Chapter 4

Linear Dynamical Systems: Part 1

In this chapter we review some basic results concerning linear dynamical systems, which are geared toward the main topic of this book, namely, approximation of large-scale systems. General references for the material in this chapter are [280], [304], [370], [371], [76]. For an introduction to linear systems from basic principles, see the book by Polderman and Willems [270]. Here it is assumed that the external variables have been partitioned into *input variables* **u** and *output variables* **y**, and we consider *convolution systems*, i.e., systems where the relation between **u** and **y** is given by a convolution sum or integral

$$\mathbf{y} = \mathbf{h} * \mathbf{u}, \tag{4.1}$$

where **h** is an appropriate *weighting pattern*. This is called the *external description*. We are also concerned with systems in which in addition to the input and output variables, the *state* **x** has been defined as well. Furthermore, the relationship between **x** and **u** is given by means of a set of first-order difference or differential equations with constant coefficients, while that of **y** with **x** and **u** is given by a set of linear algebraic equations. It is assumed that **x** lives in a finite-dimensional space:

$$\sigma\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \;\; \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \tag{4.2}$$

where σ is the derivative operator or shift operator and **A**, **B**, **C**, **D** are linear constant maps. This is called the *internal description*.

The first section is devoted to the discussion of systems governed by (4.1), while the next section investigates some structural properties of systems described by (4.2), namely, reachability and observability. Closely related is the concept of gramians for linear systems, which is central in subsequent developments. The third section discusses the equivalence of the external and internal descriptions. As it turns out, going from the latter to the former involves the elimination of **x** and is thus straightforward. The converse, however, is far from trivial as it involves the *construction of state*. It is called the *realization problem*.

4.1 External description

Let $\mathbb{U} = \{\mathbf{u} : \mathbb{Z} \to \mathbb{R}^m\}$, $\mathbb{Y} = \{\mathbf{y} : \mathbb{Z} \to \mathbb{R}^p\}$. A *discrete-time linear system* $\mathbf{\Sigma}$ with m input and p output channels can be viewed as an operator between the *input space* $\mathbb{U}$ and the *output space* $\mathbb{Y}$, $\mathcal{S} : \mathbb{U} \longrightarrow \mathbb{Y}$, which is linear. There exists a sequence of matrices $\mathbf{h}(i, j) \in \mathbb{R}^{p\times m}$ such that

$$\mathbf{u} \longmapsto \mathbf{y} = \mathcal{S}(\mathbf{u}), \quad \mathbf{y}(i) = \sum_{j\in\mathbb{Z}} \mathbf{h}(i, j)\mathbf{u}(j), \qquad i \in \mathbb{Z}.$$

This relationship can be written in matrix form as follows:

$$\left(\begin{array}{c} \vdots \\ \mathbf{y}(-2) \\ \mathbf{y}(-1) \\ \hline \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \end{array}\right) = \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \iddots \\ \cdots & \mathbf{h}(-2,-2) & \mathbf{h}(-2,-1) & \mathbf{h}(-2,0) & \mathbf{h}(-2,1) & \cdots \\ \cdots & \mathbf{h}(-1,-2) & \mathbf{h}(-1,-1) & \mathbf{h}(-1,0) & \mathbf{h}(-1,1) & \cdots \\ \hline \cdots & \mathbf{h}(0,-2) & \mathbf{h}(0,-1) & \mathbf{h}(0,0) & \mathbf{h}(0,1) & \cdots \\ \cdots & \mathbf{h}(1,-2) & \mathbf{h}(1,-1) & \mathbf{h}(1,0) & \mathbf{h}(1,1) & \cdots \\ \iddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right) \left(\begin{array}{c} \vdots \\ \mathbf{u}(-2) \\ \mathbf{u}(-1) \\ \hline \mathbf{u}(0) \\ \mathbf{u}(1) \\ \vdots \end{array}\right).$$

The system $\mathbf{\Sigma}$ described by $\mathcal{S}$ is called *causal* if

$$\mathbf{h}(i, j) = \mathbf{0}, \qquad i \le j,$$

and *time-invariant* if

$$\mathbf{h}(i, j) = \mathbf{h}_{i-j} \in \mathbb{R}^{p\times m}.$$

For a time-invariant system $\mathbf{\Sigma}$ we can define the sequence of $p \times m$ constant matrices,

$$\mathbf{h} = (\ldots, \mathbf{h}_{-2}, \mathbf{h}_{-1}, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \ldots).$$

This sequence is called the *impulse response* of $\mathbf{\Sigma}$. In the *single-input, single-output* (SISO) case $m = p = 1$, it is the output obtained in response to a unit pulse,

$$\mathbf{u}(t) = \delta(t) = \begin{cases} 1, & t = 0, \\ 0, & t \ne 0. \end{cases}$$

In the *multi-input, multi-output* (MIMO) case, the subsequence of $\mathbf{h}$ composed of the kth column of each entry $\mathbf{h}_i$ is produced by applying the input $\mathbf{e}_k\delta(t)$, where $\mathbf{e}_k$ is the kth canonical unit vector (all entries are zero except the kth, which is 1). The operation of $\mathcal{S}$ can now be represented as a *convolution sum*:

$$\mathcal{S} : \mathbf{u} \longmapsto \mathbf{y} = \mathcal{S}(\mathbf{u}) = \mathbf{h} * \mathbf{u}, \quad \text{where} \quad (\mathbf{h} * \mathbf{u})(t) = \sum_{k=-\infty}^{\infty} \mathbf{h}_{t-k}\mathbf{u}(k), \qquad t \in \mathbb{Z}. \tag{4.3}$$

