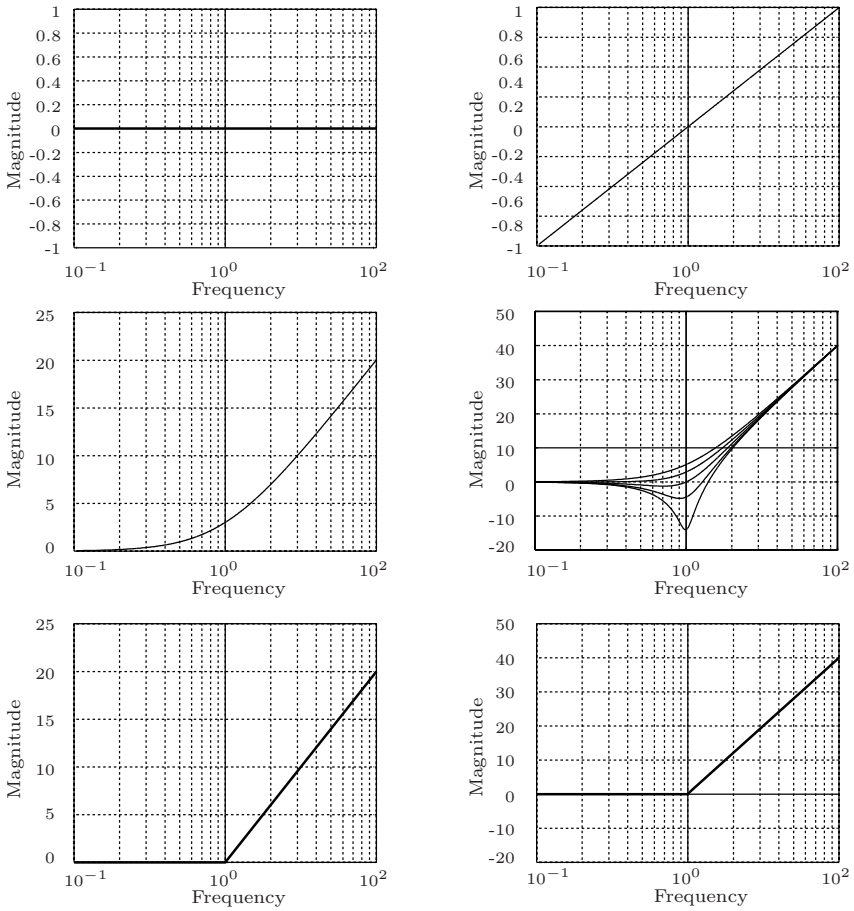


FIGURE 8.9. Bode plots of $1, s, 1+s$, and $1+2\zeta s+s^2$, and approximate Bode plots of $1+s$ and $1+2\zeta s+s^2$.

The fact that the Bode plots of the elementary factors are simple leads to the following procedure for quickly sketching the approximate magnitude part of the Bode plot of the rational transfer function $G(s)$, once it has been expressed in its elementary factors as in (8.12). Mark first on the frequency axis the points $\mid z_i \mid, \omega_j, \mid z_k \mid$, and ω_ℓ. These frequencies are called the *breakpoints* of the Bode plot. The sketch is now obtained as follows. The approximate magnitude plot is a continuous, piecewise linear graph. Between each of the breakpoints, the magnitude plot is a straight line segment. At each of the breakpoints, the slope of this straight line is modified. To the far left, the slope is 20 dB/decade. If the breakpoint corresponds to a real zero (the z_is), 20 dB/decade is added to the slope; for a complex zero (the ω_js), 40 dB/decade is added; for a real pole (the z_ks),

20 dB/decade is subtracted; and for a complex pole (the ω_ℓs), 40 dB/decade is subtracted. Finally, the origin of the magnitude scale is chosen such that for $\omega = 1$ the magnitude is $20 \log |K|$.

There exist similar techniques for sketching the phase part of the Bode plot of rational transfer functions, but we will not give details.

It should be mentioned, however, that the resulting approximations may be poor, particularly in the neighborhood of the breakpoints. In particular, if some of the poles or zeros have small damping coefficients, then the approximations will be very inaccurate around the corresponding breakpoints. Thus, in particular, when analyzing the frequency response of lightly damped mechanical systems with many oscillatory modes, these approximations should not be used.

These approximations allows one, with a bit of routine, to obtain very quickly a sketch of the magnitude (and phase) part of the Bode plot of a system whose poles and zeros have been determined. It is fair to add, however, that this sort of expertise is rapidly becoming quaint, since computer packages draw Bode plots much faster and much more accurately than any grey-haired mature expert in control.

Example 8.6.3 Consider the transfer function

$$\frac{(1 + 0.5s + s^2)}{s(1 + 10s + 100s^2)(1 + 0.14s + 0.01s^2)} \ . \tag{8.13}$$

The breakpoints corresponding to zeros are at 1; those corresponding to the poles are at 0.1 and at 10. The sketch of the Bode plot and the actual Bode plot are shown in Figure 8.10. □

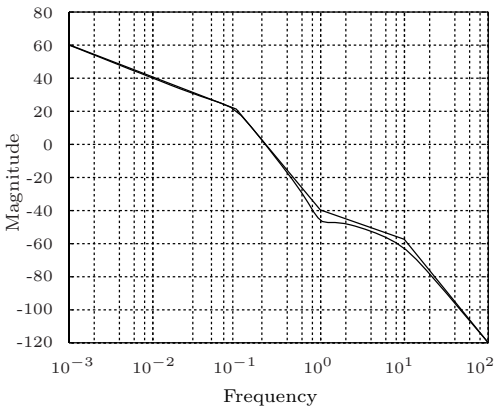


FIGURE 8.10. Bode plot and approximate Bode plot of (8.13).

8.7 Recapitulation

In this chapter we described some characteristic time and frequency-domain features of linear time-invariant systems. The main points are the following:

- A linear time-invariant system processes periodic (more generally, exponential) components of signals individually, without interference among the different frequencies. This leads to the notions of transfer function and frequency response. The transfer function expresses how exponential inputs are transformed into exponential outputs (Section 8.2).
- The transfer function specifies only the controllable part of a system, and uncontrollable modes are not represented by it. This limits the potential of transfer function methods for the description of systems (Theorem 8.2.7).
- The step response is the response of a system to a step input. Many useful time domain characteristics (such as overshoot, settling time, rise time) of a system can be read of from its step response (Section 8.3).
- A useful way of representing the frequency response of a system is by its Bode plot. Many useful frequency-domain characteristics (such as bandwidth, resonant frequencies) can be read off from the Bode plot (Section 8.4.1).
- The parameters of a first-order system have an immediate interpretation in terms of the timeconstant and the steady-state gain. For second-order systems, the characteristic frequency and the damping coefficient are the important parameters (Section 8.5).
- Rational transfer functions occur very frequently in applications. Their characteristic features can readily be deduced by their pole/zero diagram. The poles and zeros specify the breakpoints of the Bode plot. The Bode plot can readily be sketched from the steady-state gain and the pole/zero diagram (Section 8.6).

8.8 Notes and References

The material covered in this chapter forms bread and butter mathematical techniques underlying classical control theory. In this text, we cover but some essential features of this very useful theory. There are numerous textbooks, mostly with an engineering emphasis, covering these topics. The underlying mathematics is that of Fourier and Laplace transforms. The result of Theorem 8.2.7, implying that the transfer function specifies only the controllable part of a system, appeared in [59]. On the level of generality presented here, this result was obtained in the above reference for the first time. However, for state space systems, a similar result has been known for some time (see, for example, [27] and [15]).

8.9 Exercises

As a simulation exercise illustrating the material covered in this chapter we suggest A.5.

8.1 Assume that $w : \mathbb{R} \to \mathbb{R}^q$ is periodic with period T. Define $f : [0,T] \to \mathbb{R}^q$ by $f(t) = w(t)$ for $0 \leq t \leq T$. Assume that $f \in \mathfrak{L}_1([0,T],\mathbb{R}^q)$. Let $\{\hat{f}_n, n \in \mathbb{Z}\}$ denote the Fourier series of f. Consider (8.1). Prove that f belongs to the behavior of this system if and only if $\hat{f}_n \in \mathfrak{E}(i\frac{2\pi n}{T})$ for all $n \in \mathbb{Z}$. Obtain an analogous result for input/output systems (8.3) applied to periodic inputs.

8.2 Let $G(\xi) \in \mathbb{R}(\xi)$. Assume for simplicity that $G(\xi)$ is strictly proper. Expand $G(\xi)$ in partial fractions, yielding an expression of the form

$$G(\xi) = \sum_{k=1}^{N} \sum_{\ell=1}^{n_k} \frac{a_{k\ell}}{(\xi - \lambda_k)^\ell},$$

with $\lambda_1, \lambda_2, \ldots, \lambda_N$ the poles of $G(\xi)$, and $n_1, n_2, \ldots, n_N$ their multiplicities. Consider the associated impulse response

$$h(t) = \begin{cases} \sum\limits_{k=1}^{N} \sum\limits_{\ell=1}^{n_k} a_{k\ell} t^{\ell-1} e^{\lambda_k t} & t \geq 0, \\ 0 & t \leq 0. \end{cases}$$

Prove that $G(s)$ is the Laplace transform of h and that its domain of convergence equals $\{s \in \mathbb{C} \mid \mathrm{Re}(s) > \mathrm{Re}(\lambda_k), k = 1, 2, \cdots, N\}$. Use this result to prove in the multivariable case that the transfer function of the initially-at-rest system (8.3) viewed as a convolution system equals $G(s) = P^{-1}(s)Q(s)$. Specify the domain of convergence of this transfer function. Prove that $P^{-1}(s)Q(s)$ is the transfer function of the initially-at-rest system (8.3) by considering the Laplace transform of $P(\frac{d}{dt})u$ and $Q(\frac{d}{dt})y$ in terms of the Laplace transforms of u and y, and considering (8.3).

8.3 Let $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathfrak{B})$ be a linear time-invariant system (not necessarily described by a differential equation). Prove that for each $\lambda \in \mathbb{C}$ the set $\{b \in \mathbb{C}^q \mid b \exp_\lambda \in \mathfrak{B}\}$ defines a linear subspace of $\mathbb{C}^q$.

8.4 Let $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$, $b \in \mathbb{C}^q$, and $s \in \mathbb{C}$. Prove that $R(\frac{d}{dt})b \exp_s = R(s)b \exp_s$. Deduce Lemma 8.2.4 from this.

8.5 Let $\mathfrak{E}(s)$ be the exponential behavior of (8.3), as obtained in (8.7). Prove that (8.1) defines a controllable system if and only if the dimension of $\mathfrak{E}(s)$ is independent of s for $s \in \mathbb{C}$.

8.6 Does the exponential response (8.7) determine the behavior (8.1) uniquely? If not, give a counterexample. Does it, if it is known that the system is controllable?

8.7 Give examples of controllable systems $G_1(s), G_2(s)$ but for which the series interconnection, the parallel interconnection, or the feedback interconnection is not controllable. Comment on the limited validity of the transfer function as describing the behavior of an interconnected system.

8.8 Plot the step response of the following systems:

(a) $y + \frac{d^2}{dt^2}y = u$.

(b) $y - \frac{d^2}{dt^2}y = u$.

(c) $y + \frac{d}{dt}y = u - \frac{d}{dt}u$.

(d) $y + \frac{d}{dt}y + 4\frac{d^2}{dt^2}y = u - \frac{d}{dt}u$.

8.9 Compute the steady state gain of the system with transfer function

$$\frac{\beta s + \alpha}{s^3 + 3s^2 + 3s + 1}.$$

8.10 Consider the single-input/single-output system

$$y(t) = \int_{-\infty}^{t} H(t - t')u(t')dt'.$$

Assume that $H(t) \geq 0$ for $t \geq 0$ and $\int_0^\infty H(t)dt < \infty$. Prove that this system has no overshoot. Give a "formula" for the 5% settling time, the rise time, and the deadtime.

8.11 Give some real-life verbal examples of systems with an adverse response.

8.12 Sketch the Bode and the Nyquist plots of

$$\frac{s + 1}{(s + 2)(s + 3)(s + 4)}$$

and

$$\frac{s^2 + 0.5s + 1}{s(s^2 + s + 1)}.$$

8.13 Consider the electrical circuit of Example 3.3.27 Take the resistor and the capacitor values equal to one. Sketch the step response and the Bode plot of the transfer function from V to I. Repeat for the the transfer function from V_{in} to V_{out}. Are these systems low-pass, band-pass, or high-pass filters?

8.14 Estimate the peak frequency, the pass-band, and the bandwidth of the system with transfer function

$$\frac{1}{(s^2 + 0.2s + 1)(s^2 + s + 1)}.$$

8.15 Consider the system described by $y(t) = u(t - T), T > 0$. This is a pure delay. Plot its step response, and its Bode and Nyquist plots.

8.16 Consider the system with behavioral equation

$$P(\frac{d}{dt})y = Q(\frac{d}{dt})\tilde{u}; \quad \tilde{u} = \Delta u,$$

with Δ the delay operator: $(\Delta u)(t) = u(t-T)$. Compare its step response to that of $P(\frac{d}{dt})y = Q(\frac{d}{dt})u$. Same question for the Bode plot.

8.17 Compute the impulse and the step responses of the series, parallel, and feedback interconnections of two single-input/single-output systems in terms of the impulse responses of the interconnected systems.

8.18 Consider the first-order system

$$\alpha y + \frac{d}{dt}y = \beta u + \frac{d}{dt}u.$$

Assume that $\alpha > 0$. Draw the step response, using as time scale $\frac{t}{\alpha}$, for a range of positive and negative values of β. Draw the Bode plot. Discuss the filtering characteristics of this system.

8.19 Consider the system

$$y + \frac{d}{dt}y + \frac{d^2}{dt^2}y = \beta u + \frac{d}{dt}u.$$

Sketch the step response for a range of positive and negative values of β.

8.20 Consider the transfer functions of Exercise 8.12. Sketch the magnitude of the Bode plot using the technique explained in Section 8.6.3. Plot the exact magnitude plot (using, for example, a computer package such as MATLAB©). Comment on the difference with the approximation.

8.21 Consider the single-input/single-output system (8.3). Assume that $P(\xi)$ is Hurwitz. Prove that if this system has an adverse response, then $Q(\xi)$ cannot be Hurwitz. Relate this to the minimum phase property discussed in Exercise 8.23. Prove that if $P(\xi)$ and $Q(\xi)$ are both first-order polynomials and if $P(\xi)$ is Hurwitz, then the system has no adverse response if and only if $Q(\xi)$ is also Hurwitz.

8.22 Two transfer functions $G_1(s)$ and $G_2(s)$ are said to be *gain equivalent* if $|G_1(i\omega)| = |G_2(i\omega)|$ for all $\omega \in \mathbb{R}$. Consider the transfer functions $G_1(s) = K_1P_1^{-1}(s)Q_1(s)$ and $G_2(s) = K_2P_2^{-1}(s)Q_2(s)$ with $P_i(\xi)$,$Q_i(\xi)$ monic polynomials, $(P_i(\xi), Q_i(\xi))$ coprime, and $K_i \neq 0$, $i = 1, 2$. Prove that $G_1(s)$ is gain equivalent to $G_2(s)$ if and only if $P_1(-\xi)P_1(\xi) = P_2(-\xi)P_2(\xi)$, $Q_1(-\xi)Q_1(\xi) = Q_2(-\xi)Q_2(\xi)$, and $|K_1| = |K_2|$. Interpret these conditions in terms of the poles and zeros of $G_1(s)$ and $G_2(s)$.

8.23 An i/o system with transfer function G is said to be *minimum phase* if whenever $G'(s)$ is gain equivalent to $G(s)$, then the phase of $G(i\omega)$ is less than that of $G'(i\omega)$ for all $\omega \in \mathbb{R}$. Consider the single-input single output system $G(s) = KP^{-1}(s)Q(s)$ with $P(\xi), Q(\xi) \in \mathbb{R}[\xi]$ and monic and coprime, and $K \in (0, \infty)$. Assume, moreover, that $P(\xi)$ is Hurwitz. Prove that $G(s)$ is minimum phase if and only if $Q(\xi)$ is also Hurwitz.

9

Pole Placement by State Feedback

In this chapter we discuss an important control design question: that of choosing a control law such that the closed loop system is stable (*stabilization*) or, more generally, such that it has a certain degree of stability reflected, for example, in a requirement on the location of the closed loop poles (*pole placement*).

We consider state feedback, and in the next chapter we study output feedback. Thus, we consider linear time-invariant dynamical systems in state form described by

$$\frac{d}{dt}x = Ax + Bu, \tag{9.1}$$

where $x : \mathbb{R} \to \mathbb{R}^n$ and $u : \mathbb{R} \to \mathbb{R}^m$ denote respectively the state and the input trajectory, and where the matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ denote the parameter matrices specifying the dynamics of the system under consideration.

9.1 Open Loop and Feedback Control

Feedback is one of the central concepts from system theory. It is one of the concepts that, together with input and output, has become part of our daily vocabulary. In order to explain the underlying idea clearly, we will discuss it first in connection and in contrast to *open loop control*. Note that the dynamical system defined by (9.1) is a special case of a system such as

(2.1), but one in which the variable u is a free input and in which x is the state. Such systems have been studied in detail in Chapter 4. If we think of (9.1) as describing a physical engineering or an economic system, then we should think of the input u as being chosen by a designer, by someone who is trying to achieve a desired behavior of the state trajectory x through a judicious choice of the input trajectory u. In this context, where we think of u as a variable that can be manipulated, it is natural to call the input u the *control*. In other situations, the input could be a disturbance imposed by nature, in which case it would not be appropriate to view u as a control.

How should a designer choose the control u in order to achieve a certain task? We have to distinguish clearly between two types of control:

1. *Open loop control.*
2. *Feedback control.*

This distinction is an extremely important one in applications.

In *open loop control* one chooses u as an explicit function of time. In other words, $u : \mathbb{R} \to \mathbb{R}^m$ is designed so as to achieve a certain goal, for example to transfer the state from x_0 to x_1. In this context the terms *motion planning* or *trajectory optimization* are often appropriate. In the Russian literature, open loop control is called *program control*. This is an appropriate term: just as in the theater, where the program announces that a particular piece of music will be played at a particular time, an open loop control announces what control action will be taken at what time. In Chapter 5, in the section about controllability, we have seen that if (9.1) is state controllable, equivalently, if the pair of matrices (A, B) is controllable, then for any $x_1, x_2 \in \mathbb{R}^n$ and any $T > 0$ it is possible to choose $u \in \mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^m)$ such that u transfers the system from state x_1 at time 0 to state x_2 at time T. Inspection of the expressions (5.30, 5.31) shows that once x_1 and x_2 and the system parameters A and B are specified, an input u can be computed that drives x_1 to x_2. This is the essence of open loop control. The fact that controllability allows this transfer to be executed at all is an important starting point in (optimal) motion planning questions.

However, in this book, we are mainly interested in another type of control, referred to as *feedback control*, and we do not pursue open loop control problems. In feedback control the value of the control input is chosen not as an explicit function of time, but on the basis of an observed output. To be specific, let us consider the dynamical system (9.1) and assume that the state x is observed. Then the choice of the value of the control is based on the observed state trajectory. Thus the control law can be thought of as a map that associates with the observed state trajectory $x : \mathbb{R} \to \mathbb{R}^n$ the chosen control input $u : \mathbb{R} \to \mathbb{R}^m$. Denote this map by F. Of course, for obvious reasons, this map has to be nonanticipating, meaning that $(Fx)(t)$ depends only on the values taken on by $x(t')$ for $t' \leq t$. The map F may be

required to have other properties: linear time invariant, or without memory. In the present chapter we discuss memoryless linear time invariant control laws. We will soon explain in detail what this means.

But before we do that, we pause for a few moments in order to emphasize two important aspects of the present discussion: firstly, the distinction between open loop and feedback control and secondly, the fact that feedback control leads to implicit equations, and to situations in which the distinction between cause and effect is blurred.

Example 9.1.1 In pondering the difference between open loop control and feedback control, an example may help. Suppose you are about to climb a flight of stairs. You can decide to do this with your eyes open or with your eyes closed. In the latter case, you will take a careful look at the stairs and the railing, count the number of stairs, process in your head a motion plan, and execute it, hoping for the best. This is open loop control. In feedback control, you keep your eyes open. By observing at each moment where your feet are with respect to the stairs, where your hands are with respect to the railing, etc., you continuously plan and adjust your movements. It should be clear from this example that feedback control in general leads to superior performance. Unexpected events, small disturbances, or miscalculations due to uncertain parameters can be taken into consideration by feedback control, but not by open loop control. □

Example 9.1.2 Consider the scalar input/state system

$$\frac{d}{dt}x + ax = u, \quad a < 0. \tag{9.2}$$

Suppose that at time $t = 0$ the system is in state x_0 and that we want to choose the input in such a way that the state is transferred to the zero state as time tends to infinity. If we want to do this in an open loop fashion, we could choose u as

$$u(t) = (a-1)e^{-t}x_0. \tag{9.3}$$

It is easy to check by substituting (9.3) into (9.2) that the resulting trajectory x is given by

$$x(t) = e^{-t}x_0, \tag{9.4}$$

so that indeed $x(t)$ converges to zero asymptotically. Notice that the input depends on the initial state x_0 and the system parameter a. In practical situations, both x_0 and a will not be known with infinite accuracy, so their nominal values contain small errors, let's say that their assumed values are $a+\epsilon$ and $x_0+\delta$ instead of the real values a and x_0, so that we use instead of (9.3),

$$u(t) = (a+\epsilon-1)e^{-t}(x_0+\delta).$$

The resulting state trajectory is then

$$x(t) = e^{-t}x_0 + (\frac{\epsilon}{a-1}(x_0+\delta)+\delta)(e^{-t}-e^{-at}).$$

Since by assumption $a < 0$, we conclude that small errors in the initial state or in the system parameter may cause instability.

If, on the other hand, we take the input at time t as a function of the state at time t, then small errors need not have such disastrous consequences. Take, for example,

$$u(t) = (a-1)x(t). \tag{9.5}$$

In the idealized case where there are no errors in a and x_0 we obtain from substituting (9.5) into (9.2)

$$x(t) = e^{-t}x_0,$$

which is identical to (9.4). However, in the case that a and x_0 contain a small error, we get

$$x(t) = e^{-(1+\epsilon)t}(x_0+\delta),$$

which is still stable if $\epsilon > -1$. Of course, if x cannot be measured exactly, then the measurement of $x(t)$ also contains an error, so instead of (9.5), the input will be of the form

$$u(t) = (a+\epsilon-1)(x(t)+\gamma(t)) \quad |\gamma(t)| < \gamma$$

for some (small) γ. The response to this input can of course not be determined exactly, but an upper bound is easily obtained:

$$\begin{aligned}
|x(t)| &= |e^{(\epsilon-1)t}x(0) + \int_0^t e^{(\epsilon-1)(t-\tau)}\gamma(\tau)d\tau| \\
&\leq |e^{(\epsilon-1)t}x(0)| + \int_0^t e^{(\epsilon-1)(t-\tau)}|\gamma(\tau)|d\tau \\
&\leq |e^{(\epsilon-1)t}x(0)| + \gamma\int_0^t e^{(\epsilon-1)(t-\tau)}d\tau = |e^{(\epsilon-1)t}x(0)| + \frac{\gamma}{1-\epsilon}(1-e^{(\epsilon-1)t}).
\end{aligned} \tag{9.6}$$

The inequality (9.6) shows that if we make explicit use of the state at time t, even if there is a small error in $x(t)$, then the resulting state trajectory is still bounded by a constant that is proportional to the upper bound of the error in the measurement of the state. This is in contrast to the open loop control strategy, where we have seen that even the smallest error may destabilize the system. □

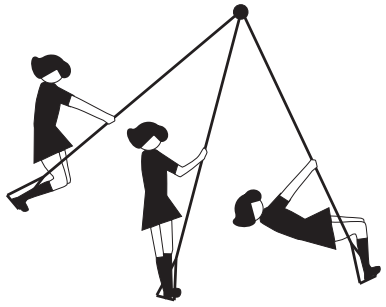


FIGURE 9.1. A child on a swing.

Example 9.1.3 A concrete example through which the difference between open loop and feedback control can be made very clear is that of a child on a swing (see Figure 9.1). By standing up when the swing moves upward and squatting when it moves downward, the child continuously switches the center of gravity of the swing up and down. In this way, it manages to pump up the amplitude of oscillation.

Let us derive the equations of motion for the swing, even though they are somewhat peripheral for the present discussion. We do this using Lagrangian mechanics. Readers not familiar with these ideas can proceed to equation (9.7), taking them for granted. Let φ denote the angle of the pendulum with respect to the vertical axis. Model the child as a point mass with mass M and denote by L the distance of the child to the pivot of the swing. Now, neglecting the mass of the swing (as compared to that of the child) yields for the kinetic energy T in terms of φ, $\dot{\varphi}$, the rate of change of φ, L, and $\dot{L}$, the rate of change of L,

$$T(\varphi, \dot{\varphi}, L, \dot{L}) = \frac{M}{2}(L^2(\dot{\varphi})^2 + (\dot{L})^2),$$

and for the potential energy

$$U(\varphi, \dot{\varphi}, L, \dot{L}) = -MgL\cos\varphi,$$

where g denotes the gravitational constant. The Euler–Lagrange equations of mechanics yield the following differential equation for the motion of φ:

$$\frac{d}{dt}\frac{\partial}{\partial\dot{\varphi}}(T-U) + \frac{\partial}{\partial\varphi}(T-U) = 0.$$

For the case at hand this yields

$$\frac{d}{dt}ML^2\frac{d}{dt}\varphi + MgL\sin\varphi = 0. \tag{9.7}$$

This equation should be viewed as the behavioral equation relating the manifest variables L and φ.

How should we explain that a child is able to pump up the motion of the swing? Mathematics texts often suggest that L will be chosen as a periodic function of t. Indeed, it can be shown that by choosing for $L : \mathbb{R} \to \mathbb{R}$ an appropriate periodic function of time, then (9.7) becomes unstable, resulting in the desired increase in amplitude of the swing. This type of instability is known as *parametric resonance*, because it requires that the amplitude of the "parameter" L be chosen in resonance with the natural frequency of the system.

Is parametric resonance really the appropriate way of explaining the way a child pumps up the motion of a swing? The answer is no. Parametric resonance implies that the length L of the swing is chosen as an explicit function of time. It suggests that the child is looking at its watch in order to decide whether to squat or to stand up at time t, or that the child moves up and down in a predetermined periodic motion. This is, of course, not what happens in reality. The choice of L is made as a function of ϕ and $\frac{d}{dt}\phi$. Typically, L will be chosen to be large if ϕ and $\frac{d}{dt}\phi$ have opposite signs and small if they have the same sign.

The parametric resonance explanation of a swing suggests that open loop control is used: the decision to squat or to stand up is taken as an explicit function of time. The explanation in which this decision depends on the observed values of ϕ and $\frac{d}{dt}\phi$ suggests feedback control. In essentially all applications of control, feedback enters in one way or another. *Feedback, not trajectory planning, is the central idea in control theory.* □

Feedback control leads to a blurring of cause and effect. In (9.1) it is logical to view the input trajectory u as the cause and the state trajectory x as the effect. However, the control law $F : x \mapsto u$ uses the state trajectory in order to decide on the control trajectory: the roles of cause and effect are now reversed. Consequently, in the controlled system there is no way of telling what causes what. A feedback system acts like a dog chasing its tail. Feedback leads to implicit equations, one of the complications characteristic of the mathematics of feedback control. For example, if the control law $u = Nx$ is used in (9.1), then the behavioral equations become

$$\frac{d}{dt}x = Ax + Bu, u = Nx. \tag{9.8}$$

Assuming that $x(0)$ is given, then the first equation in (9.8) defines x in terms of u, while the second defines u in terms of x, whence the implicit nature of the equations for x and u.

The dynamical system to be controlled (e.g., (9.1)) is usually called the *plant*. It is typically a physical, a chemical, or an economic system. The system that implements the control law is called the *feedback processor*. It processes the observations in order to obtain the control input. Nowadays, a feedback processor is often implemented in hardware as a microprocessor. In

older times, it was common to implement feedback processors by means of mechanical or pneumatic devices. The behavior of the plant in conjunction with the control law leads to the *closed loop system*. This can conveniently be illustrated by means of the signal flow graphs shown in Figure 9.2.

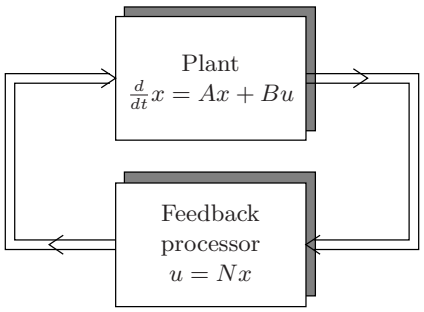


FIGURE 9.2. State feedback.

9.2 Linear State Feedback

Although in principle the control is allowed to depend also on the past of x, feedback laws are often memoryless; i.e., the control at time t depends on the state at the present time t only. A memoryless linear time-invariant linear control law for (9.1) is thus defined by

$$u = Nx, \tag{9.9}$$

where the matrix $N \in \mathbb{R}^{m\times n}$ is called the *feedback gain matrix*. The control law (9.9) functions as follows. If the value of x is observed, then the control value is specified by (9.9). The law (9.9) is called *memoryless* because the value of x at time t determines the value of u at time t. Sometimes the term *static* is used instead of memoryless. In Chapter 10 we consider feedback laws with memory: the present value of the control is then also influenced by the strict past of the measurements. Such control laws are called *dynamic*.

Now consider (9.9) in conjunction with (9.1). This leads to the equations

$$\frac{d}{dt}x = (A + BN)x \ ; \ u = Nx. \tag{9.10}$$

Equations (9.10) tell us how the state trajectory of (9.1) evolves when the control law (9.9) is applied. These equations are the *closed loop equations* for the case at hand. Observe that (9.10) defines an autonomous dynamical system.

The problem of controlling the plant (9.1) by means of a feedback processor (9.9) thus leads to a seemingly straightforward question in matrix theory. Indeed, it comes down to choosing, for given matrices (A, B) appearing in (9.1), the feedback matrix N in (9.9) such that the matrix pair $(A+BN, N)$ appearing in (9.10) has desirable properties. For example, the question may be to choose N such that all solutions x of (9.10) satisfy $x(t) \to 0$ as $t \to \infty$ (and hence $u(t) \to 0$ as $t \to \infty$)—this is the problem of *feedback stabilization*; or the question may be to choose N such that the average value of $\int_0^\infty (\|u(t)\|^2 + \|x(t)\|^2)dt$ is as small as possible, with *average* suitably interpreted. This is a problem in *optimal feedback control*, which we do not address in this book.

9.3 The Pole Placement Problem

Consider the system (9.1). We call the eigenvalues (counting multiplicity) of the matrix A the *poles* of the dynamical system governed by (9.1). Similarly, we call the eigenvalues of $A + BN$ the poles of (9.10). This nomenclature stems from considering as input to (9.1) $u = Nx + v$, with v a new input. This yields the system $\frac{d}{dt}x = (A+BN)+Bv$, which has $(I\xi - A - BN)^{-1}B$ as transfer function from v to x. If the pair (A, B) is controllable, then the poles of this matrix of rational functions are equal to the eigenvalues of $A + BN$. This explains the nomenclature, which is, strictly speaking, somewhat confusing, since we assume that the input v is absent. In order to distinguish between the poles of (9.1) and those of (9.10) we speak of the eigenvalues of A as the *open loop poles* and of those of $A + BN$ as the *closed loop poles*. Similarly, we call the characteristic polynomial of A the *open loop characteristic polynomial*, and that of $A + BN$ the *closed loop characteristic polynomial*. Of course, the open loop and closed loop poles are the roots of the corresponding characteristic polynomials.

The poles of the closed loop system are very important features for judging the behavior of the closed loop system (9.10). In Chapter 7 we have seen that the asymptotic stability of (9.10) can be decided from the location of the eigenvalues of $A+BN$ with respect to the imaginary axis. When all poles lie in the open left half plane, the system is asymptotically stable. However, it follows from Theorem 3.2.16 that much more can be concluded from the poles, as, for example, the exponential decay (or growth) of the solutions of (9.10), their frequency of oscillation, and the presence of polynomial factors in the solutions. This leads to the following (compelling) question:

What closed loop pole locations are achievable by choosing the feedback gain matrix N?

This problem is known as the *pole placement problem*. Since the closed loop poles of (9.10) are the roots of the characteristic polynomial of $A + BN$,

it is possible to reformulate the pole placement problem in linear algebra terms as follows:

Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given matrices. Choose $N \in \mathbb{R}^{m\times n}$, and let $\chi_{A+BN}(\xi)$ denote the characteristic polynomial of the matrix $A + BN$. What is the set of polynomials $\chi_{A+BN}(\xi)$ obtainable by choosing the matrix $N \in \mathbb{R}^{m\times n}$?

Of course, if $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ are the desired poles, then the desired closed loop characteristic polynomial is $r(\xi) = \prod_{k=1}^{n} (\xi - \lambda_k)$. Note that for the coefficients of $r(\xi)$ to be real, we obviously need that $(\lambda_k \in \Lambda) \Leftrightarrow (\bar{\lambda}_k \in \Lambda)$. The main result on pole placement states that the closed loop poles (equivalently the closed loop characteristic polynomial) can be chosen to be arbitrary if and only if the system (9.1) is controllable, i.e., if and only if $\text{rank}[B, AB, \ldots, A^{n-1}B] = n$. This result is proven in the next section.

Theorem 9.3.1 (Pole placement) *Consider the system (9.1). For any real monic polynomial $r(\xi)$ of degree n there exists a feedback gain matrix $N \in \mathbb{R}^{m\times n}$ such that $\chi_{A+BN}(\xi) = r(\xi)$ if and only if (9.1) is controllable.*

Recall that a polynomial is called *monic* if the coefficient of its leading term is one. In Chapter 5 we have seen the open loop interpretation of controllability in terms of the possibility of steering the state of (9.1) from any initial to any final value. Theorem 9.3.1 gives controllability a closed loop, feedback control significance in terms of the possibility of being able to choose a feedback gain matrix that achieves an arbitrary pole location.

9.4 Proof of the Pole Placement Theorem

We give a detailed proof of Theorem 9.3.1. This proof is a bit intricate, even though each of its steps is logical and straightforward. The necessity part of the proof uses the notion of system similarity. We start therefore with a section on system similarity and its relation to pole placement. This leads readily to the conclusion that controllability is a necessary condition for pole placement.

It is the sufficiency part that takes most of the work. The proof that controllability implies pole placement is structured as follows. First we consider the single-input case ($m = 1$). For such systems there is an algorithm that shows how the feedback gain should be chosen so as to achieve a desired closed loop characteristic polynomial. We subsequently turn to the multi-input case ($m > 1$). By using a clever lemma (Lemma 9.4.4), we can reduce this case to the single-input case.

9.4.1 System similarity and pole placement

Denote by $\Sigma_{n,m}$ the family of systems such as (9.1) with n state and m input variables. Thus each element of $\Sigma_{n,m}$ is parametrized by a pair of matrices (A, B) with $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Denote this by $(A, B) \in \Sigma_{n,m}$. Let $(A_1, B_1), (A_2, B_2) \in \Sigma_{n,m}$. In line with Section 4.6, we call (A_1, B_1) and (A_2, B_2) *similar* if there exists a nonsingular matrix $S \in \mathbb{R}^{n\times n}$ such that

$$A_2 = SA_1S^{-1} \quad , \quad B_2 = SB_1. \tag{9.11}$$

Note that this notion of similarity is a generalization to systems of the type (9.1) of the notion of similarity of square matrices, A_1 and A_2 being called *similar* if the first equation of (9.11) holds. If in the state space of (9.1) we change the coordinate basis by defining $z(t) = Sx(t)$, then it is clear that the dynamics of z are governed by

$$\frac{d}{dt}z = SAS^{-1}z + SBu.$$

Hence similarity as defined by (9.11) corresponds to changing the basis in the state space.

Note that if (A_1, B_1) is controllable and if (A_2, B_2) is similar to (A_1, B_1), then (A_2, B_2) is also controllable. To see this, simply compute the controllability matrix of (A_2, B_2). We obtain

$$[B_2, A_2B_2, \ldots, A_2^{n-1}B_2] = S[B_1, A_1B_1, \ldots, A_1^{n-1}B_1].$$

Since $\operatorname{rank}[B_1, A_1B_1, \ldots, A_1^{n-1}, B_1] = n$, and S is invertible, controllability of (A_2, B_2) follows.

The following lemma shows that for two similar systems the closed loop characteristic polynomials that are achievable by state feedback coincide.

Lemma 9.4.1 *Assume that $(A_1, B_1), (A_2, B_2) \in \Sigma_{n,m}$ are similar. Let $r(\xi) \in \mathbb{R}[\xi]$ be a monic polynomial. Then there exists a matrix $N_1 \in \mathbb{R}^{m\times n}$ such that $\chi_{A_1+B_1N_1}(\xi) = r(\xi)$ if and only if there exists a matrix $N_2 \in \mathbb{R}^{m\times n}$ such that $\chi_{A_2+B_2N_2}(\xi) = r(\xi)$.*

Proof Compare the effect of using the feedback matrix N_1 on (A_1, B_1) with that of using $N_2 = N_1S^{-1}$ on (A_2, B_2). The resulting closed loop system matrices are $A_1 + B_1N_1$ and $A_2 + B_2N_2 = S(A_1 + B_1N_1)S^{-1}$. Hence $A_1 + B_1N_1$ and $A_2 + B_2N_2$ are similar, and therefore they have the same characteristic polynomial. The lemma follows. □

It follows from this lemma that in order to prove pole placement for (9.1), we may as well consider a system that is similar to it.

9.4.2 Controllability is necessary for pole placement

The proof of the necessary part of Theorem 9.3.1 is based on the decomposition of systems into a controllable and a noncontrollable part (the so-called *Kalman decomposition*, see Corollary 5.2.25). The following lemma was proven already in Chapter 5 and is repeated here for easy reference, and in order to make the proof of the pole placement theorem self-contained and independent of the material in Chapter 5.

Lemma 9.4.2 *The system (9.1) is similar to a system* $(A', B') \in \Sigma_{n,m}$ *with* A', B' *of the form*

$$A' = \begin{bmatrix} A'_{11} & A'_{12} \\ 0 & A'_{22} \end{bmatrix}, \qquad B' = \begin{bmatrix} B'_1 \\ 0 \end{bmatrix} \tag{9.12}$$

and with (A'_{11}, B'_1) *controllable.*

Proof See Corollary 5.2.25. □

This lemma immediately shows that controllability is a necessary condition for pole placement. Indeed, assume that (9.1) is not controllable. Then it is similar to a system (A', B') of the form (9.12) with n_1, the dimension of A'_{11}, less than n. Now consider the effect of a feedback matrix $N' = [N'_1\ N'_2]$, with $N'_1 \in \mathbb{R}^{m\times n_1}$ and $N'_2 \in \mathbb{R}^{m\times(n-n_1)}$, on this system. The matrix $(A' + B'N')$ is given by

$$\begin{bmatrix} A'_{11} + B'_1N'_1 & A'_{12} + B'_1N'_2 \\ 0 & A'_{22} \end{bmatrix}.$$

Hence its characteristic polynomial is given by $\chi_{A'_{11}+B'_1N'_1}(\xi)\chi_{A'_{22}}(\xi)$ (see Exercise 9.23). Therefore, the characteristic polynomial $\chi_{A'+B'N'}(\xi)$, and hence $\chi_{A+BN}(\xi)$, has, regardless of what N is chosen, $\chi_{A'_{22}}(\xi)$ as a factor. Hence $\chi_{A+BN}(\xi)$ cannot be made equal to any characteristic polynomial. In the language of pole placement, this means that pole placement does not hold if (A, B) is not controllable.

9.4.3 Pole placement for controllable single-input systems

The proof of the sufficiency part of Theorem 9.3.1 for the case $m = 1$ follows from the following theorem, that actually provides an algorithm for choosing N from A, B and $r(\xi)$.

Theorem 9.4.3 *Assume that (9.1) is controllable, and that* $m = 1$. *Let* $F \in \mathbb{R}^{1\times n}$ *be the solution of the system of linear equations*

$$F[B\ AB \cdots A^{n-2}B\ A^{n-1}B] = [0\ 0 \cdots 0\ 1]. \tag{9.13}$$

Then

$$N = -Fr(A) \tag{9.14}$$

yields

$$\chi_{A+BN}(\xi) = r(\xi). \tag{9.15}$$

The notation $r(A)$ *signifies the* $n \times n$ *matrix* $r_0 I + r_1 A + \cdots + r_{n-1}A^{n-1} + A^n$.

Proof Let N be given by (9.14) and denote the characteristic polynomial of $A + BN$ by $\chi_{A+BN}(\xi) = \alpha_0 + \alpha_1\xi + \cdots + \alpha_{n-1}\xi^{n-1} + \xi^n$. From (9.13) it follows that $FB = \cdots = FA^{n-2}B = 0$ and $FA^{n-1}B = 1$, hence

$$\begin{array}{ll} F(A+BN)^k = FA^k & \text{for } k = 0, 1, \ldots, n-1, \\ F(A+BN)^n = FA^n + N. & \end{array} \tag{9.16}$$

By the Cayley–Hamilton theorem, $\chi_{A+BN}(A + BN) = 0$. Hence $F\chi_{A+BN}(A+BN) = 0$. Multiplying both sides of (9.16) by the appropriate coefficients of $\chi_{A+BN}(\xi)$ and adding yields

$$\begin{array}{rcl} N & = & -F\chi_{A+BN}(A) \\ & = & -\alpha_0 F - \alpha_1 FA - \cdots - \alpha_{n-1}FA^{n-1} - FA^n. \end{array}$$

Hence

$$N = -[\alpha_0 \ \alpha_1 \ \cdots \ \alpha_{n-1}] \underbrace{\begin{bmatrix} F \\ FA \\ \vdots \\ FA^{n-1} \end{bmatrix}}_{M} - FA^n. \tag{9.17}$$

From

$$\begin{bmatrix} F \\ FA \\ \vdots \\ FA^{n-1} \end{bmatrix} [A^{n-1}B \ A^{n-2}B \ldots B] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ * & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ * & \cdots & * & 1 \end{bmatrix}$$

it follows that M is an invertible matrix. Combining (9.14) and (9.17) yields

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{n-1} \end{bmatrix} M = \begin{bmatrix} r_0 & r_1 & \cdots & r_{n-1} \end{bmatrix} M.$$

Since M is invertible it follows that

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{n-1} \end{bmatrix} = \begin{bmatrix} r_0 & r_1 & \cdots & r_{n-1} \end{bmatrix}.$$

This completes the proof of Theorem 9.3.1 for the single-input case. □

Observe that it follows from the proof of Theorem 9.4.3, which gives N uniquely in terms of $r(\xi)$, that in the single-input case, the feedback gain N that achieves $r(\xi)$ is unique.

9.4.4 Pole placement for controllable multi-input systems

We now proceed towards the proof of Theorem 9.3.1 in the multi-input case $m > 1$. Denote by $\Sigma_{n,m}^{\text{cont}}$ the set of pairs $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ that are controllable. Consider the pair of matrices (A, B). Let B_k denote the kth column of B. If there existed a k such that the single-input system (A, B_k) were controllable, then the pole placement problem would immediately be solvable by considering feedback laws $u = Nx$ with N of the form

$$N = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} N', \tag{9.18}$$

with the 1 in the kth entry of the first matrix on the right-hand side of (9.18). Indeed, since $A + BN = A + B_k N'$, we see that the problem would then be reduced to the single-input case. Note that (9.18) is a feedback gain matrix that uses only the kth input channel for feedback, with the other inputs set to zero. More generally, if there existed a $K \in \mathbb{R}^{m\times 1}$ such that (A, BK) were controllable, the control law $u = Nx$ with N of the form $N = KN'$ would similarly reduce the problem to the single-input case. However, the system $(I, I) \in \Sigma_{n,n}^{\text{cont}}$ shows that such a $K \in \mathbb{R}^{m\times 1}$ may not exist. Thus it appears not to be possible to reduce the problem to the single-input case by simply taking a linear combination of the inputs. The next lemma shows that we can reduce the problem to the single-input case by combining preliminary feedback with a linear combination of the inputs.

Lemma 9.4.4 *Let $(A, B) \in \Sigma_{n,m}^{\text{cont}}$, and assume that $K \in \mathbb{R}^{m\times 1}$ is such that $BK \neq 0$. Then there exists a matrix $N' \in \mathbb{R}^{m\times n}$ such that $(A+BN', BK) \in \Sigma_{n,1}^{\text{cont}}$.*

Proof (i) Let us first prove that there exist $v_1, \ldots, v_{n-1} \in \mathbb{R}^m$ such that the algorithm

$$x_0 = 0 \ ; \ v_0 = K, x_{t+1} = Ax_t + Bv_t \tag{9.19}$$

generates vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ that are linearly independent. (In Exercise 9.8 an interpretation of (9.19) is given in terms of discrete-time systems.) The proof goes by induction. Note that $x_1 = BK \neq 0$. Assume that $x_1, x_2, \ldots, x_t$, with $t < n$, are linearly independent. We need to prove that there exists a $v_t \in \mathbb{R}^m$ such that the vectors $x_1, x_2, \cdots, x_t, x_{t+1} = Ax_t + Bv_t$ are also linearly independent. Assume to the contrary that for all v_t, $Ax_t + Bv_t \in \mathfrak{L} := \text{span}\{x_1, x_2, \ldots, x_t\}$. Note that since $t < n$, $\mathfrak{L}$ is a

proper subspace of $\mathbb{R}^n$. We now demonstrate that $\mathfrak{L}$ must satisfy

$$\operatorname{im} B \subseteq \mathfrak{L} \text{ and } A\mathfrak{L} \subseteq \mathfrak{L}, \tag{9.20}$$

and subsequently that this contradicts controllability. Indeed, Theorem 5.2.24 on controllability implies that $\mathbb{R}^n$ is the smallest A-invariant subspace that contains $\operatorname{im} B$.

To prove (9.20), note that since $Ax_t + Bv_t \in \mathfrak{L}$ for all $v_t \in \mathbb{R}^m$, there must hold $Ax_t \in \mathfrak{L}$ (take $v_t = 0$) and $\operatorname{im} B \subseteq \mathfrak{L}$. Further, since for $k = 0, 1, \ldots, t-1$, there exist $v_0, v_1, \ldots, v_{t-1}$ such that $x_{k+1} = Ax_k + Bv_k$, it follows that $Ax_k \in \mathfrak{L}$ for $k = 1, 2, \ldots, t-1$. Hence $Ax_k \in \mathfrak{L}$ for $k = 1, 2, \ldots, t$. This yields $A\mathfrak{L} \subset \mathfrak{L}$.

To show that (9.20) contradicts controllability, observe that $\operatorname{im} A^k B = A^k \operatorname{im} B \subseteq \mathfrak{L}$ for $k = 0, 1, \ldots$. Consequently, $\operatorname{im}[B, AB, \ldots, A^{n-1}B] \subseteq \mathfrak{L}$. This implies that $\mathfrak{L} = \{x_1, x_2, \ldots, x_t\} = \mathbb{R}^n$, contradicting the fact that $\mathfrak{L}$ is a proper subspace of $\mathbb{R}^n$. Hence $t = n$.

(ii) It follows from (i) that there exist $v_0, v_1, \ldots, v_{n-1} \in \mathbb{R}^m$ such that x_1,x_2,...,$x_n \in \mathbb{R}^n$ defined by (9.19) are linearly independent. Also, it follows that we can take $v_0 = K$, and hence $x_1 = BK$. Now define the matrix N' by $[v_1, \ldots, v_{n-1}, v_n] = N'[x_1, \ldots, x_{n-1}, x_n]$ (with $v_n \in \mathbb{R}^m$ arbitrary). Note that this defines N', since $[x_1, \ldots, x_{n-1}, x_n] \in \mathbb{R}^{n \times n}$ is nonsingular. This yields $x_{t+1} = (A + BN')^t x_1$ for $t = 0, 1, \ldots, n-1$. Since, $x_1 = BK$, this implies $[BK, (A+BN')BK, \ldots, (A+BN')^{n-1}BK] = [x_1, x_2, \ldots, x_n]$. Since $[x_1, x_2, \ldots, x_n]$ is nonsingular, it follows that the pair $(A + BN', BK)$ is indeed controllable. □

We are now ready to deliver the *coup de grâce*.

Proof of the sufficiency part of Theorem 9.3.1 The proof of the sufficiency of Theorem 9.3.1 in the case $m > 1$ is as follows. First choose $K \in \mathbb{R}^{m \times 1}$ such that $BK \neq 0$, and $N' \in \mathbb{R}^{m \times 1}$ such that $(A + BN', BK)$ is controllable. By controllability, $B \neq 0$, and hence such a K exists. By Lemma 9.4.4 such an N' exists. Next, use Theorem 9.4.3 in the case $m = 1$, applied to $(A+BN', BK)$, to obtain $N'' \in \mathbb{R}^{1 \times n}$ such that $A+BN'+BKN''$ has the desired characteristic polynomial $r(\xi)$. Finally, observe that the feedback law $u = Nx$ with $N = N' + KN''$ achieves $\chi_{A+BN}(\xi) = r(\xi)$. This yields the desired characteristic polynomial with feedback applied to the original system (9.1). □

We now review briefly the key points of the proof of Theorem 9.3.1. First we showed that pole placement is invariant under system similarity (cf. Lemma 9.4.1). Using the transformation of (A, B) into the similar system (9.12) and examining the effect of feedback on this similar system immediately yields the conclusion that controllability is a necessary condition for pole placement. To prove the converse, i.e., that controllability implies pole

placement, observe that Lemma (9.4.4) reduces the multi-input case to the single-input case. The single-input case is proven in Theorem 9.4.3. Several alternative ideas for elements of the proofs are explored in Exercises 9.11 to 9.13.

Observe that if we view the equation for pole placement $\chi_{A+BN}(\xi) = r(\xi)$ as n real equations (the coefficients of the polynomials) in mn real unknowns (the elements of the matrix $N \in \mathbb{R}^{m\times n}$), then we have shown that these equations are solvable for all $r(\xi)$ if and only if (A, B) is controllable. If $m = 1$, then the number of equations is equal to the number of unknowns, and indeed (see the comment at the end of Section 9.4.3) the solution is unique. In the multi-input case, there are more unknowns than equations, and there will be multiple solutions (we have not formally shown this, but the nonuniqueness of the matrices K and N' constructed in Lemma 9.4.4 makes it at least intuitively reasonable). It is interesting to note that it was harder to prove solvability of an equation that has multiple solutions than one that has a unique solution.

9.5 Algorithms for Pole Placement

Theorem 9.3.1 implies that if (9.1) is controllable and if $r(\xi) \in \mathbb{R}[\xi]$ is any monic polynomial of degree n, then there exists an $N \in \mathbb{R}^{m\times n}$ such that the closed loop matrix $A + BN$ has characteristic polynomial $r(\xi)$. Our proof, while in principle constructive, does not really provide an algorithm for computing an N from the data (A, B) and $r(\xi)$ in the multi-input case. In this section we discuss some algorithmic aspects of the computation of N.

The following conceptual algorithm may be deduced from Theorem 9.4.3 and the proof of Theorem 9.3.1.

Algorithm 9.5.1 Pole placement by state feedback

□

Data: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, with (A, B) controllable; $r(\xi) \in \mathbb{R}[\xi]$ with $r(\xi)$ monic and of degree n.

Required: $N \in \mathbb{R}^{m\times n}$ such that $\chi_{A+BN}(\xi) = r(\xi)$.

Algorithm:

1. Find $K \in \mathbb{R}^{m\times 1}$ and $N' \in \mathbb{R}^{m\times n}$ such that $(A + BN', BK)$ is controllable. Lemma 9.4.4 shows that such K, N' exist. We shall see in Theorem (9.5.2) that in fact a "random" choice produces such a pair (K, N').

2. Put $A' = A + BN'$, $B' = BK$, and compute F from

$$F[B', A'B', \ldots, (A')^{n-1}B'] = [0\ 0 \ldots\ 0\ 1].$$

3. Compute $N'' = -Fr(A')$.

4. Compute $N = N' + KN''$.

Result: N is the desired feedback matrix.

Note that step 1 of the above algorithm may be skipped for single-input systems, since $K = 1$ and $N' = 0$ will do in this case. Even for multi-input systems this step is a great deal easier than the explicit construction carried out in the proof of Lemma 9.4.4 suggests. The procedure for finding the matrices K and N' given in Lemma 9.4.4 is, in a sense, constructive. However, it turns out that if the matrices K and N' are chosen using a random number generator, then we can be *sure* that the resulting matrices $(A + BN', BK)$ form a controllable pair. Well, formally speaking, we can only be *almost sure*. We now explain this.

Let S be a subset of $\mathbb{R}^N$. Think of $\mathbb{R}^N$ as parametrizing a family of concrete objects and of S as those objects that enjoy a certain desired property. For example, S could consist of those $(K, N') \in \mathbb{R}^{m\times 1} \times \mathbb{R}^{m\times n}$ such that $(A+BN', BK)$ is controllable. We call S an *algebraic variety* if there exists a polynomial $p(\xi_1, \xi_2, \ldots, \xi_N) \in \mathbb{R}[\xi_1, \xi_2, \ldots, \xi_N]$ (that is, a polynomial with real coefficients in N variables $\xi_1, \xi_2, \ldots, \xi_N$) such that

$$S = \{\mathrm{col}(z_1, z_2, \ldots, z_N) \in \mathbb{R}^N \mid p(z_1, z_2, \ldots, z_N) = 0\}.$$

If an algebraic variety S is not equal to all of $\mathbb{R}^N$ (hence if the coefficients of $p(\xi_1, \xi_2, \ldots, \xi_N)$ are not all zero), then we call S a *proper* algebraic variety. It can be shown that a proper algebraic variety must be a "*very small*" set. Specifically, it can be shown that if S is a proper algebraic variety, then

1. $S^{\text{complement}}$ is open and dense in $\mathbb{R}^N$.

2. S has zero Lebesgue measure. This means that for all $\epsilon > 0$ there exists a countable sequence $a_k \in \mathbb{R}^N$, $k = 1, 2, \ldots$, such that

$$S \subseteq \bigcup_{k=1}^{\infty} (a_k - \frac{\epsilon'}{2^k}, a_k + \frac{\epsilon'}{2^k}), \tag{9.21}$$

   with ϵ' the vector $\mathrm{col}(\epsilon, \epsilon, \ldots, \epsilon) \in \mathbb{R}^N$. Note that the volume of the set on the right-hand side of (9.21) goes to zero as ϵ goes to zero. Hence (2) states that S is contained in a set of arbitrarily small volume.

Intuitively these two properties mean that if we choose a point $x \in \mathbb{R}^N$ "*at random*," then it essentially never belongs to S. Mathematicians often

call $S^{\text{complement}}$ *generic*, or in *general position* It is useful to think that consequently, *typical* elements of $\mathbb{R}^N$ enjoy property $S^{\text{complement}}$. As an illustration of the situation at hand, draw in $\mathbb{R}^2$ the familiar curve defined by $z_1^2 + z_2^2 = 1$. Observe that it is an algebraic variety, and indeed, a randomly chosen point in $\mathbb{R}^2$ does not lie on the unit circle.

We now show that matrices K, N' generically have the property required in Lemma 9.4.4.

Theorem 9.5.2 *Let* $(A, B) \in \Sigma_{m,n}^{\text{cont}}$. *Then the set* $\{(K, N') \in \mathbb{R}^{m\times 1} \times \mathbb{R}^{m\times n} \mid (A + BN', BK)$ *is controllable*$\}$, *viewed as a subset of* $\mathbb{R}^{m(n+1)}$, *is the complement of a proper algebraic variety.*

Proof Define $M = A + BN'$, and observe that

$$\{(K, N') \in \mathbb{R}^{m\times 1} \times \mathbb{R}^{m\times n} \mid \det[BK, MBK, \ldots, M^{n-1}BK] = 0\} \quad (9.22)$$

is an algebraic variety, since the equation expressing that the determinant in (9.22) is zero obviously defines a polynomial in the components of the matrices K and N'. That it is a proper algebraic variety is the content of Lemma 9.4.4. □

We can conclude from this theorem that the first step of Algorithm 9.5.1 can be carried out by choosing the elements of K and N' by means of a random number generator.

As a final comment regarding computation of an N such that

$$\chi_{A+BN}(\xi) = r(\xi), \quad (9.23)$$

observe that as already remarked before, it requires solving n equations (obtained by equating the n coefficients of the monic polynomials on the left- and right-hand sides of (9.23)) with mn unknowns (the entries of the matrix $N \in \mathbb{R}^{m\times n}$). Actually, if $m = 1$, then the solution N of (9.23), if it exists at all, is unique (and it always exists in the controllable case). However, if $m > 1$, there are less equations than unknowns, and indeed, the solution is nonunique. This feature can be exploited to obtain solutions that are "better" than others. For example, the linear algebra and control systems package MATLAB© uses this nonuniqueness in order to find an N such that the sensitivity of $\chi_{A+BN}(\xi)$ under changes in N is minimized in some appropriate sense.

9.6 Stabilization

Theorem 9.3.1 can be refined so that it gives a complete answer to the pole placement question for systems (A, B) that are not necessarily controllable.

This refinement is based on Lemma 9.4.2, which basically provides a canonical form for $\Sigma_{n,m}$. This canonical form puts the controllability structure into evidence. Consider the matrix A'_{22} in (9.12). This matrix characterizes the noncontrollable behavior of the system (9.1). Its characteristic polynomial $\chi_{A_{22}}(\xi)$ is called the *uncontrollable polynomial* of the system (9.1), equivalently of (A, B), and its roots are called the *uncontrollable poles*, often called the *uncontrollable modes*. This allows us to state the following refinement of Theorem 9.3.1.

Theorem 9.6.1 *Consider the system (9.1), and assume that $\chi_{\rm u}(\xi)$ is its uncontrollable polynomial. There exists a feedback matrix $N \in \mathbb{R}^{n\times m}$ such that $\chi_{A+BN(\xi)} = r(\xi)$ if and only if $r(\xi)$ is a real monic polynomial of degree n that has $\chi_{\rm u}(\xi)$ as a factor.*

Proof Observe that by Lemma 9.4.1, if (A, B) and (A', B') are similar, then there exists an $N \in \mathbb{R}^{n\times m}$ such that $\chi_{A+BN(\xi)} = r(\xi)$ if and only if there exists an $N' \in \mathbb{R}^{n\times m}$ such that $\chi_{A'+B'N'(\xi)} = r(\xi)$. Now take (A', B') as in (9.12). Partition N' conformably as $N' = [N'_1\ N'_2]$. Then

$$A' + B'N' = \begin{bmatrix} A'_{11} + B'_1N'_1 & A'_{12} + B'_1N'_2 \\ 0 & A'_{22} \end{bmatrix}.$$

Obviously, $\chi_{A'+B'N'(\xi)} = \chi_{A'_{11}+B'_1N'_1(\xi)}\chi_{A_{22}}(\xi) = \chi_{A'_{11}+B'_1N'_1(\xi)}\chi_{\rm u}(\xi)$. Now, since (A'_{11}, B'_1) is controllable, $\chi_{A'_{11}+B'_1N'_1}(\xi)$ can, by Theorem 9.3.1, be made equal to any real monic polynomial of degree n_1. The result follows. □

Consider the system (9.1) with the control law (9.9). The closed loop system (9.10) is asymptotically stable if and only if $A + BN$ is a Hurwitz matrix. The question thus arises whether for a given system $(A, B) \in \Sigma_{m,n}$, there exists a feedback matrix $N \in \mathbb{R}^{m\times n}$ such that $(A + BN)$ is Hurwitz.

Corollary 9.6.2 *There exists a feedback law (9.9) for (9.1) such that the closed loop system (9.10) is asymptotically stable if and only if the uncontrollable polynomial of (9.1) is Hurwitz.*

From the canonical form (9.12) it follows, in fact, that for the existence of a feedback control law of any kind (linear/nonlinear time-invariant/time-varying, static/dynamic) such that the closed loop system is asymptotically stable, it is simply always necessary that $\chi_{\rm u}(\xi)$ is Hurwitz. Indeed, the second equation of (9.12) shows that x'_2 is not influenced in any way by the control. Hence the matrix A'_{22} has to be Hurwitz to start with if we want all solutions to go to zero after control is applied. Motivated by this discussion, we call the system (9.1), or, equivalently, the pair (A, B), *stabilizable* if its uncontrollable polynomial is Hurwitz (see also Section 5.2.2 and Exercise 9.15).

9.7 Stabilization of Nonlinear Systems

The result on pole placement and stabilization by state feedback can immediately be applied in order to stabilize a *nonlinear* system around an equilibrium by using a linear feedback law.

Consider the system

$$\frac{d}{dt}x = f(x,u), \quad x \in \mathbb{R}^n, u \in \mathbb{R}^m, \tag{9.24}$$

with $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ continuously differentiable. Assume that (u^*, x^*) is an equilibrium, i.e., that $f(x^*, u^*) = 0$. Linearization around this equilibrium (see Section 4.7) yields

$$\frac{d}{dt}\Delta_x = A\Delta_x + B\Delta_u, \tag{9.25}$$

with $A = \frac{\partial f}{\partial x}(x^*, u^*), B = \frac{\partial f}{\partial u}(x^*, u^*)$. If (A, B) is controllable, then, following Theorem 9.3.1, there exists an $N \in \mathbb{R}^{m\times n}$ such that $A + BN$ has preassigned eigenvalues, in particular such that $A + BN$ is Hurwitz. The control law $\Delta_u = N\Delta_x$, of course, stabilizes the linear system (9.25). However, our aim is to stabilize the *nonlinear* system (9.24) around the equilibrium x^*.

In order to achieve this, consider the following control law for (9.24) $u = u^* + N(x - x^*)$. Combined with (9.24) this yields the controlled system

$$\frac{d}{dt}x = f(x, u^* + N(x - x^*)). \tag{9.26}$$

Denote the function on the right-hand side of (9.25) by g; i.e., $g(x) := f(x, u^* + N(x - x^*))$. The system (9.26) can thus be written as

$$\frac{d}{dt}x = g(x). \tag{9.27}$$

Since obviously, x^* satisfies $g(x^*) = 0$, x^* is an equilibrium of the autonomous system (9.27). Linearization around this equilibrium yields

$$\frac{d}{dt}\Delta_x = \frac{\partial g}{\partial x}(x^*)\Delta_x.$$

Using the chain rule yields

$$\frac{\partial g}{\partial x}(x^*) = \frac{\partial f}{\partial x}(x^*) + \frac{\partial f}{\partial u}(u^*)N = A + BN.$$

Hence $\frac{\partial g}{\partial x}(x^*)$ is Hurwitz. Therefore, by Theorem 7.5.2, x^* is an asymptotically stable equilibrium of (9.26). This shows that if the system (9.24) linearized around the equilibrium (x^*, u^*), is controllable, it is always possible to stabilize the nonlinear system around the equilibrium x^*.

Example 9.7.1 Consider the pendulum (see Examples 7.1.1 and 9.1.3). Assume that in contrast to Example 9.1.3, the length is constant and that an external force acts on the pendulum, leading to the differential equation

$$\frac{d^2}{dt^2}\theta + \frac{g}{L}\sin\theta = \frac{1}{ML^2}F.$$

Use the state variables $x_1 = \theta, x_2 = \frac{d}{dt}\theta$, and denote the external force F, which will be the control, by u. This leads to the state equations

$$\begin{aligned} \frac{d}{dt}x_1 &= x_2, \\ \frac{d}{dt}x_2 &= -\frac{g}{L}\sin x_1 + \frac{1}{ML^2}u. \end{aligned} \tag{9.28}$$

Both $x^* = (0,0), u^* = 0$ and $x^* = (\pi,0), u^* = 0$ are equilibria. Linearization around the first equilibrium yields

$$\begin{aligned} \frac{d}{dt}\Delta_{x_1} &= \Delta_{x_2}, \\ \frac{d}{dt}\Delta_{x_2} &= -\frac{g}{L}\Delta_{x_1} + \frac{1}{ML^2}\Delta_u, \end{aligned} \tag{9.29}$$

and around the second equilibrium

$$\begin{aligned} \frac{d}{dt}\Delta'_{x_1} &= \Delta'_{x_2}, \\ \frac{d}{dt}\Delta'_{x_2} &= \frac{g}{L}\Delta'_{x_1} + \frac{1}{ML^2}\Delta'_u. \end{aligned} \tag{9.30}$$

These linearized systems are both controllable. Hence both equilibrium points can be made asymptotically stable. For the first equilibrium point this can, for example, be achieved by using the control law $u = -Kx_2$, with $K > 0$. This corresponds to introducing damping in the system. For the second equilibrium point, stabilization can be achieved using the control law $u = -K_1(x_1 - \pi) - K_2x_2$, with $K_1 > LMg$ and $K_2 > 0$. □

Example 9.7.2 Consider the motion of a spinning body. This motion has been studied before in Exercise 7.32; see equations (7.40). However, we now assume that the spinning around the principal axis can be accelerated by torques N_1, N_2, N_3. The equations then become

$$\begin{aligned} I_1\frac{d\omega_1}{dt} &= (I_2 - I_3)\omega_2\omega_3 + N_1, \\ I_2\frac{d\omega_2}{dt} &= (I_3 - I_1)\omega_3\omega_1 + N_2, \\ I_3\frac{d\omega_3}{dt} &= (I_1 - I_2)\omega_1\omega_2 + N_3, \end{aligned} \tag{9.31}$$

with $0 < I_1 < I_2 < I_3$. Here ω_1, ω_2, ω_3 denote the rate of spinning of the body around its principal axes, and the acceleration torques N_1, N_2, N_3 are inputs that can be achieved for example by electro-motors mounted on the main axis of the spinning body.

We have seen in Exercise 7.32 that in the absence of the inputs $N_1 = N_2 = N_3 = 0$, this body cannot spin in an asymptotically stable mode. We now examine whether asymptotic stability can be achieved by exerting control. We do this by considering the linearized system, and stabilize the system around the equilibrium point $\mathrm{col}(0, \omega_2^*, 0)$ with $\omega_2^* > 0$. Linearization yields

$$\begin{aligned} I_1 \frac{d\Delta_{\omega_1}}{dt} &= (I_2 - I_3)\omega_2^* \Delta_{\omega_3} + N_1, \\ I_2 \frac{d\Delta_{\omega_2}}{dt} &= N_2, \\ I_3 \frac{d\Delta_{\omega_3}}{dt} &= (I_1 - I_2)\omega_2^* \Delta_{\omega_1} + N_3. \end{aligned} \tag{9.32}$$

Note that this system is not stabilizable if we use only one torque. Hence in order for this system to be stabilizable, we need to use at least two controls: ΔN_1 and ΔN_2, or ΔN_2 and ΔN_3. The open loop poles of (9.32) are at $0, \pm\omega_2^* \sqrt{\frac{(I_3 - I_2)(I_2 - I_1)}{I_1 I_3}}$. Thus, when the system runs open loop, this equilibrium point is unstable. We look for a feedback control law using the torques N_1, N_2 that puts all three closed loop poles in the left half plane at $-\omega_2^* \sqrt{\frac{(I_3 - I_2)(I_2 - I_1)}{I_1 I_3}}$. The feedback law

$$\begin{aligned} N_1 &= -\omega_2^* \sqrt{\frac{I_1}{I_3}} \sqrt{(I_3 - I_2)(I_2 - I_1)} \Delta_{\omega_1} - 2\omega_2^* (I_3 - I_2) \Delta_{\omega_3}, \\ N_2 &= -\omega_2^* \sqrt{\frac{I_2^2}{I_1 I_3}} \sqrt{(I_3 - I_2)(I_2 - I_1)} \Delta_{\omega_3} \end{aligned} \tag{9.33}$$

puts the closed loop poles in the desired locations. It follows from the discussion at the beginning of this section that the control law (9.33) with $\Delta_{\omega_1}, \Delta_{\omega_3}$ replaced by ω_1, ω_3 makes $(0, \omega_2^*, 0)$ an asymptotically stable equilibrium of the controlled nonlinear system (9.31). □

Example 9.7.3 Theorem 9.3.1 shows that a controllable system can be stabilized by means of a memoryless state feedback law (9.9). This feedback law assumes that all the state variables are available for feedback. In the next chapter we will discuss how one can proceed when only output measurements are available. However, as we shall see, the resulting control laws are dynamic. In the present example we illustrate the limitations that can result from the use of a *memoryless output feedback* control law. This issue is also illustrated in simulation exercise A.1.

Consider the motion of a point mass in a potential field with an external force as control. Let q denote the position of the point mass with respect

to some coordinate system, and F the external force exerted on it. This leads to the equation of motion

$$\frac{d^2}{dt^2}q + G(q) = F,$$

where the internal force $G(q)$ is due to the potential field. Let us consider the one-dimensional case, i.e., motion along a line. Thus $q : \mathbb{R} \to \mathbb{R}$. As examples, we can think of Newton's second law ($G = 0$), the motion of a mass in a mass–spring combination (G linear), and, interpreting q as the angle and F as an external torque, the motion of a pendulum (see Example 9.7.1). Assume that $G(0) = 0$, and let us consider the question of how to stabilize the equilibrium point 0 for this system. For the sake of concreteness, assume that G is linear. In that case we can, with a mild abuse of notation, write $G(q)$ as $G\,q$, with G now a constant parameter. Let us try to stabilize this system by means of a memoryless control law that uses q as measurements. For example, one could hope to achieve stabilization by always pushing the point mass back to the origin, thus by taking $F < 0$ if $q > 0$, and $F > 0$ if $q < 0$. It can be shown that this does not result in asymptotic stability, no matter how subtly F may be chosen. In order to see this, try first a linear feedback law $u = Nq$. Then the closed loop system

$$\frac{d^2}{dt^2}q + (G - N)q = 0$$

is never asymptotically stable. It is stable if $G > N$, its solutions are sinusoidal, and unstable if $G \leq N$. If we choose a nonlinear feedback law $F = N(q)$ instead, then we end up with a system of the form

$$\frac{d^2}{dt^2}q + \phi(q) = 0, \tag{9.34}$$

where $\phi(q) = G\,q - N(q)$. Can this system be asymptotically stable? We have already seen that the linearized system cannot be made asymptotically stable, but could one perhaps choose $N(q)$ cleverly such that 0 is an asymptotically stable equilibrium of the nonlinear system? *The answer is no.* In order to see this, consider the function

$$V(q, \frac{d}{dt}q) = \frac{1}{2}\left(\frac{d}{dt}q\right)^2 + \int_0^q \phi(\mu)d\mu. \tag{9.35}$$

Its derivative along solutions of (9.34) is zero. The value of (9.35) is thus constant along solutions of (9.34). Hence, if we start with an initial condition $(q(0), (\frac{d}{dt}q)(0))$ such that $V(q(0), (\frac{d}{dt}q)(0)) \neq V(0, 0)$, then by continuity, we simply cannot have that $\lim_{t\to\infty}(q(t), \frac{d}{dt}q(t)) = (0, 0)$, excluding the possibility that 0 is an asymptotically stable equilibrium.

So in order to stabilize this very simple mechanical system, we either have to build memory into the feedback processor or measure more than only the position. Actually, for the case at hand (and assuming $G > 0$), asymptotic stability can be obtained by velocity feedback

$$F = -D\frac{d}{dt}q. \tag{9.36}$$

This control law can be implemented by means of a simple damper, or by a tachometer (a device that measures the velocity) that generates the required force by means of a transducer (a device that transforms the output of the tacho into a force). Note that when $G = 0$ (a free point mass) even (9.36) does not stabilize, and a combination of position and velocity feedback is required.

This example shows the limitations of memoryless output feedback and underscores the need for state or dynamic output feedback. □

9.8 Recapitulation

In this chapter we studied pole placement and stabilization of state space systems. The main ideas are the following:

- Feedback is one of the basic concepts of control theory (Section 9.1). In feedback control, the control input is chosen as a function of the past and the present of the observed output. In this chapter, we studied what can be achieved by means of memoryless linear state feedback control.
- The main result obtained is the pole placement theorem. This theorem states that controllability is equivalent to the existence of a memoryless feedback gain matrix such that the closed characteristic polynomial is equal to an arbitrary preassigned one (Theorem 9.3.1).
- There are effective algorithms for computing this feedback gain matrix (Algorithm 9.5.1).
- For noncontrollable systems, the closed loop characteristic polynomial always has the uncontrollable polynomial of the plant as a factor. Thus a system is stabilizable if and only this uncontrollable polynomial is Hurwitz (Theorem 9.6.1).
- An equilibrium point of a nonlinear system can be stabilized if the linearized system is controllable or, more generally, stabilizable (Section 9.7).

9.9 Notes and References

Theorem 9.3.1 is one of the most important and elegant results in control theory. In the single-input case, the result seems to have been more or less known around

1960 (see [29]). The first proof appeared in [47]. The multivariable case was obtained by [45] and [35] in the complex case, which proved to be considerably easier than the real case. The latter was proven by [64]. Lemma 9.4.4 is known as Heymann's lemma [22]. Our proof follows the one in [21]. The algorithm that is used in the proof of Theorem 9.4.3 is known as Ackermann's algorithm [2].

9.10 Exercises

9.1 Discuss the distinction between open loop and feedback control as applied to the scheduling of the red/green settings of a traffic light.

Discuss the distinction between open and feedback control as used by the player that serves and the one that returns the serve in a tennis match.

9.2 Consider Newton's second law, $M\frac{d^2}{dt^2}y = u$. Let $y(0) = a$ and $(\frac{d}{dt}y)(0) = b$ be given. Compute a control $u : [0, 1] \to \mathbb{R}$ such that $y(1) = 0$ and $\frac{d}{dt}y(1) = 0$. This control law obviously stabilizes the system. Now assume that you use the same control u with slightly different values of a, b, or M. Will this control still bring the system to rest?

Now assume that both the position y and the velocity $\frac{d}{dt}y$ are measured, and consider the control law $u = -K_p y - K_v \frac{d}{dt}y$, with $K_p, K_v > 0$. Will this control law drive the system to rest? Does this depend on a, b, or M?

Discuss by means of this example some of the advantages of feedback control versus open loop control.

9.3 Consider the feedback structure shown in Figure 9.3. View K and μ as

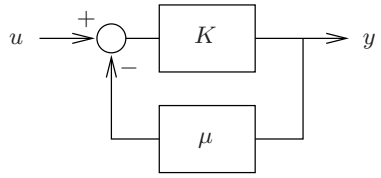


FIGURE 9.3. Feedback structure with static gains.

simple static gains.

(a) Prove that y is given in terms of u by

$$y = \frac{K}{1 + \mu K}u.$$

Call the resulting gain K'. Thus $K' = \frac{K}{1+\mu K}$.

(b) Prove that if K is large, then $K' \approx \frac{1}{\mu}$. Compute the % change of K' in terms of the % change of K. Conclude that K' is relatively insensitive to changes in K for large K.

Note: This seemingly trivial result has far-reaching consequences in amplifier design. This can be explained as follows. Let K be the gain of an active device (for example, a transistor). Typically, K is very sensitive to operating conditions (such as the temperature and the load). On the other hand, it is possible to produce simple passive devices that are insensitive; for example, voltage dividers using passive resistors. Let $\mu < 1$ be the gain of such a device. Now, using the (sensitive) active device in the forward loop of a feedback system and the (insensitive) passive device in the feedback loop results in *an insensitive amplifier* with gain approximately equal to $1/\mu$. This principle is the basic idea behind the *operational amplifier* as invented by Black [12]. For the history surrounding this invention see [10]. For a narrative account of his invention see [13]. That one can make an insensitive amplifier using a sensitive one sounds like a *perpetuum mobile*, but it isn't: it is one of the ingredients that made reliable long-distance telephone communication possible. See also the preface to this book.

9.4 Consider a harmonic oscillator with an external force

$$\frac{d^2}{dt^2}y + y = u.$$

Consider the control law $u = f_1 y + f_2 \frac{d}{dt}y$. Is this a linear memoryless state feedback law? Explain the possible implementation of both terms of the control law physically (assuming $f_1 \leq 0$ and $f_2 \leq 0$).