The convolution sum is also known as a *Laurent operator* in the theory of Toeplitz matrices (see, e.g., [70]). Moreover, the matrix representation of $\mathcal{S}$ in this case is a (doubly infinite) block Toeplitz matrix,

$$\begin{pmatrix} \vdots \\ \mathbf{y}(-2) \\ \mathbf{y}(-1) \\ \hline \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \end{pmatrix} = \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \iddots \\ \cdots & \mathbf{h}_0 & \mathbf{h}_{-1} & \mathbf{h}_{-2} & \mathbf{h}_{-3} & \cdots \\ \cdots & \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_{-1} & \mathbf{h}_{-2} & \cdots \\ \hline \cdots & \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_{-1} & \cdots \\ \cdots & \mathbf{h}_3 & \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right) \begin{pmatrix} \vdots \\ \mathbf{u}(-2) \\ \mathbf{u}(-1) \\ \hline \mathbf{u}(0) \\ \mathbf{u}(1) \\ \vdots \end{pmatrix}. \tag{4.4}$$

In what follows we will restrict our attention to causal and time-invariant linear systems. The matrix representation of $\mathcal{S}$ in this case is lower triangular and Toeplitz. In this case

$$\mathbf{\Sigma}: \quad \mathbf{y}(t) = \mathbf{h}_0\mathbf{u}(t) + \mathbf{h}_1\mathbf{u}(t-1) + \cdots + \mathbf{h}_k\mathbf{u}(t-k) + \cdots, \qquad t \in \mathbb{Z}.$$

The first term, $\mathbf{h}_0\mathbf{u}(t)$, denotes the *instantaneous* action of the system. The remaining terms denote the *delayed* or *dynamic* action of $\mathbf{\Sigma}$.

In analogy to the discrete-time case, let $\mathbb{U} = \{\mathbf{u} : \mathbb{R} \to \mathbb{R}^m\}$, $\mathbb{Y} = \{\mathbf{y} : \mathbb{R} \to \mathbb{R}^p\}$. A *continuous-time linear system* $\mathbf{\Sigma}$ with m input and p output channels can be viewed as an operator $\mathcal{S}$ mapping the input space $\mathbb{U}$ to the output space $\mathbb{Y}$, which is linear. In particular, we will be concerned with systems for which $\mathcal{S}$ can be expressed by means of an integral

$$\mathcal{S} : \mathbf{u} \longmapsto \mathbf{y}, \ \mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t,\tau)\mathbf{u}(\tau)\, d\tau, \qquad t \in \mathbb{R},$$

where $\mathbf{h}(t,\tau)$ is a matrix-valued function called the *kernel* or *weighting pattern* of $\mathbf{\Sigma}$. The system just defined is *causal* if

$$\mathbf{h}(t,\tau) = 0, \qquad t \le \tau,$$

and *time-invariant* if $\mathbf{h}$ depends on the difference of the two arguments,

$$\mathbf{h}(t,\tau) = \mathbf{h}(t-\tau).$$

In this case $\mathcal{S}$ is a *convolution operator*

$$\mathcal{S}: \ \mathbf{u} \mapsto \mathbf{y} = \mathcal{S}(\mathbf{u}) = \mathbf{h} * \mathbf{u}, \ \text{ where } \ (\mathbf{h} * \mathbf{u})(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\tau)\mathbf{u}(\tau)\, d\tau, \qquad t \in \mathbb{R}. \tag{4.5}$$

It is assumed from now on that $\mathcal{S}$ is both causal and time-invariant, which means that the upper limit of integration can be replaced by t. In addition, as in the discrete-time case, we will distinguish between *instantaneous* and purely *dynamic* action, that is, we will express the output as a sum of two terms, the first being the instantaneous and the second the dynamic action:

$$\mathbf{y}(t) = \mathbf{h}_0\mathbf{u}(t) + \int_{-\infty}^{t} \mathbf{h}_a(t-\tau)\mathbf{u}(\tau)\, d\tau,$$

where $\mathbf{h}_0 \in \mathbb{R}^{p\times m}$ and $\mathbf{h}_a$ is a smooth kernel. In particular, this requirement implies that $\mathbf{h}$ can be expressed as

$$\mathbf{h}(t) = \mathbf{h}_0\delta(t) + \mathbf{h}_a(t), \qquad t \ge 0, \tag{4.6}$$

where δ denotes the δ-distribution. It readily follows that $\mathbf{h}$ is the response of the system to the impulse δ and is therefore termed the *impulse response* of $\mathbf{\Sigma}$.

In what follows we will assume that $\mathbf{h}_a$ is an *analytic* function. This assumption implies that $\mathbf{h}_a$ is uniquely determined by the coefficients of its Taylor series expansion at $t = 0^+$:

$$\mathbf{h}_a(t) = \mathbf{h}_1 + \mathbf{h}_2 \frac{t}{1!} + \mathbf{h}_3 \frac{t^2}{2!} + \cdots + \mathbf{h}_k \frac{t^{k-1}}{(k-1)!} + \cdots, \qquad \mathbf{h}_k \in \mathbb{R}^{p \times m}.$$

It follows that if (4.6) is satisfied, the output $\mathbf{y}$ is at least as smooth as the input $\mathbf{u}$, and $\mathbf{\Sigma}$ is consequently called a *smooth* system. Hence, just like in the case of discrete-time systems, smooth continuous-time linear systems can be described by means of the infinite sequence of $p \times m$ matrices $\mathbf{h}_i$, $i \geq 0$. We formalize this conclusion next.

Definition 4.1. *The* external description *of a time-invariant, causal, and smooth continuous-time system and that of a time-invariant, causal, discrete-time linear system with m inputs and p outputs is given by an infinite sequence of $p \times m$ matrices,*

$$\mathbf{\Sigma} = (\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k, \ldots), \qquad \mathbf{h}_k \in \mathbb{R}^{p \times m}. \tag{4.7}$$

The matrices $\mathbf{h}_k$ are often referred to as the Markov parameters *of the system $\mathbf{\Sigma}$.*

Notice that by abuse of notation, we use $\mathbf{\Sigma}$ to denote both the system operator and the underlying sequence of Markov parameters. It should be clear from the context which of the two cases applies.