9.5 Consider the system

$$\frac{d^3}{dt^3}y = u.$$

Write this system in state form. Is this system open loop stable? Is it controllable? Find (without using the algorithms discussed in this chapter) a state feedback control law such that the closed loop characteristic polynomial is $1 + 2\xi + \xi^2 + \xi^3$. Is the resulting controlled system asymptotically stable?

9.6 Find for the systems

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

and

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

matrices $K \in \mathbb{R}^m$ and $N' \in \mathbb{R}^{m \times n}$ such that $(A + BN', BK)$ is controllable.

9.7 Give an example of a system $(A, B) \in \Sigma_{4,2}^{\text{cont}}$ for which there exist no $K \in \mathbb{R}^{2 \times 1}$ such that $(A, BK) \in \Sigma_{4,1}^{\text{cont}}$.

Hint: A matrix $M \in \mathbb{R}^{n \times m}$ is said to be *cyclic* if there exists $x \in \mathbb{R}^n$ such that the vectors $x, Mx, \ldots, M^{n-1}x$ are linearly independent; M is cyclic if

and only if its characteristic polynomial is equal to its minimal polynomial. Prove that if $(A, B) \in \Sigma_{n,1}^{\text{cont}}$, then A must be cyclic. Use this to construct the example.

9.8 Consider the discrete-time analogue of (9.1):

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

Define controllability analogously as in the continuous-time case. Prove that this system is controllable if and only if for any $0 \neq x_0 \in \mathbb{R}^n$ there exists a state trajectory x such that $x(0) = x_0$ and such that $x(0), x(1), \ldots, x(n-1)$ are linearly independent. Show that this result implies part (i) of the proof of Lemma 9.4.4.

9.9 Use the algorithm of Theorem 9.4.3 in order to find a state feedback control gain matrix for the system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u$$

such that the closed loop system has a pole of multiplicity 3 at the point -1. Repeat this for the system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & -2 & 0 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} u$$

and the closed loop characteristic polynomial $1 + 2\xi + \xi^2 + 2\xi^3 + \xi^4$.

9.10 Find a state feedback control gain matrix for the system

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} u$$

such that the closed loop characteristic polynomial is $1+3\xi+4\xi^2+3\xi^3+\xi^4$. Is the controlled system asymptotically stable?
Repeat this for the system

$$\frac{d}{dt}x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} u$$

and the closed loop eigenvalues $\{-1, -2, -3, -4\}$.

9.11 Consider single-input systems (9.1) defined by (A, B) matrices of the following special form:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & \cdots & -p_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

The resulting system is in controller canonical form (see Section 6.4.2). Prove (again) that it is controllable. Verify that the pole placement is basically trivial for systems in controller canonical form. Indeed, prove that the feedback gain $N = [N_1 \ N_2 \cdots N_n]$, applied to (9.1), yields as closed loop characteristic polynomial $r(\xi) = r_0 + r_1\xi + \cdots + r_{n-1}\xi^{n-1} + \xi^n$, if you choose

$$N_k = p_{k-1} - r_{k-1}. \tag{9.37}$$

Prove that you also obtain expression (9.37) from the formula in Theorem 9.4.3 applied to the case at hand.

9.12 Let (A, B) be controllable, with $m = 1$. Recall from Corollary 6.5.5 that (A, B) is then similar to a system that is in controller canonical form. Now use Lemma 9.4.1 and the result of Exercise 9.11 to derive the pole placement result in the single-input case. Note that this yields an alternative proof to Theorem 9.4.3, without invoking algorithm (9.13). The present proof is in a sense also algorithmic, in that it requires computing the similarity matrix S that brings (A, B) into controller canonical form, followed by formula (9.37).

9.13 Use the ideas of Exercises 9.11 and 9.12 to obtain an alternative proof of Theorem 9.4.3. Proceed as follows. First prove it using Exercise 9.11 when (9.1) is in controller canonical form. For clarity denote this pair by (A_c, B_c). Compute F_c as in (9.13) by $F_c[B_c \ A_cB_c \cdots A_c^{n-2}B_c \ A_c^{n-1}B_c] = [0 \ 0 \cdots 0 \ 1]$. Prove that

$$N_c = -F_c r(A_c). \tag{9.38}$$

The right-hand side is given by $r_0F_c - r_1F_cA_c - \cdots - r_{n-1}F_cA_c^{n-1} - F_cA_c^n$. The first $(n-1)$ terms of (9.38) yield $[-r_0 - r_1 \cdots - r_{n-1}]$. Let $\chi_{A_c}(\xi) = p_0 + p_1\xi + \cdots + p_{n-1}\xi^{n-1} + \xi^n$. Observe that by the Cayley–Hamilton theorem $F_cA_c^n = -p_0F_c - p_1F_cA_c - \cdots - p_{n-1}F_cA_c^{n-1}$. Hence the last term of (9.38) yields $[p_0 \ p_1 \ \cdots \ p_{n-1}]$. Consequently, (9.38 yields

$$N_c = \begin{bmatrix} p_0 - r_0 & p_1 - r_1 & \cdots & p_{n-1} - r_{n-1} \end{bmatrix}.$$

Now turn to the general case. Let S be the nonsingular matrix that brings (A, B) into control canonical form $SAS^{-1} = A_c, SB = B_c$. Now prove that $N = N_cS = -F_cr(A_c)S = -F_cSr(A)$ yields $\chi_{A+BN}(\xi) = r(\xi)$. Therefore, defining F by (9.13) yields $F[B \ AB \cdots A^{n-1}B] = FS^{-1}[B_c \ A_cB_c \cdots A_c^{n-1}B_c] = [0 \cdots 0 \ 1]$. Therefore $FS^{-1} = F_c$. Hence $N = N_cS = -Fr(A)$ yields (9.15) in the general case.

9.14 Determine exactly the closed loop characteristic polynomials achievable by memoryless linear state feedback for the following pairs (A, B):

(a) $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$.

(b) $A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $B = \text{col}(b_1, b_2, \ldots, b_n)$.

9.15 Consider the notion of stabilizability as introduced in Section 9.6. Prove that (9.1) is stabilizable if and only if for all $x_0 \in \mathbb{R}^n$ there exists a (smooth) input $u : \mathbb{R} \to \mathbb{R}^m$ such that the solution of

$$\frac{d}{dt}x = Ax + Bu(t), x(0) = x_0$$

satisfies $x(t) \to 0$ as $t \to \infty$.

9.16 Call (9.1) *marginally stabilizable* if there exists a feedback gain matrix $N \in \mathbb{R}^{m\times n}$ such that all solutions of (9.10) are bounded on $[0, \infty)$. Give necessary and sufficient conditions for marginal stabilizability assuming that the uncontrollable polynomial of (9.1) has simple roots. Note that following Chapter 7 it may have been better to use the term *asymptotically stabilizable* for what we have called *stabilizable* in Chapter 9, and *stabilizable* for what we just now decided to call *marginally stabilizable*.

9.17 Let $(A, B) \in \Sigma^{\text{cont}}_{n,m}$, the set of controllable systems with n states and m inputs. Consider the map $\gamma : \mathbb{R}^{m\times n} \to \mathbb{R}^n$ defined by $\gamma(N) := (r_0, r_1, \ldots, r_{n-1})$, where $r_0 + r_1\xi + \cdots + r_{n-1}\xi^{n-1} + \xi^n$ is the characteristic polynomial of $A + BN$. Is γ surjective? Injective? Linear? Affine? Treat the cases $m = 1$ and $m > 1$ separately.

9.18 (a) Prove that the set $\{M \in \mathbb{R}^{n\times n} \mid \det\ M = 0\}$ defines a proper algebraic variety of $\mathbb{R}^{n^2}$.

(b) Let $S_1, S_2 \subset \mathbb{R}^N$ be proper algebraic varieties. Prove that $S_1 \cap S_2$ and $S_1 \cup S_2$ are also proper algebraic varieties.

(c) Prove that if S is a proper algebraic variety in $\mathbb{R}^N$, then $S^{\text{complement}}$ is open and dense.

(d) (For mathematically advanced readers.) Prove that if S is a proper algebraic variety in $\mathbb{R}^N$, then S has Lebesgue measure zero.

Hint: You may use the following fact. Let $\mathfrak{L}$ be an $(N-1)$-dimensional subspace of $\mathbb{R}^N$, and z an element of $\mathbb{R}^N, z \in \mathfrak{L}$. Now consider the linear variety $\mathfrak{L}_\alpha := \alpha z + \mathfrak{L}$ with $\alpha \in \mathbb{R}$. Then S has zero Lebesgue measure if for all but a finite number of αs, $\mathfrak{L}_\alpha \cap S$, viewed as a subset of $\mathbb{R}^{N-1}$ in the obvious way, has zero Lebesgue measure.

9.19 Does Theorem 9.3.1 hold for discrete-time systems $x(t+1) = Ax(t)+Bu(t)$? Does it hold for discrete-time systems with $A \in \mathbb{F}^{n\times n}$, $B \in \mathbb{F}^{n\times m}$, and $r \in \mathbb{F}[\xi]$, with $\mathbb{F}$ an arbitrary field?

9.20 Consider $\Sigma_{n,m} \cong \mathbb{R}^{n^2+nm}$. Prove that the following classes of systems are generic in the sense that the systems that do not have this property are contained in a proper algebraic variety:

(a) The controllable systems.

(b) The systems (A, B) such that (A, B_k) is controllable for all $k = 1, 2, \ldots, m$; B_k denotes the kth column of B.

(c) The systems with A semisimple.

9.21 Theorem 9.3.1 may leave the impression that since for a controllable system (A, B) the eigenvalues of $A + BN$ can be chosen arbitrarily (in particular, all with arbitrarily large negative real parts), the transient response of

$$\frac{d}{dt}x = (A + BN)x$$

can be made arbitrarily small. This impression is erroneous. In fact, it can be shown that

$$\inf_{N \in \mathbb{R}^{n \times m}} \int_0^\infty \|e^{(A+BN)t}x_0\|^2 dt \tag{9.39}$$

is zero if and only if $x_0 \in \operatorname{im} B$. Thus a fast settling time ($e^{(A+BN)t}x_0$ small for $t \geq \epsilon > 0$ with ϵ small) must be at the expense of a large overshoot ($e^{(A+BN)t}x_0$ large for $0 \leq t \leq \epsilon$).

Consider the system $y + \frac{d^2}{dt^2}y = u$. Write it in state form as

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u.$$

Compute (9.39) for $x_0 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$. Interpret this result in terms of a mass–spring combination in which you are allowed to add friction and to modify the spring constant.

9.22 In Example 9.7.3 we have seen that it is impossible to stabilize the system $\frac{d^2}{dt^2}y = u$ by means of a memoryless (linear or nonlinear) control law $u = N(y)$. The question arises whether this can be done by means of time varying control law.

Consider therefore the differential equation

$$\frac{d^2}{dt^2}y + N(t)y = 0.$$

Prove that whatever N is (but assumed locally integrable), it cannot happen that the solutions with initial conditions $y(0) = 1, \frac{d}{dt}y(0) = 0$ and $y(0) = 0, \frac{d}{dt}y(0) = 1$, both go to zero as $t \to \infty$.

Hint: You may use the following fact from the theory of differential equations. Assume that $\Phi : \mathbb{R} \to \mathbb{R}^{n \times n}$ and $A : \mathbb{R} \to \mathbb{R}^{n \times n}$ satisfy

$$\frac{d}{dt}\Phi(t) = A(t)\Phi(t), \quad \Phi(0) = I.$$

Then $\det \Phi(t) = \exp \int_0^t \operatorname{Tr} A(t')dt'$.

9.23 Consider a partitioned matrix of the form

$$M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix},$$

with $M_{11} \in \mathbb{R}^{n_1 \times n_1}, M_{12} \in \mathbb{R}^{n_1 \times n_2}, M_{22} \in \mathbb{R}^{n_2 \times n_2}$. Prove that $\chi_M(\xi)$ factorizes as $\chi_M(\xi) = \chi_{M_{11}}(\xi)\chi_{M_{22}}(\xi)$. Generalize this to a partitioned matrix of the form

$$M = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix}.$$

9.24 We call a matrix *partially specified* if certain elements are fixed, while the others can be chosen. Denote the fixed elements by $*$s, and those that can be chosen by ?s. The following question arises: Can, for given $*$s, the ?s be chosen such that the resulting matrix (assumed square) has preassigned eigenvalues? Use the pole placement result to obtain necessary and sufficient conditions in terms of the $*$s for the following two cases:

$$\begin{bmatrix} * & * & \cdots & * \\ * & * & \cdots & * \\ \vdots & \vdots & \vdots\!\vdots\!\vdots & \vdots \\ * & * & \cdots & * \\ ? & ? & \cdots & ? \end{bmatrix}, \quad \begin{bmatrix} * & * & \cdots & * & ? \\ * & * & \cdots & * & ? \\ \vdots & \vdots & \vdots\!\vdots\!\vdots & \vdots & \vdots \\ * & * & \cdots & * & ? \end{bmatrix}.$$

Other cases of interest, but beyond the scope of this book, are:

$$\begin{bmatrix} ? & \cdots & ? & * & \cdots & * \\ \vdots & \vdots\!\vdots\!\vdots & \vdots & \vdots & \vdots\!\vdots\!\vdots & \vdots \\ ? & \cdots & ? & * & \vdots & * \\ * & \cdots & * & * & \cdots & * \\ \vdots & \vdots\!\vdots\!\vdots & \vdots & \vdots & \vdots\!\vdots\!\vdots & \vdots \\ * & \cdots & * & * & \cdots & * \end{bmatrix}, \quad \begin{bmatrix} ? & * & \cdots & * \\ * & ? & \cdots & * \\ & & \ddots & \vdots \\ * & \cdots & * & ? \end{bmatrix}.$$

9.25 Consider the following scalar nonlinear systems:

(a) $\frac{d}{dt}x = \sin x + u$.
(b) $\frac{d}{dt}x = xu$.
(c) $\frac{d}{dt}x = u^2$.

For each of these systems, $x^* = 0$, $u^* = 0$ is an equilibrium point. Linearize around this equilibrium. Check the controllability of the resulting linearized systems and of the original nonlinear ones. If the linearized system is controllable, find a linear state feedback law such that the equilibrium $x^* = 0$ becomes asymptotically stable. For the other cases, find a nonlinear control law such that $x^* = 0$ becomes asymptotically stable, or prove that no such (nonlinear) control law exists.

10

Observers and Dynamic Compensators

10.1 Introduction

In Chapter 9 we have seen how feedback control can be applied to a dynamical system when the state is measured. The salient result that we obtained states that with this type of control, stabilization—in fact, pole placement—is always possible for controllable systems.

In real-life applications it is often not feasible to measure the complete state vector. Each measurement requires an additional sensor, and some of the state variables (temperatures inside ovens, concentrations of chemical products, velocities of masses, etc.) may be difficult to measure directly in real time. We shall see, however, that it is not necessary to measure all the state variables in order to use the ideas of the previous chapter for the design of a stabilizing feedback controller. By appropriate signal processing, we are often able to obtain good estimates of all state variables from the measured outputs. The algorithm that performs this signal processing is called an *observer*. The observers that we obtain in this chapter possess many appealing features, in particular, the recursivity of the resulting signal processing algorithm. By this we mean that the state estimate is continuously updated. Once this updating has been done, the past observations can be deleted from the observer memory.

As we have seen in the previous chapter, controllability is the crucial property that enables us to choose the state feedback gains so as to achieve pole placement or stabilization. For observers, it is observability that plays this

role: for observable systems, the state can be deduced from the measured output with error dynamics whose poles can be chosen arbitrarily.

By combining a state observer with a static control law, we subsequently obtain a feedback controller, often called a *compensator*, that processes the measured outputs in order to compute the required control input. We will see that the design of a good feedback compensator requires the combined properties of controllability and observability.

The observer and feedback compensator algorithms that we develop are based on a number of appealing cybernetic principles. The first one is the interaction of an *internal model* and of *error feedback*. This principle states that the estimate of the state can be constructed by implementing the following idea. If the new observations do not give an indication that our current estimate is incorrect, then we let the state estimate evolve according to the model of the plant. The error between the observed output and the expected observed output produces a signal that is fed back in order to correct the state evolution as suggested by the model. Thus the observer consists of an internal model corrected by error feedback. The design of the feedback compensator is based on the combination of two principles: *separation* and *certainty equivalence*. The feedback compensator uses an estimate of the state in order to compute the control action. The observer produces an estimate of the state of the plant. The certainty equivalence principle states that for the control action we proceed as if the estimate were equal to the exact value of the state. Thus the controller acts equivalently as if it were certain of the value of the state. The controller gains that act on the estimate of the state are computed as if this estimate is correct. This is the content at the separation principle: the design of the observer and of the controller gains are carried out separately.

Example 10.1.1 Before plunging into the observer question, let us illustrate the difficulties involved by means of an example. Consider a mass moving under the influence of an external force. For simplicity, assume that the motion is one-dimensional, yielding the behavioral equations

$$M\frac{d^2}{dt^2}q = F,$$

with $M > 0$ the mass, q the position, and F the external force. We know that in this case the state is given by the position combined with the velocity

$$x = \begin{bmatrix} q \\ \dfrac{d}{dt}q \end{bmatrix}.$$

Assume that we can measure the position q and the force F, and that we want to estimate the state x. In other words, we need to estimate the velocity $\frac{d}{dt}q$ from q and F. This sounds easy: just differentiate q. However,

differentiation can be very inaccurate due to measurement noise. In order to see this, assume that q is measured by a device that is influenced by some high-frequency vibration, yielding the measurement $\tilde{q}$ that is the sum of q and a high-frequency signal. It is easy to see that the derivative of $\tilde{q}$ will be a very corrupted version of the derivative of q. So, numerical differentiation is ill-advised (see Exercise 10.1).

Since we don't like differentiation, let us turn to integration. An alternative way of getting hold of $\frac{d}{dt}q$ would be to integrate $\frac{F}{m}$, which equals $\frac{d^2}{dt^2}q$, i.e., use

$$\hat{v}(t) = (\frac{d}{dt}q)(0) + \int_0^t \frac{F(t')}{m}dt' \tag{10.1}$$

as the estimate of the velocity $\frac{d}{dt}q(t)$. Since in (10.1) the (possibly noisy) measurement F is integrated, we can indeed expect a certain noise immunity. As compared to differentiation, there is a different but equally serious problem with the estimate (10.1). It gives a perfectly accurate estimate of $\frac{d}{dt}q$ provided that we know $(\frac{d}{dt}q)(0)$ exactly. However, if the initial condition in (10.1) is taken to be $(\frac{d}{dt}q)(0) + \Delta$ instead of $(\frac{d}{dt}q)(0)$, with $\Delta \neq 0$, then (10.1) gives an estimate that is not even stable, i.e., the estimation error

$$\frac{d}{dt}q(t) - \hat{v}(t)$$

equals Δ for all $t \leq 0$ and does not converge to zero.

The type of observer that we learn to design in this chapter yields the following type of algorithm for obtaining $\hat{v}$, the estimate of $\frac{d}{dt}q$:

$$\frac{d}{dt}z = -z + \frac{F}{m} - q, \quad \hat{v} = z + q. \tag{10.2}$$

Note that $\frac{d}{dt}q - \hat{v}$ is now governed by

$$\frac{d}{dt}(\frac{d}{dt}q - \hat{v}) = -(\frac{d}{dt}q - \hat{v}),$$

yielding

$$(\frac{d}{dt}q - \hat{v})(t) = e^{-t}((\frac{d}{dt}q)(0) - \hat{v}(0)).$$

Hence, even if our estimate $\hat{v}(0)$ of $(\frac{d}{dt}q)(0)$ is inaccurate, we always have

$$\lim_{t\to\infty}(\frac{d}{dt}q - \hat{v})(t) = 0.$$

Further, (10.2) shows (via the variation of constants formula) that both the measurements F and q are integrated in the observer, guaranteeing also a certain noise immunity in addition to good convergence properties. □

10.2 State Observers

In this section we explain the structure of the observer algorithms. In the next section we show how to "tune" the observer, how to choose the gains of the observer. Consider the following *plant*:

$$\frac{d}{dt}x = Ax + Bu \ , \ y = Cx, \tag{10.3}$$

where x is the state, u the input, and y the output. The system parameters are given by the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. We denote the class of systems (10.3) by $\Sigma_{n,m,p}$, where the subscripts denote the number of state, input, and output variables, respectively. Since each element of $\Sigma_{n,m,p}$ is parametrized by a triple of matrices (A, B, C), we can also write $(A, B, C) \in \Sigma_{n,m,p}$. In (10.3) we assume that the external (*manifest*) signals u and y are measured and that we would like to deduce the internal (*latent*) signal x for these measurements. An algorithm that estimates x from u and y is called a (*state*) *observer*. Let us denote the *estimate* of x by $\hat{x}$, and define the *estimation error* as $e := x - \hat{x}$. Thus an observer is a dynamical system with u and y as input, $\hat{x}$ as output, and that makes $e = x - \hat{x}$ small in some sense. In this chapter we focus on the asymptotic behavior of $e(t)$ for $t \to \infty$. The signal flow graph of an observer is shown in Figure 10.1. In Section 5.3.1, we have actually considered the

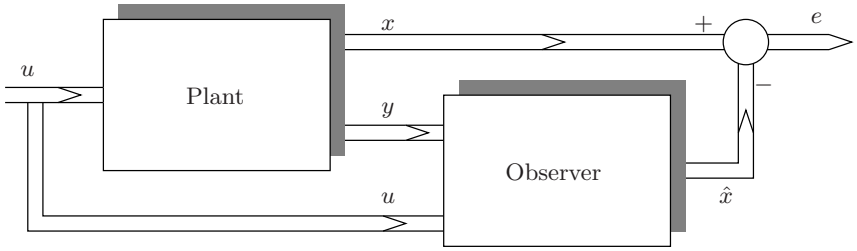


FIGURE 10.1. The signal flow graph of an observer.

problem of deducing the state x from (u, y). In fact, if (A, C) is observable, then knowledge of (u, y) allows x to be reconstructed. Indeed, consider (10.3), and repeatedly differentiate y. Substituting $\frac{d}{dt}x = Ax + Bu$, we obtain

$$\begin{bmatrix} y \\ \frac{d}{dt}y \\ \frac{d^2}{dt^2}y \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}}y \end{bmatrix} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} x + \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ CB & 0 & \cdots & 0 & 0 \\ CAB & CB & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{n-2}B & CA^{n-3}B & \cdots & CB & 0 \end{bmatrix} \begin{bmatrix} u \\ \frac{d}{dt}u \\ \frac{d^2}{dt^2}u \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}}u \end{bmatrix}. \tag{10.4}$$

Now, since the system is observable, the matrix $\mathrm{col}(C, CA, \ldots, CA^{n-1})$ has a left inverse. Premultiplying (10.4) by this left inverse yields an expression of the form

$$x = M_y(\frac{d}{dt})y + M_u(\frac{d}{dt})u, \tag{10.5}$$

with $M_y(\xi) \in \mathbb{R}^{n\times p}[\xi]$ and $M_u(\xi) \in \mathbb{R}^{n\times m}[\xi]$ polynomial matrices that can be computed from (A, B, C), but whose exact values do not matter. This formula shows that if (A, C) is observable, x can indeed be obtained from (u, y). However, (10.5) is not a suitable observer because it implies repeatedly differentiating (u, y), that suffers from the lack of noise immunity discussed in Example 10.1.1.

How then should we choose the equations governing a state observer? The design that we put forward has a very appealing logic. The two central ideas are:

1. the observer contains a copy of the plant, called an *internal model*.

2. the observer is driven by the *innovations*, by the error feedback, that is, by a signal that expresses how far the actual observed output differs from what we would have expected to observe.

This logic functions not unlike what happens in daily life. Suppose that we meet a friend. How do we organize our thoughts in order to deduce his or her mood, or other latent properties, from the observed manifest ones? Based on past experience, we have an "internal model" of our friend in mind, and an estimate of the "associated state" of his/her mood. This tells us what reactions to expect. When we observe an action or hear a response, then this may cause us to update the state of this internal model. If the observed reaction agrees with what we expected from our current estimate, then there is no need to change the estimate. The more the reaction differs from our expectations, the stronger is the need to update. The difference between what we actually observe and what we had expected to observe is what we call the innovations. Thus it is logical to assume that the updating algorithm for the estimate of the internal model is driven by the innovations. We may also interpret the innovations as the *surprise factor*.

Returning to (10.3), it is clear that if our current estimate of the state is $\hat{x}(t)$, then the innovation at time t equals $i(t) = y(t) - C\hat{x}(t)$. Indeed, at time t, we observe $y(t) = Cx(t)$, and on the basis of our estimate of the state, $\hat{x}(t)$, we would have expected to observe $C\hat{x}(t)$. Let us denote the expected observation by $\hat{y}$. Hence $\hat{y} = C\hat{x}$, and $i = y - \hat{y}$. Coupling the

internal model with the innovations leads to the observer equations

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \underbrace{A\hat{x} + Bu}_{\text{internal model}} + \underbrace{Li}_{\text{innovations correction}}, \\ \hat{y} &= C\hat{x}, \\ i &= y - \hat{y}. \end{aligned} \tag{10.6}$$

The structure of this state observer is shown in Figure 10.2. The only matrix

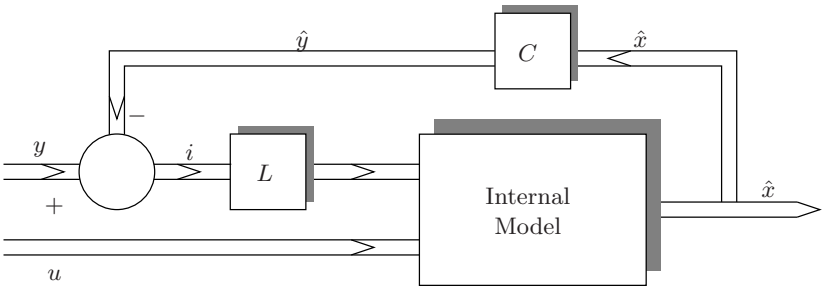


FIGURE 10.2. The structure of the state observer.

in (10.6) that is not given by the system equations (10.3) is the matrix $L \in \mathbb{R}^{n\times p}$. This matrix is called the *observer gain matrix*. It expresses the relative faith that the observer algorithm has in its memory, in the current estimate $\hat{x}$, versus the current observation, y. If the values of the elements of L are small, then the observer gives much weight to the memory $\hat{x}$, and relatively little to the most recent observation y. If L is large, then the situation is the other way around.

10.3 Pole Placement in Observers

In order to capture the role of the observer gain matrix L, let us consider the dynamics of the estimation error $e := x - \hat{x}$. Combining equations (10.3) with (10.6) immediately leads to the following equation for e:

$$\frac{d}{dt}e = (A - LC)e.$$

Note the striking simplicity of this equation. Indeed, as a consequence of the structure of the observer, consisting of an internal model driven by the innovations, the error evolves in a completely autonomous way. Consistent with the definition of system poles introduced in Section 9.3, we call the eigenvalues of the matrix $A - LC$ the *observer poles*, and its characteristic

polynomial $\chi_{A-LC}(\xi)$ the *observer characteristic polynomial*. If we rewrite (10.6) as

$$\frac{d}{dt}\hat{x} = (A - LC)\hat{x} + Bu + Ly, \tag{10.7}$$

then we can see that the observer poles are in fact the poles of the transfer function $(I\xi - A + LC)^{-1}(B\ L)$ from (u, y), the inputs to the observer, to $\hat{x}$, the output of the observer.

Of course, we want that $e(t) \to 0$ as $t \to \infty$; i.e., $A - LC$ must be Hurwitz. However, often we need a certain rate of convergence. This leads to the following question:

What observer pole locations are achievable by choosing the observer gain matrix L?

In linear algebra terms, this question becomes

Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$ be given matrices. What is the set of polynomials $\chi_{A-LC}(\xi)$ obtainable by choosing the matrix $L \in \mathbb{R}^{n\times p}$?

Of course, this question looks like déjà vu: it is completely analogous to the pole location problem studied in Chapter 9. In fact, the result that we shall obtain states that the observer poles can be chosen arbitrarily if and only if the system (10.3) is observable, i.e., if and only if $\operatorname{rank}\operatorname{col}(C, CA, \ldots, CA^{n-1}) = n$. The following result is often called the *observer pole placement theorem*.

Theorem 10.3.1 *Consider the system (10.3). There exists for every real monic polynomial $r(\xi)$ of degree n an observer gain matrix L such that the characteristic polynomial for the observer poles $\chi_{A-LC}(\xi)$ equals $r(\xi)$ if and only if the system (10.3) is observable.*

Proof The easiest way to prove this theorem is by duality. Indeed, by Exercise 5.19, (A, C) is an observable pair if and only if (A^T, C^T) is a controllable pair. Hence by Theorem 9.3.1, for all $r(\xi)$ there exists a matrix $N \in \mathbb{R}^{p\times n}$ such that $\chi_{A^T+C^TN}(\xi) = r(\xi)$ if and only if (A^T, C^T) is controllable. This shows that observability of (A, C) is a necessary condition. To show the converse, note that for any real square matrix M, $\chi_M(\xi) = \chi_{M^T}(\xi)$. Therefore, $\chi_{A+N^TC}(\xi) = \chi_{A^T+C^TN}(\xi)$. Now $L = -N^T$ yields $\chi_{A-LC}(\xi) = r(\xi)$. □

Algorithm 10.3.2 The proof of the above theorem shows that in order to compute an $L \in \mathbb{R}^{n\times p}$ such that $\chi_{A-LC}(\xi) = r(\xi)$, we can immediately apply Algorithm 9.5.1. Substitute in this algorithm A^T for A, C^T for B, and p for m. Compute, using this algorithm, a matrix N such that $\chi_{A^T+C^TN}(\xi) = r(\xi)$. Then $L = -N^T$ gives the desired observer gain matrix. □

Example 10.3.3 Consider a mass–spring–damper combination. The distance of the mass from its equilibrium position under the influence of the external force F is assumed to be governed by the (linear or linearized) behavioral equation

$$Kq + D\frac{d}{dt}q + M\frac{d^2}{dt^2}q = F, \tag{10.8}$$

with K the spring coefficient, D the friction coefficient, and M the mass. Assume that the position q and the external force F are measured. The problem is to build an observer, in particular, to obtain an estimate of the velocity $\frac{d}{dt}q$.

Writing (10.8) in state variable form with $x_1 = q$ and $x_2 = \frac{d}{dt}q$ yields

$$\begin{aligned} \frac{d}{dt}x &= \begin{bmatrix} 0 & 1 \\ -\frac{K}{M} & -\frac{D}{M} \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix} F, \\ q &= \begin{bmatrix} 1 & 0 \end{bmatrix} x. \end{aligned} \tag{10.9}$$

Let us construct an observer that puts both observer poles at $-\lambda$, with $\lambda > 0$ a design parameter. In order to compute the observer gain for the case at hand, it is not necessary to invoke Algorithm 10.3.2. The computation of $L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$ such that

$$\det \begin{bmatrix} \xi + L_1 & -1 \\ \frac{K}{M} + L_2 & \xi + \frac{D}{M} \end{bmatrix} = (\xi + \lambda)^2$$

can be carried out directly, and it yields

$$L_1 = 2\lambda - \frac{D}{M}, \quad L_2 = (\lambda - \frac{D}{M})^2 - \frac{K}{M}.$$

The observer algorithm becomes

$$\begin{aligned} \frac{d}{dt}\hat{x} &= \begin{bmatrix} -2\lambda + \frac{D}{M} & 1 \\ (\lambda - \frac{D}{M})^2 & -\frac{D}{M} \end{bmatrix} \hat{x} + \begin{bmatrix} 2\lambda - \frac{D}{M} \\ (\lambda - \frac{D}{M})^2 - \frac{K}{M} \end{bmatrix} q + \begin{bmatrix} 0 \\ \frac{1}{M} \end{bmatrix} F, \\ \hat{q} &= \hat{x}_1, \quad \frac{\hat{dq}}{dt} = \hat{x}_2. \end{aligned} \tag{10.10}$$

The error dynamics in this example are given by

$$\frac{d}{dt}e = \begin{bmatrix} -2\lambda + \frac{D}{M} & 1 \\ (\lambda - \frac{D}{M})^2 & -\frac{D}{M} \end{bmatrix} e.$$

Note that this problem is only slightly more complicated than the motivational Example 10.1.1. The resulting observer (10.10) is two-dimensional,

while (10.2) was one-dimensional. The possibility of reducing the order of (10.10) is discussed in Section 10.6.

The important feature of (10.10) is that only integrations need to be employed in constructing $\hat{x}$, while guaranteeing (in the noise free case) that $x(t) - \hat{x}(t) \to 0$ as $t \to \infty$: both noise immunity and asymptotic tracking are guaranteed by our observer. □

10.4 Unobservable Systems

Recall from Section 4.6 that the dynamical systems $(A_1, B_1, C_1) \in \Sigma_{n,m,p}$ and $(A_2, B_2, C_2) \in \Sigma_{n,m,p}$ are called *similar* if there exist a nonsingular matrix S such that $A_1 = SA_2S^{-1},\ B_1 = SB_2,\ C_1 = C_2S^{-1}$. Just as with similarity of matrices, or the type of system similarity introduced in Section 9.4.1, similar systems differ only in that the state coordinates are expressed with respect to a different basis. The following lemma provides a canonical form under similarity that puts the observability structure of a system into evidence.

Lemma 10.4.1 *The system (10.3) is similar to a system of the form*

$$\frac{d}{dt}x' = A'x' + B'u\ ;\ y = C'x$$

in which A' and C' have the following structure:

$$A' = \begin{bmatrix} A_{11} & 0 \\ A_{12} & A_{22} \end{bmatrix}, \quad C' = [C_1 \ \ 0], \tag{10.11}$$

with (A_{11}, C_1) observable. All such decompositions lead to matrices A_{22} that have the same characteristic polynomial.

Proof See Corollary 5.3.14, or apply Lemma 9.4.2 to (A^T, C^T), and take the transpose of the result. □

The polynomial $\chi_{A_{22}}(\xi)$ identified in Lemma 10.4.1 is called the *unobservable polynomial* of (10.3), and its roots are called the *unobservable poles*, or often the *unobservable modes*. This lemma allows us to state the following refinement of Theorem 10.3.1.

Theorem 10.4.2 *Consider the system (10.3) and assume that $\chi_0(\xi)$ is its unobservable polynomial. Then there exists an observer gain matrix $L \in \mathbb{R}^{n\times p}$ such that $\chi_{A-LC}(\xi) = r(\xi)$ if and only if $r(\xi)$ is a real monic polynomial of degree n that has $\chi_0(\xi)$ as a factor.*

Proof Apply Theorem 9.6.1 to (A^T, C^T). □

An immediate consequence of Theorem 10.4.2 is that there exists an observer (10.3) such that

$$\lim_{t\to\infty} \hat{x}(t) - x(t) = 0, \tag{10.12}$$

in other words, such that $A - LC$ is Hurwitz, if and only if the unobservable polynomial $\chi_0(\xi)$ of (10.3) is Hurwitz. Actually, from the canonical form (10.11), it follows that this condition is in fact necessary for the existence of any (linear/nonlinear, time- invariant/time-varying, finite-dimensional/infinite-dimensional) observer such that (10.12) holds. Hence $\chi_0(\xi)$ being Hurwitz is a necessary and sufficient condition for the existence of an observer that asymptotically reconstructs the state x. Systems (or, equivalently, pairs (A, C)) that have this property that $\chi_0(\xi)$ is Hurwitz are called *detectable* (see also Section 5.3.2). It follows from the results obtained there that (A, C) is detectable if and only if

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

for all $\lambda \in \mathbb{C}$ with $\operatorname{Re}\lambda > 0$. It is for these systems that the state is asymptotically reconstructible from the observations of u and y.

10.5 Feedback Compensators

We are now ready to discuss the main design algorithm of this book: that of choosing a feedback compensator for the dynamical system (10.3) such that the closed loop system has a desirable transient response. In the case of output measurements, we need to use feedback laws with memory. In other words, rather than having a feedback controller in which the value of the control at time t depends only on the measured output at time t, we use a controller such that the control input also uses the past values of the measured output. Thus, rather than having a memoryless control law of the type $u = Ny$, we generate u from y through a feedback compensator that has a state of its own. This state captures the dependence of u on the past of y. In Chapter 9, we have seen how using a memoryless state feedback law can be used to stabilize a system. This feedback law is called memoryless because the value of the control input at time t depends only on the value of the measured output. In Chapter 9 we assumed that the whole state was measured. However, when only output measurements are available, the situation becomes more involved, and in general it is not possible to stabilize a system (even a controllable one) by a memoryless control law. In order to cope with this difficulty, we use dynamic control laws. Thus

the controllers that we consider themselves have memory, they have their own state. The input to the controller is the measured output of the plant; the output of the controller is the control input to the plant. Since the controller is dynamic, the control input depends on the past observations. Note that we use the terminology memoryless and static as synonymous; and similarly, we use the terms dynamic system, state system, and system with memory as synonymous.

Consider the plant

$$\frac{d}{dt}x = Ax + Bu \ , \ y = Cx \tag{10.13}$$

and the linear time-invariant feedback processor with memory, expressed by the controller state z:

$$\frac{d}{dt}z = Kz + Ly \ , \ u = Mz + Ny, \tag{10.14}$$

with $z : \mathbb{R} \to \mathbb{R}^d$ the state of the feedback processor, and where the matrices $K \in \mathbb{R}^{d\times d}$, $L \in \mathbb{R}^{d\times p}$, $M \in \mathbb{R}^{m\times d}$, and $N \in \mathbb{R}^{m\times p}$ denote the parameter matrices specifying the feedback processor. The controller state dimension $d \in \mathbb{N}$ is called the *order* of the compensator. It is a design parameter. Typically, we want d to be small, since this requires simple logic for the compensator. Note that the memoryless feedback control laws studied in Chapter 9 correspond to feedback compensators of order zero, compensators with an extremely simple logic. However, this limited logic entails high measurement requirements (in Chapter 9 all the state components needed to be measured).