The (continuous- or discrete-time) Laplace transform of the impulse response yields the transfer function of the system

$$\mathbf{H}(\xi) = (\mathcal{L}\mathbf{h})(\xi). \tag{4.8}$$

The Laplace variable is denoted by ξ for both continuous- and discrete-time systems. It readily follows that $\mathbf{H}$ can be expanded in a formal power series in ξ:

$$\mathbf{H}(\xi) = \mathbf{h}_0 + \mathbf{h}_1\xi^{-1} + \mathbf{h}_2\xi^{-2} + \cdots + \mathbf{h}_k\xi^{-k} + \cdots.$$

This can also be regarded as a Laurent expansion of $\mathbf{H}$ around infinity. Consequently, (4.3) and (4.5) can be written as

$$\mathbf{Y}(\xi) = \mathbf{H}(\xi)\mathbf{U}(\xi).$$

Remark 4.1.1. *The behavioral framework.* In the classical framework (see, e.g., Kalman, Falb, and Arbib [192, Chapter 1]), a dynamical system is viewed as a mapping which transforms inputs $\mathbf{u}$ into outputs $\mathbf{y}$. Two basic considerations express the need for a framework at a more fundamental level. First, in many cases (think, for example, of electrical circuits), the distinction between inputs and outputs is not a priori clear; instead, it should follow as a consequence of the modeling. Second, it is desirable to be able to treat the different representations of a given system (for example, input-output and state-space representations) in a unified way.

In the behavioral setting, the basic variables considered are the external or manifest variables **w**, which consist of **u** and **y**, without distinguishing between them. The collection of trajectories describing the evolution of **w** over time defines a dynamical system. It turns out that this definition provides the right level of abstraction, necessary for accommodating the two considerations laid out above. This establishes the foundations of a parameter-free theory of dynamical systems, the advantages of representation-independent results—or, vice versa, the disadvantages of representation-dependent results—being well recognized. The resulting central object is the most powerful unfalsified model (MPUM) derived from the data, which, again, is a space of trajectories. Subsequently, inputs and outputs can be introduced and the corresponding input-output operator recovered. For details on the behavioral framework, see [270]; see also [272]. ■

4.2 Internal description

Alternatively, we can characterize a linear system via its *internal description*, which in addition to the input **u** and the output **y** uses the state **x**. Again, for a first-principles treatment of the concept of state, see the book by Willems and Poldeman [270]. For our purposes, three linear spaces are given: the *state space* $\mathbb{X}$, the *input space* $\mathbb{U}$, and the *output space* $\mathbb{Y}$, containing functions taking values in $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$, respectively. The *state equations* describing a linear system are a set of first-order linear *differential* or *difference* equations, according to whether we are dealing with a continuous- or a discrete-time system:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad t \in \mathbb{R}, \quad \text{or} \tag{4.9}$$

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad t \in \mathbb{Z}. \tag{4.10}$$

In both cases, $\mathbf{x} \in \mathbb{X}$ is the *state* of the system, while $\mathbf{u} \in \mathbb{U}$ is the input function. Moreover,

$$\mathbf{B} : \ \mathbb{R}^m \to \mathbb{R}^n, \ \ \mathbf{A} : \ \mathbb{R}^n \to \mathbb{R}^n$$

are (constant) linear maps; the first is called the *input* map, while the second describes the *dynamics* or *internal evolution* of the system. Equations (4.9) and (4.10) can be written in a unified way,

$$\sigma\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \tag{4.11}$$

where σ denotes the derivative operator for continuous-time systems and the (backward) shift operator for discrete-time systems.

The *output equations*, for both discrete- and continuous-time linear systems, are composed of a set of linear algebraic equations,

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \tag{4.12}$$

where **y** is the output function (response) and

$$\mathbf{C} : \ \mathbb{R}^n \to \mathbb{R}^p, \ \ \mathbf{D} : \ \mathbb{R}^m \to \mathbb{R}^p$$

are (constant) linear maps; **C** is called the *output* map. It describes how the system interacts with the outside world.

In what follows, the term *linear system* in the internal description is used to denote a linear, time-invariant, continuous- or discrete-time system which is finite-dimensional. Linear means that $\mathbb{U}$, $\mathbb{X}$, $\mathbb{Y}$ are linear spaces, and $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are linear maps; finite-dimensional means that m, n, p are all finite positive integers; time-invariant means that $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ do not depend on time; their matrix representations are constant $n \times n$, $n \times m$, $p \times n$, $p \times m$ matrices. By a slight abuse of notation, the linear maps $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ as well as their matrix representations (in some appropriate basis) are denoted with the same symbols. We are now ready to give a definition.

Definition 4.2. **(a)** *A* linear system *in* internal *or* state space *description is a quadruple of linear maps (matrices)*

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{D} \in \mathbb{R}^{p\times m}. \tag{4.13}$$

The dimension *of the system is defined as the dimension of the associated state space:*

$$\dim \Sigma = n. \tag{4.14}$$

(b) $\boldsymbol{\Sigma}$ *is called* stable *if the eigenvalues of* $\mathbf{A}$ *have negative real parts (for continuous-time systems) or lie inside the unit disk (for discrete-time systems).*

The concept of stability is introduced formally in the above definition. For a more detailed discussion, see section 5.8. We will also use the notation

$$\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ \mathbf{C} \in \mathbb{R}^{p\times n}. \tag{4.15}$$

It denotes a linear system where either $\mathbf{D} = \mathbf{0}$ or $\mathbf{D}$ is irrelevant for the argument pursued.

Example 4.3. We consider the dynamical system $\boldsymbol{\Sigma}$ shown in Figure 4.1. The external variables are the voltage applied at the terminals denoted by $\mathbf{u}$ and the voltage across the resistor denoted by $\mathbf{y}$. The former is the *input* or *excitation function* and the latter the *output* or *measured variable* of $\boldsymbol{\Sigma}$. One choice for the internal or state variables is to pick the *current* through the inductor, denoted by $\mathbf{x}_1$, and the voltage across the capacitor, denoted by $\mathbf{x}_2$. The state equations are thus $\mathbf{u} = R\mathbf{x}_1 + L\dot{\mathbf{x}}_1 + \mathbf{x}_2$ and $C\dot{\mathbf{x}}_2 = \mathbf{x}_1$, while the output equation is $\mathbf{y} = R\mathbf{x}_1$. Consequently, in (4.9), $\mathbf{x} = (\mathbf{x}_1,\ \mathbf{x}_2)^*$,

$$\mathbf{A} = \begin{pmatrix} -\frac{R}{L} & -\frac{1}{L} \\ \frac{1}{C} & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \frac{1}{L} \\ 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} R & 0 \end{pmatrix}, \quad \mathbf{D} = 0.$$

The system has dimension $n = 2$, and assuming that R, L, C are positive, it is stable since the characteristic polynomial $\chi_{\mathbf{A}}(s) = s^2 + \frac{R}{L}s + \frac{1}{CL}$ of $\mathbf{A}$ has roots with negative real parts.

Solution of the state equations

We will now give the solution of (4.11). For this we will need the *matrix exponential*; given $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}$, we define the matrix exponential by means of the same series

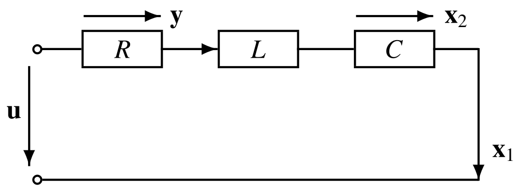

Figure 4.1. *An RLC circuit.*

representation as the scalar exponential, namely,

$$e^{\mathbf{M}t} = \mathbf{I}_n + \frac{t}{1!}\mathbf{M} + \frac{t^2}{2!}\mathbf{M}^2 + \cdots + \frac{t^k}{k!}\mathbf{M}^k + \cdots . \tag{4.16}$$

Let $\phi(\mathbf{u}; \mathbf{x}_0; t)$ denote the solution of the state equations (4.11), i.e., the state of the system at time t attained from the initial state $\mathbf{x}_0$ at time t_0, under the influence of the input $\mathbf{u}$. In particular, for the continuous-time state equations (4.9),

$$\phi(\mathbf{u}; \mathbf{x}_0; t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}_0 + \int_{t_0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau, \qquad t \geq t_0, \tag{4.17}$$

while for the discrete-time state equations (4.10),

$$\phi(\mathbf{u}; \mathbf{x}_0; t) = \mathbf{A}^{t-t_0}\mathbf{x}_0 + \sum_{j=t_0}^{t-1} \mathbf{A}^{t-1-j}\mathbf{B}\mathbf{u}(j), \qquad t \geq t_0. \tag{4.18}$$

For both discrete- and continuous-time systems, it follows that the output is given by

$$\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{u}; \mathbf{x}(t_0); t) + \mathbf{D}\mathbf{u}(t) = \mathbf{C}\phi(\mathbf{0}; \mathbf{x}(t_0); t) + \mathbf{C}\phi(\mathbf{u}; \mathbf{0}; t) + \mathbf{D}\mathbf{u}(t). \tag{4.19}$$

If we compare the above expressions for $t_0 = -\infty$ and $\mathbf{x}_0 = 0$ with (4.3) and (4.5), it follows that the *impulse response* $\mathbf{h}$ has the form below. For continuous-time systems,

$$\mathbf{h}(t) = \begin{cases} \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \delta(t)\mathbf{D}, & t \geq 0, \\ \mathbf{0}, & t < 0, \end{cases} \tag{4.20}$$

where δ denotes the δ-distribution. For discrete-time systems,

$$\mathbf{h}(t) = \begin{cases} \mathbf{C}\mathbf{A}^{t-1}\mathbf{B}, & t > 0, \\ \mathbf{D}, & t = 0, \\ \mathbf{0}, & t < 0. \end{cases} \tag{4.21}$$

Finally, by (4.8) the Laplace transform of the impulse response, which is called the transfer function of $\mathbf{\Sigma}$, is

$$\mathbf{H}_{\Sigma}(\xi) = \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}, \tag{4.22}$$

where $\xi = s$ (continuous-time Laplace transform) and $\xi = z$ (discrete-time Laplace or $\mathcal{Z}$-transform). Expanding the transfer function in a Laurent series for large ξ, i.e., in the neighborhood of infinity, we get

$$\mathbf{H}_\Sigma(\xi) = \mathbf{D} + \mathbf{CB}\,\xi^{-1} + \mathbf{CAB}\,\xi^{-2} + \cdots + \mathbf{CA}^{k-1}\mathbf{B}\,\xi^{-k} + \cdots,$$

and the corresponding external description given by the Markov parameters (4.7) is

$$\boldsymbol{\Sigma} = (\mathbf{D}, \mathbf{CB}, \mathbf{CAB}, \mathbf{CA}^2\mathbf{B}, \ldots, \mathbf{CA}^{k-1}\mathbf{B}, \ldots). \tag{4.23}$$

Sometimes it is advantageous to describe the system from a point of view different from the original one. In our case, since the external variables (i.e., the input and the output) are fixed, only the state variables can be transformed. In particular, if the new state is

$$\widetilde{\mathbf{x}} = \mathbf{Tx}, \quad \det \mathbf{T} \neq 0,$$

the corresponding matrices describing the system will change. More precisely, given the *state transformation* $\mathbf{T}$, (4.11) and (4.12) become

$$\sigma\widetilde{\mathbf{x}} = \underbrace{\mathbf{TAT}^{-1}}_{\widetilde{\mathbf{A}}}\widetilde{\mathbf{x}} + \underbrace{\mathbf{TB}}_{\widetilde{\mathbf{B}}}\mathbf{u}, \quad \mathbf{y} = \underbrace{\mathbf{CT}^{-1}}_{\widetilde{\mathbf{C}}}\widetilde{\mathbf{x}} + \underbrace{\mathbf{D}}_{\widetilde{\mathbf{D}}}\mathbf{u},$$

where $\mathbf{D}$ remains unchanged. The corresponding system triples are called *equivalent*. Put differently, $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$ are equivalent if