In the plant (10.13), u is the control, and y is the observed output. The feedback processor (10.14), on the other hand, is a dynamical system that has the observations y as input and the control u as output. This reverse input/output structure is characteristic for feedback loops.

By substituting (10.14) in (10.13) we obtain the closed loop system

$$\begin{aligned} &\frac{d}{dt}\begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} A + BNC & BM \\ LC & K \end{bmatrix}\begin{bmatrix} x \\ z \end{bmatrix}, \\ &y = Cx, \quad u = Mz + Ny. \end{aligned} \tag{10.15}$$

If we write this in compact form with $x_\mathrm{e} := \mathrm{col}(x, z)$ (the "*extended state*") and with A_e, C_e, and H_e defined in the obvious way, we obtain the closed loop system equations

$$\frac{d}{dt}x_\mathrm{e} = A_\mathrm{e}x_\mathrm{e} \ , \ y = C_\mathrm{e}x_\mathrm{e} \ , \ u = H_\mathrm{e}x_\mathrm{e}.$$

From this it is clear that the closed loop system is an autonomous dynamical system. We call the eigenvalues of A_e the *closed loop poles* and

$\chi_{A_e}(\xi)$ the *closed loop characteristic polynomial*. Denote the plant (10.13) by $(A, B, C) \in \Sigma_{n,m,p}$ and the feedback processor (10.14) by $(K, L, M, N) \in \Sigma_{d,p,m}$. Note that our notation is a bit sloppy, since we have used the same notation for systems such as (10.13) without a feedthrough term, and for systems such as (10.14) with a feedthrough term. However, this does not cause confusion.

The following question thus arises:

What closed loop pole locations are achievable by choosing (K, L, M, N)?

In linear algebra terms this question becomes,

Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, *and* $C \in \mathbb{R}^{p\times n}$ *be given matrices. Determine the set of polynomials* $\chi_{A_e}(\xi)$ *obtainable by choosing an integer* $d \in \mathbb{N}$ *and matrices* $K \in \mathbb{R}^{d\times d}$, $L \in \mathbb{R}^{d\times p}$, $M \in \mathbb{R}^{m\times d}$, $N \in \mathbb{R}^{m\times p}$, *and where* A_e *is given by (cf. (10.15))*

$$A_e = \begin{bmatrix} A + BNC & BM \\ LC & K \end{bmatrix}. \tag{10.16}$$

The full solution to this problem is unknown at the time of writing. However, we obtain a very useful partial result in the remainder of this chapter.

In Chapter 9 we have seen how to proceed when $C = I$, i.e., when the full state vector is measured. Let

$$u = N'x$$

be a memoryless state feedback control law obtained this way. In Sections 10.2–10.4 we have seen how we can estimate the state x of (10.3) from (u, y). Let

$$\frac{d}{dt}\hat{x} = (A - L'C)\hat{x} + Bu + L'y \tag{10.17}$$

be a suitable observer. Now use the *separation principle* and the *certainty equivalence principle*, introduced in Section 10.1. The separation principle tells us to combine an observer with a state controller and use the same controller gains as in the case in which the state is measured), and the certainty equivalence principle tells us to consider $\hat{x}$ as being exact. This yields the following natural feedback controller:

$$\frac{d\hat{x}}{dt} = (A - L'C)\hat{x} + BN'\hat{x} + L'y, \qquad u = N'\hat{x}. \tag{10.18}$$

This is, of course, a feedback processor like (10.14), with $d = n$, $K = A - L'C + BN'$, $L = L'$, $M = N'$, and $N = 0$. These formulas may seem a bit complicated, but they have been obtained by two very logical steps: a state feedback law and an observer combined by separation and certainty equivalence. The observer (10.17), in turn, was obtained by similar very logical steps: an internal model driven by the innovations as error feedback.

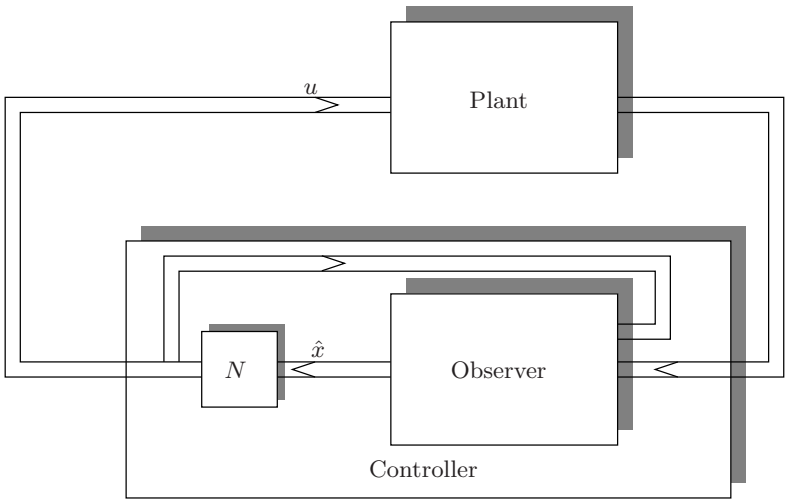


FIGURE 10.3. Dynamic feedback processor.

The resulting dynamic feedback processor is shown in Figure 10.3. Let us analyze the closed loop system obtained by using (10.18) as a feedback processor for (10.13). The closed loop system is governed by

$$\frac{d}{dt}\begin{bmatrix} x \\ \hat{x} \end{bmatrix} = \begin{bmatrix} A & BN' \\ L'C & A - L'C + BN' \end{bmatrix}\begin{bmatrix} x \\ \hat{x} \end{bmatrix}, \tag{10.19}$$

$$u = N'\hat{x}, \qquad y = Cx.$$

We are interested in the characteristic polynomial of the system matrix of (10.19). This polynomial can be easily calculated by transforming (10.19) into a similar system. For that purpose, define as new state vector in the closed loop system

$$\begin{bmatrix} x \\ e \end{bmatrix} = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix}\begin{bmatrix} x \\ \hat{x} \end{bmatrix}.$$

Note that this corresponds to choosing the observer estimation error $e := x - \hat{x}$ instead of $\hat{x}$ as the second part of the closed loop state vector. The dynamics of $\mathrm{col}(x, e)$ are easily derived from (10.19). This yields

$$\begin{bmatrix} \frac{d}{dt}x \\ \frac{d}{dt}e \end{bmatrix} = \begin{bmatrix} A + BN' & -BN' \\ 0 & A - L'C \end{bmatrix}\begin{bmatrix} x \\ e \end{bmatrix}. \tag{10.20}$$

Equation (10.20) shows that the closed loop characteristic polynomial equals the product of $\chi_{A+BN'}(\xi)$ and $\chi_{A-L'C}(\xi)$. Hence, by choosing the feedback compensator based on the separation principle and the certainty equivalence principle, we have obtained a closed loop system whose

characteristic polynomial is the product of the characteristic polynomial $\chi_{A+BN'}(\xi)$ of the state controlled system (using N' as the controller gain matrix) and the observer characteristic polynomial $\chi_{A-L'C}(\xi)$ (using L' as the observer gain matrix).

Combining Theorems 9.3.1 and 10.3.1 immediately leads to the following important result.

Theorem 10.5.1 *Consider the system (10.3) and assume that (A, B) is controllable and that (A, C) is observable. Then for every real monic polynomial $r(\xi)$ of degree $2n$ that is factorizable into two real polynomials of degree n, there exists a feedback compensator (K, L, M, N) of order n such that the closed loop system (10.16) has characteristic polynomial $r(\xi)$.*

Proof Follow the preamble and take $d = n$, $K = A - L'C + BN'$, $L = L'$, $M = N'$, and $N = 0$. Choose N' such that $\chi_{A+BN'}(\xi) = r_1(\xi)$ and L' such that $\chi_{A-L'C}(\xi) = r_2(\xi)$, where $r_1(\xi)$ and $r_2(\xi)$ are real factors of $r(\xi)$ such that $r(\xi) = r_1(\xi)r_2(\xi)$. □

Note that this proof also provides an algorithm for computing the compensator (K, L, M, N). Because of its importance in applications we spell it out in detail.

Algorithm 10.5.2 for pole placement by dynamic compensation:

Data: $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$, with (A, B) controllable and (A, C) observable;

$r(\xi) \in \mathbb{R}[\xi]$ with $r(\xi)$ monic of degree $2n$, factored as $r(\xi) = r_1(\xi)r_2(\xi)$ with $r_1(\xi), r_2(\xi) \in \mathbb{R}[\xi]$ both of degree n.

Required: $K \in \mathbb{R}^{n\times n}$, $L \in \mathbb{R}^{n\times p}$, $M \in \mathbb{R}^{m\times n}$, and $N \in \mathbb{R}^{m\times p}$ such that $\chi_{A_e}(\xi) = r(\xi)$ where A_e is given by (10.16).

Algorithm:

1. Apply Algorithm 9.5.1 with data A, B, and $r_1(\xi)$. Call the result of this computation N_1.
2. Apply Algorithm 9.5.1 with data A^T, C^T, and $r_2(\xi)$. Call the result of this computation N_2.
3. Compute $K = A + N_2^T C + BN_1$, $L = -N_2^T$, $M = N_1$, $N = 0$.

Result: $(K, L, M, N) \in \Sigma_{n,p,m}$ is the desired feedback compensator for the plant $(A, B, C) \in \Sigma_{n,m,p}$.

□

Theorem 10.5.1 is one of the important results in linear system theory. It shows in particular that plants (10.3) that are controllable and observable

can always be stabilized. In fact, the theorem tells us that in a certain sense we can achieve arbitrary pole location in the closed loop system. Note that the required factorizability of $r(\xi)$ into $r_1(\xi)r_2(\xi)$ induces a slight loss of generality in the achievable $r(\xi)$s when n, the dimension of the state of the plant, is odd. However, we remark, without proof, and without entering into details, that it is possible to avoid this restriction and that for any $r(\xi)$ of order $2n$ (not necessarily factorizable into two real factors of order n) a compensator exists. It is easy to sharpen Theorem 10.5.1 so that it yields a necessary and sufficient condition for stabilization. In the theorem that follows, the uncontrollable polynomial of (10.3) should be understood relative to (A, B), while the unobservable polynomial should be understood relative to (A, C).

Theorem 10.5.3 *Consider the system (10.3) and assume that $\chi_u(\xi)$ is its uncontrollable polynomial, and that $\chi_0(\xi)$ is its unobservable polynomial. Then*

(i) For any real monic polynomials $r_1(\xi)$ and $r_2(\xi)$ of degree n such that $r_1(\xi)$ has $\chi_u(\xi)$ as a factor and $r_2(\xi)$ has $\chi_0(\xi)$ as a factor, there exists a feedback compensator (K, L, M, N) of order n such that the closed loop system (10.16) has characteristic polynomial $r(\xi) = r_1(\xi)r_2(\xi)$.

(ii) There exists a feedback compensator (K, L, M, N) as in (10.14) such that the closed loop system (10.15) is asymptotically stable if and only if both $\chi_u(\xi)$ and $\chi_0(\xi)$ are Hurwitz, i.e., if and only if (A, B) is stabilizable and (A, C) is detectable.

Proof (i) Let (K, L, M, N) be as in the proof of Theorem 10.5.1 and use Theorems 9.6.1 and 10.4.2.

(ii) The "if" part follows from (i). To prove the "only if" part, observe first that if (A, B, C) and (A', B', C') are similar systems, then the closed loop systems obtained by using the same compensator (K, L, M, N) are also similar. Hence the achievable closed loop characteristic polynomials remain unchanged after we change the basis in the state space of the plant. Assume therefore that (A, B) is in the canonical form (9.12). Then, whatever be the compensator (K, L, M, N), the component x_2 is always governed by $\frac{d}{dt}x_2 = A_{22}x_2$. This implies that $\chi_u(\xi) = \chi_{A_{22}}(\xi)$ is a factor of $\chi_{A_e}(\xi)$.

Next, observe that if we use the compensator (K^T, M^T, L^T, N^T) on the plant (A^T, C^T, B^T), then the resulting closed loop characteristic polynomial is A_e^T, with A_e given by (10.15). Since $\chi_0(\xi)$ is the uncontrollable polynomial for (A^T, C^T), it follows from the first part of our proof that $\chi_0(\xi)$ is also a factor of $\chi_{A_e^T}(\xi) = \chi_{A_e}(\xi)$.

We conclude that for $\chi_{A_e}(\xi)$ to be Hurwitz, both $\chi_u(\xi)$ and $\chi_0(\xi)$ need to be Hurwitz. □

It follows from the above results that for (10.3) to be stabilizable (in the sense that A_e is Hurwitz) by means of a dynamic compensator (K, L, M, N), it is necessary and sufficient that (A, B) be stabilizable and (A, C) detectable. Actually, it is easily seen that this is, in fact, a necessary condition for the existence of any (linear/nonlinear, time-invariant/time-varying, finite-dimensional/infinite-dimensional) stabilizing feedback compensator.

Example 10.5.4 Consider the equation of a pendulum as in Example 9.1.3 with an external torque:

$$ML^2 \frac{d^2}{dt^2}\phi + MLg \sin \phi = T,$$

where M is the mass of the pendulum, L its length, g the constant of gravity, ϕ the angle, and T the torque. Consider the following equilibria:

1. $\phi^* = 0$ and $T^* = 0$;
2. $\phi^* = \pi$ and $T^* = 0$.

Linearizing around these equilibria yields respectively:

1. $\dfrac{d^2}{dt^2}\Delta_\varphi + \dfrac{g}{L}\Delta_\varphi = \dfrac{1}{ML^2}\Delta_T$;
2. $\dfrac{d^2}{dt^2}\Delta_\varphi - \dfrac{g}{L}\Delta_\varphi = \dfrac{1}{ML^2}\Delta_T$.

Writing state equations with $x_1 = \Delta_\varphi$ and $x_2 = \frac{d}{dt}\Delta_\varphi$, $u = \Delta_T$, and $y = \Delta_\varphi$, yields

1. $\dfrac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -g/L & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix} u, \;\; y = \begin{bmatrix} 1 & 0 \end{bmatrix} x;$
2. $\dfrac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ g/L & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix} u, \;\; y = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$

It is easy to verify that both these systems are controllable and observable. The first linearized system is stable but not asymptotically stable. Actually, with $T = 0$, the first equilibrium point is stable but not asymptotically stable (in the linear as well as in the nonlinear system). With $T = 0$, the second equilibrium is unstable (in the linear as well as in the nonlinear system). Hence, in order to stabilize (in the sense of achieving asymptotic stability), control has to be exercised for either of these equilibria. Further, Example 9.7.3 shows that a successful controller has to be dynamic.

Let us first consider the first equilibrium point. The motions without control ($T = 0$) are periodic motions with period $2\pi\sqrt{\frac{L}{g}}$. Let us strive for a closed loop behavior with a settling time of the same order of magnitude as the period of the oscillation of the uncontrolled system. Since the plant is of order two, we will, following the theory that we have developed, design a compensator also of order two, and obtain a closed loop system of order four. Choose $(-1 \pm \frac{i}{2})\sqrt{\frac{g}{L}}$ to be the controller poles and $-\frac{1}{2}\sqrt{\frac{g}{L}}$, $-\sqrt{\frac{g}{L}}$ to be the observer poles. Hence $r_1(\xi) = \frac{5}{4}\frac{g}{L} + 2\sqrt{\frac{g}{L}}\xi + \xi^2$ and $r_2(\xi) = \frac{1}{2}\frac{g}{L} + \frac{3}{2}\sqrt{\frac{g}{L}}\xi + \xi^2$.

In order to compute the controller gain matrix, we should choose a $\begin{bmatrix} N_1' & N_2' \end{bmatrix} \in \mathbb{R}^{1\times 2}$ such that the characteristic polynomial of

$$\begin{bmatrix} 0 & 1 \\ \frac{-g}{L} + \frac{N_1'}{ML^2} & \frac{N_2'}{ML^2} \end{bmatrix}$$

is $r_1(\xi)$. For the case at hand, there is no need to invoke Algorithm 9.5.1 in order to figure out N_1' and N_2'. The characteristic polynomial of the above matrix is

$$\frac{g}{L} - \frac{N_1'}{ML^2} - \frac{N_2'}{ML^2}\xi + \xi^2.$$

Hence $N_1' = -\frac{1}{4}MLg$, $N_2' = -2ML\sqrt{gL}$. In order to compute the observer gain matrix, we should choose $\begin{bmatrix} L_1' \\ L_2' \end{bmatrix} \in \mathbb{R}^2$ such that the characteristic polynomial of

$$\begin{bmatrix} -L_1' & 1 \\ -\frac{g}{L} - L_2' & 0 \end{bmatrix} \tag{10.21}$$

is $r_2(\xi)$. The characteristic polynomial of (10.21) is $\frac{g}{L} + L_2' + L_1'\xi + \xi^2$. Hence $L_1' = \frac{3}{2}\sqrt{\frac{g}{L}}$, $L_2' = -\frac{1}{2}\frac{g}{L}$. The resulting observer is given by

$$\frac{d}{dt}\hat{x} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L} & 0 \end{bmatrix}\hat{x} + \begin{bmatrix} \frac{3}{2}\sqrt{\frac{g}{L}} \\ -\frac{1}{2}\frac{g}{L} \end{bmatrix}(y - \hat{x}_1) + \begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix}u.$$

Equivalently,

$$\frac{d}{dt}\hat{x} = \begin{bmatrix} -\frac{3}{2}\sqrt{\frac{g}{L}} & 1 \\ -\frac{1}{2}\frac{g}{L} & 0 \end{bmatrix}\hat{x} + \begin{bmatrix} \frac{3}{2}\sqrt{\frac{g}{L}} \\ -\frac{1}{2}\frac{g}{L} \end{bmatrix}y + \begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix}u.$$

The resulting feedback compensator, using equations (10.19), is given by

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \begin{bmatrix} -\frac{3}{2}\sqrt{\frac{g}{L}} & 1 \\ -\frac{3}{4}\frac{g}{L} & -2\sqrt{\frac{g}{L}} \end{bmatrix}\hat{x} + \begin{bmatrix} \frac{3}{2}\sqrt{\frac{g}{L}} \\ -\frac{1}{2}\frac{g}{L} \end{bmatrix}y, \\ u &= \begin{bmatrix} -\frac{1}{4}MLg & -2ML\sqrt{gL} \end{bmatrix}\hat{x}. \end{aligned}$$

A similar calculation for the second, the unstable, equilibrium leads to

$$\begin{aligned} N_1' &= -\tfrac{9}{4}MLg, \qquad N_2' = -2ML\sqrt{gL}, \\ L_1' &= \tfrac{3}{2}\sqrt{\tfrac{g}{L}}, \qquad L_2' = \tfrac{3}{2}\tfrac{g}{L}. \end{aligned}$$

The observer is now

$$\frac{d}{dt}\hat{x} = \begin{bmatrix} -\frac{3}{2}\sqrt{\frac{g}{L}} & 1 \\ -\frac{1}{2}\frac{g}{L} & 0 \end{bmatrix}\hat{x} + \begin{bmatrix} \frac{3}{2}\sqrt{\frac{g}{L}} \\ \frac{3}{2}\frac{g}{L} \end{bmatrix} y + \begin{bmatrix} 0 \\ \frac{1}{ML^2} \end{bmatrix} u,$$

and the resulting feedback compensator becomes

$$\begin{aligned} \frac{d\hat{x}}{dt} &= \begin{bmatrix} -\frac{3}{2}\sqrt{\frac{g}{L}} & 1 \\ -\frac{11}{4}\frac{g}{L} & -2\sqrt{\frac{g}{L}} \end{bmatrix}\hat{x} + \begin{bmatrix} \frac{3}{2}\sqrt{\frac{g}{L}} \\ \frac{3}{2}\frac{g}{L} \end{bmatrix} y, \\ u &= \begin{bmatrix} -\frac{9}{4}MLg & -2ML\sqrt{gL} \end{bmatrix}\hat{x}. \end{aligned}$$

□

10.6 Reduced Order Observers and Compensators

Let us take a new look at the design of a state observer as discussed in Section 10.1. The observer

$$\frac{d}{dt}\hat{x} = A\hat{x} + Bu + L(y - C\hat{x})$$

has the curious property that it estimates even the output, i.e., the components Cx of the state that are directly observed are reestimated in the observer. For instance, in Example 10.3.3, we constructed the observer (10.10) that computes $\hat{x}_1$ as an estimate of q even though q was observed. The question arises whether it is possible to avoid this inefficiency and design an observer that yields an estimate $\hat{x}$ of x with at least $C\hat{x} = y = Cx$. This is indeed possible. The order of the resulting observer will actually be smaller than the order of the plant. We briefly outline how such a *reduced order observer* may be obtained. Assume that C has full column rank. Note that this assumption entails no real loss of generality. For otherwise, simply consider the image of C as the output space. If C has full column rank, then the system $(A, B, C) \in \Sigma_{n,m,p}$ is similar to one of the form

$$A' = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B' = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, C' = [0 \quad I]. \tag{10.22}$$

Assume that the basis in the state space was chosen such that $(A, B, C) \in \Sigma_{n,m,p}$ is in this canonical form (10.22). Partition x conformably as $\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$.

Then the behavioral equations are given by

$$\begin{aligned} \frac{d}{dt}x_1 &= A_{11}x_1 + A_{12}x_2 + B_1u, \\ \frac{d}{dt}x_2 &= A_{21}x_1 + A_{22}x_2 + B_2u, \\ y &= x_2. \end{aligned} \tag{10.23}$$

These equations can be rewritten as

$$\frac{d}{dt}x_1 = A_{11}x_1 + A_{12}y + B_1u, \tag{10.24}$$

$$\frac{d}{dt}y = A_{21}x_1 + A_{22}y + B_2u. \tag{10.25}$$

We want to obtain an observer that estimates x_1. Together with $y = x_2$, which is measured directly, this provides an estimate for the full state vector x. Equation (10.25) shows the information about x_1 that is present in the observations (u, y). It tells us that we can basically consider

$$A_{21}x_1 = \frac{d}{dt}y - A_{22}y - B_2u \tag{10.26}$$

as being directly observed. Now, it is easy to show that (A, C) is observable if and only if (A_{11}, A_{21}) is (see Exercise 10.13). Use equation (10.24) and the internal model/innovation idea explained in Section 10.2 in order to obtain the following observer for x_1:

$$\frac{d}{dt}\hat{x}_1 = A_{11}\hat{x}_1 + A_{12}y + B_1u + L_1(A_{21}x_1 - A_{21}\hat{x}_1).$$

Equivalently,

$$\frac{d}{dt}\hat{x}_1 = A_{11}\hat{x}_1 + A_{12}y + B_1u + L_1(\frac{d}{dt}y - A_{22}y - B_2u - A_{21}\hat{x}_1). \tag{10.27}$$

It is easily verified that the estimation error $e_1 = x_1 - \hat{x}_1$ is then governed by

$$\frac{d}{dt}e_1 = (A_{11} - L_1A_{21})e_1. \tag{10.28}$$

Consequently, by Theorem 10.3.1, if (A, C) is observable (equivalently, if (A_{11}, A_{21}) is observable), there exists for any desired real monic polynomial of degree $(n - \operatorname{rank} C)$, a matrix L_1 such that the characteristic polynomial $\chi_{A_{11}-L_1A_{21}}$ is equal to this polynomial.

As it stands, equation (10.27) has a serious drawback as a dynamic algorithm for the estimation of x_1, since it uses the derivative $\frac{d}{dt}y$ of the observations on the right-hand side of the equation. Differentiating observations is not a good idea in applications, since differentiation has the tendency

to amplify noise, particularly the all-too-common high-frequency noise (see Exercise 10.1). However, it is possible to modify (10.27) so as to avoid differentiation. Introduce therefore

$$v = \hat{x}_1 - L_1 y$$

and rewrite (10.28) to obtain

$$\begin{aligned} \frac{dv}{dt} &= A_{11}(v + L_1 y) + A_{12} y + B_1 u \\ &\quad -L_1(A_{22} y + B_2 u + A_{21}(v + L_1 y)). \end{aligned}$$

This yields the *reduced-order observer*

$$\begin{aligned} \frac{dv}{dt} &= (A_{11} - L_1 A_{21})v + (B_1 - L_1 B_2)u \\ &\quad +(A_{11} L_1 + A_{12} - L_1 A_{22} - L_1 A_{21} L_1)y, \end{aligned} \tag{10.29}$$

$$\begin{aligned} \hat{x}_1 &= v + L_1 y, \\ \hat{x}_2 &= y. \end{aligned} \tag{10.30}$$

Of course, since $y = x_2$, the estimate $\hat{x}_2$ is error free, while the error $e_1 = x_1 - \hat{x}_1$ is governed by (10.28). Whence (10.29, 10.30) defines an observer of dimension $n-$ dim(C) that estimates x_2 error free, and x_1 with a preassigned observer error characteristic polynomial.

This reduced-order observer can also be used in feedback controllers. The result obtained in Theorem 10.5.1 shows that for nth-order systems (10.3) that are controllable and observable, there exists, for any preassigned monic polynomial of degree $2n$, an nth-order compensator such that the closed loop characteristic polynomial is equal to this preassigned polynomial at least when it is factorizable into the product of two real polynomials of degree n. If we measure the complexity of a dynamical system by its dynamic order, i.e., by the dimension of the state space, then the construction of Theorem 10.5.1 yields a feedback compensator whose complexity is equal to that of the plant. Most classical control schemes use much simpler control algorithms. Often, in fact, a simple proportional output gain is used. Nowadays, the use of simple controllers is less important, since microprocessors provide an inexpensive and reliable way of implementing complex control algorithms with a high degree of intelligence. The question nevertheless arises whether this complexity is really necessary. In particular, is it possible to obtain a lower order compensator when the number of actuators (i.e., the dimension m of the control input space of the plant (10.3)) or the number of sensors (i.e., the dimension p of the measurement output space of the plant (10.3)) is increased?

One way of obtaining a stabilizing feedback compensator of lower order is to use a reduced-order observer combined with the separation and the certainty equivalence principles used before. Thus it is easy to prove (see Exercise

10.14) that if (A, B) is controllable and (A, C) is observable, the reduced-order observer of Section 10.6 yields a controller of order $(n - \operatorname{rank} C)$ that actually achieves the desired closed loop characteristic polynomial, provided that it is factorizable into a real polynomial of degree n and one of degree $(n-\operatorname{rank} C)$. Using the same idea on the dual system (see Exercise 10.15), we can obtain a controller of order $(n - \operatorname{rank}\ B)$. Such feedback compensators are called *reduced order compensators.*

Example 10.6.1 Return to Example 10.1.1. Assume that we wish to stabilize the point mass in the equilibrium state $q = 0$ using a first-order position output controller. Let $(1+\xi)(1+\xi+\xi^2)$ be the desired closed loop characteristic polynomial.

It is easily verified that the state feedback law

$$F = -M(q + \frac{d}{dt}q) \tag{10.31}$$

achieves the closed loop characteristic polynomial $(1 + \xi + \xi^2)$. However, the control law (10.31) uses the velocity $\frac{d}{dt}q$, which is not measured. We need therefore to replace in this expression $\frac{d}{dt}q$ by its estimate obtained from an observer.

We can estimate $\frac{d}{dt}q$ using a reduced-order observer. The equations of motion written in state form with the state chosen as $x = \operatorname{col}(\frac{d}{dt}q, q)$, so that (10.22) is satisfied, are

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} x + \begin{bmatrix} \frac{1}{M} \\ 0 \end{bmatrix} F, \quad q = \begin{bmatrix} 0 & 1 \end{bmatrix} x. \tag{10.32}$$

Equations (10.24) and (10.26) for the case at hand are

$$\frac{d}{dt}x_1 = \frac{1}{M}F; \quad x_1 = \frac{d}{dt}q. \tag{10.33}$$

The observer (10.27) corresponding to the characteristic polynomial $(1+\xi)$ becomes

$$\frac{d}{dt}\hat{x}_1 = \frac{1}{M}F + \frac{d}{dt}q - \hat{x}_1. \tag{10.34}$$

Rewriting in terms of $v = \hat{x}_1 - q$, so as to avoid differentiation, yields

$$\frac{d}{dt}v = v - q + \frac{1}{M}F; \quad \hat{x}_1 = v + q. \tag{10.35}$$

Since $x_1 = \frac{d}{dt}q$, this yields an estimate $\hat{x}_1$ for $\frac{d}{dt}q$ from q and F that does not require differentiation. Using this observer with the control law (10.31) yields the controller

$$\begin{aligned} \frac{d}{dt}v &= v - q + \frac{F}{m}, \\ \hat{x} &= v + q, && \text{(10.36a)} \\ F &= -M(q + \hat{x}_1). && \text{(10.36b)} \end{aligned}$$

After substitution, this expression can be simplified to

$$\begin{aligned} \frac{d}{dt}v &= -2v - 3q, \\ F &= -2q - v. \end{aligned} \tag{10.37a}$$

This controller can thus be written as

$$2F + \frac{d}{dt}F = -M(q + 2\frac{d}{dt}q).$$

Combining this with

$$M\frac{d^2}{dt^2}q = F,$$

we obtain a closed loop characteristic polynomial that is indeed $(1 + 2\xi + 2\xi^2 + \xi^3)$, as we set out to make it be. □

The following theorem (which we merely state, without proof) shows that Theorem 10.5.1 can be improved in such a way that it gives a controller of lower complexity, which at the same time avoids the factorizability of $r(\xi)$ into the product of two real factors. Consider the system (10.3), and assume that it is controllable and observable. Define its *controllability index* as

$$\kappa = \min\{k \in \mathbb{N} \mid \operatorname{rank}[B, AB, \ldots, A^{k-1}B] = n\}$$

and its *observability* index as

$$\nu = \min\{k \in \mathbb{N} \mid \operatorname{rank}\operatorname{col}\,[C, CA, \ldots, CA^{k-1}] = n\}.$$

Theorem 10.6.2 *Consider the plant (10.3) and assume that it is controllable and observable. Let κ be its controllability index and ν its observability index. Then for any $n' \geq \min(\kappa, \nu) - 1$, and any real monic polynomial $p(\xi)$ of degree $n + n'$, there exists a feedback compensator (10.14) such that the closed loop characteristic polynomial $\chi_{A_{\mathrm{cl}}}(\xi)$ equals $p(\xi)$, with A_{e} defined by (10.16).*

Proof See [64]. □

10.7 Stabilization of Nonlinear Systems

The results of Theorem 10.5.3, combined with the stability properties of the linearized system, lead to stabilization of equilibrium points of nonlinear systems, using output feedback. The idea is completely analogous to what has been explained in Section 9.7. Consider the nonlinear system

$$\begin{aligned} \tfrac{d}{dt}x &= f(x, u), \\ y &= h(x, u), \end{aligned} \tag{10.38}$$

with equilibrium point (x^*, u^*, y^*); i.e., assume that $f(x^*, u^*) = 0$ and $h(x^*, u^*) = y^*$. In Section 4.7 we have introduced the linearization of this system around this equilibrium point. This leads to the system

$$\begin{array}{rcl} \dfrac{d\Delta_x}{dt} & = & A\Delta_x + B\Delta_u, \\ \Delta_y & = & C\Delta_x + D\Delta_u, \end{array} \tag{10.39}$$

with (A, B, C, D) computed from f and h as given by formulas (4.60) and (4.61). The relation between these two systems is that (10.39) describes the behavior of (10.38) in the neighborhood of the equilibrium (x^*, u^*, y^*), up to first-order terms in $x - x^*$, $u - u^*$, and $y - y^*$.

In Section 7.5 we learned that for the nonlinear system

$$\frac{d}{dt}x = \tilde{f}(x)$$

the equilibrium point x^* (hence $\tilde{f}(x^*) = 0$) is asymptotically stable if the linearized system

$$\frac{d}{dt}\Delta_x = \tilde{f}'(x^*)\Delta_x$$

is asymptotically stable, i.e., if the Jacobi matrix $\tilde{f}'(x^*)$ is a Hurwitz matrix. See (7.26) for the definition of $\tilde{f}'$.

This result and the theory developed in Section 10.6 allows us to stabilize a nonlinear system around an equilibrium. Indeed, assume that (10.39) is controllable and observable. Then by Theorem 10.5.1 there exist a feedback compensator

$$\begin{array}{rcl} \dfrac{d}{dt}z & = & Kz + L\Delta_y, \\ \Delta_u & = & Mz + N\Delta_y \end{array}$$

such that the closed loop (linear) system is asymptotically stable, that is, such that

$$A_{\mathrm{e}} = \begin{bmatrix} A + BNC & BM \\ LC & K \end{bmatrix}$$

is a Hurwitz matrix. Now use this linear control law for the nonlinear system, using the interpretation $\Delta_y \approx y - y^*$ and $\Delta_u \approx u - u^*$. This leads to the control law

$$\begin{array}{rcl} \dfrac{d}{dt}z & = & Kz + L(y - y^*), \\ u & = & u^* + Mz + N(y - y^*). \end{array} \tag{10.40}$$

The resulting nonlinear closed loop system is given by

$$\begin{aligned} \frac{d}{dt}x &= f(x,u), \\ \frac{d}{dt}z &= Kz + L(h(x,u) - y^*), \\ u &= u^* + Mz + N(h(x,u) - y^*), \\ y &= h(x,u). \end{aligned} \tag{10.41}$$

Obviously, $x = x^*, z^* = 0, u = u^*, y = y^*$ defines an equilibrium point of (10.41). Also, it is easy to verify that the linearization of (10.41) around this equilibrium point equals

$$\frac{d}{dt}\begin{bmatrix} \Delta_x \\ \Delta_z \end{bmatrix} = A_{\rm e}\begin{bmatrix} \Delta_x \\ \Delta_z \end{bmatrix}.$$

Since $A_{\rm e}$ is Hurwitz, this implies, by Theorem 7.5.2, that the control law (10.40) stabilizes the equilibrium point x^*, u^*, y^* of the nonlinear system (10.38).

The condition of controllability and observability of the linearized system (10.39) is more serious than may appear at first glance. Often, in fact, the nonlinear terms are essential for achieving controllability and observability.

10.8 Control in a Behavioral Setting

10.8.1 Motivation

The purpose of this section is to briefly introduce a novel way of looking at controllers and at control problems. In the preface we explained that it is customary to view a controller as a signal processor that processes measured outputs in order to compute control inputs. In Chapter 9 (see in particular Section 9.1) and in the previous sections of Chapter 10, we have used this very idea of intelligent control, of a controller viewed as a signal processor, in order to obtain feedback controllers that stabilize linear systems or equilibrium points of nonlinear systems. The signal flow graph underlying such controllers is shown in Figure 10.4. The controller processes the sensor outputs in order to compute the actuator inputs that control the plant. In this section we look at controllers from a different vantage point: instead of considering a controller as a signal processor, we view a controller as a dynamical system that is interconnected to the plant. Before this interconnection takes place, the trajectories generated by the plant variables are constrained only to belong to the behavior of the plant. However, when the controller is put into place, after the interconnection, the plant variables are constrained to obey the laws of *both* the plant and the

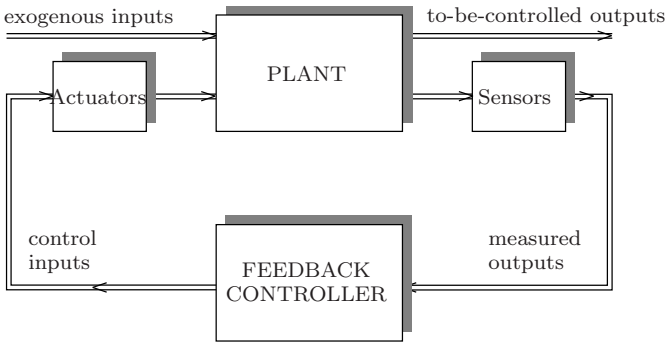


FIGURE 10.4. Intelligent control.

controller. In this way we can hope that the controller retains from all the trajectories in the plant behavior only desirable ones, and rejects those that are not. The idea of controller interconnection is illustrated in Figure 10.5.

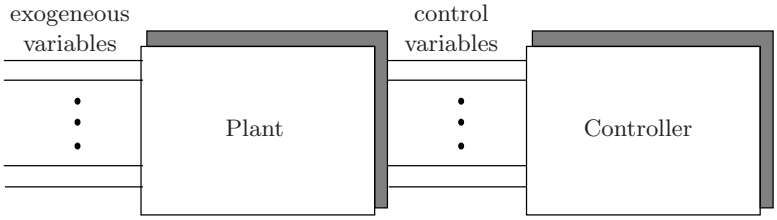


FIGURE 10.5. Controller interconnection.

In this figure the control terminals mean the variables that are available to the controller to interface with. The remaining terminals signify exogenous variables that are not available to the controller to interact with, although the controller of course influences them indirectly through the plant.

Example 10.8.1 In order to focus these ideas, we now analyze a very widespread automatic control mechanism, namely the traditional devices that ensure the automatic closing of doors. A typical such mechanism is schematically shown in Figure 10.6. This device consists of a spring in order to force the closing of the door and a damper in order to make sure that it closes gently. These mechanisms often have considerable weight, so that their mass cannot be neglected as compared to the mass of the door itself. We model such a mechanism as a mass–spring–damper combination, as shown in Figure 10.7. Neglecting friction in the hinges, we model the door as a free mass M' on which two forces are acting. The first force, F_c, is the force exerted by the door-closing device, while the second force, F_e, is an exogenous force (exerted, for example, by a person pushing the door in

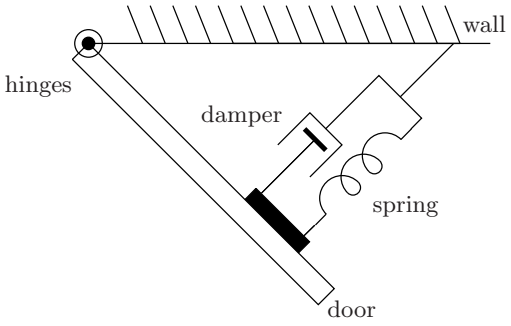


FIGURE 10.6. A door-closing mechanism.