$$\left(\begin{array}{c|c} \mathbf{T} & \\ \hline & \mathbf{I}_p \end{array}\right)\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \widetilde{\mathbf{A}} & \widetilde{\mathbf{B}} \\ \hline \widetilde{\mathbf{C}} & \widetilde{\mathbf{D}} \end{array}\right)\left(\begin{array}{c|c} \mathbf{T} & \\ \hline & \mathbf{I}_m \end{array}\right) \tag{4.24}$$

for some invertible matrix $\mathbf{T}$. If $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$ are equivalent with equivalence transformation $\mathbf{T}$, it readily follows that

$$\begin{aligned}\mathbf{H}_{\boldsymbol{\Sigma}}(\xi) &= \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{D} + \mathbf{CT}^{-1}\mathbf{T}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{T}^{-1}\mathbf{TB} \\ &= \mathbf{D} + \mathbf{CT}^{-1}(\xi\mathbf{I} - \mathbf{TAT}^{-1})^{-1}\mathbf{TB} = \widetilde{\mathbf{D}} + \widetilde{\mathbf{C}}(\xi\mathbf{I} - \widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{B}} = \mathbf{H}_{\widetilde{\boldsymbol{\Sigma}}}(\xi).\end{aligned}$$

This immediately implies that $\mathbf{h}_k = \widetilde{\mathbf{h}}_k$, $k = 1, 2, \ldots$. We have thus proved the following.

Proposition 4.4. *Equivalent triples have the same transfer function and consequently the same Markov parameters.*

Example 4.5. *Continuation of Example* 4.3. The first five Markov parameters of $\boldsymbol{\Sigma}$ are

$$0,\ \frac{R}{L},\ -\frac{R^2}{L^2},\ (CR^2 - L)\frac{R}{CL^3},\ -(CR^2 - 2L)\frac{R^2}{C^2L^5}.$$

Assuming that $R = 1$, $L = 1$, $C = 1$, the matrix exponential is

$$e^{\mathbf{A}t} = e^{-\frac{t}{2}}\cos\left[\frac{t\sqrt{3}}{2}\right]\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{3}e^{-\frac{t}{2}}\sin\left[\frac{t\sqrt{3}}{2}\right]\begin{pmatrix} -1 & -2 \\ 2 & 1 \end{pmatrix}.$$

Thus, the impulse response is $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} = e^{-\frac{t}{2}}\cos\left[\frac{t\sqrt{3}}{2}\right] - \frac{1}{3}e^{-\frac{t}{2}}\sin\left[\frac{t\sqrt{3}}{2}\right]$, $t \geq 0$, while the transfer function (in terms of R, L, C) is

$$\mathbf{H}_{\Sigma}(s) = \frac{\frac{R}{L}s}{s^2 + \frac{R}{L}s + \frac{1}{RC}}.$$

Finally, if the state is changed to $\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x}$, where $\mathbf{T} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, the new state space representation of $\mathbf{\Sigma}$ is

$$\tilde{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1} = \frac{-1}{2LC}\begin{pmatrix} RC-L+C & RC-L-C \\ RC+L+C & RC+L-C \end{pmatrix}, \quad \tilde{\mathbf{B}} = \mathbf{T}\mathbf{B} = \frac{1}{L}\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\tilde{\mathbf{C}} = \mathbf{C}\mathbf{T}^{-1} = \frac{R}{2}(1 \quad 1).$$

4.2.1 The concept of reachability

In this subsection we introduce and discuss the fundamental concept of reachability of a linear system $\mathbf{\Sigma}$. This concept allows us to identify the extent to which the state of the system $\mathbf{x}$ can be manipulated through the input $\mathbf{u}$. The related concept of *controllability* is discussed subsequently. Both concepts involve only the state equations. Consequently, for this subsection and the next, $\mathbf{C}$ and $\mathbf{D}$ will be ignored: $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$. For a survey of reachability and observability (which is introduced later), see [16].

Definition 4.6. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$. *A state* $\bar{\mathbf{x}} \in \mathbb{X}$ *is* reachable *from the zero state if there exist an input function* $\bar{\mathbf{u}}(t)$, *of finite energy, and a time* $\bar{T} < \infty$, *such that*

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; 0; \bar{T}).$$

The reachable subspace $\mathbb{X}^{\text{reach}} \subset \mathbb{X}$ *of* $\mathbf{\Sigma}$ *is the set containing all reachable states of* $\mathbf{\Sigma}$. *The system* $\mathbf{\Sigma}$ *is* (completely) reachable *if* $\mathbb{X}^{\text{reach}} = \mathbb{X}$. *Furthermore,*

$$\mathcal{R}(\mathbf{A}, \mathbf{B}) = [\mathbf{B} \quad \mathbf{A}\mathbf{B} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B} \quad \cdots\ ] \tag{4.25}$$

is the reachability matrix *of* $\mathbf{\Sigma}$.

By the Cayley–Hamilton theorem, the rank of the reachability matrix and the span of its columns are determined (at most) by the first n terms, i.e., $\mathbf{A}^t\mathbf{B}$, $t = 0, 1, \ldots, n-1$. Thus for computational purposes the following (finite) reachability matrix is of importance:

$$\mathcal{R}_n(\mathbf{A}, \mathbf{B}) = [\mathbf{B} \quad \mathbf{A}\mathbf{B} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B}]. \tag{4.26}$$

The *image* of a linear map $\mathbf{L}$ is denoted by $\operatorname{im}\mathbf{L}$. The fundamental result concerning reachability is the following.

Theorem 4.7. *Given* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ *for both the continuous- and the discrete-time case,* $\mathbb{X}^{\text{reach}}$ *is a linear subspace of* $\mathbb{X}$*, given by the formula*

$$\mathbb{X}^{\text{reach}} = \text{im } \mathcal{R}(\mathbf{A}, \mathbf{B}). \tag{4.27}$$

As mentioned in section 10.4, expression (10.7) is referred to as the *Krylov subspace* in the numerical linear algebra community, while it is known as the reachability space in the systems community.

Corollary 4.8. **(a)** $\mathbf{A}\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{reach}}$. **(b)** Σ *is (completely) reachable if and only if* $\text{rank}\mathcal{R}(\mathbf{A}, \mathbf{B}) = n$. **(c)** *Reachability is basis independent.*

Proof. We will first prove Corollary 4.8. **(a)** $\mathbf{A}\,\mathbb{X}^{\text{reach}} = \mathbf{A}\,\text{im}\,\mathcal{R}(\mathbf{A}, \mathbf{B}) = \text{im}\,\mathbf{A}\mathcal{R}(\mathbf{A}, \mathbf{B}) = \text{im}\,(\mathbf{AB}\ \mathbf{A}^2\mathbf{B}\ \cdots\ ) \subset \text{im}\,(\mathbf{B}\ \ \mathbf{AB}\ \ \cdots\ ) = \mathbb{X}^{\text{reach}}$. **(b)** The result follows by noticing that $\mathbf{\Sigma}$ is (completely) reachable if and only if $\text{im}\,\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathbb{R}^n$. **(c)** Let $\mathbf{T}$ be a nonsingular transformation in $\mathbb{X}$, i.e., $\det \mathbf{T} \neq 0$. It follows from (4.24) that the pair $\mathbf{A}, \mathbf{B}$ is transformed into the pair $\mathbf{TAT}^{-1}, \mathbf{TB}$. It is readily checked that

$$\mathcal{R}(\mathbf{TAT}^{-1}, \mathbf{TB}) = \mathbf{T}\mathcal{R}(\mathbf{A}, \mathbf{B}),$$

which shows that the ranks of the original and the transformed reachability matrices are the same. □

Before proceeding with the proof of the theorem, some remarks are in order. In general, reachability is an *analytic* concept. The above theorem, however, shows that for linear, finite-dimensional, time-invariant systems, reachability reduces to an *algebraic* concept depending only on properties of $\mathbf{A}$ and $\mathbf{B}$ and, in particular, on the rank of the reachability matrix $\mathcal{R}(\mathbf{A}, \mathbf{B})$ but *independent* of time and the input function. It is also worthwhile to notice that formula (4.27) is valid for both continuous- and discrete-time systems. This, together with a similar result on observability (4.39), has as a consequence the fact that many tools for studying linear systems are algebraic. It should be noticed, however, that the physical significance of $\mathbf{A}$ and $\mathbf{B}$ is different for the discrete- and continuous-time cases; if we discretize, for instance, the continuous-time system $\dot{\mathbf{x}}(t) = \mathbf{A}_{\text{cont}}\mathbf{x}(t) + \mathbf{B}_{\text{cont}}\mathbf{u}(t)$ to $\mathbf{x}(t+1) = \mathbf{A}_{\text{discr}}\mathbf{x}(t) + \mathbf{B}_{\text{discr}}\mathbf{u}(t)$, then $\mathbf{A}_{\text{discr}} = e^{\mathbf{A}_{\text{cont}}}$.

A very useful concept is that of the *reachability gramian*. It is used in the proof of the theorem above and extensively in later chapters.

Definition 4.9. *The finite* reachability gramians *at time* $t < \infty$ *are defined for continuous-time systems as*

$$\mathcal{P}(t) = \int_0^t e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*\tau}\, d\tau, \qquad t \in \mathbb{R}_+, \tag{4.28}$$

and for discrete-time systems as

$$\mathcal{P}(t) = \mathcal{R}_t(\mathbf{A}, \mathbf{B})\mathcal{R}_t^*(\mathbf{A}, \mathbf{B}) = \sum_{k=0}^{t-1} \mathbf{A}^k\mathbf{B}\mathbf{B}^*(\mathbf{A}^*)^k, \qquad t \in \mathbb{Z}_+. \tag{4.29}$$

A Hermitian matrix $\mathbf{X} = \mathbf{X}^*$ is called *positive semidefinite* (*positive definite*) if its eigenvalues are nonnegative (positive). The difference of two Hermitian matrices $\mathbf{X}$, $\mathbf{Y}$ satisfies $\mathbf{X} \geq \mathbf{Y}$, ($\mathbf{X} > \mathbf{Y}$) if the eigenvalues of the difference $\mathbf{X} - \mathbf{Y}$ are nonnegative (positive).

Proposition 4.10. *The reachability gramians have the following properties:* **(a)** $\mathcal{P}(t) = \mathcal{P}^*(t) \geq 0$ *and* **(b)** *their columns span the reachability subspace, i.e.,*

$$\operatorname{im}\mathcal{P}(t) = \operatorname{im}\mathcal{R}(\mathbf{A}, \mathbf{B}).$$

This relationship holds for continuous-time systems for all $t > 0$ and for discrete-time systems (at least) for $t \geq n$.

Corollary 4.11. $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ *is reachable if and only if $\mathcal{P}(t)$ is positive definite for some $t > 0$.*

Proof. Next we will prove Proposition 4.10. **(a)** The symmetry and semidefiniteness of the reachability gramian follows by definition. **(b)** To prove the second property, it is enough to show that $\mathbf{q} \perp \mathcal{P}(t)$ if and only if $\mathbf{q} \perp \mathcal{R}(\mathbf{A}, \mathbf{B})$. Since $\mathcal{P}(t)$ is symmetric and semidefinite, $\mathbf{q} \perp \mathcal{P}(t)$ if and only if $\mathbf{q}^*\mathcal{P}(t)\mathbf{q} = 0$. Moreover,

$$\mathbf{q}^*\mathcal{P}(t)\mathbf{q} \;=\; \int_0^t \|\mathbf{B}^* e^{\mathbf{A}^*(t-\tau)}\mathbf{q}\|^2\, d\tau = 0$$

is equivalent to $\mathbf{B}^* e^{\mathbf{A}^* t}\mathbf{q} = 0$ for all $t \geq 0$. Since the exponential is an analytic function, this condition is equivalent to the function and all its derivatives being zero at $t = 0$, i.e., $\mathbf{B}^*(\mathbf{A}^*)^{i-1}\mathbf{q} = 0$, $i > 0$. This in turn is equivalent to $\mathbf{q} \perp \mathbf{A}^{i-1}\mathbf{B}$, $i > 0$, i.e., $\mathbf{q} \perp \mathcal{R}(\mathbf{A}, \mathbf{B})$. The proof for discrete-time systems is similar. □