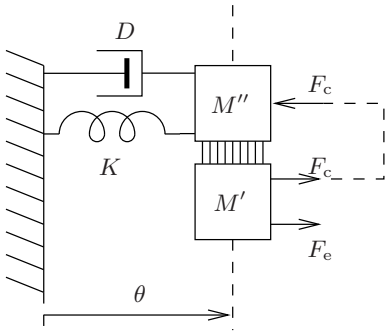


FIGURE 10.7. A mass–spring–damper representation.

order to open it). The equation of motion for the door becomes

$$M' \frac{d^2}{dt^2}\theta = F_c + F_e, \tag{10.42}$$

where θ denotes the opening angle of the door and M' its mass. The door-closing mechanism, modeled as a mass–spring–damper combination, yields

$$K\theta + D\frac{d}{dt}\theta + M'' \frac{d^2}{dt^2}\theta = -F_c. \tag{10.43}$$

Here M'' denotes the mass of the door-closing mechanism, D its damping coefficient, and K its spring constant. Combining (10.42) and (10.43) leads to

$$K\theta + D\frac{d}{dt}\theta + (M' + M'') \frac{d^2}{dt^2}\theta = F_e. \tag{10.44}$$

In order to ensure proper functioning of the door-closing device, the designer can to some extent choose M'', D, and K (all of which must, for physical reasons, be positive). The desired response requirements are small

overshoot (to avoid banging of the door), fast settling time, and a not-too-low steady-state gain from F_e to θ (in order to avoid having to use an excessive force when opening the door). A good design is achieved by choosing a light mechanism (M'' small) with a reasonably strong spring (K large), but not too strong, so as to avoid having to exert excessive force in order to open the door, and with the value of D chosen so as to achieve slightly less than critical damping (see Section 8.5.2 for an analysis of second-order systems such as (10.44)). □

10.8.2 Control as interconnection

In this section we describe mathematically how one can view control as the interconnection of a plant and a controller. We do this first very generally, in the context of the behavioral approach to dynamical systems introduced in Chapter 1. Recall that a *dynamical system* Σ is defined as a triple, $\Sigma = (\mathbb{T}, \mathbb{W}, \mathfrak{B})$, with $\mathbb{T} \subseteq \mathbb{R}$ the *time-axis*, $\mathbb{W}$ a set called the *signal space*, and $\mathfrak{B} \subseteq \mathbb{W}^{\mathbb{T}}$ the *behavior*. Thus $\mathbb{T}$ denotes the set of time instances relevant to the dynamical system under consideration. In the present section we exclusively deal with continuous-time systems with $\mathbb{T} = \mathbb{R}$. The signal space denotes the set in which the time trajectories that the system generates take on their values. The prescription of the behavior $\mathfrak{B}$ can, as we have seen, occur in many different ways. In this book we have mainly studied situations in which the behavior is defined through the solution set of a system of differential equations.

Let $\Sigma_1 = (\mathbb{T}, \mathbb{W}, \mathfrak{B}_1)$ and $\Sigma_2 = (\mathbb{T}, \mathbb{W}, \mathfrak{B}_2)$ be two dynamical systems with the same time-axis and the same signal space. The *interconnection of* Σ_1 *and* Σ_2, denoted by $\Sigma_1 \wedge \Sigma_2$, is defined as $\Sigma_1 \wedge \Sigma_2 := (\mathbb{T}, \mathbb{W}, \mathfrak{B}_1 \cap \mathfrak{B}_2)$. The behavior of the interconnection $\Sigma_1 \wedge \Sigma_2$ consists simply of those trajectories $w : \mathbb{T} \to \mathbb{W}$ that are compatible with the laws of both Σ_1 (i.e., w belongs to $\mathfrak{B}_1$) and of Σ_2 (i.e., w belongs also to $\mathfrak{B}_2$). Thus in the interconnected system, the trajectories that can be generated must be *acceptable* to both Σ_1 and Σ_2. The control problem can now be described as follows. We proceed

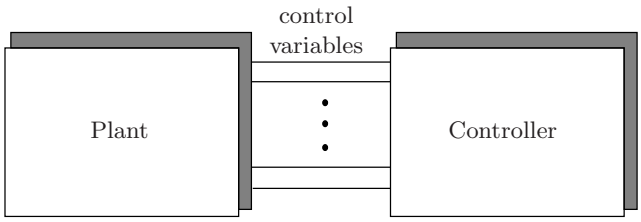


FIGURE 10.8. Controller interconnection without latent variables.

from the mental picture shown in Figure 10.8 but without the exogenous

variables. Actually, the fact that the controller can act only on certain variables, the control variables, can be taken into consideration by defining the set of admissible controllers appropriately. Assume that the *plant*, a dynamical system $\Sigma_p = (\mathbb{T}, \mathbb{W}, \mathfrak{B}_{\rm p})$, is given. Let $\mathfrak{C}$ be a family of dynamical systems, all with $\mathbb{T}$ as common time axis and $\mathbb{W}$ as common signal space. We call $\mathfrak{C}$ the set of *admissible controllers*. An element $\Sigma_c \in \mathfrak{C}$, $\Sigma_c = (\mathbb{T}, \mathbb{W}, \mathfrak{B}_c)$, is called an *admissible controller*. The interconnected system $\Sigma_p \wedge \Sigma_c$ is called the *controlled system*. The controller Σ_c should be chosen so as to make sure that $\Sigma_p \wedge \Sigma_c$ meets certain specifications. The problem of control theory is *first*, to describe the set of admissible controllers; *second*, to describe what desirable properties the controlled system should have; and, *third*, to find an admissible controller Σ_c such that $\Sigma_p \wedge \Sigma_c$ has these desired properties.

We now specialize this definition for linear time-invariant differential systems. Assume that $R(\xi) \in \mathbb{R}^{g \times q}[\xi]$ and that the plant is described by

$$R(\frac{d}{dt})w = 0. \tag{10.45}$$

This induces the dynamical system $\Sigma_R = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B}_R)$ with $\mathfrak{B}_R$ the set of weak solutions of (10.45).

Assume that the class of admissible controllers consists of the set of linear time-invariant differential systems. Thus an admissible controller is described through a polynomial matrix $C(\xi) \in \mathbb{R}^{g' \times q}[\xi]$ by

$$C(\frac{d}{dt})w = 0. \tag{10.46}$$

Let $\Sigma_C = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B}_c)$ be the dynamical system induced by (10.46). Thus Σ_C is the controller. The controlled system is hence given by $\Sigma_R \wedge \Sigma_C = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B}_R \cap \mathfrak{B}_C)$ and is obviously described by the combined equations (10.45) and (10.46), i.e., by

$$\begin{bmatrix} R(\frac{d}{dt}) \\ C(\frac{d}{dt}) \end{bmatrix} w = 0. \tag{10.47}$$

The problem then is to find, for a given $R(\xi)$, a $C(\xi)$ such that the controlled system (10.47) has certain desired properties.

Example 10.8.2 (Example 10.8.1 revisited) In Example 10.8.1 we have $q = 3$, the variables involved being $\theta, F_{\rm c}$, and $F_{\rm e}$. The plant is described by (10.42). Hence $g = 1$, and $R(\xi) = \begin{bmatrix} M'\xi^2 & -1 & -1 \end{bmatrix}$. The controller is given by (10.43). Hence $g' = 1$, and $C(\xi) = \begin{bmatrix} K + D\xi + M''\xi^2 & 1 & 0 \end{bmatrix}$. The closed loop is hence specified by the combination of (10.42) and (10.43). After elimination (in the sense of Chapter 6) of $F_{\rm c}$, this yields (10.44) as the

equation governing the manifest behavior of the variables (θ, F_e) in the controlled system.

What are, in this example, desirable properties of the controlled system? To begin with, F_e should remain free: it should still be possible to exert an arbitrary exogenous force $F_e \in \mathfrak{L}_1^{loc}(\mathbb{R}, \mathbb{R})$ on the controlled system. From the result of Section 3.3, it follows that F_e is indeed free in (10.44). Other desirable properties are asymptotic stability, slightly less than critical damping, and a small steady-state gain $\frac{1}{K}$ from F_e to θ. □

10.8.3 Pole placement

We now develop the problem of pole placement by means of controllers specified as interconnections, as explained in Section 10.8.2. For simplicity we only consider the case of a plant with two variables constrained by one equation (the single-input/single-output case, if you like to view it that way). The plant is thus assumed to be governed by

$$a(\frac{d}{dt})w_1 + b(\frac{d}{dt})w_2 = 0, \tag{10.48}$$

with $a(\xi), b(\xi) \in \mathbb{R}[\xi]$. We assume that $a(\xi)$ and $b(\xi)$ are not both zero. The set of admissible controllers consists of the linear time-invariant differential systems that are also governed by one equation (thus only one control law is imposed). Denote such a controller by

$$c(\frac{d}{dt})w_1 + d(\frac{d}{dt})w_2 = 0. \tag{10.49}$$

The controlled system is then governed by

$$\begin{bmatrix} a(\frac{d}{dt}) & b(\frac{d}{dt}) \\ c(\frac{d}{dt}) & d(\frac{d}{dt}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0. \tag{10.50}$$

The associated polynomial system matrix is

$$\begin{bmatrix} a(\xi) & b(\xi) \\ c(\xi) & d(\xi) \end{bmatrix} \tag{10.51}$$

Note that typically (since we expect the determinant of (10.51) to be nonzero), the system governed by (10.50) is autonomous. Thus, as we have seen in Section 3.2, the main properties of the behavior of (10.50), its stability, settling time, frequencies of oscillation, etc., are effectively characterized by the singularities of (10.51), i.e., by the roots of its determinant

$$e(\xi) = a(\xi)d(\xi) - b(\xi)c(\xi). \tag{10.52}$$

The question thus arises as to what polynomials $e(\xi)$ can be obtained by choosing $c(\xi)$ and $d(\xi)$ for given $a(\xi)$ and $b(\xi)$. In analogy of what we explained in Chapter 9, we call the roots of $e(\xi)$ the *poles* of the controlled system. The question just asked is hence (for single-input/single-output systems) the complete analogue in a behavioral context of the pole placement problem studied in Chapter 9 and in Sections 10.5 and 10.6. The following result is readily obtained.

Theorem 10.8.3 *Consider the plant (10.48) and assume that $a(\xi)$ and $b(\xi)$ are not both zero. Then for any $e(\xi) \in \mathbb{R}[\xi]$, there exists a controller (10.49) such that*

$$\det \begin{bmatrix} a(\xi) & b(\xi) \\ c(\xi) & d(\xi) \end{bmatrix} = e(\xi) \tag{10.53}$$

if and only if the plant is controllable (equivalently, if and only if $a(\xi)$ and $b(\xi)$ are coprime).

Proof (if): From the Bezout identity (see Section 2.5) it follows that $a(\xi)$ and $b(\xi)$ coprime implies the existence of polynomials $c_1(\xi), d_1(\xi)$ such that $a(\xi)d_1(\xi) - b(\xi)c_1(\xi) = 1$. Now take $c(\xi) = e(\xi)c_1(\xi), d(\xi) = e(\xi)d_1(\xi)$. This yields (10.53).

(only if): Assume that $a(\xi)$ and $b(\xi)$ are not coprime. Let $h(\xi)$ be a common factor. Then obviously, the determinant in (10.53) also has $h(\xi)$ as a factor. □

The above theorem can easily be refined so as to specify exactly what the achievable polynomials $e(\xi)$ are. Motivated by the discussion in Section 9.6, we call the greatest common divisor of $a(\xi)$ and $b(\xi)$ the *uncontrollable polynomial* of the plant (10.48). Denote this polynomial by $\chi_u(\xi)$.

Theorem 10.8.4 *Consider the plant (10.48) and assume that $a(\xi)$ and $b(\xi)$ are not both zero. Let $e(\xi) \in \mathbb{R}[\xi]$ be given. Then there exists a controller (10.49) such that (10.53) holds if and only if the uncontrollable polynomial $\chi_u(\xi)$ of the plant is a factor of $e(\xi)$.*

Proof (if): Write $a(\xi) = \chi_u(\xi)a'(\xi), b(\xi) = \chi_u(\xi)b'(\xi)$ with $\chi_u(\xi)$ the greatest common divisor of $a(\xi)$ and $b(\xi)$. Since $a'(\xi)$ and $b'(\xi)$ are coprime, there exist $c'_1(\xi), d'_1(\xi)$ such that $a'(\xi)d'_1(\xi) - b'(\xi)c'_1(\xi) = 1$. Since $\chi_u(\xi)$ divides $e(\xi)$, $e(\xi)$ can be written as $e(\xi) = \chi_u(\xi)e'(\xi)$. Now use $c(\xi) = e'(\xi)c'_1(\xi), d(\xi) = e'(\xi)d'_1(\xi)$.

(only if): This part of the proof is completely analogous to the "only if" part of the proof of Theorem 10.8.3. □

An important special case of the above result refers to asymptotic stability. Recall that we call the plant (10.48) *stabilizable* (see Section 5.2.2) if for all (w_1, w_2) in its behavior $\mathfrak{B}$ there exists $(w'_1, w'_2) \in \mathfrak{B}$ such that

$(w_1, w_2)(t) = (w_1', w_2')(t)$ for $t < 0$, and $(w_1', w_2')(t) \to 0$ as $t \to \infty$. We have seen that (10.48) is stabilizable if and only if $a(\xi)$ and $b(\xi)$ have no common factor with roots in the closed right half plane, equivalently, if and only if the uncontrollable polynomial $\chi_u(\xi)$ of (10.48) is Hurwitz. This yields the following result.

Corollary 10.8.5 *Consider the plant (10.48) and assume that $a(\xi)$ and $b(\xi)$ are not both zero. Then the following are equivalent:*

(i) (10.48) is stabilizable.

(ii) Its uncontrollable polynomial $\chi_u(\xi)$ is Hurwitz.

(iii) There exists a controller (10.49) for (10.48) such that the controlled system (10.50) is asymptotically stable, i.e., such that (10.52) is Hurwitz.

Remark 10.8.6 In the control problem just considered, we have assumed that the purpose of the controller is to control the dynamics of both variables (w_1, w_2) in the controlled system. Often, however, our interest in the controlled system is on only one of these variables, say w_1. For instance, in the door-closing example treated in Example 10.8.1, we are principally interested in the dynamics of θ in the controlled system. Assume therefore that we wish to design the controller (10.49) such that in the controlled system the variable w_1 is governed by

$$e(\frac{d}{dt})w_1 = 0, \tag{10.54}$$

with $e(\xi) \in \mathbb{R}[\xi]$ a given polynomial. In other words, given (10.48), we wish to find (10.49) such that the dynamics of (10.50) yield, after eliminating w_2 (in the sense explained in Chapter 6) (10.54) for the dynamics of w_1.

Assume that (10.48) is controllable. Then by Theorem 10.8.3, there exists a controller (10.49) such that

$$a(\xi)d(\xi) - b(\xi)c(\xi) = e(\xi). \tag{10.55}$$

If $b(\xi)$ and $e(\xi)$ are also coprime, then this equation implies that $b(\xi)$ and $d(\xi)$ must also be coprime. Now use the result of Exercise 6.11 to conclude that w_1 is then indeed governed by (10.54). Thus under the added assumption that the plant polynomial $b(\xi)$ and the desired polynomial $e(\xi)$ are also coprime, Theorem 10.8.3 shows how to obtain a controller that regulates the dynamics of w_1. □

10.8.4 An algorithm for pole placement

The proof of Theorem 10.8.3, while in principle constructive, does not provide a particularly effective algorithm for computing (10.49) as the solu-

tion to (10.52). In addition, the controller obtained in the proof does not have particularly good properties, since the resulting degrees of the controller polynomials $c(\xi), d(\xi)$ may turn out to be unnecessarily high. We now present an algorithm that keeps the degrees of $c(\xi), d(\xi)$ under control.

Let $a(\xi), b(\xi) \in \mathbb{R}[\xi]$ be given. Define the *Bezout map* $B_{a,b} : \mathbb{R}[\xi] \times \mathbb{R}[\xi] \to \mathbb{R}[\xi]$ by

$$c(\xi), d(\xi) \stackrel{B_{a,b}}{\longmapsto} d(\xi)a(\xi) - c(\xi)b(\xi). \tag{10.56}$$

The map $B_{a,b}$ is surjective if and only if $a(\xi)$ and $b(\xi)$ are coprime (see Exercise 2.24). More generally, the image of $B_{a,b}$ consists exactly of the polynomials that have the greatest common divisor of $a(\xi)$ and $b(\xi)$ as a factor. It is easy to see that the solution to (10.52) is not unique. We now impose restrictions on the degrees of $c(\xi)$ and $d(\xi)$ so that (10.55) does have a unique solution that satisfies certain degree properties.

Lemma 10.8.7 *Assume that the greatest common factor $g(\xi)$ of $a(\xi)$ and $b(\xi)$ is also a factor of $e(\xi)$. Then equation (10.52) has a unique solution $c(\xi), d(\xi)$ such that*

$$\deg c(\xi) < \deg \frac{a(\xi)}{g(\xi)}. \tag{10.57}$$

This solution has the following further degree properties: If for some $m \in \mathbb{Z}_+$ $\deg e(\xi) \leq \deg a(\xi) + m - 1$ and $\deg b(\xi) \leq m$, then

$$\deg d(\xi) \leq m - 1. \tag{10.58}$$

If $\deg b(\xi) < \deg a(\xi)$, and $\deg e(\xi) \geq 2 \deg a(\xi) - 1$, then

$$\deg c(\xi) \leq \deg d(\xi) = \deg e(\xi) - \deg a(\xi). \tag{10.59}$$

Proof The first part of the proof involves existence and uniqueness of the solution to (10.52) satisfying (10.57).

Consider initially the case that $a(\xi)$ and $b(\xi)$ are coprime. Denote by $\mathbb{R}_k[\xi]$ the set of elements of $\mathbb{R}[\xi]$ that have degree less than or equal to k. Note that $\mathbb{R}_k[\xi]$ is a real vector space of dimension $k+1$.

Let $\deg a(\xi) = n$ and let $m \in \mathbb{Z}_+$ be such that $e(\xi) \in \mathbb{R}_{n+m-1}[\xi]$ and $b(\xi) \in \mathbb{R}_m[\xi]$. Now consider the Bezout map (10.56) restricted to $\mathbb{R}_{n-1}[\xi] \times \mathbb{R}_{m-1}[\xi]$. This restriction obviously maps $\mathbb{R}_{n-1}[\xi] \times \mathbb{R}_{m-1}[\xi]$ into $\mathbb{R}_{n+m-1}[\xi]$. Hence $B_{a,b}$ restricted to $\mathbb{R}_{m-1}[\xi] \times \mathbb{R}_{n-1}[\xi]$ maps an $(n+m)$-dimensional real vector space into an $(n+m)$-dimensional real vector space. This map is injective. To see this assume that $a(\xi)d(\xi) - b(\xi)c(\xi) = 0$ for some nonzero polynomials $c(\xi$ and $d(\xi)$ with $\deg c(\xi) < \deg a(\xi)$. Now $a(\xi)d(\xi) - b(\xi)c(\xi) = 0$ implies that $a^{-1}(\xi)b(\xi) = c^{-1}(\xi)d(\xi)$. Since by assumption $\deg c(\xi) < \deg \xi)$, this can only be the case when $a(\xi)$ and $b(\xi)$ have a factor in common. See Exercise 10.25. This contradicts the assumption that $a(\xi)$ and $b(\xi)$ are coprime. Since a linear map from an $(n+m)$-dimensional real vector space into an $(n+m)$-dimensional real vector space

is injective if and only if it is surjective, this implies that the solution to (10.52) satisfying (10.57) exists and is unique.

To prove existence and uniqueness in the case that $a(\xi)$ and $b(\xi)$ are not coprime, factor $a(\xi) = a'(\xi)f(\xi), b(\xi) = b'(\xi)f(\xi)$, and $e(\xi) = e'(\xi)f(\xi)$, with $a'(\xi)$ and $b'(\xi)$ coprime and apply the result obtained in the coprime case for $a'(\xi), b'(\xi)$, and $e'(\xi)$.

Now consider the two other degree estimates in the lemma. The first estimate, (10.58), is an immediate consequence of the previous part of the proof. To show the second degree estimate, (10.59), observe that if $\deg b(\xi)$ and $\deg c(\xi)$ are both less than $n = \deg a(\xi)$, and if $\deg e(\xi) \geq 2n-1$, then $a(\xi)d(\xi) - b(\xi)c(\xi) = e(\xi)$ implies $\deg a(\xi)d(\xi) = \deg e(\xi)$. Hence $\deg d(\xi) \geq n-1$ and $\deg c(\xi) \leq n-1 \leq \deg d(\xi) = \deg e(\xi) - \deg a(\xi)$, as claimed. □

The above lemma only gives conditions for the existence of solutions to (10.52), but its proof also provides an algorithm for computing a solution. Moreover, equation (10.52), viewed as an equation with the polynomials $a(\xi), b(\xi), e(\xi)$ given and $c(\xi), d(\xi)$ unknown is, of course, a linear one. But without an estimate on the degrees of $c(\xi), d(\xi)$, this equation is infinite-dimensional. However, the lemma provides such an estimate. This allows to reduce the problem of finding a solution $c(\xi), d(\xi)$ to a finite-dimensional linear equation. For the case that $a(\xi)$ and $b(\xi)$ are coprime this can be done as follows. Denote as in the proof of Lemma 10.8.7 $n = \deg a(\xi), m = \deg b(\xi)$, and assume that $\deg e(\xi) < n+m$. Now express the polynomials in terms of their coefficients, using the fact that, by the above lemma, we may assume $\deg c(\xi) < n, \deg d(\xi) < m$. Then equate equal powers of ξ on the left- and right hand sides of (10.52), and write the resulting expressions in vector/matrix notation. Then equation (10.52) becomes the *Sylvester equation*

$$[e_0 \;\; e_1 \dots e_{n+m-1}] = [d_0 \;\; d_1 \dots d_{m-1} \;\; -c_{n-1} \dots -c_1 \;\; -c_0]S,$$

where S is the *Sylvester resultant*

$$S = \begin{bmatrix} a_0 & a_1 & \cdots & a_n & & & \\ & a_0 & \cdots & a_{n-1} & a_n & & \\ & & \ddots & & & \ddots & \\ & & & a_0 & a_1 & \cdots & a_n \\ & & & b_0 & b_1 & \cdots & b_m \\ & & b_0 & b_1 & \cdots & b_m & \\ & \iddots & & & \iddots & & \\ b_0 & b_1 & \cdots & \cdots & b_m & & \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\\a\\a\end{matrix}}\right\} m \text{ rows} \\ \left.\vphantom{\begin{matrix}a\\a\\a\\a\\a\\a\\a\end{matrix}}\right\} n \text{ rows} \end{matrix}$$

and where the a_ks, b_ks, c_ks, d_ks, and e_ks are the coefficients of the polynomials $a(\xi), b(\xi), c(\xi), d(\xi)$, and $e(\xi)$.

The second degree estimate in Lemma 10.8.7, (10.59), has important consequences for the control problem at hand. We have argued that controllers need not be endowed with the signal flow graph structure suggested by the intelligent control paradigm of Figure 10.4. Indeed, the door-closing mechanism of Figure 10.7 and many other industrial control mechanisms do not function as signal processors that process observed outputs in order to generate control inputs. In Theorem 10.8.4 we have seen that useful controller design procedures can be set up while ignoring the input/output structure of the plant (10.48) and the controller (10.49). Nevertheless, it is of interest to determine conditions under which Theorem 10.8.4 allows us to recover a controller such that the input/output structure of the plant and controller are reversed. More precisely, we say that the plant (10.48) and the controller (10.49) have *reversed input/output structure* if either w_1 is the input to the plant and w_2 the output, and w_2 is the input to the controller and w_1 the output; or, vice versa, if w_2 is the output to the plant and w_1 the output, and w_1 is the input to the controller and w_2 the output.

In Chapters 2 and 3, this input/output structure has been studied extensively in terms of the degrees of the polynomials involved in the behavioral differential equations. Thus the reversed input/output structure holds if and only if either $a^{-1}(\xi)b(\xi)$ and $d^{-1}(\xi)c(\xi)$ are both proper, or $b^{-1}(\xi)a(\xi)$ and $c^{-1}(\xi)d(\xi)$ are both proper.

The second degree estimate in Lemma 10.8.7, (10.59), leads to the following important refinement of Theorem 10.8.4 and Corollary 10.8.5.

Theorem 10.8.8 *Consider the plant (10.48) and assume that $a^{-1}(\xi)b(\xi)$ is strictly proper, i.e., that in the plant w_2 is input and w_1 is output. Assume further that $e(\xi) \in \mathbb{R}[\xi]$ has $\deg e(\xi) \geq 2 \deg a(\xi) - 1$, and that the greatest common factor of $a(\xi)$ and $b(\xi)$ is also a factor of $e(\xi)$. Then there exists a controller (10.49) such that (10.53) holds and such that $d^{-1}(\xi)c(\xi)$ is proper, i.e., such that the plant and the controller have reversed input/output structure.*

Proof Lemma 10.8.7 implies that (10.52) has a solution with $\deg c(\xi) \leq \deg d(\xi)$. □

Controllers that have a reversed input/output structure can in principle be implemented by means of sensors and actuators, and still have a reasonable robustness with respect to sensor noise, since no differentiators are required for their implementation. Note, finally, that the controllers obtained in Theorem 10.8.8 yield the (reduced order) compensators obtained in the SISO case in Sections 10.5 and 10.6.

Example 10.8.9 Consider the electrical circuit shown in Figure 10.9. Assume that $L, C > 0$. The differential equation relating the port variables V

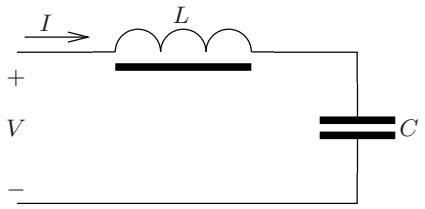


FIGURE 10.9. Electrical circuit: plant.

and I is

$$C\frac{d}{dt}V = (1 + LC\frac{d^2}{dt^2})I.$$

This system is controllable. For simplicity, take $L = 1$ and $C = 1$. Assume that we would like to stabilize this circuit and obtain as controlled dynamics for the voltage V,

$$2V + 4\frac{d}{dt}V + 3\frac{d^2}{dt^2}V + \frac{d^3}{dt^3}V = 0. \tag{10.60}$$

Thus the desired controlled characteristic polynomial is $2 + 4\xi + 3\xi^2 + \xi^3$. The roots of this polynomial are $-1, -1 \pm i$. Note that since $1 + \xi^2$ and $2 + 4\xi + 3\xi^2 + \xi^3$ are coprime, we can use Remark 10.8.6. The controller

$$c(\frac{d}{dt})V = d(\frac{d}{dt})I$$

leads to the desired controlled dynamics, provided that $c(\xi)$ and $d(\xi)$ are chosen such that $c(\xi)(1 + \xi^2) - d(\xi)\xi = 2 + 4\xi + 3\xi^2 + \xi^3$. The polynomials $c(\xi) = 2+\xi, d(\xi) = -(3+\xi)$ solve this equation. The resulting controller can be implemented in a number of ways, for example by terminating the circuit of Figure 10.9 by the circuit shown in Figure 10.10. The differential equation

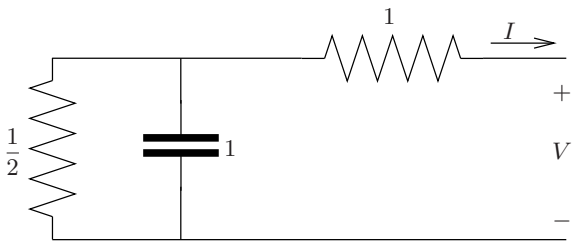


FIGURE 10.10. Electrical circuit: controller.

relating V and I in this circuit is indeed given by $(2 + \frac{d}{dt})V = -(3 + \frac{d}{dt})I$, and hence it yields the desired control law. The controlled circuit is shown in Figure 10.11. One can readily verify that in this circuit the voltage V

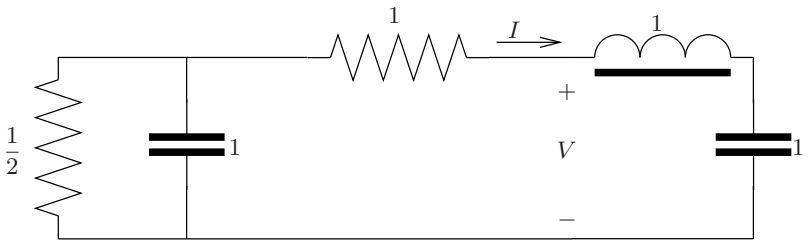


FIGURE 10.11. Electrical circuit: controlled system.

is indeed governed by (10.60). Actually, for the case at hand, it turns out that also the other voltages and currents in this circuit are governed by the same differential equation. □

10.9 Recapitulation

In this chapter we studied the construction of observers and of output feedback compensators. The main results were the following:

- An observer is an algorithm that uses the system equations in order to deduce from the measured signals (e.g. inputs and outputs) the state of a system, or, more generally, another nonobserved output.
- The observers that we considered are a combination of an internal model and error feedback. This error, the innovation, is the difference between the measured output and its estimate (Section 10.2).
- There exists an observer with a prescribed characteristic polynomial for the error dynamics if and only if the plant is observable (Theorem 10.3.1).
- The combination of the separation principle and the certainty equivalence principle leads to an effective way for obtaining an output feedback controller. This design is based on a state observer combined with a state feedback law (Theorem 10.5.1).
- There exists an output feedback compensator with a prescribed characteristic polynomial for the closed loop system if and only if the plant is controllable and observable (Theorem 10.5.3).
- These linear system techniques applied to the linearized system can be used to stabilize a nonlinear system around an equilibrium point (Section 10.7).
- Instead of viewing controllers as feedback processors, it is sometimes more appropriate to view them as an interconnected subsystem. This point of view leads to design principles in which the behavioral point of view becomes essential. The algorithms for pole placement can be generalized to this setting. (Sections 10.8.1 and 10.8.2).

10.10 Notes and References

Observers were formally introduced by Luenberger in his Ph.D. dissertation in 1963 (see [37] for an expository paper on the subject). The observers discussed in this chapter are actually nonoptimal, nonstochastic versions of the celebrated Kalman filter. The Kalman filter has been the topic of numerous papers and texts (see [26, 30] for historical references and [33, 4] for textbooks which treat this topic in detail). It is somewhat surprising that using the measurement of one variable combined with the system dynamics, one can actually design an algorithm that functions as a reliable sensor for another variable. The full technological impact of this is only being felt recently, and these techniques go under a variety of names, such as: soft sensors, sensor fusion, smart sensors. The combination of state estimators with state feedback control laws originated in the context of the linear-quadratic-gaussian problem (see [5] for some early survey papers on this topic and [33, 3] for texts that treat this topic in detail). It is interesting to note that observers and the certainty equivalence and the separation principles originated in the context of optimality questions and stochastic systems first, and that the nonoptimal, nonstochastic versions were discovered only later. A very elegant geometric way of looking at observers, and at many other questions surrounding the problems discussed in Chapters 9 and 10, is developed in [65]. A proof of Theorem 10.6.2 can be found there. Looking at controllers as interconnections of subsystems, as discussed in Section 10.8, originates with the second author of the present book [61].

10.11 Exercises

As a simulation exercise illustrating the material covered in this chapter we suggest A.6.

10.1 Assume that a sinusoidal signal is measured in the background of additive high-frequency sinusoidal noise, yielding

$$y(t) = \underbrace{A_s \sin \omega_s t}_{\text{signal}} + \underbrace{A_n \sin \omega_n t}_{\text{noise}}.$$

Define the signal-to-noise ratio to be $\frac{|A_s|}{|A_n|}$. Compute the signal-to-noise ratio of the derivative $\frac{d}{dt}y$. Prove that for a given signal-to-noise ratio of the observation, the signal-to-noise ratio of its derivative goes to zero as $\omega_n \to \infty$. Conclude that it is not a good idea to differentiate a signal when there is high-frequency noise present in the measurement.

10.2 Consider the system (10.3) with

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

Construct an observer with error dynamics characteristic polynomial $r(\xi) = 1 + \xi + \xi^2$. Repeat for

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

and $r(\xi) = 1 + 2\xi + 2\xi^2 + \xi^3$.

10.3 Assume that for the system

$$\frac{d}{dt}x = Ax + Bu, \quad y = Cx, \quad z = Hx$$

we wish to estimate z from observations of (u, y). Construct an observer for z, assuming that (A, C) is observable.

10.4 Extend the observer (10.7) and Theorem 10.3.1 to systems with a feedthrough term

$$\frac{d}{dt}x = Ax + Bu, \quad y = Cx + Du.$$

10.5 Consider the discrete-time analogue of (10.3),

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t). \tag{10.61}$$

The linear system

$$\hat{x}(t+1) = K\hat{x}(t) + Ly(t) + Mu(t)$$

is said to be a *deadbeat observer* for (10.61) if there exists a $T \in \mathbb{Z}_+$ such that for all $x(0), \hat{x}(0)$, and u there holds $\hat{x}(t) = x(t)$ for $t \geq T$. Assume that (A, C) is observable. Construct a deadbeat observer for (10.61). Prove that T can be taken to be equal to n, the dimension of the state space of (10.61).

10.6 Let $0 = \omega_0 < \omega_1 < \cdots < \omega_N$. Consider the linear system

$$\begin{aligned} \frac{d}{dt}x &= \begin{bmatrix} \omega_0 & & & & & & \\ & 0 & \omega_1 & & & & \\ & -\omega_1 & 0 & & & & \\ & & & \ddots & & & \\ & & & & & 0 & \omega_N \\ & & & & & -\omega_N & 0 \end{bmatrix} x, \\ y &= \begin{bmatrix} 1 & 0 & 1 & \cdots & 0 & 1 \end{bmatrix} x. \end{aligned} \tag{10.62}$$

Assume that the output y is observed at $t = 0, \Delta, 2\Delta, \ldots$. Define $y_k = y(k\Delta)$. Prove that y_k is governed by the discrete-time system

$$x_{k+1} = e^{A\Delta}x_k, \quad y_k = Cx_k, \tag{10.63}$$

with A and C defined in terms of (10.62) in the obvious way. Deduce necessary and sufficient conditions in terms of $\omega_0, \omega_1, \ldots, \omega_N$ and Δ for (10.63) to be an observable discrete-time system. Deduce from there the best $\Delta_{\max}$, with $\Delta_{\max}$ a function of $\omega_1, \omega_2, \ldots, \omega_N$, guaranteeing observability for all $0 < \Delta < \Delta_{\max}$. Readers familiar with the *sampling theorem* may wish to interpret this condition in these terms.

Assuming that (10.63) is observable, construct an observer that asymptotically reconstructs $x(0)$ from $y_0, y_1, y_2, \ldots$. Refine this algorithm using the ideas of Exercise 10.5 in order to compute $x(0)$ from a finite sample $y_0, y_1, \ldots, y_T$. What is the minimum required T?

10.7 Consider the system $\omega^2 y + \frac{d^2}{dt^2} y = 0$. The general solution to this differential equation is $A\cos(\omega t + \varphi)$, with $A \in [0, \infty)$ and $\varphi \in [0, 2\pi)$. The problem considered in this exercise is to design an observer that estimates A and φ by observing y. Introduce the state $x = \mathrm{col}(y, \frac{d}{dt} y)$. Construct an asymptotically stable observer for x. Deduce from there an estimate $\hat{A}(t), \hat{\varphi}(t)$ such that $(\hat{A}(t), \hat{\varphi}(t)) \underset{t \to \infty}{\longrightarrow} (A, \varphi)$.

10.8 Consider the electrical circuit shown in Figure 10.12 Assume that the port

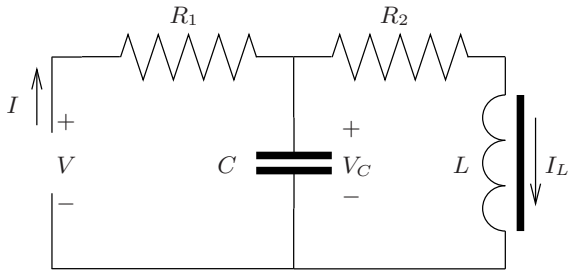


FIGURE 10.12. An electrical circuit.

variables V and I are observed. Construct an observer that estimates V_C and I_L with asymptotically convergent estimates. Take for simplicity $R_1 = R_2 = 1$, $L = 1$, and $C = 1$.

10.9 For what values of the parameters is the following system (a) controllable, (b) observable, (c) stabilizable, (d) detectable?

$$\begin{aligned} \frac{d}{dt} x &= \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} x + \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} u, \\ y &= \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} x. \end{aligned}$$

10.10 Is the system

$$\frac{d}{dt} x = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} x, \qquad y = \begin{bmatrix} 0 & 1 \end{bmatrix} x$$

detectable? If it is, construct an asymptotically convergent state observer.

10.11 Prove that (A, C) is detectable if and only if (A^T, C^T) is stabilizable.

10.12 Consider the dynamical system described by the behavioral equation $\frac{d^4}{dt^4}y = u$. Write state equations with $x_1 = y$, $x_2 = \frac{d}{dt}y$, $x_3 = \frac{d^2}{dt^2}y$, $x_4 = \frac{d^3}{dt^3}y$. Find a feedback compensator such that the closed loop characteristic polynomial equals $(1 + \xi + \xi^2)^2(1 + 2\xi + \xi^2)^2$.

10.13 Consider the dynamical system described by (10.23). Prove that this system is observable if and only if (A_{11}, A_{21}) is an observable pair. Prove that it is detectable if and only if (A_{11}, A_{21}) is.

10.14 Consider the system (10.3). Use the controller $u = N\hat{x}$ with $\hat{x}$ the output of the observer (10.29, 10.30). Prove that the closed loop characteristic polynomial is given by $\chi_{A+BN}(\xi)\chi_{A_{11}-L_1A_{22}}(\xi)$. Use this to show that if (10.3) is controllable and observable, then for any real monic polynomial $r(\xi)$ of deg $2n -$ rank C that is factorizable into a real polynomial of degree n and one of degree $n -$ rank C, there exists a feedback compensator such that $r(\xi)$ is the closed loop characteristic polynomial.