Proof. We now turn our attention to Theorem 4.7. First we show that $\mathbb{X}^{\text{reach}}$ is a linear space, i.e.,

$$\text{if } \mathbf{x}_i = \phi(\mathbf{u}_i; \mathbf{0}; T_i) \in \mathbb{X}^{\text{reach}},\; i = 1, 2, \;\text{ then }\; \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 \in \mathbb{X}^{\text{reach}}$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$. Let $T_1 \geq T_2$. Define the input function

$$\hat{\mathbf{u}}_2(t) \;=\; \begin{cases} \mathbf{0}, & t \in [0, T_1 - T_2], \\ \mathbf{u}_2(t - T_1 + T_2), & t \in [T_1 - T_2, T_1]. \end{cases}$$

It is readily checked that for both continuous- and discrete-time systems,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = \phi(\alpha_1\mathbf{u}_1 + \alpha_2\hat{\mathbf{u}}_2; 0; T_1).$$

Next we prove (4.27) for discrete-time systems. Consider

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T}) = \sum_{j=0}^{\bar{T}-1} \mathbf{A}^{\bar{T}-1-j}\mathbf{B}\bar{\mathbf{u}}(j).$$

Clearly, $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}_{\bar{T}}(\mathbf{A}, \mathbf{B}) \subset \operatorname{im}\mathcal{R}(\mathbf{A}, \mathbf{B})$. Conversely, consider an element $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}(\mathbf{A}, \mathbf{B})$. By the Cayley–Hamilton theorem, this implies $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}_n(\mathbf{A}, \mathbf{B})$; thus, there exist elements

$\bar{\mathbf{u}}(j) \in \mathbb{R}^m$, $j = 0, 1, \ldots, n-1$, such that $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; n-1)$. To prove (4.27) for continuous-time systems, we make use of the expansion (4.16), i.e., $e^{\mathbf{A}t} = \sum_{i>0} \frac{t^{i-1}}{(i-1)!}\mathbf{A}^{i-1}$. Let $\bar{\mathbf{x}} \in \mathbb{X}^{\text{reach}}$. Then, for some $\bar{\mathbf{u}}$, $\bar{T}$ we have

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T}) = \sum_{i>0} \mathbf{A}^{i-1}\mathbf{B} \int_0^t \frac{(t-\tau)^{i-1}}{(i-1)!}\mathbf{u}(\tau)\, d\tau,$$

which shows that $\bar{\mathbf{x}} \in \operatorname{im} \mathcal{R}(\mathbf{A}, \mathbf{B})$.

For the converse inclusion, we use the proposition given above, which asserts that for every $\bar{\mathbf{x}} \in \operatorname{im} \mathcal{R}(\mathbf{A}, \mathbf{B})$, there exists $\bar{\xi}$ such that

$$\bar{\mathbf{x}} = \mathcal{P}(\bar{T})\bar{\xi}, \tag{4.30}$$

where $\mathcal{P}$ is the reachability gramian and $\bar{T}$ is any positive real number for the continuous-time case and at least n for the discrete-time case. Choose

$$\bar{\mathbf{u}}(t) = \mathbf{B}^* e^{\mathbf{A}^*(\bar{T}-t)}\bar{\xi} \tag{4.31}$$

for the continuous-time case and

$$\bar{\mathbf{u}}(t) = \mathbf{B}^*(\mathbf{A}^*)^{(\bar{T}-t)}\bar{\xi} \tag{4.32}$$

for the discrete-time case. (Recall that * denotes transposition if the matrix or vector is real and denotes complex conjugation and transposition if the matrix or vector is complex.) It follows that $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; 0; \bar{T}) \in \mathbb{X}^{\text{reach}}$. This concludes the proof of the theorem. □

Remark 4.2.1. *A formula for the matrix exponential.* Consider the square matrix $\mathbf{A} \in \mathbb{R}^{\nu \times \nu}$, with eigenvalues λ_i, $i = 1, \ldots, \nu$. One way to compute the matrix exponential of $\mathbf{A}$ given by (4.16) is

$$\begin{aligned} e^{\mathbf{A}t} &= f_\nu(t)\mathbf{A}^{\nu-1} + f_{\nu-1}(t)\mathbf{A}^{\nu-2} + \cdots + f_2(t)\mathbf{A} + f_1(t)\mathbf{I}_\nu, \quad \text{where} \\ [f_1(t) \ \cdots \ f_\nu(t)] &= [\phi_1(t) \ \cdots \ \phi_\nu(t)]\, \mathbf{V}(\lambda_1, \ldots, \lambda_\nu)^{-1}. \end{aligned}$$

If the eigenvalues of $\mathbf{A}$ are distinct, the functions ϕ_i are $\phi_i(t) = e^{\lambda_i t}$, $i = 1, \ldots, \nu$, and $\mathbf{V}$ is the Vandermonde matrix,

$$\mathbf{V}(\lambda_1, \ldots, \lambda_\nu) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & & \vdots \\ \lambda_1^{\nu-1} & \lambda_2^{\nu-1} & \cdots & \lambda_n^{\nu-1} \end{pmatrix}.$$

If the eigenvalues are not distinct, the functions ϕ_i are a fundamental set of solutions of the autonomous differential equation $q(D)f = 0$, where $D = \frac{d}{dt}$ and q is the characteristic polynomial of $\mathbf{A}$. In this case the Vandermonde matrix has to be modified accordingly.

From a numerical viewpoint, the computation of the matrix exponential is a challenging proposition. A method known as *scaling and squaring* yields the best results for a wide variety of matrices $\mathbf{A}$. A survey on this topic can be found in Moler and Van Loan [241].