10.15 Consider the system (10.3), called the *primal* system, and its *dual*

$$\frac{d}{dt}\bar{x} = A^T\bar{x} + C^T\bar{u} \ ; \ \bar{y} = B^T\bar{x}.$$

Design a controller

$$\frac{d}{dt}\bar{w} = K\bar{w} + L\bar{y} \ ; \ \bar{u} = M\bar{w} + N\bar{y}$$

for the dual system. Let $\bar{\chi}_{\rm cl}(\xi)$ denote the resulting closed loop characteristic polynomial for the controlled dual system. Now use the dual controller (K^T, M^T, L^T, N^T) on the primal system. Let $\chi_{\rm cl}(\xi)$ denote the closed loop characteristic polynomial obtained by using this controller on the primal system. Prove that $\chi_{\rm cl}(\xi) = \bar{\chi}_{\rm cl}(\xi)$. Use this and Exercise 10.14 to show that if (10.3) is controllable and observable, then for any real monic polynomial $r(\xi)$ of degree $2n -$ rank B that is factorizable into a real polynomial of degree n and one of degree $n -$ rank B, there exists a feedback compensator such that $r(\xi)$ is the closed loop characteristic polynomial.

10.16 Consider the system (10.3). Let $\chi_v(\xi)$ be the least common multiple of the uncontrollable polynomial $\chi_u(\xi)$ and the unobservable polynomial $\chi_0(\xi)$. Let (K, L, M, N) be any compensator (of any order) and let $\chi_{A_{\rm e}}(\xi)$ be the resulting closed loop characteristic polynomial. Prove that $\chi_v(\xi)$ is a factor of $\chi_{A_{\rm e}}(\xi)$.

10.17 This exercise is concerned with *robustness*. Loosely speaking, we call a controlled system robust if small errors in the model or in the controller have small effects on the controlled behavior. In this exercise, we consider robustness both with respect to measurement errors and with respect to parameter uncertainty. Consider the i/o system

$$6y - 5\frac{d}{dt}y + \frac{d^2}{dt^2}y = u. \tag{10.64}$$

(a) Show that this system is open-loop ($u = 0$) unstable.

Assume that we want to stabilize the system using feedback control. Our first attempt is

$$u = -5\frac{d}{dt}y + \frac{d^2}{dt^2}y. \tag{10.65}$$

It appears that this yields an extremely fast and accurate controller, since the system output is

$$y = 0.$$

We now investigate whether the proposed controller is indeed such a superior controller. If we were able to implement the controller with infinite precision, then, there seems to be no problem. Suppose, however, that this controller is implemented by means of a sensor that does not measure y exactly. Assume that the sensor output is $y + v$, where v is a noise term. The controller is then given by

$$u = -5\frac{d}{dt}(y + v) + \frac{d^2}{dt^2}(y + v).$$

(b) Determine the output y for the case that $v(t) = \epsilon \sin(2\pi f t)$, $\epsilon > 0$, and $f \in \mathbb{R}$. Conclude that an arbitrarily small disturbance can have a significant impact if f is sufficiently large. Thus, the controller (10.65) is not robust with respect to measurement noise.

(c) Determine the controller canonical form for the system (10.64).

Determine a state feedback that assigns the closed-loop poles to $-1, -2$. Design an observer with observer poles equal to $-3, -4$. Combine the controller and the observer to obtain a feedback compensator with poles at $-1, -2, -3, -4$.

(d) Suppose that this observer has the noisy sensor output as input. The observer equation then becomes

$$\frac{d}{dt}\hat{x} = A\hat{x} + bu + k(y + v) - kc\hat{x}.$$

Does this observer lead to an acceptable controlled system? Compare your conclusion with the one obtained in part 10.17b.

(e) Another inaccuracy with respect to which the controller (10.65) is very sensitive is parameter uncertainty. Suppose that the plant parameters deviate from their nominal values and that the plant is given by

$$(6 + \epsilon_2)y - (5 + \epsilon_1)\frac{d}{dt}y + \frac{d^2}{dt^2}y = u. \tag{10.66}$$

Determine behavioral equations for the output y of the controlled system (10.66, 10.65). Conclude that the controller (10.65) is again not robust, this time with respect to parameter uncertainty.

(f) Determine the output when the system (10.66) is controlled by the compensator obtained in part 10.17c.

10.18 Consider the discrete-time linear system

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t). \tag{10.67}$$

Formulate the analogue of Theorems 10.3.1 and 10.5.1. Are the theorems valid when the variables in (10.67) and of the required observer and output feedback controller take values in an arbitrary field?

10.19 Construct reduced order observers for the systems given in Exercise 10.2 with the desired error characteristic polynomials given by $1+\xi$ and $1+2\xi+\xi^2$ respectively.

10.20 Consider the plant

$$\frac{d}{dt}x = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad y = \begin{bmatrix} 1 & 0 \end{bmatrix} x.$$

Construct, using the theory developed in Section 10.6, a first-order compensator such that the closed loop system has characteristic polynomial $1+2\xi+2\xi^2+\xi^3$.

10.21 Find a linear output feedback compensator that stabilizes the equilibrium point $y^* = 0, u^* = 0$ of the nonlinear system

$$y^3 + \frac{d^2}{dt^2}y + \sin y = u(u-1).$$

10.22 Extend the theory explained in Section 10.7 in which you use a reduced order compensator. Use this to find a first-order linear compensator that stabilizes the pendulum of Example 10.5.4 in its upright position. Choose reasonable values of M and L and simulate the linearized and nonlinear closed loop systems for a number of initial conditions for the plant and the compensator.

10.23 Use the theory developed in Section 10.8 and equation (10.48) to find a control law that makes the following plant asymptotically stable:

$$(1 - \frac{d}{dt} + \frac{d^2}{dt^2})w_1 = w_2.$$

Repeat for

$$(1 + \frac{d^3}{dt^3})w_1 = (1 - \frac{d}{dt})w_2.$$

10.24 Let $a(\xi), b(\xi) \in \mathbb{R}[\xi]$ be coprime. Prove that there exist polynomials $c(\xi), d(\xi) \in \mathbb{R}[\xi]$ such that the control law

$$c(\frac{d^2}{dt^2})w_1 = d(\frac{d^2}{dt^2})w_2$$

stabilizes (in the sense that all solutions are bounded on $[0,\infty)$) the plant

$$a(\frac{d^2}{dt^2})w_1 = b(\frac{d^2}{dt^2})w_2.$$

Assume that $a^{-1}(\xi^2)b(\xi^2)$ is proper. Prove that $d^{-1}(\xi^2)c(\xi^2)$ can be taken to be proper as well.

10.25 In the proof of Lemma 10.8.7 we claimed that if $a^{-1}(\xi)b(\xi) = c^{-1}(\xi)d(\xi)$ with $\deg c(\xi) < \deg a(\xi)$, then $a(\xi)$ and $b(\xi)$ cannot be coprime. Prove this by viewing $a^{-1}(s)b(s)$ and $c^{-1}(s)d(s)$ as complex functions and counting their poles. This property may, however, also be proven by purely algebraic means, i.e., by using the fact that $\mathbb{R}[\xi]$ is a Euclidean ring. Can you find a proof in this spirit? See also Exercise 2.24.

10.26 Prove the following generalization of Theorem 10.8.4. Let $R(\xi) \in \mathbb{R}^{p\times(p+m)}[\xi]$ and assume that rank $R(\lambda) = p$ for all $\lambda \in \mathbb{C}$. Note that this implies that the plant

$$R(\frac{d}{dt})w = 0 \tag{10.68}$$

is controllable. Prove that for any $r(\xi) \in \mathbb{R}[\xi]$ there exists $C(\xi) \in \mathbb{R}^{m\times(p+m)}[\xi]$, inducing the control law

$$C(\frac{d}{dt})w = 0 \tag{10.69}$$

such that the controlled system

$$\begin{bmatrix} R(\frac{d}{dt}) \\ C(\frac{d}{dt}) \end{bmatrix} w = 0$$

satisfies $\det(\begin{bmatrix} R(\xi) \\ C(\xi) \end{bmatrix}(\xi)) = r(\xi)$. Generalize this result to noncontrollable systems and obtain necessary and sufficient conditions for stabilizability of (10.68) by means of (10.69).

Hint: Consider the Smith form of (10.68).

10.27 Consider the nonlinear system

$$\frac{d^2}{dt^2}w_1 = f(w_1, \frac{d}{dt}w_1, w_2, \frac{d}{dt}w_2, \frac{d^2}{dt^2}w_2). \tag{10.70}$$

All variables are assumed to be scalar, and $f : \mathbb{R}^5 \to \mathbb{R}$. Assume that $w_1^*, w_2^* \in \mathbb{R}$ is such that $f(w_1^*, 0, w_2^*, 0, 0) = 0$. Define in what sense you can consider (w_1^*, w_2^*) to be an equilibrium point of (10.70). Linearize (10.70) around this equilibrium, obtaining

$$\frac{d^2}{dt^2}\Delta_1 = a_0\Delta_1 + a_1\frac{d}{dt}\Delta_1 + b_0\Delta_2 + b_1\frac{d}{dt}\Delta_2 + b_2\frac{d^2}{dt^2}\Delta_2, \tag{10.71}$$

where $a_0 = f_1'(w_1^*, 0, w_0^*, 0, 0)$, $a_1 = f_2'(w_1^*, 0, w_2^*, 0, 0)$, $b_0 = f_3'(w_1^*, 0, w_2^*, 0, 0)$, $b_1 = f_4'(w_1^*, 0, w_2^*, 0, 0)$, $b_2 = f_4'(w_1^*, 0, w_2^*, 0, 0)$, and f_k' denotes the derivative of f with respect tot the kth variable.

Let $a(\xi) = a_0 + a_1\xi - \xi^2$, and $b(\xi) = b_0 + b_1\xi + b_2\xi^2$. Assume that $a(\xi)$ and $b(\xi)$ are coprime polynomials. Prove that there exist first-order polynomials $c(\xi), d(\xi)$ such that $a(\xi)d(\xi) + b(\xi)c(\xi) = 1 + 2\xi + 2\xi^2 + \xi^3$. Prove that the control law

$$c(\frac{d}{dt})\Delta_1 = d(\frac{d}{dt})\Delta_2$$

makes (10.71) asymptotically stable, and that the control law

$$c(\frac{d}{dt})(w_1 - w_1^*) = d(\frac{d}{dt})(w_2 - w_2^*)$$

makes the equilibrium (w_1^*, w_2^*) of (10.70) asymptotically stable. Explain clearly what you mean by this last statement.

Apply these ideas to stabilize a pendulum in its upright position.

Appendix A
Simulation Exercises

In this appendix, we present some exercises that require the aid of a computer. They aim at giving the student insight into some simple modeling examples and control problems. These exercises are most easily carried out using a numerical software package such as MATLAB© with its CONTROL SYSTEM© toolbox for the design of the controller parameters, and its SIMULINK© toolbox for dynamic simulation. In the last exercise, a formula manipulation package, such as Mathematica©, is also required.

A.1 Stabilization of a Cart

This is a very simple exercise. We recommend that it be assigned at the very beginning of the course, before much theory development has been done, in order to familiarize the students with MATLAB©, and to give them a feeling for the difficulties that already emerge in a very simple stabilization problem. This exercise treats stabilization of the position of a mass by means of a linear control law. The same problem is treated in Example 9.7.3 for nonlinear control laws.

Consider a cart that moves horizontally under the influence of an external force. The relevant geometry is shown in Figure A.1. The system variables are u, the external force acting on the cart; and y, the horizontal displacement of the cart relative to some reference point. The system has only one parameter: M the mass of the cart (take $M = 1$ throughout).

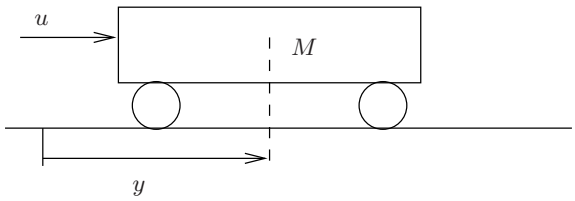


FIGURE A.1. A cart.

The dynamic relation between the variables u and y has been known since the beginning of (modern) scientific time, and it is given by Newton's second law:

$$M\frac{d^2y}{dt^2} = u. \tag{A.1}$$

Assume that we want to keep this mass at the position $y = 0$. Of course, one could nail the mass at the position $y = 0$, but we frown on taking such crude measures. What we would like to do is to achieve this goal by choosing the force u judiciously as a function of the measured position y of the cart.

1. What is more natural than simply to push the mass back to its desired equilibrium? That is, if the mass is to the (far) right of the equilibrium, push it (hard) to the left, and if it is to the (far) left, push it (hard) to the right. So, use a position sensor and apply the control law

   $$u = -K_p y \tag{A.2}$$

   with $K_p > 0$ a to-be-chosen constant.

   Substitute (A.2) into (A.1) and examine, using MATLAB©, the performance of this control law. Take $y(0) = 1$ and $\frac{dy}{dt}(0) = 0$ and simulate the response y for various values of K_p, say $K_p = 0.1$, $K_p = 1$, and $K_p = 10$. This control law does not seem to work. Explain mathematically why this is so.

2. Challenged by the failure of the obvious, let's try something more subtle. Perhaps it is more logical to counter the velocity and push the cart back to its equilibrium against the direction in which it is moving away from it. So, use a tachometer (a device that measures velocity) that senses $\frac{dy}{dt}$, and apply the control law

   $$u = -K_v\frac{dy}{dt} \tag{A.3}$$

   with $K_v > 0$ a to-be-chosen constant.

   Substitute (A.3) in (A.1) and examine, using MATLAB©, the performance of this control law. Take $y(0) = 0$, $\frac{dy}{dt}(0) = 1$, and simulate the response y for various values of K_v, say $K_v = 0.1, K_v = 1, K_v = 10$. Still not perfect, but better, it seems. In fact, for high K_v, this control law seems to work very well. So take K_v very large, say $K_v = 100$, and simulate the response again, but now with the different initial condition $y(0) = 1$ and $\frac{dy}{dt}(0) = 1$. Conclude that also this control law is no good. Explain mathematically why this is so.

3. Annoyed by these failures, let's think even harder. Maybe we should use a combination of the laws (A.2) and (A.3)? So, let's apply the control law

$$u = -K_p y - K_v \frac{dy}{dt} \tag{A.4}$$

with $K_p > 0$ and $K_v > 0$ to be chosen.

Substitute (A.4) into (A.1) and simulate y with the initial conditions $y(0) = 1$, $\frac{dy}{dt}(0) = 0$, and with $y(0) = 0$, $\frac{dy}{dt}(0) = 1$, for the following values of K_p and K_v: $K_p = 0.1, 1, 10$; $K_v = 0.1, 1, 10$. In total there are thus 9 combinations of the gains (K_p, K_v). Which one do you like best? Why? Try some more values, and reach the conclusion that $K_p \approx (K_v)^2$ means small *overshoot* (explain what you mean by this), while K_p and K_v both large means fast *settling time* (explain what you mean by this). So simulate again with $K_p = 100, K_v = 14$, and be happy with what you see.

4. In practice, of course, the position sensor and the tachometer may not work perfectly. So let us assume that the result of the measurement at time t is

$$\eta(t) = y(t) + \epsilon(t), \tag{A.5}$$

with $\epsilon(t)$ measurement noise. Take for ϵ the omnipresent 50– (or 60–) cycle noise

$$\epsilon(t) = |A \cos(50.2\pi t)|. \tag{A.6}$$

Take the amplitude of the noise small (in comparison to the distances we have been using), say $A = 0.01$.

Now (A.4) leads to the control law

$$u = -K_p \eta - K_v \frac{d\eta}{dt}. \tag{A.7}$$

Substitute (A.7) into (A.1), with the gains $K_p = 100, K_v = 14$ found in the previous part of this exercise, with η given by (A.5) and ϵ by (A.6), and simulate the response y with $y(0) = 1$, $\frac{dy}{dt}(0) = 0$, and with $y(0) = 0$, $\frac{dy}{dt}(0) = 1$. Comment on the quality of the response.

A.2 Temperature Control of a Container

This simulation exercise is also a simple one. We recommend that it be assigned as soon as the student has some familiarity with input/state/output equations. It aims at familiarizing the students with MATLAB© and to give them a feeling for control problems. It also illustrates some of the concepts of Chapter 8.

Consider the control of the temperature in a container (or a room). The relevant geometry is shown in Figure A.2. The purpose is to analyze a controller that regulates, by means of a valve, the amount of heat supplied to the container. The decision of how to set the valve is based on the measurements performed by a thermometer. The temperature of the container is also influenced by the ambient temperature. The thermometer is assumed to be a mercury thermometer. The

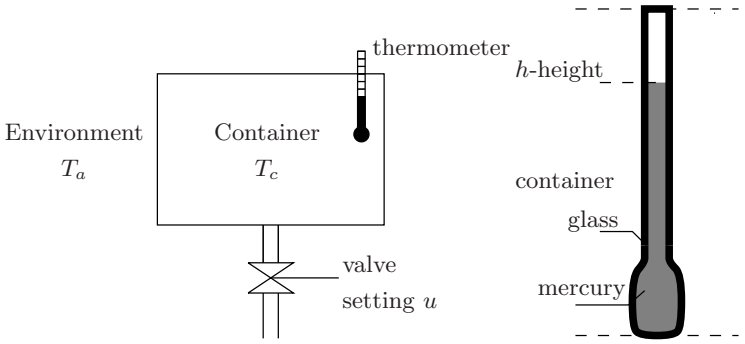


FIGURE A.2. Temperature control.

measurement taken is thus not the temperature of the container directly, but the height of the mercury column which is, of course, related to the temperature in the container. We now set out to model these dynamic relations.

1. The following are the *manifest* variables, the primary variables whose dynamics we wish to model:
   u: the valve setting (the control variable)
   T_{a}: the ambient temperature (a disturbance)
   T_{c}: the temperature in the container (the to-be-controlled variable)
   h: the height of the mercury column (the measured variable)
   In order to obtain a model for the relation between the manifest variables, it is useful to introduce a number of auxiliary (*latent*) variables. In particular, introduce
   q: the rate of heat supplied to the container by the heat valve
   q_{a}: the rate of heat supplied to the container from the environment
   q_{g}: the rate of heat supplied to the thermometer glass from the container
   q_{Hg}: the rate of heat supplied to the thermometer mercury from the thermometer glass
   T_{g}: the temperature of the thermometer glass
   T_{Hg}: the temperature of the mercury
   A_{g}: the internal cross section of the thermometer glass
   V_{g}: the volume inside the thermometer glass
   V_{Hg}: the volume of the mercury
   Consider the following relations:

$$\begin{aligned} q &= a_0 u, \\ q_{\mathrm{a}} &= a_1(T_{\mathrm{a}} - T_{\mathrm{c}}), \quad q_{\mathrm{g}} = a_2(T_{\mathrm{c}} - T_{\mathrm{g}}), \quad q_{\mathrm{Hg}} = a_3(T_{\mathrm{g}} - T_{\mathrm{Hg}}), \\ \frac{dT_{\mathrm{c}}}{dt} &= b_1(q_{\mathrm{a}} + q - q_{\mathrm{g}}), \quad \frac{dT_{\mathrm{g}}}{dt} = b_2(q_{\mathrm{g}} - q_{\mathrm{Hg}}), \quad \frac{dT_{\mathrm{Hg}}}{dt} = b_3 q_{\mathrm{Hg}}. \end{aligned}$$

   Explain each of these relations, noting that it is reasonable to assume that the heat flow through a boundary is proportional to the difference of the temperatures on both sides of the boundary, and that the temperature change of a medium is proportional to the heat supplied to it.

We still need to set up the equation for h. Assume that materials expand by an amount proportional to their temperatures. Hence $A_{\rm g}$ is proportional $T_{\rm g}^2$, $V_{\rm g}$ is proportional to $T_{\rm g}^3$, while $V_{\rm Hg}$ is proportional to $T_{\rm Hg}^3$. Therefore,

$$A_{\rm g} = c_1 T_{\rm g}^2, \quad V_{\rm g} = c_2 T_{\rm g}^3, \quad V_{\rm Hg} = c_3 T_{\rm Hg}^3, \quad h = \frac{V_{\rm Hg} - V_{\rm g}}{A_{\rm g}}. \tag{A.9}$$

The system *parameters* $a_0, a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3$, are positive constants depending on the geometry and the material properties. In a thermometer that works well, c_3 must be larger than c_2, for when both $T_{\rm Hg}$ and $T_{\rm g}$ reach a higher steady state value, we want h to increase. This is the reason why materials such as mercury are used in thermometers: their temperature expansion coefficient is large.

The above equations can be put into i/s/o form, with u and $T_{\rm a}$ as input variables, h and $T_{\rm c}$ as output variables, and by keeping the latent variables $T_{\rm c}$, $T_{\rm g}$, and $T_{\rm Hg}$ as state variables and eliminating the other latent variables. Observe that for $u^* = 0$, all the temperatures equal 273 degrees Kelvin, and h^* the corresponding value given by (A.9) is an equilibrium point. Linearizing around this equilibrium yields a system of equations of the following form:

$$\begin{aligned}
\frac{dT_{\rm c}}{dt} &= \alpha_1(T_{\rm a} - T_{\rm c}) + \beta_1 u, \\
\frac{dT_{\rm g}}{dt} &= \alpha_2(T_{\rm c} - T_{\rm g}) + \alpha_3(T_{\rm Hg} - T_{\rm g}), \\
\frac{dT_{\rm Hg}}{dt} &= \alpha_4(T_{\rm g} - T_{\rm Hg}), \\
h &= \gamma_1 T_{\rm Hg} - \gamma_2 T_{\rm g}.
\end{aligned}$$

The parameters appearing in the above equation are all positive, with $\gamma_1 > \gamma_2$. In the remainder of this exercise, take the following values for these parameters: $\beta_1 = 0.1, \alpha_1 = 0.5, \alpha_2 = 1, \alpha_3 = 0.1, \alpha_4 = 0.2, \gamma_1 = 0.7, \gamma_2 = 0.05$.

2. Let us first consider the system without control: Take $u = 0$. Assume that the ambient temperature increases by a unit amount; i.e., take for $T_{\rm a}$ the unit step

$$T_{\rm a}(t) = \begin{cases} 1 & \text{for } t \geq 0, \\ 0 & \text{for } t < 0. \end{cases} \tag{A.10}$$

Take $T_{\rm c}(0) = T_{\rm g}(0) = T_{\rm Hg}(0) = 0$. Plot (all on the same graph) the responses $T_{\rm c}, T_G, T_{\rm Hg}$. Plot on a separate graph the response h. Zoom in on the small time behavior of h. Explain physically why this happens. The phenomenon that you observe is called an *adverse response*.

3. Let us now use control. It is logical to use as control law a proportional law:

$$u = K_P h. \tag{A.11}$$

So if the measured temperature goes down, we supply more heat, etc. Now study the performance of this control law for several values of K_P. Put

$T_a = 0$, $T_c(0) = 1$, $T_G(0) = 0$, $T_{Hg}(0) = 0$, and plot the response T_c for a wide range of values of K_P. Show that for K_P small, the performance, in particular the *settling time* (explain what you mean by this), is not very good. Show that for K_P large the performance becomes awful (the poor performance obtained from such an all-too-enthusiastic controller is called *overcompensation*). This is due to the adverse response. You are welcome to try to explain this. Show the response for the K_P that has your preference.

4. Now use this controller (A.11) and examine the step response of the controlled system. Take for T_a again the unit step (A.10) and $T_c(0) = T_g(0) = T_{Hg}(0) = 0$, and simulate the response T_c again for a wide range of values of K_P. Pay special attention to the steady-state value. Plot the steady-state as a function of K_P. Now, taking into account settling time, overshoot, steady-state error, what value of K_P do you now prefer? Plot the step response from T_a to T_c for this controller.
5. You should be unhappy still, since your control was unable to compensate the effect of a steady-state increase in the ambient temperature. One way (a very clever idea of early—1920s—control theory) is to control not only on the basis of h, but also on the basis of the integral of h. Think about it: a controller that acts on the integral of h cannot accept a steady-state error. It generates an unbounded correction signal if a steady-state error is present.

   Consider therefore the following control law:

$$\frac{dz}{dt} = h, \quad u = K_P h + K_I z.$$

   Now take for T_a step (6), and for the initial conditions $T_c(0) = T_g(0) = T_{Hg}(0) = 0$, $z(0) = 0$. Take for K_P the value obtained in the previous part of this exercise, and simulate for a wide range of K_Is. Settle on a preferred value of K_I. Show the resulting step response from T_a to T_c.

A.3 Autonomous Dynamics of Coupled Masses

The purpose of this simulation exercise is to study some interesting oscillatory phenomena using the theory of autonomous behaviors as developed in Section 3.2. Consider the mass–spring system of Figure A.3. The two masses are taken to be unity. For a theoretical analysis see Exercise 3.13.

1. Derive the behavioral equations for this system and write them in the form $P(\frac{d}{dt})w = 0$ for a suitable matrix $P(\xi) \in \mathbb{R}^{2\times 2}[\xi]$.
2. Determine the characteristic polynomial and the characteristic values of this system.
3. Derive the general form of the trajectories in the behavior of the system. Write it in trigonometric form as follows:

$$\begin{aligned} w_1(t) &= \alpha \cos\sqrt{k_1}t + \beta \sin\sqrt{k_1}t + \gamma \cos\sqrt{k_1 + 2k_2}t + \delta \sin\sqrt{k_1 + 2k_2}t, \\ w_2(t) &= \alpha \cos\sqrt{k_1}t + \beta \sin\sqrt{k_1}t - \gamma \cos\sqrt{k_1 + 2k_2}t - \delta \sin\sqrt{k_1 + 2k_2}t. \end{aligned} \tag{A.12}$$

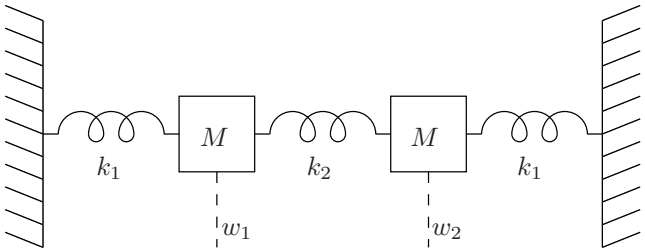


FIGURE A.3. Mass–spring system.

4. Take $k_1 = 25$ and $k_2 = 1$. Physically this means that the masses are connected to the walls by means of relatively strong springs, whereas their mutual spring connection is rather weak. Assume that at time $t = 0$ both masses have velocity zero, with the first mass displaced from its equilibrium by one unit, and the second mass in its equilibrium. Simulate the behavior of (w_1, w_2) with this initial condition. Observe that it appears as if the periodic motion of the two masses is periodically exchanged between them.

5. Use the formula $\cos p + \cos q = 2\cos\frac{p-q}{2}\cos\frac{p+q}{2}$ to explain that the behavior of w_1 and w_2 given by (A.12) can be seen as a fast oscillation modulated by a slowly oscillating amplitude. Determine the slow and the fast frequencies.

6. Show for the solution derived above that the slowly oscillating amplitudes of w_1 and w_2 are in *antiphase*. Explain what you mean by this.

7. Now take $k_1 = 1$ and $k_2 = 25$. This corresponds to the situation where the two masses are connected to each other by means of a strong spring and are connected to the walls by weak springs. Simulate the behavior of (w_1, w_2), and observe that the masses appear to oscillate in antiphase with relatively high frequency about "equilibria" that are themselves slowly oscillating in phase with each other.

8. Explain this behavior mathematically, and determine the slow and the fast frequencies.

A.4 Satellite Dynamics

The purpose of this exercise is to illustrate the modeling of the dynamics of a satellite and the determination of a desired equilibrium. Subsequently, the motion is linearized, and stability, controllability, and observability are analyzed. Through simulation, we finally also illustrate the extent to which linearized equations approximate the nonlinear ones. The exercise illustrates the theory covered in Chapter 5.

A.4.1 Motivation

The most important civilian spin-off of the space program is undoubtedly communication via satellites. The idea is a simple one: instead of transmitting a message over a wire, it is beamed up from a transmitter to a satellite, and from there the message is routed further to a receiver. For obvious reasons it is desirable that the satellite be located in a position in the firmament that is fixed for an observer on Earth. Not only does this avoid having to track the satellite continuously, but it also results in the fact that the satellite can be used at all times, since it never disappears below the horizon. Satellites that sit at a fixed position in the sky are employed for telephone and TV communication, in navigation for ships and airplanes, for weather prediction, etc. Such satellites are called *geostationary* satellites. In principle, a geostationary orbit could be achieved by exerting forces on the satellite such that it remains in the desired orbit. Such forces can be produced by means of small jets that are mounted on the satellites. However, it is undesirable to require that forces be exerted continuously. These jets get their energy from fuel that the satellite must take along at launch (present space programs aim at refueling satellites) or from solar panels. However, energy is a scarce resource far up in the sky, and as such it is desirable that the jets not be used for continuously steering the satellite, but only for unavoidable orbit corrections and other maneuvers.

The question thus arises, Is there a geostationary equilibrium orbit for a satellite when the only force exerted on it is the gravitational force field of the earth? Is this orbit stable? If not, what controls are required in order to keep the satellite in its orbit?

If we assume that the satellite is only influenced by the gravitational field of the earth, then its orbit obeys Kepler's laws. Hence the satellite moves in an elliptical orbit with the center of the earth at one of the foci. Thus, as a consequence of the fact that the earth turns around its North Pole/South Pole axis at a rate of 2π radians/day, a geostationary orbit must be circular and in the equatorial plane. Kepler's laws also imply that there is a relation between the diameter of the circular orbit and its period of revolution, which for the satellite to be geostationary, must be the same as that of the earth. Our first order of business is to determine the height of such a circular orbit.

A.4.2 Mathematical modeling

We now derive the dynamical equations of the motion of the satellite. It is subject to four kinds of forces:

1. The inertial force, $\vec{F}_{\text{in}}$.
2. The gravitational pull of the earth, $\vec{F}_{\text{g}}$.
3. External forces due to the jets $\vec{F}_{\text{jet}}$: these are our controls.
4. Other external forces such as the gravitational pull of the moon and the sun and solar wind. We denote these forces by $\vec{F}_{\text{d}}$. These are disturbances, and it is precisely the aim of the controls to compensate for the unpredictable influence of these disturbances.

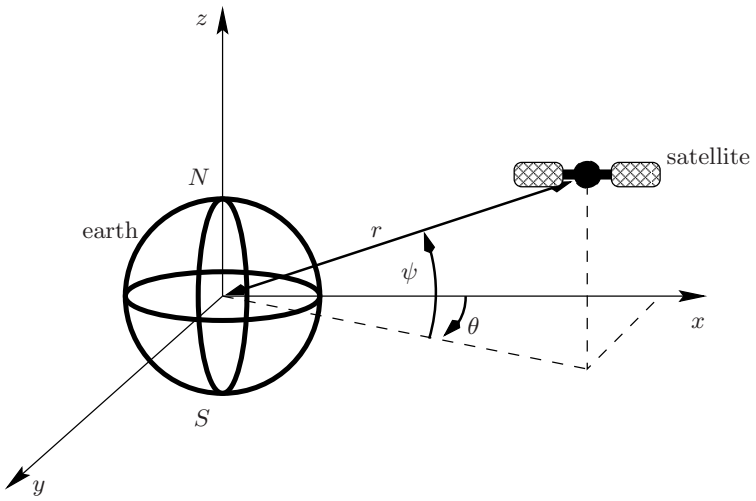


FIGURE A.4. Satellite in Earth orbit.

The position of the satellite can be described by its polar coordinates (r, ψ, θ). Figure A.4 shows the relevant geometry. In principle, the motions of these coordinates are coupled. However, we study the motion in the equatorial plane only. The resulting geometry is shown in Figure A.5.

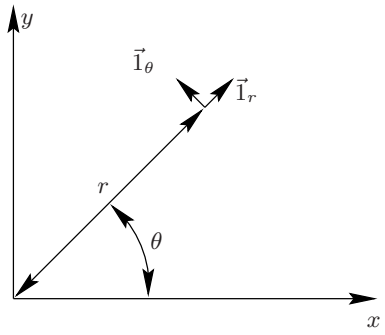


FIGURE A.5. The vectors $\vec{1}_\theta$ and $\vec{1}_r$.

We now set up the equations of motion by expressing equality of forces. Denote the unit vectors in the radial and tangential direction by $\vec{1}_r$ and $\vec{1}_\theta$ respectively.

Then $\vec{r} = r \cdot \vec{1}_r$. Next, observe that

$$\begin{aligned}
&\frac{d}{dt}\vec{1}_r = \frac{d\theta}{dt}\vec{1}_\theta \text{ and } \frac{d}{dt}\vec{1}_\theta = -\frac{d\theta}{dt}\vec{1}_r,\\
&\frac{d}{dt}\vec{r} = \frac{dr}{dt}\vec{1}_r + r\frac{d\theta}{dt}\vec{1}_\theta,\\
&\frac{d^2}{dt^2}\vec{r} = (\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2)\vec{1}_r + (r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt})\vec{1}_\theta.
\end{aligned} \tag{A.13}$$

The inertial force $\vec{F}_{\text{in}}$ is hence given by

$$m\frac{d^2}{dt^2}\vec{r} = m(\frac{d^2r}{dt^2} - r\left(\frac{d^2\theta}{dt}\right)^2)\vec{1}_r + m(r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt})\vec{1}_\theta,$$

where m denotes the mass of the satellite. The gravitational pull of the earth on the satellite $\vec{F}_{\text{g}}$ is given by the inverse square law

$$\vec{F}_{\text{g}} = -k\frac{m}{r^2}\vec{1}_r,$$

with $k = 4.10^{14}$ m^3/sec^2, which is obtained by multiplying the gravitational constant by the mass of the earth.

We assume that the jets of the satellite exert a force $\vec{F}_{\text{jet}}$. Decompose this force into a control force u_r in the radial direction and a control force u_θ in the tangential direction. The disturbance force $\vec{F}_{\text{d}}$ is similarly decomposed into a force d_r in the radial direction and a force d_θ in the tangential direction. Hence $\vec{F}_{\text{jet}} = u_r\vec{1}_r + u_\theta\vec{1}_\theta$ and $\vec{F}_d = d_r\vec{1}_r + d_\theta\vec{1}_\theta$.

Expressing that the sum of the forces acting on the satellite is zero yields

$$m(\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2)\vec{1}_r + m(r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt})\vec{1}_\theta + \frac{km}{r^2}\vec{1}_r = \vec{F}_{\text{jet}} + \vec{F}_{\text{d}}.$$

Taking the components in the radial and tangential directions yields

$$\begin{aligned}
\frac{d^2r}{dt^2} &= r\left(\frac{d\theta}{dt}\right)^2 - \frac{k}{r^2} + \frac{u_r}{m} + \frac{d_r}{m},\\
\frac{d^2\theta}{dt^2} &= -\frac{2\frac{dr}{dt}\frac{d\theta}{dt}}{r} + \frac{u_\theta}{r\cdot m} + \frac{d_\theta}{r\cdot m}.
\end{aligned} \tag{A.14}$$

These differential equations give us the dynamical equations linking the variables r, θ to the control inputs u_r, u_θ and the disturbance inputs d_r and d_θ.

These equations of motion can also be derived from the Euler–Lagrange equations. In order to do that, we should first express the potential and the kinetic energy of the satellite. Introduce the variables $\dot{r}$, the radial velocity of the satellite, and $\dot{\theta}$, the rate of change of θ. The energy is a function of r, θ, $\dot{r}$, and $\dot{\theta}$. The potential energy P and the kinetic energy K are given by

$$P(r,\theta,\dot{r},\dot{\theta}) = -\frac{km}{r}, \quad K(r,\theta,\dot{r},\dot{\theta}) = \frac{1}{2}m(\dot{r}^2 + r^2(\dot{\theta})^2).$$

The Lagrangian L is $K-P$. According to the principle of Lagrange, the equations of motion are then given by (please take careful note of the notation)

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{r}}(r,\theta,\frac{dr}{dt},\frac{d\theta}{dt}) - \frac{\partial L}{\partial r}(r,\theta,\frac{dr}{dt},\frac{d\theta}{dt}) = u_r + d_r,$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}}(r,\theta,\frac{dr}{dt},\frac{d\theta}{dt}) - \frac{\partial L}{\partial \theta}(r,\theta,\frac{dr}{dt},\frac{d\theta}{dt}) = u_\theta + d_\theta.$$

Derive equations (A.14) from Lagrange's equations.

A.4.3 Equilibrium Analysis

Prove that there exists an R such that the trajectory $r(t) = R, \theta(t) = \Omega t$, with $\Omega = 2\pi/\text{day}$, $u_r(t) = 0, u_\theta(t) = 0, d_r(t) = 0, d_\theta(t) = 0$, is a solution of the equations of motion. Explain why this corresponds to a geostationary orbit. Compute R in meters.

This calculation lets you conclude that a geostationary orbit sits at a distance of 42,000 km from the center of the earth, hence approximately 36,000 km above the surface of the earth. In two-way telephone communication this yields a delay of at least $4 \times 36,000$ km/speed of light $\approx \frac{1}{2}$ sec. This delay can be cumbersome in two-way intercontinental voice communication.

Introduce the new variable $\phi(t) = \theta(t) - \Omega t$. The equations for r, ϕ become

$$\begin{aligned} \frac{d^2 r}{dt^2} &= r\left(\frac{d\phi}{dt} + \Omega\right)^2 - \frac{k}{r^2} + \frac{u_r}{m} + \frac{d_r}{m}, \\ \frac{d^2\phi}{dt^2} &= -\frac{2\frac{dr}{dt}(\frac{d\phi}{dt} + \Omega)}{r} + \frac{u_\theta}{r \cdot m} + \frac{d_\theta}{r \cdot m}. \end{aligned} \tag{A.15}$$

In order to complete the model, we now specify the observed output. Take the *sighting angle*, $\phi(t) = \theta(t) - \Omega t$, as the measured output. Note that we have thus obtained a system with control inputs u_r, u_θ; disturbance inputs d_r, d_θ; variables $r, \phi, \dot{r}, \dot{\phi}$; measured output ϕ; and to-be-controlled output ϕ. Verify that $r^* = R, \phi^* = 0, u_r^* = 0, u_\theta^* = 0, d_r^* = 0, d_\theta^* = 0$ is an equilibrium solution of (A.15). It is in this position that we hope to find the satellite at all times. We now verify the stability of this equilibrium.

Prove that $r(t) = \sqrt[3]{k/(\Omega+\Delta)^2}, \phi(t) = \Delta t$, with $\Delta \in \mathbb{R}$ satisfying $\Omega + \Delta > 0$, is also a solution of (A.15). Show that this implies that the geostationary orbit is not stable without controls. In other words, a small deviation from the equilibrium may cause the sighting angle to drift further and further away. The uncontrolled motion is hence unstable, and control action is required in order to keep a satellite in a geostationary orbit.

A.4.4 Linearization

Introduce the state variables $r, \frac{dr}{dt}, \phi, \frac{d\phi}{dt}$, and write (A.15) in state space form. Linearize the dynamical equations around the geostationary equilibrium. For ver-

ification, you should now have obtained the following linearized system:

$$\frac{d}{dt}\begin{bmatrix}\Delta r\\ \Delta\dot{r}\\ \Delta\phi\\ \Delta\dot{\phi}\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0\\ 3\Omega^2 & 0 & 0 & 2R\Omega\\ 0 & 0 & 0 & 1\\ 0 & -\frac{2\Omega}{R} & 0 & 0\end{bmatrix}\begin{bmatrix}\Delta r\\ \Delta\dot{r}\\ \Delta\phi\\ \Delta\dot{\phi}\end{bmatrix} + \begin{bmatrix}0\\ \frac{1}{m}\\ 0\\ 0\end{bmatrix} u_r + \begin{bmatrix}0\\ 0\\ 0\\ \frac{1}{Rm}\end{bmatrix} u_\theta + \begin{bmatrix}0\\ \frac{1}{m}\\ 0\\ 0\end{bmatrix} d_r + \begin{bmatrix}0\\ 0\\ 0\\ \frac{1}{Rm}\end{bmatrix} d_\theta,$$

$$\Delta\varphi = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}\Delta r\\ \Delta\dot{r}\\ \Delta\phi\\ \Delta\dot{\phi}\end{bmatrix},$$

where $\Omega = 7.3 \times 10^{-5}$ rad/sec, $m =$ the mass of the satellite, and $R = 4.2 \times 10^7$ meters. If you were unable to derive these as the equations, continue with these equations of motion and try to figure out what went wrong.

A.4.5 Analysis of the model

Is the linearized system stable, asymptotically stable, or unstable? Explain the eigenvalues of the linearized A-matrix using Kepler's laws. Is the linearized system controllable with control u_r, or u_θ, or both? Is it observable?

A.4.6 Simulation

The final part of this exercise consists in simulating typical responses in order to obtain a feeling for the dynamics of this system and for the difference between the behavior of the nonlinear and the linearized equations of motion.

Simulate the response for r and ϕ for both the linearized and the nonlinear system for the following situations.

1. Take a small (1%) initial disturbance for $r(0)$. Repeat with a small initial disturbance for $\phi(0)$.

2. Plot responses for a radial step disturbance input equal to 5% of the gravitational pull of the earth. You could think of this disturbance as being due to the gravitational pull of the moon.

3. Repeat with a radial disturbance of that magnitude, but assume it to be periodic with period equal to the period of revolution of the earth.

A.5 Dynamics of a Motorbike

The purpose of this exercise is to analyze the dynamics of a motorbike riding over a rough road. This exercise illustrates the theory developed in Chapter 8.

Consider a motorbike that rides over a rough road, as shown in Figure A.6. We

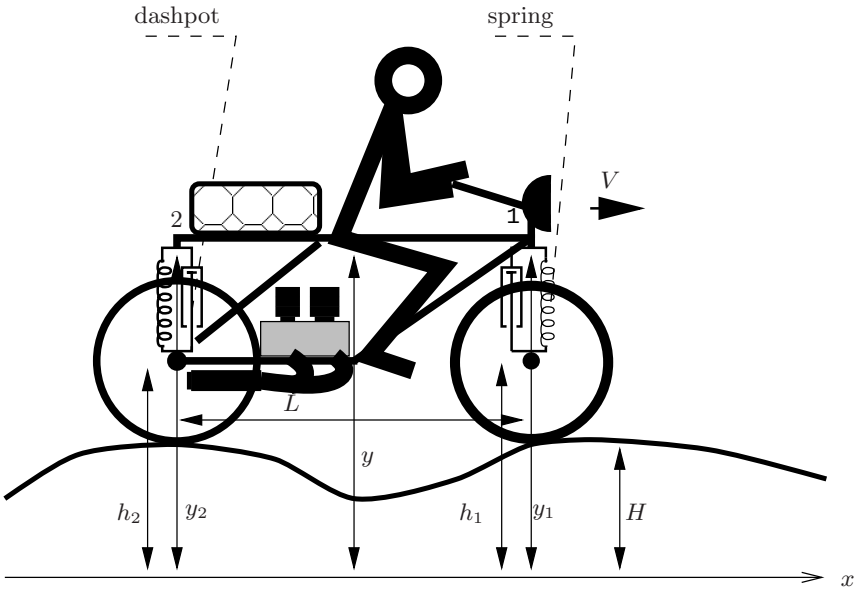


FIGURE A.6. A motorbike.

are interested in the relation between the road profile and the vertical motion of the driver. We assume that the motorbike has constant forward velocity. Consequently, we can assume that the height of the wheels is a certain function of time, with the height of the rear wheel a delayed version of the height of the front wheel. Initially, we ignore this time delay. Throughout, we neglect rotational motions of the motorbike that would occur in reality.

1. $L = 0$. When we neglect the length of the bike as compared to the rate of change in the road profile, we obtain the following equations of motion:

$$M\frac{d^2}{dt^2}y = -2K(y-h) - 2D(\frac{d}{dt}y - \frac{d}{dt}h), \qquad h(t) = H(Vt), \qquad \text{(A.16)}$$

where M denotes the mass of the motorbike plus driver, K the spring constant of each of the springs, D the friction coefficient of the dashpots, V the forward velocity of the bike, and $H(x)$ the height of the road at distance x from a reference point.

2. $L \neq 0$. When we do not neglect the length of the bike as compared to the rate of change in the road profile, we obtain the equations

$$\begin{aligned}
\frac{M}{2}\frac{d^2}{dt^2}y_1 &= -K(y_1 - h_1) - D(\frac{d}{dt}y_1 - \frac{d}{dt}h_1),\\
\frac{M}{2}\frac{d^2}{dt^2}y_2 &= -K(y_2 - h_2) - D(\frac{d}{dt}y_2 - \frac{d}{dt}h_2),\\
y &= \frac{1}{2}(y_1 + y_2), \quad h_1(t) = H(Vt), \quad h_2(t) = h(Vt - L).
\end{aligned}$$

   Explain each of these equations. Explain why it is logical to consider H and y as the manifest variables and h, h_1, h_2, y_1, y_2 as latent variables. The system parameters are M, K, D, L, and V. Take as values for the system parameters $M = 300$ kg, $K = 10,000$ kg/sec^2, $D = 3,000$ kg/sec, $L = 1$ meter, $V = 90$ km/hour. Argue that these figures are in the correct ballpark by reasoning about what sort of value you would expect for the natural frequency, for the steady-state gain obtained by putting a weight on the bike, and for the damping coefficient as observed from the overshoot after taking the weight back off.

3. *Simulation.* Plot the step response in the case $L = 0$. Determine the resonant frequency, the peak gain, and the pass-band. Repeat when L is not neglected. What happens to these plots when the forward velocity V changes? Repeat this for the case that the bike has a defective damper so that its damping coefficient is first reduced to 50%, and subsequently to 10% of its original value. Repeat this again for the case that the bike has a defective spring so that its spring coefficient is first reduced to 50%, and subsequently to 10% of its original value.

A.6 Stabilization of a Double Pendulum

The purpose of this exercise is to illustrate the full extent of the theory developed in Chapters 9 and 10. The exercise uses many of the concepts introduced in this book (modeling, controllability, observability, stability, pole placement, observers, feedback compensation). We recommend that it be assigned after all the theory has been covered, as a challenging illustration of it. This exercise requires extensive use of computer aids: Mathematica© for formula manipulation, and MATLAB© for control system design and numerical simulation.

A.6.1 Modeling

We study the stabilization of a double pendulum mounted on a movable cart. The relevant geometry is shown in Figure A.7. It is assumed that the motion takes place in a vertical plane. The significance of the system parameters is as follows:

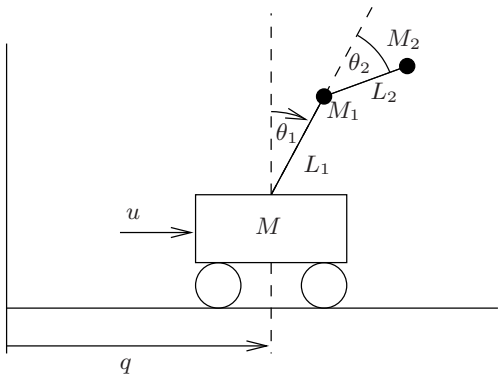


FIGURE A.7. A double pendulum on a cart.

M : mass of the cart
M_1: mass of the first pendulum
M_2: mass of the second pendulum
L_1: length of the first pendulum
L_2: length of the second pendulum

The cart and the pendula are all assumed to be point masses, with the masses of the pendula concentrated at the top. It is instructive, however, to consider how the equations would change if the masses of the pendula are uniformly distributed along the bars.

The significance of the system variables is as follows:

u: the external force on the cart
q: the position of the cart
θ_1: the inclination angle of the first pendulum
θ_2: the inclination angle of the second pendulum

For the output y we take the 3-vector consisting of the horizontal positions of the cart and of the masses at the top of the pendula.

The purpose of this exercise is to develop and test a control law that holds the cart at a particular position with the pendula in upright position. We assume that all three components of the output y are measured and that the force u is the control input. Our first order of business is to find the dynamical relation between u and y. For this, we use Lagrange's equations. In order to express the energy of this system, introduce also the variables

$\dot{q}$: the velocity of the cart
$\dot{\theta}_1$: the rate of change of θ_1
$\dot{\theta}_2$: the rate of change of θ_2

The kinetic energy is given by

$$\begin{aligned} K(q,\theta_1,\theta_2,\dot{q},\dot{\theta}_1,\dot{\theta}_2) \;=\; & \tfrac{1}{2}M\dot{q}^2 + \tfrac{1}{2}M_1[(\dot{q}+L_1\dot{\theta}_1\cos\theta_1)^2 + (L_1\dot{\theta}_1\sin\theta_1)^2] \\ & + \tfrac{1}{2}M_2[(\dot{q}+L_1\dot{\theta}_1\cos\theta_1 + L_2(\dot{\theta}_1+\dot{\theta}_2)\cos(\theta_1+\theta_2))^2 \\ & + (L_1\dot{\theta}_1\sin\theta_1 + L_2(\dot{\theta}_1+\dot{\theta}_2)\sin(\theta_1+\theta_2))^2]. \end{aligned}$$

The potential energy is given by

$$P(q,\theta_1,\theta_2,\dot{q},\dot{\theta}_1,\dot{\theta}_2) = M_1 g L_1 \cos\theta_1 + M_2 g[L_1\cos\theta_1 + L_2\cos(\theta_1+\theta_2)].$$

Lagrange's principle lets us write the equations of motion directly from K and P. In other words, once we have modeled K and P, we have the dynamical equations that we are looking for. Lagrange's principle is a truly amazingly effective modeling tool for mechanical systems. An alternative but much more cumbersome way of obtaining the equations of motion would be to express equality of forces for each of the masses involved. Define the Lagrangian $L := K - P$ and obtain the equations of motion as (please take note of the notation)

$$\begin{aligned}
&\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) - \frac{\partial L}{\partial q}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) = u,\\
&\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}_1}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) - \frac{\partial L}{\partial \theta_1}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) = 0,\\
&\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}_2}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) - \frac{\partial L}{\partial \theta_2}(q,\theta_1,\theta_2,\frac{dq}{dt},\frac{d\theta_1}{dt},\frac{d\theta_2}{dt}) = 0.
\end{aligned}$$

Note that these equations contain many partial derivatives of the functions K and P, which are rather complex expressions of their arguments. Carrying out such differentiations by hand is not something one looks forward to. However, there are computer tools that do this for us. Use Mathematica© to derive the dynamical equations. You should obtain

$$\begin{aligned}
&-(L_1M_1 + L_1M_2)(\tfrac{d\theta_1}{dt})^2\sin\theta_1 - L_2M_2\tfrac{d\theta_1}{dt}\;\tfrac{d\theta_2}{dt}\sin(\theta_1+\theta_2)\\
&-L_2M_2(\tfrac{d\theta_2}{dt})^2\sin(\theta_1+\theta_2) + (M+M_1+M_2)\tfrac{d^2q}{dt^2}\\
&+(L_1M_1+L_1M_2)\tfrac{d^2\theta_1}{dt^2}\cos\theta_1 + L_2M_2\tfrac{d^2\theta_2}{dt^2}\cos(\theta_1+\theta_2) \qquad = u,
\end{aligned}$$

$$\begin{aligned}
&-gL_1(M_1+M_2)\sin\theta_1 - gL_2M_2\sin(\theta_1+\theta_2)\\
&+L_2M_2\tfrac{dq}{dt}\;\tfrac{d\theta_2}{dt}\sin(\theta_1+\theta_2) - L_1L_2M_2(\tfrac{d\theta_2}{dt})^2\sin\theta_2\\
&+(L_1M_1+L_1M_2)\tfrac{d^2q}{dt^2}\cos\theta_1 + L_1^2(M_1+M_2)\tfrac{d^2\theta_1}{dt^2}\\
&+L_1L_2M_2\tfrac{d^2\theta_2}{dt^2}\cos\theta_2 \qquad = 0,
\end{aligned} \tag{A.19a}$$

$$\begin{aligned}
&-gL_2M_2\sin(\theta_1+\theta_2) - L_2M_2\tfrac{dq}{dt}\;\tfrac{d\theta_1}{dt}\sin(\theta_1+\theta_2)\\
&+L_2M_2\tfrac{d^2q}{dt^2}\cos(\theta_1+\theta_2) + L_1L_2M_2\tfrac{d^2\theta_1}{dt^2}\cos\theta_2 + L_2^2M_2\tfrac{d^2\theta_2}{dt^2} \;=\; 0.
\end{aligned} \tag{A.19b}$$

Completed with the output equation

$$y = \begin{bmatrix} q \\ g + L_1\sin\theta_1 \\ q + L_1\sin\theta_1 + L_2\sin(\theta_1+\theta_2) \end{bmatrix}, \tag{A.20}$$

we obtain a full system of equations relating the input to the output.

A.6.2 Linearization

Prove that $u^* = 0, q^* = 0, \theta_1^* = 0, \theta_2^* = 0, y^* = 0$ is an equilibrium. Explain physically that this is as expected. Do you see other equilibria?

Introduce as state variables $x_1 = q, x_2 = \theta_1, x_3 = \theta_2, x_4 = \dot{q}, x_5 = \dot{\theta}_1, x_6 = \dot{\theta}_2$. Derive the input/state/output equations; i.e., write the equations in the form

$$\frac{dx}{dt} = f(x,u), y = h(x). \tag{A.21}$$

Note that in order to do this, you have to invert a matrix. It is recommended that you use Mathematica©: who wants to invert matrices by hand if you can let a computer do this for you? Use Mathematica© to linearize the nonlinear input/state/output equations around the equilibrium that you derived. You should obtain the following equations:

$$(M + M_1 + M_2)\frac{d^2\Delta q}{dt^2} + (L_1M_1 + L_1M_2)\frac{d^2\Delta\theta_1}{dt^2} + L_2M_2\frac{d^2\Delta\theta_2}{dt^2} = \Delta u,$$

$$\begin{aligned}(L_1M_1 + L_1M_2)\tfrac{d^2\Delta q}{dt^2} - g(L_1(M_1 + M_2) + L_2M_2)\Delta\theta_1 \\ +L_1^2(M_1 + M_2)\tfrac{d^2\Delta\theta_1}{dt^2} - gL_2M_2\Delta\theta_2 + L_1L_2M_2\tfrac{d^2\Delta\theta_2}{dt^2} \;&= 0,\end{aligned} \tag{A.22a}$$

$$\begin{aligned}L_2M_2\tfrac{d^2\Delta q}{dt^2} - gL_2M_2\Delta\theta_1 + L_1L_2M_2\tfrac{d^2\Delta\theta_1}{dt^2} \\ -gL_2M_2\Delta\theta_2 + L_2^2M_2\tfrac{d^2\Delta\theta_2}{dt^2} \;&= 0,\end{aligned} \tag{A.22b}$$

$$\begin{bmatrix}\Delta q \\ \Delta q + L_1\Delta\theta_1 \\ \Delta q + (L_1 + L_2)\Delta\theta_1 + L_2\Delta\theta_2\end{bmatrix} = \Delta y. \tag{A.22c}$$

Or in state space form,

$$\tfrac{d\Delta x}{dt} = A\Delta x + B\Delta u, \qquad \Delta y = C\Delta x,$$

with

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -\frac{g(L_1(M_1+M_2)+L_2M_2)}{L_1M} & -\frac{gL_2M_2}{L_1M} & 0 & 0 & 0 \\ 0 & \frac{g(L_1M_1(M+M_1+M_2)+L_2M_2(M+M_1))}{L_1^2MM_1} & \frac{gM_2(-L_1M+L_2(M+M_1))}{L_1^2MM_1} & 0 & 0 & 0 \\ 0 & -\frac{gM_2}{L_1M_1} & \frac{g(L_1(M_1+M_2)-L_2M_2)}{L_1L_2M_1} & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{M} \\ -\frac{1}{L_1M} \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & L_1 & 0 & 0 & 0 & 0 \\ 1 & L_1 + L_2 & L_2 & 0 & 0 & 0 \end{bmatrix}.$$

A.6.3 Analysis

For what values of the system parameters M, M_1, M_2, L_1, L_2 (all > 0) is this linearized system stable/asymptotically stable/unstable? Controllable? Observable? Is the equilibrium a stable/asymptotically stable/unstable equilibrium of the nonlinear system?

Assume henceforth the following reasonable choices for the system parameters: $M = 100$ kg, $M_1 = 10$ kg, $M_2 = 10$ kg, $L_1 = 2$ m, $L_2 = 1$ m.

Use MATLAB© to compute the eigenvalues of the resulting system matrix A and plot them in the complex plane. Plot the Bode diagrams, with u as input and y as output. Note that you should have three diagrams, one for each of the output components.

A.6.4 Stabilization

- We first stabilize the system using state feedback. The system is sixth order, the control is a scalar. Thus we have to choose six eigenvalues, $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$, in the left half plane and compute the six components of the feedback gain such that the closed loop system matrix has the desired eigenvalues. In order to pick the λs (and from there the feedback gain matrix), you should experiment a bit, using the linearized system. Use the following initial conditions in your experiment: $x_1(0) = -5\text{m}, x_2(0) = 0, x_3(0) = 0, x_4(0) = 0, x_5(0) = 0, x_6(0) = 0$. This corresponds to making a maneuver: the cart is moved from one equilibrium position to the desired one, with the cart at the origin. You should choose the λs such that the transient response does not have excessive overshoot and a reasonable settling time. We suggest that you try the following λs: $-7.5 \pm 0.3i, -6.5 \pm 0.9i, -3.3 \pm 2.3i$. Plot the transient responses x_1, x_2, x_3 for the linearized system, and subsequently for the nonlinear system, with your chosen λs. Explain why you liked your λs better than the others that you tried.

  Note that you obtained a good transient response notwithstanding a rather high initial disturbance. Observe in particular the interesting small time behavior of $x_1 = q$.

- Obtain a state observer based on the measured output y and the input u. Choose the eigenvalues of the error dynamics matrix $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ by considering the initial estimation error $e_1(0) = 5\text{m}, e_2(0) = 0, e_3(0) = 0, e_4(0) = 0, e_5(0) = 0, e_6(0) = 0$, and tuning the μs so that the resulting error transients e_1, e_2, e_3 show a reasonable settling time without excessive overshoot. Plot these transients for the μs that you selected, for the linearized system, and subsequently for the nonlinear system. Note that the observer gains are not unique in this case, since the observed output is three-dimensional. In this case, MATLAB© optimizes the chosen gains in a judicious way: it minimizes the sensitivity of the error dynamics eigenvalues.

  It appears not easy to obtain a reasonable performance for the observer. The following μs gave us some of the best results: $-10, -10, -5, -3, -1, -1$.

- Combine the state feedback gains and the observer gains obtained before in order to obtain a controller from y to u. Test this controller by plotting the transient responses of x_1, x_2, x_3 for the linearized system, and subsequently for the nonlinear system, with the initial disturbances: $x_1(0) = -5\text{m}, x_2(0) = 0, x_3(0) = 0, x_4(0) = 0, x_5(0) = 0, x_6(0) = 0$. This corresponds to the same maneuver used before. The initial state estimates

are $\hat{x}(0) = x(0)$, $\hat{x}(0) = x(0)+$ a small error, and $\hat{x}(0) = [1,\ 0,\ 0,\ 0,\ 0,\ 0]$. The results for the first two initial conditions are good (explain), but not for the third. Conclude that in order to use this controller, one should always reset the observer so that its initial state estimate is accurate.

- Test the robustness of your controller against parameter changes. More concretely, you have obtained a controller that stabilizes the equilibrium for specific values of M, M_1, M_2, L_1, L_2. Now keep the controller fixed, and compute the range of values of M for which this controller remains stabilizing.

A.7 Notes and References

The advent of easy-to-use software packages such as MATLAB© and Mathematica© greatly enhances the applicability of mathematical methods in engineering. There are many recent texts (for example [36]) that aim at familiarizing students with MATLAB©, applied to the analysis of linear systems and the design of control systems. The impossibility of stabilizing a point mass using memoryless position feedback in A.1 is a well-known phenomenon. In [52] it is also used as an example motivating the need for control theory. The occurrence of an adverse response in thermal systems, demonstrated in A.2, is a typical non minimum phase phenomenon. It implies, for example, that high-gain feedback leads to instability and illustrates the need for careful tuning of controller gains. The interesting dynamical response of (weakly) coupled oscillators illustrated in A.3 was already observed by Huygens, and has been the subject of numerous analyses since. The need for control in order to stabilize a geostationary satellite in its station-keeping equilibrium position explained in A.4 is a convincing and very relevant example of a control problem. There is a large literature on this and related topics. See [14] for a recent reference and an entry into the literature. In A.5 we discuss only some very simple aspects of the dynamics of a motorbike. Designing an autonomous device (for example, a robot) that stably rides a bicycle is one of the perennial challenges for control engineering laboratories. Stabilization of a double pendulum in its upright positions (see A.6) is a neat application of the theory of stabilization of a nonlinear system around a very unstable equilibrium. Many control laboratories have an experimental setup in which such a control law is implemented. Note that our results only discuss local stability. Recent papers [6] and experimental setups implement also the swing-up of a double pendulum. Such control laws must, of course, be nonlinear: the double pendulum starts in an initial position in which both pendula hang in a downward position, and by exerting a force on the supporting mass, the pendula swing up to the stabilized upright equilibrium.

Appendix B
Background Material

B.1 Polynomial Matrices

In this section we have collected some results on polynomial matrices that we have used in the book. We state the results for polynomial matrices with real coefficients. However, all the results hold for matrices with entries in $\mathbb{C}$ or in a more general Euclidean ring, and the reader is encouraged to translate these results for matrices with entries in the integers. In the latter case *degree* should be replaced by *absolute value*. Unless otherwise stated, $R(\xi)$ is a fixed polynomial matrix in $\mathbb{R}^{g\times q}[\xi]$. To streamline the proofs we will sometimes cover only the square case ($g = q$) in full detail. The general case is then left as an exercise.

Theorem B.1.1 (upper triangular form) *There exists a unimodular matrix* $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ *such that* $U(\xi)R(\xi) = T(\xi)$ *and* $T_{ij}(\xi) = 0$ *for* $i = 1, \ldots, n,\ j < i$.

Proof Consider the first nonzero column of $R(\xi)$. Choose in that column a nonzero element of minimal degree and use that element to carry out division with remainder on the other elements in that column. More precisely, let j_1 be the index of the first nonzero column. If necessary interchange rows (premultiplication by a permutation matrix) so as to achieve that the $(1, j_1)$ element is nonzero and has minimal degree within the j_1th column. Call this element $R_{1,j_1}(\xi)$. Division with remainder yields

$$R_{i,j_1}(\xi) = Q_{i,j_1}(\xi)R_{1,j_1}(\xi) + r_{i,j_1}(\xi), \quad \deg r_{i,j_1}(\xi) < \deg R_{i,j_1}(\xi),\ i = 2, \ldots, g.$$

Consider

$$\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -Q_{2,j_1}(\xi) & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -Q_{g,j_1}(\xi) & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad R_{\cdot,j_1}(\xi) = \begin{bmatrix} R_{1,j_1}(\xi) \\ r_{2,j_1}(\xi) \\ \vdots \\ \vdots \\ r_{g,j_1}(\xi) \end{bmatrix}, \tag{B.1}$$

where $R_{\cdot,j_1}(\xi)$ denotes the j_1th column of $R(\xi)$. Obviously, the matrix in (B.1) is unimodular. Search for the nonzero entry of minimal degree in the right-hand side of (B.1). Interchange the first row and the row in which this entry appears. Again, this is achieved by premultiplication by a permutation matrix. Repeat the division with remainder procedure. Every time that we apply this procedure, the minimal degree decreases by at least one. Also, we can apply the procedure as long as more than one entry in the j_1th column is nonzero. Since degrees are always nonnegative, this process stops within a finite number of steps. We have then transformed the j_1th column into a column consisting of a nonzero element in the top entry and all the other elements being zero. Remember that this has been achieved by premultiplication by unimodular matrices. Call the product of these unimodular matrices (from right to left in the order in which they appear) $U_1(\xi)$. We then have obtained

$$U_1(\xi)R(\xi) = \begin{bmatrix} 0 & \cdots & 0 & \tilde{R}_{1,j_1}(\xi) & * & \cdots & * \\ \vdots & & \vdots & 0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{bmatrix}$$

(an asterisk denotes an arbitrary element).

If $j_1 = q$ or $g = 1$, we are done. Otherwise, we consider the submatrix of the transformed matrix $U_1(\xi)R(\xi)$ consisting of the last $q - j_1$ columns and the second through the last row. Repeat the whole procedure. It follows that there exists a unimodular matrix $U_2(\xi)$ such that

$$U_2(\xi)U_1(\xi)R(\xi) = \begin{bmatrix} 0 & \cdots & 0 & \tilde{R}_{1,j_1}(\xi) & * & \cdots & \cdots & \cdots & \cdots & \cdots & * \\ \vdots & & \vdots & 0 & 0 & \cdots & 0 & \tilde{R}_{2,j_2}(\xi) & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & 0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & * & \cdots & * \end{bmatrix}.$$

Finally, after a finite number of steps, we end up with a matrix of the required form. □

Remark B.1.2 The *lower triangular form* is defined analogously. □

Theorem B.1.3 *A unimodular matrix $U(\xi) \in \mathbb{R}^{g\times g}[\xi]$ can be written as the product of elementary unimodular matrices.*

Proof Define the unimodular matrix $V(\xi)$ as $V(\xi) := U^{-1}(\xi)$. Inspection of the proof of Theorem B.1.1 yields that the triangular form of $V(\xi)$ is obtained by premultiplication by unimodular matrices of the form (all entries that are not indicated explicitly or by dots are zero)

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -q_{lj}(\xi) & 1 & & & \\ & & \vdots & & \ddots & & \\ & & \vdots & & & \ddots & \\ & & \vdots & & & & \ddots \\ & & -q_{gj}(\xi) & & & & 1 \end{bmatrix} \tag{B.2}$$

$$= \prod_{i=l}^{g} \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & 0 & 1 & & & \\ & & \vdots & & \ddots & & \\ & & -q_{ij}(\xi) & & & \ddots & \\ & & \vdots & & & & \ddots \\ & & 0 & & & & 1 \end{bmatrix}$$

and permutation matrices (to interchange two rows). Furthermore, every factor on the right-hand side of (B.2) is the product of elementary unimodular matrices, since (without loss of generality we assume that the polynomial term appears in the first column)

$$\begin{bmatrix} 1 & & & \\ & \ddots & & \\ \sum_{i=0}^{k} q_i \xi^i & & 1 & \\ & & & \ddots \\ & & & 1 \end{bmatrix} = \prod_{i=0}^{k} \begin{bmatrix} 1 & & & \\ & \ddots & & \\ q_i \xi^i & & 1 & \\ & & & \ddots \\ & & & 1 \end{bmatrix},$$

and if $q_i \neq 0$, we have

$$\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ q_i\xi^i & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & q_i & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ \xi^i & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \frac{1}{q_i} & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix},$$

The conclusion is that the upper triangular form of $V(\xi)$ is obtained by premultiplication by elementary unimodular matrices, say $N_k(\xi)D_kM_k \cdots N_1(\xi)D_1M_1V(\xi) = T(\xi)$. Since $T(\xi)$ is unimodular, its diagonal elements are nonzero constants. Without loss of generality we may, in fact, assume that the elementary factors have been chosen such that the diagonal elements of $T(\xi)$ are all equal to 1. Now consider the last column of $T(\xi)$. Since its last element is 1, we can annihilate the upper part of that column by premultiplication by matrices of the form (B.2). Analogously, we can use the remaining diagonal elements to annihilate the elements above them. All these operations correspond to premultiplication by matrices of the form (B.2). We have seen that these matrices can be written as the product of elementary factors. Combining this with the triangularization part of the proof, we conclude that by premultiplication of $V(\xi)$ by elementary unimodular matrices, we obtain the identity matrix:

$$N_m(\xi)D_mM_m \cdots N_1(\xi)D_1M_1V(\xi) = I.$$

Since $V(\xi)$ is the inverse of $U(\xi)$,it follows that

$$U(\xi) = N_m(\xi)D_mM_m \cdots N_1(\xi)D_1M_1.$$

□

Theorem B.1.4 (Smith form, square case) *Let $R(\xi) \in \mathbb{R}^{g\times q}[\xi]$. Assume $g = q$. There exist unimodular matrices $U(\xi), V(\xi) \in \mathbb{R}^{g\times g}$ such that*

1. *$U(\xi)R(\xi)V(\xi) = \operatorname{diag}(d_1(\xi), \ldots, d_g(\xi))$.*
2. *$d_i(\xi)$ divides $d_{i+1}(\xi)$; i.e., there exist (scalar) polynomials $q_i(\xi)$ such that $d_{i+1}(\xi) = q_i(\xi)d_i(\xi), i = 1, \ldots, g-1$.*

Proof The proof is an algorithm. Assume that $R(\xi)$ is nonzero. Apply row and column permutations so as to achieve that the nonzero element of minimal degree of $R(\xi)$ appears at the $(1,1)$ spot. Use this element to carry out division with

remainder on both the first column (premultiplication by unimodular matrices) and the first row (postmultiplication by unimodular matrices). Repeat the whole procedure as many times as possible. Notice that every time, the degree of the nonzero element of minimal degree decreases by at least one. Also, as long as the first row or column contains at least two nonzero elements, we can apply the procedure once more. Since degrees are nonnegative, this implies that within a finite number of steps we reach the following situation:

$$\begin{bmatrix} * & 0 & \cdots & 0 \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{bmatrix}. \tag{B.3}$$

Either the $(1,1)$ element in (B.3) divides all the other elements in the matrix, or there exists a column that contains an element that is *not* a multiple of the $(1,1)$ element. If the latter is true, add this column to the first column of (B.3) and start all over again. Again after a finite number of steps we arrive at a matrix of the form (B.3), but with a $(1,1)$ element of strictly smaller degree. As long as there is an element in the matrix that is not divisible by the $(1,1)$ element, we can repeat this process. As a consequence, we obtain, in a finite number of steps, a matrix of the form (B.3) where the $(1,1)$ element divides all the other elements. Then we move on to the $(g-1)\times(g-1)$ right-lower submatrix and apply the whole procedure to that matrix. Of course, the $(1,1)$ element from the previous step keeps dividing the elements of the $(g-1)\times(g-1)$ right-lower submatrix, and hence we obtain a matrix of the form

$$\begin{bmatrix} * & 0 & \cdots & \cdots & 0 \\ 0 & * & 0 & \cdots & 0 \\ \vdots & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{bmatrix}. \tag{B.4}$$

The $(1,1)$ element of (B.4) divides all the other elements of the matrix, and the $(2,2)$ element divides all the elements of the $(g-2)\times(g-2)$ right-lower matrix. Next we move on the $(g-3)\times(g-3)$ right-lower matrix, and so on.

Finally, we end up with the desired diagonal matrix. □

Remark B.1.5 If $R(\xi)$ is not square, then the Smith form can also be defined, and it is obtained via the same algorithm. If $R(\xi)$ is wide ($g < q$) or if $R(\xi)$ is tall ($g > q$), the Smith forms are given by

$$\begin{bmatrix} d_1(\xi) & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & d_g(\xi) & 0 & \cdots & 0 \end{bmatrix}, \quad \begin{bmatrix} d_1(\xi) & & \\ & \ddots & \\ & & d_q(\xi) \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

respectively. □

Theorem B.1.6 *Let $r_1(\xi), \ldots, r_k(\xi) \in \mathbb{R}[\xi]$. Assume that $r_1(\xi), \ldots, r_k(\xi)$ have no common factor. Then there exists a unimodular matrix $U(\xi) \in \mathbb{R}^{k\times k}[\xi]$ such that the last row of $U(\xi)$ equals $[r_1(\xi), \ldots, r_k(\xi)]$.*

Proof Define the vector $r(\xi) := [r_1(\xi) \cdots r_k(\xi)]$. Take an entry of $r(\xi)$ of minimal degree. Carry out "division with remainder" on the other entries. This comes down to right multiplication by a unimodular matrix $V_1(\xi)$. Now consider the result $r(\xi)V_1(\xi)$. Take an entry of $r(\xi)V_1(\xi)$ of minimal degree and repeat the procedure of the first step. Again, this can be seen as postmultiplication by a unimodular matrix $V_2(\xi)$. After repeating this procedure as many times as possible, we get the following result:

$$r(\xi)V_1(\xi)V_2(\xi)\cdots V_\ell(\xi) = \begin{bmatrix} 0 & \cdots & 0 & g(\xi) & 0 & \cdots & 0 \end{bmatrix}. \tag{B.5}$$

It is easy to see that $g(\xi)$ is a common divisor of $r_1(\xi), \ldots, r_k(\xi)$, and by the coprimeness assumption it follows that we may assume that in fact $g(\xi) = 1$; see Exercise B.1. Finally, by postmultiplication by a suitable unimodular matrix $V_{l+1}(\xi)$, we get

$$r(\xi)V_1(\xi)V_2(\xi)\cdots V_{\ell+1}(\xi) = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{B.6}$$

Define $V(\xi) := V_1(\xi)V_2(\xi)\cdots V_{\ell+1}(\xi)$. Then $V(\xi)$ is unimodular, and hence $U(\xi) := V^{-1}(\xi)$ is also unimodular. From (B.6) it follows that

$$r(\xi) = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} U(\xi).$$

This implies that the last row of $U(\xi)$ is $r(\xi)$, and the proof is finished. □

As a bonus we obtain from the proof of Theorem B.1.6 the following result, called the Bezout equation.

Corollary B.1.7 (Bezout) *Let $r_1(\xi), \ldots, r_k(\xi) \in \mathbb{R}[\xi]$. Assume that $r_1(\xi), \ldots, r_k(\xi)$ have no common factor. Then there exist polynomials $a_1(\xi), \ldots, a_k(\xi) \in \mathbb{R}[\xi]$, such that*

$$r_1(\xi)a_1(\xi) + \cdots + r_k(\xi)a_k(\xi) = 1.$$

Proof From (B.6) it follows that there exists a unimodular matrix $V(\xi)$ such that

$$r(\xi)V(\xi) = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}.$$

This shows that we can take

$$a_1(\xi) := V_{1k}(\xi), \ldots, a_k(\xi) := V_{kk}(\xi).$$

□

B.2 Partial Fraction Expansion

Theorem B.2.1 (Partial fraction expansion, scalar case) *Let* $p(\xi), q(\xi) \in \mathbb{R}[\xi]$ *and* $\deg q(\xi) = m \leq n = \deg p(\xi)$. *Suppose* $p(\xi) = \prod_{i=1}^{N}(\xi - \lambda_i)^{n_i}$, *with* $\lambda_i \neq \lambda_j$ *for* $i \neq j$. *Then there exist* a_0 *and* $a_{ij} \in \mathbb{C}$ *such that*

$$\frac{q(\xi)}{p(\xi)} = a_0 + \sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{a_{ij}}{(\xi - \lambda_i)^j}.$$

The proof of Theorem B.2.1 is divided into two parts.

Lemma B.2.2 *Let* $p(\xi), q(\xi) \in \mathbb{C}[\xi]$, $p(\xi) = \prod_{i=1}^{N}(\xi - \lambda_i)^{n_i}$, *and* $\deg q(\xi) < \deg p(\xi)$. *(Notice the strict inequality.) Then there exist polynomials* $q_i(\xi) \in \mathbb{C}[\xi]$, *with* $\deg q_i(\xi) < n_i$ *for* $i = 1, \ldots N$, *such that*

$$\frac{q(\xi)}{p(\xi)} = \sum_{i=1}^{N}\frac{q_i(\xi)}{(\xi - \lambda_i)^{n_i}}.$$

Proof The proof goes by induction on N. For $N = 1$ there is nothing to prove. Suppose that the statement is true for all $q(\xi), p(\xi)$ for which $p(\xi)$ has at most N distinct roots, and let $p(\xi)$ have $N + 1$ distinct roots. Factorize $p(\xi)$ as

$$p(\xi) = p_1(\xi)p_2(\xi),$$

where $p_1(\xi)$ and $p_2(\xi)$ have no common factor, and the number of distinct roots of both $p_1(\xi)$ and $p_2(\xi)$ is at most equal to N. By Corollary B.1.7 there exist polynomials $a_1(\xi)$ and $a_2(\xi)$ such that

$$a_1(\xi)p_1(\xi) + a_2(\xi)p_2(\xi) = 1.$$

Define $b_i(\xi) := q(\xi)a_i(\xi)$ $(i = 1, 2)$. Then

$$b_1(\xi)p_1(\xi) + b_2(\xi)p_2(\xi) = q(\xi).$$

By (2.25), there exist $r_1(\xi)$ and $c_1(\xi)$ such that

$$b_1(\xi) = c_1(\xi)p_2(\xi) + r_1(\xi), \text{ with } \deg r_1(\xi) < \deg p_2(\xi).$$

This implies

$$\underbrace{r_1(\xi)}_{\tilde{b}_1(\xi)}p_1(\xi) + \underbrace{(b_2(\xi) + c_1(\xi)p_1(\xi))}_{\tilde{b}_2(\xi)}p_2(\xi) = q(\xi).$$

Suppose that $\deg \tilde{b}_2(\xi) \geq \deg p_1(\xi)$. Then

$$\begin{aligned}
\deg q(\xi) &= \deg(r_1(\xi)p_1(\xi) + (b_2(\xi) + c_1(\xi)p_1(\xi))p_2(\xi)) \\
&= \deg((b_2(\xi) + c_1(\xi)p_1(\xi))p_2(\xi)) \\
&= \deg(b_2(\xi) + c_1(\xi)p_1(\xi)) + \deg p_2(\xi) \\
&\geq \deg(p_1(\xi)) + \deg(p_2(\xi)) = \deg p(\xi).
\end{aligned}$$

This contradicts the assumption that $\deg q(\xi) < \deg p(\xi)$, and hence $\deg \tilde{b}_2(\xi) < \deg p_1(\xi)$. Define $q_2(\xi) := \tilde{b}_1(\xi)$ and $q_1(\xi) := \tilde{b}_2(\xi)$. Then

$$q_2(\xi)p_1(\xi) + q_1(\xi)p_2(\xi) = q(\xi), \quad \deg q_i(\xi) < \deg p_i(\xi), \quad i = 1, 2.$$

Now,

$$\frac{q_1(\xi)}{p_1(\xi)} + \frac{q_2(\xi)}{p_2(\xi)} = \frac{q_2(\xi)p_1(\xi) + q_1(\xi)p_2(\xi)}{p_1(\xi)p_2(\xi)} = \frac{q(\xi)}{p(\xi)}.$$

By the induction hypothesis, $\dfrac{q_1(\xi)}{p_1(\xi)}$ and $\dfrac{q_2(\xi)}{p_2(\xi)}$ can be expanded in the desired form. The statement follows. □

Lemma B.2.3 *Let $p(\xi) = (\xi - \lambda)^n$ and $q(\xi)$ be a polynomial of degree smaller than n. There exist $a_1, \ldots, a_n \in \mathbb{C}$ such that*

$$\frac{q(\xi)}{p(\xi)} = \sum_{j=1}^{n} \frac{a_j}{(\xi - \lambda)^j}. \tag{B.7}$$

Proof Let $a(\xi) = a_1(\xi - \lambda)^{n-1} + a_2(\xi - \lambda)^{n-2} + \cdots + a_n$. This is a polynomial of degree $\leq n - 1$. For (B.7) to be true we should have

$$q(\xi) = \sum_{j=1}^{n} a_j(\xi - \lambda)^{(n-j)} = a(\xi). \tag{B.8}$$

By equating the coefficients of the left-hand side and the right-hand side of (B.8), we obtain the values of the a_js. □

Proof of Theorem B.2.1 By (2.25) there exist polynomials $r(\xi), a(\xi)$ such that $q(\xi) = p(\xi)a(\xi) + r(\xi)$, with $\deg r(\xi) < \deg p(\xi)$. This implies that $\frac{q(\xi)}{p(\xi)} = a(\xi) + \frac{r(\xi)}{p(\xi)}$. It is easily seen that since $\deg q(\xi) \leq \deg p(\xi)$, $a(\xi)$ is a constant, say a_0. The theorem now follows by first applying Lemma B.2.2 to $\frac{r(\xi)}{p(\xi)}$ and subsequently Lemma B.2.3 to the result. □

B.3 Fourier and Laplace Transforms

In order to make this book reasonably self-contained, we provide in this section of the appendix for easy reference the basics about the Fourier and Laplace transforms. We assume, however, that the reader has some previous acquaintance with these ideas. We start by explaining the notation. In this appendix, and throughout the book, we use the following notation. Let A be a (possibly infinite) interval in $\mathbb{R}$, and $B = \mathbb{R}^n$ or $\mathbb{C}^n$ for some $n \in \mathbb{N}$. An important family of maps are the $\mathfrak{L}_p$-functions. For $1 \leq p < \infty$, $\mathfrak{L}_p(A, B)$ denotes the set of maps from A to B such that

$$(\int_A \|f(t)\|^p dt)^{1/p} < \infty. \tag{B.9}$$

If (B.9) holds, then the left-hand side is defined to be the $\mathfrak{L}_p$-norm of f, denoted by $\|f\|_{\mathfrak{L}_p}$. The space $\mathfrak{L}_\infty(A,B)$ denotes the set of maps from A to B with the following property: $f \in \mathfrak{L}_\infty(A,B)$ if $f : A \to B$ and if there exists $M < \infty$ such that

$$\|f(t)\| \leq M \text{ for almost all } t \in A. \tag{B.10}$$

The smallest M for which (B.10) holds is defined to be the $\mathfrak{L}_\infty$-norm of f, denoted by $\|f\|_{\mathfrak{L}_\infty}$. The spaces $\mathfrak{L}_p(A,B)$ for $1 \leq p \leq \infty$ are normed linear spaces (in fact, Banach spaces, but we do not need this property). The space $\mathfrak{L}_2(A,B)$ has even more structure. It is a Hilbert space with the inner product defined as

$$< f_1, f_2 >_{\mathfrak{L}_2} := \int_A \bar{f}_1^T(t) f_2(t) dt.$$

The fact that it is a Hilbert space implies that if f_k is a Cauchy sequence in in $\mathfrak{L}_2(A,B)$, i.e., for $k \in \mathbb{N}$ and if for all $\epsilon > 0$, there exists an N such that $\|f_{k'} - f_{k''}\|_{\mathfrak{L}_2} < \epsilon$ for $k', k'' > N$, then there exists an $f \in \mathfrak{L}_2(A,B)$ such that $f_k \underset{k\to\infty}{\longrightarrow} f$, with convergence understood in the sense of $\mathfrak{L}_2$; i.e.,

$$\int_A \|f(t) - f_k(t)\|^2 dt \underset{k\to\infty}{\longrightarrow} 0.$$

B.3.1 Fourier transform

Let $f : \mathbb{R} \to \mathbb{C}$. Assume first that f is integrable; i.e., $f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$. Let $\omega \in \mathbb{R}$, and define

$$\hat{f}(i\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt. \tag{B.11}$$

Obviously, $|\hat{f}(i\omega)| \leq \|f\|_{\mathfrak{L}_1}$, and hence $\hat{f} : \mathbb{R} \to \mathbb{C}$ is bounded; i.e., $\hat{f} \in \mathfrak{L}_\infty(\mathbb{R},\mathbb{C})$. The function $\hat{f}$ is called the *Fourier transform* of f. Sometimes, in order to emphasize that (B.11) is defined for $\mathfrak{L}_1$-functions, it is called the $\mathfrak{L}_1$-*Fourier transform*. The Fourier transform can nicely be generalized to $\mathfrak{L}_2$-functions as follows. Let $f \in \mathfrak{L}_2(\mathbb{R},\mathbb{C})$. Define $f_T : \mathbb{R} \to \mathbb{C}$ as

$$f_T(t) = \begin{cases} f(t) & \text{for} \quad |t| \leq T, \\ 0 & \text{for} \quad |t| > T. \end{cases}$$

Then it can be shown that for all $T \in \mathbb{R}, f_T \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$. Hence its $\mathfrak{L}_1$-Fourier transform, $\hat{f}_T$, is well-defined. The Fourier transforms $\hat{f}_T$ have the following interesting behavior for $T \to \infty$. It can be shown that there exists a function $\hat{f} \in \mathfrak{L}_2(\mathbb{R},\mathbb{C})$ such that

$$\lim_{T\to\infty} \int_{-\infty}^{+\infty} |\hat{f}(i\omega) - \hat{f}_T(i\omega)|^2 d\omega = 0.$$

This limit function $\hat{f}$ is called the *l.i.m. (limit-in-the-mean)*, or the $\mathfrak{L}_2$-*Fourier transform* of f. Since for $f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C}) \cap \mathfrak{L}_2(\mathbb{R},\mathbb{C})$, the $\mathfrak{L}_1$- and $\mathfrak{L}_2$-Fourier transforms coincide, the same notation, $\hat{f}$, is used for both.

The advantage of using the $\mathfrak{L}_2$-Fourier transform instead of the $\mathfrak{L}_1$-Fourier transform lies in the fact that it maps $\mathfrak{L}_2(\mathbb{R},\mathbb{C})$ into $\mathfrak{L}_2(\mathbb{R},\mathbb{C})$, which makes it possible to define the inverse transform. Indeed, for the $\mathfrak{L}_1$-Fourier transform, the inverse transform presents difficulties, since the $\mathfrak{L}_1$-transform of $f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$ belongs to $\mathfrak{L}_\infty(\mathbb{R},\mathbb{C})$, but in general not to $\mathfrak{L}_1(\mathbb{R},\mathbb{C})$ nor to $\mathfrak{L}_2(\mathbb{R},\mathbb{C})$. Thus if in turn we want to calculate the transform of $\hat{f}$, we run into difficulties, since

$$\int_{-\infty}^{+\infty} \hat{f}(i\omega)e^{i\omega t}d\omega \tag{B.12}$$

may not be well-defined, at least not as an ordinary integral, when $\hat{f} \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$. A second feature of the $\mathfrak{L}_2$-Fourier transform is the following. Let $f, g \in \mathfrak{L}_2(\mathbb{R},\mathbb{C})$. Denote by $\hat{f}, \hat{g}$ their $\mathfrak{L}_2$-Fourier transform. Let $< \cdot, \cdot >_{\mathfrak{L}_2}$ denote the $\mathfrak{L}_2$-inner product. Then

$$\int_{-\infty}^{+\infty} \hat{f}(i\omega)e^{i\omega t}d\omega = \frac{1}{2\pi}f(t) \qquad \text{a.e.,} \tag{B.13}$$

$$< f, g >_{\mathfrak{L}_2} = \frac{1}{2\pi} < \hat{f}, \hat{g} >_{\mathfrak{L}_2}, \tag{B.14}$$

$$\|f\|_{\mathfrak{L}_2} = \frac{1}{\sqrt{2\pi}}\|\hat{f}\|_{\mathfrak{L}_2}. \tag{B.15}$$

These formulas have nice interpretations. The left-hand side of (B.13) is called the inverse Fourier transform. Note that it differs from the Fourier transform only in that the term $e^{i\omega t}$ instead of $e^{-i\omega t}$ appears in the integral. Formula (B.13) states that the inverse $\mathfrak{L}_2$-Fourier transform of $\hat{f}$ equals f up to the factor $\frac{1}{2\pi}$. Formula (B.14) states that the $\mathfrak{L}_2$-Fourier transform preserves the $\mathfrak{L}_2$-inner product up to the factor $\frac{1}{2\pi}$. Applying (B.14) with $g = f$ yields (B.15), which states that the $\mathfrak{L}_2$-Fourier transform preserves the $\mathfrak{L}_2$-norm up to a factor $\frac{1}{\sqrt{2\pi}}$.

One of the reasons for the importance of Fourier transforms is their interplay with convolutions. Thus, let $h, f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$. Define their *convolution* by

$$(h * f)(t) := \int_{-\infty}^{+\infty} h(t-t')f(t')dt'. \tag{B.16}$$

Then it is easy to show that $h * f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$, and that

$$\widehat{h * f} = \hat{h}\hat{f}. \tag{B.17}$$

In the above formula, all the transforms should be interpreted as $\mathfrak{L}_1$-Fourier transforms. However, if $h \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$ and $f \in \mathfrak{L}_2(\mathbb{R},\mathbb{C})$, then (B.16) is still well-defined. Indeed, $h * f \in \mathfrak{L}_2(\mathbb{R},\mathbb{C})$, and in fact, (B.17) holds. Note, however, that now $\widehat{h * f}$ and $\hat{f}$ are $\mathfrak{L}_2$-Fourier transforms, whereas $\hat{h}$ is the $\mathfrak{L}_1$-Fourier transform.

Similarly as with convolutions, the Fourier transform also acts very conveniently on differential operators. Thus if $f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$ is such that $\frac{d}{dt}f \in \mathfrak{L}_1(\mathbb{R},\mathbb{C})$, then

$$\widehat{\frac{d}{dt}f}(i\omega) = i\omega\hat{f}(i\omega).$$

More generally, if $p(\xi) \in \mathbb{R}[\xi]$ is such that $p(\frac{d}{dt})f \in \mathfrak{L}_1(\mathbb{R}, \mathbb{C})$, then

$$\widehat{p(\frac{d}{dt})f}(i\omega) = p(i\omega)\hat{f}(i\omega).$$

Similar expressions hold for the $\mathfrak{L}_2$-Fourier transform.

The Fourier transform is easily extended to vector- or matrix-valued functions, and there are obvious analogues of the above formulas for convolutions and differential operators, but there is no need to show the formulas explicitly.

B.3.2 Laplace transform

Let $s \in \mathbb{C}$. Denote by $\exp_s : \mathbb{R} \to \mathbb{C}$ the exponential function defined by $\exp_s(t) := e^{st}$. Let $f \in \mathbb{R} \to \mathbb{C}$, and consider the formula

$$\hat{f}(s) = \int_{-\infty}^{+\infty} f(t)e^{-st}dt. \tag{B.18}$$

Note that the improper integral in (B.18) converges if $f \exp_s \in \mathfrak{L}_1(\mathbb{R}, \mathbb{C})$. If such an $s \in \mathbb{C}$ exists, then we call the function f *Laplace transformable*. In general, (B.18) defines a function from a *subset* of $\mathbb{C}$ to $\mathbb{C}$. Such a function (a map from a subset of $\mathbb{C}$ into $\mathbb{C}$) is called a complex function; the particular function $\hat{f}$ is called the *two-sided Laplace transform* of f. The set of $s \in \mathbb{C}$ for which (B.18) exists is called the *region of convergence* of $\hat{f}$. Note that the $s \in \mathbb{C}$ such that $f \exp_s \in \mathfrak{L}_1(\mathbb{R}, \mathbb{C})$ defines a vertical strip in $\mathbb{C}$, but this strip may be closed, (half-) open, a half-space, or all of $\mathbb{C}$ (or even empty, in which case f is not Laplace transformable).

Laplace transforms have properties that are very similar to those of Fourier transforms. In particular, for the convolution $h * f$, there holds

$$\widehat{h * f}(s) = h(s)f(s),$$

and the region of convergence of $\widehat{h * f}$ contains the intersection of those of $\hat{h}$ and $\hat{g}$.

The one-sided Laplace transform is defined for functions $f : [0, \infty) \to \mathbb{C}$ and is given by

$$\hat{f}(s) = \int_{0}^{\infty} f(t)e^{-st}dt.$$

The region of convergence is now a half-plane $\{s \in \mathbb{C} | Re(s) \geq \sigma\}$ or $\{s \in \mathbb{C} | Re(s) > \sigma\}$ or all of $\mathbb{C}$.

B.4 Notes and References

The theory of polynomial matrices is a special case of the theory of rings. There are numerous books in mathematics and in systems theory that treat these topics. For

introductory mathematics texts, see for example [34, 43], and for systems theory references, see [8, 48, 63]. Fourier and Laplace transforms are central techniques in systems and control theory. See [50] for a mathematical introduction to Fourier analysis.

B.5 Exercises

B.1 Prove that $g(\xi)$ in (B.5) is equal to the greatest common divisor of $r_1(\xi), \ldots, r_k(\xi)$. (You have to prove that $g(\xi)$ divides $r_1(\xi), \ldots, r_k(\xi)$ and that moreover, every common divisor of $r_1(\xi), \ldots, r_k(\xi)$ divides $g(\xi)$).

B.2 Give functions $f : \mathbb{R} \to \mathbb{C}$ such that $\frac{1}{1+s}$ is its two-sided Laplace transform with region of convergence

1. $\{s \in \mathbb{C} | \mathrm{Re}(s) < -1\}$.
2. $\{s \in \mathbb{C} | \mathrm{Re}(s) > -1\}$.

B.3 Prove Theorem 8.2.1, Part (ii)$'$. Note that the transforms involved are the $\mathfrak{L}_2$-Fourier transforms for $\hat{u}$ and $\hat{y}$, and the $\mathfrak{L}_1$-Fourier transform for H.

B.4 Consider the system (8.5). Let H be its impulse response. Compute its Fourier transform. Note that since $H \in \mathfrak{L}_1(\mathbb{R}, \mathbb{R}) \cap \mathfrak{L}_2(\mathbb{R}, \mathbb{R})$, the Fourier transform is both the $\mathfrak{L}_1$- and the $\mathfrak{L}_2$-Fourier transform. Prove that the Fourier transform $\hat{H}(i\omega) = 2\frac{\sin \omega \Delta}{\omega}$ does not belong to $\mathfrak{L}_1(\mathbb{R}, \mathbb{C})$, but that it belongs to $\mathfrak{L}_2(\mathbb{R}, \mathbb{C})$. Compute the $\mathfrak{L}_2$-Fourier transform of $\hat{H}$. This exercise illustrates the need of $\mathfrak{L}_2$-transforms.

Notation

Symbol	Short description	Page
$\mathbb{R}_+$	set of nonnegative real numbers	
$\mathbb{Z}_+$	set of nonnegative integers	
$\mathbb{R}^{n_1 \times n_2}$	set of real $n_1 \times n_2$ matrices	
$\mathbb{N}$	set of positive integers	
$\mathbb{Z}$	set of integers	
$\mathbb{Q}$	set of rational numbers	
$\mathbb{R}$	set of real numbers	
$\mathbb{C}$	set of complex numbers	
$\mathbb{W}^{\mathbb{T}}$	set of functions $\mathbb{T} \to \mathbb{W}$	9
σ, σ^t	shift-operator	16
$\mathcal{C}^k(\mathbb{R}, \mathbb{R}^q)$	set of k times continuously differentiable functions	22
$\mathbb{R}[\xi]$	set of polynomials with real coefficients	29
$\mathbb{R}^{n_1 \times n_2}[\xi]$	set of real polynomial $n_1 \times n_2$ matrices	29
$\mathbb{R}^{\bullet \times n}[\xi]$	set of real polynomial matrices with n columns	29
$\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$	set of infinitely differentiable functions	34
$\mathfrak{L}_1^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}^q)$	set of locally integrable functions	34
$\phi * w$	convolution product	40
deg	degree of polynomial	44
det	determinant	45
$P^{(k)}(\xi)$	kth formal derivative of polynomial	72
$\binom{j}{\ell}$	Binomial coefficient	72
Tr	trace	151
$\mathfrak{C}$	controllabilty matrix	168
im	image of linear map	172

$\mathfrak{O}$	observability matrix	182
ker	kernel of linear map	184
$\chi_A(\xi)$	characteristic polynomial of the square matrix A	278
$\exp_s$	exponential function	289
$\mathfrak{L}_1(\mathbb{R}, \mathbb{R}^q)$	set of integrable functions	289
$\mathfrak{L}_2(\mathbb{R}, \mathbb{R}^q)$	set of square integrable functions	290
$\mathfrak{E}(s)$	exponential behavior	292
$\Sigma_{n,m}$	set of systems with m inputs and n states	326
$\Sigma_{n,m}^{\text{cont}}$	set of controllable systems with m inputs and n states	329
$\Sigma_{n,m,p}$	set of systems, n states, m inputs, and p outputs	350
$\mathfrak{L}_p(\mathbb{R}, \mathbb{R}^q)$		418
$\mathfrak{L}_\infty(\mathbb{R}, \mathbb{R}^q)$	set of bounded functions	419

References

[1] R. Abraham and J.E. Marsden. *Foundations of Mechanics*. The Benjamin/Cummings Publishing Company, London, 2nd edition, 1978.

[2] J. Ackermann. Der Entwurf linearer Reglungssysteme in Zustandsraum. *Regelungstechnik*, 20:297–300, 1972.

[3] B.D.O. Anderson and J.B. Moore. *Linear Optimal Control*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[4] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice Hall, Englewood Cliffs, NJ, 1979.

[5] M. Athans (*Guest Editor*). Special issue on the linear-quadratic-gaussian estimation and control problem. *IEEE Transactions on Automatic Control*, 16, no. 6, 1971.

[6] K.J. Åström and K. Furuta. Swing up a pendulum by energy control. *Proceedings World Congress International Federation on Automatic Control*, E:37–42, 1996.

[7] S. Barnett. *Polynomials and Linear Control Systems*. Marcel Dekker, New York, NY, 1983.

[8] V. Belevich. *Classical Network Theory*. Holden Day, San Francisco, CA, 1968.

[9] R. Bellman and R. Kalaba. *Selected Papers on Mathematical Trends in Control Theory*. Dover, New York, NY, 1964.

[10] S. Bennett. *A History of Control Engineering 1800 – 1930*. Peter Peregrinus, London, 1979.

[11] S. Bennett. *A History of Control Engineering 1930 – 1955*. Peter Peregrinus, London, 1993.

[12] H.S. Black. Stabilized feedback amplifiers. *Bell System Technical Journal*, 13:1–18, 1934.

[13] H.S. Black. Inventing the negative feedback amplifier. *IEEE Spectrum*, 14:54–60, 1977.

[14] A.M. Bloch, P.S. Krishnaprasad, J.E. Marsden, and G. Sánchez de Alvarez. Stabilization of rigid body dynamics by internal and external torques. *Automatica*, 28:745–756, 1992.

[15] R.W. Brockett. *Finite Dimensional Linear Systems*, volume 17. John Wiley & Sons, New York, NY, 1970.

[16] J.C. Doyle, K. Glover, P.P. Khargonekar, and B.A. Francis. State-space solutions to standard H_2 and H_∞ control problems. *IEEE Transactions on Automatic Control*, 34:831–847, 1989.

[17] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon. Sur les systèmes nonlinéaires différentiellements plats. *Comptes Rendues de l´Académie des Sciences de Paris*, I-315:619–624, 1992.

[18] K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their $\mathfrak{L}_\infty$-error bounds. *International Journal of Control*, 39:1115–1193, 1984.

[19] H. H. Goldstine. *A History of the Calculus of Variations from the 17th to the 19th Century*. Springer Verlag, New York, NY, 1981.

[20] M.L.J. Hautus. Controllability and observability conditions of linear autonomous systems. *Proceedings Nederlandse Akademie van Wetenschappen Serie A*, 72:443–448, 1969.

[21] M.L.J. Hautus. A simple proof of Heymann's lemma. *IEEE Transactions on Automatic Control*, 22:885–886, 1977.

[22] M. Heymann. Comments on 'on pole assignment in multi-input controllable linear systems'. *IEEE Transactions on Automatic Control*, 13:748–749, 1968.

[23] W. Hirsch and S. Smale. *Differential Equations and Linear Algebra*. Academic Press, New York, NY, 1974.

[24] A. Hurwitz. Über die Bedingungen unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt. *Mathematische Annalen*, 46:273–284, 1895.

[25] T. Kailath. *Linear Systems*. Prentice Hall, Englewood Cliffs, NJ, 1980.

[26] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82D:35–45, 1960.

[27] R.E. Kalman. On the general theory of control systems. In *Proceedings of the 1st World Congress of the International Federation of Automatic Control*, 481–493, Moscow, 1960.

[28] R.E. Kalman. Mathematical description of linear dynamical systems. *SIAM Journal on Control*, 1:152–192, 1963.

[29] R.E. Kalman. Lectures on controllability and observability. In *CIME Lecture Notes*, Bologna, Italy, 1968.

[30] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering*, 83D:95–108, 1961.

[31] R.E. Kalman, P.L. Falb, and M.A. Arbib. *Topics in Mathematical System Theory*. McGraw-Hill, New York, NY, 1981.

[32] V.L. Kharitonov. Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsial'nye Uravneniya*, 14:2086–2088, 1978.

[33] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems*. John Wiley & Sons, New York, NY, 1972.

[34] S. Lang. *Algebra*. Addison-Wesley, Menlo Park, CA, 2nd edition, 1984.

[35] C.E. Langenhop. On stabilization of linear systems. *Proceedings of the American Mathematical Society*, 15:748–749, 1964.

[36] N.E. Leonard and W.S. Levine. *Using MATLAB© to Analyze and Design Control Systems*. Addison Wesley, Menlo Park, CA; Reading, MA; New York; Don Mills, Ontario; Wolkingham, UK; Amsterdam; Bonn; Paris; Milan; Madrid; Sydney; Singapore; Seoul; Taipei; Mexico City; San Juan, Puerto Rico, 1995.

[37] D.G. Luenberger. An introduction to observers. *IEEE Transactions on Automatic Control*, 16:569–603, 1971.

[38] D.G. Luenberger. *Introduction to Dynamical Systems: Theory, Models, & Applications*. John Wiley & Sons, New York, NY, 1979.

[39] J.C. Maxwell. On governors. *Proceedings of the Royal Society of London*, 16:270–283, 1868.

[40] O. Mayr. *The Origins of Feedback Control*. MIT Press, Cambridge, MA, 1970.

[41] G. Meinsma. Elementary proof of the Routh – Hurwitz test. *Systems & Control Letters*, 25:237–242, 1995.

[42] C. Moler and C. van Loan. Nineteen dubious ways to compute the exponential of a matrix. *Siam Review*, 20:801–836, 1978.

[43] M. Newman. *Integral Matrices*. Academic Press, New York, NY, 1972.

[44] J.W. Polderman. Proper elimination of latent variables. *Systems & Control Letters*, 1997. To appear.

[45] V.M. Popov. Hyperstability and optimality of automatic systems with several control functions. *Revue Roumaine Sci. Techn., Serie Electrotech. Energ.*, 9:629–690, 1964.

[46] V.M. Popov. *Hyperstability of Control Systems*. Springer Verlag, Berlin, 1969.

[47] J. Rissanen. Control system synthesis by analogue computer based on the generalized linear feedback concept. In *Proceedings of the Symposium on Analog Computation Applied to the Study of Chemical Processes*, 1–31, Brussels, Belgium, 1961. Presses Académiques Européennes.

[48] H.H. Rosenbrock. *State-Space and Multivariable Theory.* John Wiley, New York, NY, 1970.

[49] E.J. Routh. A treatise on the stability of a given state of motion. Adams Prize Essay, Cambridge University, 1877.

[50] W. Rudin. *Real and Complex Analysis.* McGraw–Hill, New York, St. Louis, San Francisco, Auckland, Bogotá, Hamburg, London, Madrid, Mexico, Milan, Montreal, New Delhi, Panama, Paris, Sao Paulo, Singapore, Sydney, Tokyo, Toronto, 1966.

[51] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill Book Company, London, 1987.

[52] E.D. Sontag. *Mathematical Control Theory.* Springer Verlag, 1990.

[53] H.J. Sussmann and J.C. Willems. 300 years of optimal control: From the brachystochrone to the maximum principle. *IEEE Control Systems Magazine*, 1997.

[54] M. Vidyasagar. *Nonlinear Systems Analysis.* Prentice Hall, Englewood Cliffs, NJ, 1978.

[55] J.C. Willems. System theoretic models for the analysis of physical systems. *Ricerche di Automatica*, 10:71–106, 1981.

[56] J.C. Willems. From time series to linear system - part I. Finite dimensional linear time invariant systems. *Automatica*, 22:561–580, 1986.

[57] J.C. Willems. From time series to linear system - part II. Exact modelling. *Automatica*, 22:675–694, 1986.

[58] J.C. Willems. From time series to linear system - part III. Approximate modelling. *Automatica*, 23:87–115, 1987.

[59] J.C. Willems. Models for dynamics. *Dynamics Reported*, 2:171–269, 1989.

[60] J.C. Willems. Paradigms and puzzles in the theory of dynamical systems. *IEEE Transactions on Automatic Control*, 36:259–294, 1991.

[61] J.C. Willems. On interconnections, control, and feedback. *IEEE Transactions on Automatic Control*, 42:326–339, 1997.

[62] J.L. Willems. *Stability Theory of Dynamical Systems.* Nelson, London, 1970.

[63] W.A. Wolovich. *Linear Multivariable Systems.* Springer Verlag, New York, NY, 1974.

[64] W.M. Wonham. On pole assignment in multi-input controllable linear systems. *IEEE Transactions on Automatic Control*, 12:660–665, 1967.

[65] W.M. Wonham. Linear multivariable control: a geometric approach. In *Lecture Notes in Economic and Mathematical Systems 101.* Springer Verlag, Berlin, 1974.

[66] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26:301–320, 1981.

Index

A

A-invariant subspace, 168
actuator, xv, 371
admissible controllers, 374
adverse response, 299
affine subspace, 90
algebraic curve in $\mathbb{R}^2$, 210
algebraic multiplicity, 136
algebraic set, 210
algebraic variety, 332
almost
    all, 35
    everywhere, 35
annihilate, 278
anticipating
    strictly non-, 93
anticipating, non-, 93
asymptotic stabilizability, 344
asymptotically stable, 268
attractor, 268
autonomous
    behavior, 69
    system, 68, 79
auxiliary variables, 2

B

backward shift, 16
band-pass, 303
bandwidth, 303
behavior, xviii, 2, 3, 9
behavioral
    difference equation, 17
    differential equation, 18
    equation representation, 4
    equations, 2, 4
    inequalities, 5
behavioral approach, xviii
Bezout
    equation, 54, 64, 416
    generalization, 54, 416
    identity, 54
    map, 378
black box, xv
Bode plot, 302
Bohl function, 103, 288
bounded input–bounded output-stability, 272
brachystochrone, xiii

C

calculus of variations, xiv
cancellation, 308
canonical form, 230
Cayley–Hamilton, 169, 173

certainty equivalence principle, 348, 358
characteristic frequency, 305
characteristic polynomial, 72, 278
    closed loop , 324
    of autonomous behavior, 81
characteristic time, 305
characteristic values of autonomous behavior, 72, 81
classical control theory, xiii
closed loop characteristic polynomial, 358
closed loop equations, 323
closed loop poles, 324
column rank, 58
common factor, 54
compact support, 98
compensator, 348
continuous-time systems, 9
control input, xv
controllability, xvi, xix, 156, 157
controllability index, 368
controllability matrix, 168
controllable pair, 168
controller
    PI, 396
    PID, xi
controller canonical form, 225, 226
convergence in $\mathfrak{L}_1^{\mathrm{loc}}$, 38
convolution product, 40
convolution systems, 99, 289
coprime, 54
Cramer's rule, 48
critically damped, 306
cut-off frequency, 303

D

damping coefficient, 305
dB, 302
deadbeat observer, 384
deadtime, 299
decade, 302
decibel, 302
detectable, 187, 356
detectable pair, 188
determinism, 132
dimension of state space representation, 124
discrete-event systems, 9
discrete-time systems, 9
distributed systems, 9
disturbance
    attenuation, xvi
    rejection, 300
division with remainder, 44
dual system, 386
duality, 182
dynamic control law, 356
dynamic feedback, 323
dynamical, 8
    system, 1, 9
        with latent variables, 10

E

elementary operations, 49
elementary unimodular matrix, 51
elimination
    exact, 213
    procedure, 210
equating space, 4, 17, 18
equilibrium, 24
    point, 144, 268
    solution, 144
equivalence
    class, 230
    relation, 230
equivalent differential equations, 45
error feedback, 348
estimate, 350
estimation error, 350
Euclidean norm, 34
Euclidean ring, 45
Euler equations, 285
Euler–Lagrange equations, 321, 400
exact elimination, 213
exclusion law, 1
exogenous inputs, xv
exponential behavior, 292
exponential of a matrix, 128, 133
extended state, 357
external
    behavior, 7
    variable, 7

F

feedback, x, xv, 318
    amplifier, xii
    compensator, 356

controller, xv
gain matrix, 323
interconnection, 238
processor, 322
stabilization, 324
feedthrough term, 126
free variable, 84
frequency response, 288
full behavior, 7, 8, 10, 120
full column rank, 58
full row rank, 58
representation, 59, 106
future, 79

G

gain, 301
gain equivalent, 316
geometric multiplicity, 136
governor, x
greatest common divisor, 54

H

H_∞ problem, xvii
Hautus tests, 184
Heaviside step function, 297
high-frequency noise, 383
high-frequency roll-off, 303
high-pass, 303
Hurwitz
matrix, 254
polynomial, 254
test, 257

I

i/o stability, 272
image representation, 235
impulse response, 99, 288, 297
independence
of functions, 74
of polynomials, 58
initially at rest system, 101
innovations, 351
input, xviii, 84
input/output partition, 84
input/output equivalent, 142
input/output form, 91
input/state/output systems, 126
instability, 251
integral representation, 34
interconnection, 373
elimination in, 214
feedback, 296
parallel, 238, 296
series, 215, 242, 296
transfer function of, 296
internal model, 348, 351
internal variable, 7
invariant subspace, 168

J

Jordan form, 135

K

Kalman decomposition, 189
kernel representation, 234
kernel, convolution, 98, 99
Kirchhoff's laws, 5, 6

L

$\mathfrak{L}_\infty$-i/o-stable, 272
$\mathfrak{L}_p$-i/o-stable, 272
lag, 17
latent
variable, xix, 2, 5, 206
variable model, 7
variable representation, 7, 10
leading principal minor, 257
left unimodular transformation, 55
linear time-invariant differential system, xviii, 18, 28, 31
linearity, 16, 43
linearization, 143, 147, 268, 335, 369
locally integrable, 34
locally specified, 19
low-pass, 303
lower triangular form, 412
LQG problem, xvi
lumped systems, 9
Lyapunov
equation, 259, 263, 266
function, 260
function, quadratic, 262

M

manifest
behavior, 7, 10, 120
dynamical system, 10
mathematical model, 7

variable, xix, 2, 7, 206
marginal stabilizability, 344
mathematical model, 1, 3
matrix exponential, 128, 133
matrix of proper rational functions, 85
maximally free, 84, 90
maximum principle, xiv
measured outputs, xv
memory, 119, 123
memoryless feedback, 323, 357
minimal polynomial, 278
minimal representation, 59, 106
minimal state representation, 234
minimum phase system, 316
minor, 257
modern control theory, xvii
monic polynomial, 54
multiplicity
algebraic, 136
geometric, 136

N

negative definite, 262
nonanticipating, 93, 99
strictly, 93
nonnegative definite, 262
nonpositive definite, 262
Nyquist plot, 303
Nyquist stability criterion, xiii

O

observability, xvi, xix, 178
index, 368
matrix, 182
of i/s/o systems, 181
observable pair, 182
observer, 347, 350
canonical form, 221, 222
characteristic polynomial, 353
gain matrix, 352
pole placement theorem, 353
poles, 352
octave, 302
open loop characteristic polynomial, 324
open loop control, 318
open loop poles, 324
optimal control, 324
order of compensator, 357
order of state space representation, 124
output, xviii, 84
overcompensation, xi
overdamped, 306
overshoot, 298

P

parallel interconnection, 238
partial fraction expansion
multivariable, 87
scalar, 86, 417
pass-band, 303
peak frequency, 303
peak gain, 303
permutation matrix, 50
phase, 301
PI-controller, 396
PID controller, xi
plant, ix, xv, 322
pole, 308
pole placement
algorithm
behavioral, 377
dynamic, 360
static, 331
in observers, 352
problem, 324
pole/zero diagram, 308
poles, 324
poles of controlled system, 376
polynomial matrix, 44
positive definite, 262
primal system, 386
principal minor, 257
proper, 85
strictly, 85
proper algebraic variety, 332
property of state, 123, 132
Proportional–Integral–Differential controller, xi

Q

quadratic form, 262

R

rank of polynomial matrix, 58
rational function, 48

reachable subspace, 174
read-out map, 144
recursivity, 347
reduced order compensator, 367
reduced order observer, 364
regulation, ix
relative degree, 116
relative gain, 303
resonance, 275
resonant frequency, 275, 303
reversed input/output structure, 380
right unimodular transformation, 55
    static, 55
ring, 44
robustness, xvi, 258, 386
Routh test, 255
Routh–Hurwitz conditions, 254, 257
row rank, 58

S

sampling theorem, 385
Schur polynomial, 282
semisimple, 251, 253
sensor, xv, 371
separation principle, 348, 358
series interconnection, 214
set of measure zero, 35
settling time, 298
shift-invariance, 16
signal space, 9
similarity
    of pairs, 326
    of quadruples, 143, 230
    of triples, 355
simple root, 251
singular value, 160
singularities, 251
SISO systems, 68
Smith form, 56, 414
solution of differential equation, 31
stability, 251, 268
stabilizable, 376
    behavior, 176
    pair, 177, 334
stabilization, xvi, 333
    of nonlinear systems, 368
state
    controllable, 168
    evolution function, 144
    observability, 181
    observer, 350
    system, 357
    transition matrix, 137
state space, 124
    representation problem, 221
    model, xvi, 123
    transformations, 142
static feedback, 323, 357
static gain, 298
static right unimodular transformation, 55
steady state, 298
step response, 297
strictly nonanticipating, 93
strictly proper
    input/output system, 94
    matrix of rational functions, 85
strong solution, 33
structural stability, 248
superposition principle, 16
Sylvester
    equation, 379
    resultant, 196, 379
system with memory, 357

T

three-term controller, xi
time axis, 9
time-invariant, 16, 43
time-reversible, 25
timeconstant, 299
to-be-controlled-outputs, xv
tracking, xvi, 300
trajectory optimization, ix, xiii
transfer function, xvi, 288, 292
transfer matrix, 86
transition matrix, 137
trim canonical form, 230

U

uncontrollable
    mode, 334
    pole, 334
    polynomial, 334, 361, 376
uncontrollable modes, 176
underdamped, 306
undershoot, 299
unimodular matrix, 47

unit step, 297
universum, 3
unobservable
    mode, 355
    pole, 355
    polynomial, 355, 361
unstable, 251, 268
upper triangular form, 56, 411

V

Vandermonde matrix, 79, 112
variation of the constants formula, 131

W

weak solution, 34, 35

Z

zero, 308
zero measure, 35