The definition of the inner product for vector-valued sequences and functions is given in (5.12). The *energy* or *norm* of the sequence or function $\mathbf{f}$ denoted by $\| \mathbf{f} \|$ is thus defined as its 2-norm, i.e.,

$$\| \mathbf{f} \|^2 = \langle \mathbf{f}, \mathbf{f} \rangle = \int_0^T \mathbf{f}^*(t)\mathbf{f}(t)\, dt.$$

The input function $\bar{\mathbf{u}}$ defined by (4.30) and (4.31) is a *minimal energy input* which steers the system to the desired state at a given time.

Proposition 4.12. *Consider* $\bar{\mathbf{u}}$ *defined by* (4.30) *and* (4.31), *and let* $\hat{\mathbf{u}}$ *be any input function which reaches the state* $\bar{\mathbf{x}}$ *at time* $\bar{T}$*, i.e.,* $\phi(\hat{\mathbf{u}}; \mathbf{0}; \bar{T}) = \bar{\mathbf{x}}$. *Then*

$$\| \hat{\mathbf{u}} \| \geq \| \bar{\mathbf{u}} \| . \tag{4.33}$$

Furthermore, the minimal energy required to reach the state $\bar{\mathbf{x}}$ *at time* $\bar{T}$ *is equal to the energy of the input function* $\bar{\mathbf{u}}$, *which is equal to* $\bar{\xi}^*\mathcal{P}(\bar{T})\bar{\xi}$; *if the system is reachable, this formula becomes*

$$\| \bar{\mathbf{u}} \|^2 = \bar{\mathbf{x}}^*\mathcal{P}(\bar{T})^{-1}\bar{\mathbf{x}}. \tag{4.34}$$

Proof. The proof is based on the fact that the inner product of $\bar{\mathbf{u}}$ with $\hat{\mathbf{u}} - \bar{\mathbf{u}}$ is zero. □

From the above considerations, we can quantify the time needed to arrive at a given reachable state.

Proposition 4.13. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. **(a)** *For discrete-time systems, every reachable state can be reached in at most n time steps.* **(b)** *For continuous-time systems, every reachable state can be reached arbitrarily fast.*

The second part of the proposition implies that the delay in reaching a given state can be attributed to the nonlinearities present in the system.

Proof. Part **(a)** follows immediately from the Cayley–Hamilton theorem together with (4.27). In the latter part of the proof of Theorem 4.7, we showed that for any $\bar{\mathbf{x}} \in \mathbb{X}^{\text{reach}}$ we have $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T})$, where $\bar{\mathbf{u}}$ is defined by (4.30) and (4.31), while $\bar{T}$ is an arbitrary positive real number. This establishes claim **(b)**. □

Next we show that a nonreachable system can be decomposed in a canonical way into two subsystems: one whose states are all reachable and a second whose states are all unreachable.

Lemma 4.14. Reachable canonical decomposition. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}, \mathbf{B}$ *have the following matrix representations:*

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right) = \left(\begin{array}{cc|c} \mathbf{A}_r & \mathbf{A}_{r\bar{r}} & \mathbf{B}_r \\ \mathbf{0} & \mathbf{A}_{\bar{r}} & \mathbf{0} \\ \hline & & \end{array}\right), \tag{4.35}$$

where the subsystem $\mathbf{\Sigma}_r = \left(\begin{array}{c|c} \mathbf{A}_r & \mathbf{B}_r \\ \hline & \end{array}\right)$ *is reachable.*

Proof. Let $\mathbb{X}' \subset \mathbb{X}$ be such that $\mathbb{X} = \mathbb{X}^{\text{reach}} + \mathbb{X}'$ and $\dim \mathbb{X}' = n - q$. Choose a basis $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $\mathbb{X}$ so that $\mathbf{x}_1, \ldots, \mathbf{x}_q$ is a basis of $\mathbb{X}^{\text{reach}}$ and $\mathbf{x}_{q+1}, \ldots, \mathbf{x}_n$ is a basis for $\mathbb{X}'$. Since $\mathbb{X}^{\text{reach}}$ is $\mathbf{A}$-invariant, the matrix representation of $\mathbf{A}$ in the above basis has the form given by formula (4.35). Moreover, since $\operatorname{im} \mathbf{B} \subset \mathbb{X}^{\text{reach}}$, the matrix representation of $\mathbf{B}$ in the above basis has the form given by the same formula. Finally, to prove that $\mathbf{A}_r \in \mathbb{R}^{q\times q}$, $\mathbf{B}_r \in \mathbb{R}^{q\times m}$ is a reachable pair, it suffices to notice that

$$\operatorname{rank} \mathcal{R}(\mathbf{A}_r, \mathbf{B}_r) = \operatorname{rank} \mathcal{R}(\mathbf{A}, \mathbf{B}) = \dim \mathbb{X}^{\text{reach}} = q.$$

This concludes the proof of the lemma. □

Thus every system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ can be decomposed in a subsystem $\mathbf{\Sigma}_r = \left(\begin{array}{c|c} \mathbf{A}_r & \mathbf{B}_r \\ \hline & \end{array}\right)$, which is reachable, and in a subsystem $\mathbf{\Sigma}_{\bar r} = \left(\begin{array}{c|c} \mathbf{A}_{\bar r} & \mathbf{0} \\ \hline & \end{array}\right)$, which is completely unreachable, i.e., it cannot be influenced by outside forces. The interaction between $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_{\bar r}$ is given by $\mathbf{A}_{r\bar r}$. Since $\mathbf{A}_{\bar r r} = 0$, it follows that the unreachable subsystem $\mathbf{\Sigma}_{\bar r}$ influences the reachable subsystem $\mathbf{\Sigma}_r$ but not vice versa. It should be noticed that although $\mathbb{X}'$ in the proof above is not unique, the form (block structure) of the reachable decomposition (4.35) is unique.

We conclude this subsection by stating various equivalent conditions for reachability.

Theorem 4.15. Reachability conditions. *The following are equivalent:*

1. *The pair* $(\mathbf{A}, \mathbf{B})$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, *is reachable.*
2. *The rank of the reachability matrix is full:* $\operatorname{rank} \mathcal{R}(\mathbf{A}, \mathbf{B}) = n$.
3. *The reachability gramian is positive definite* $\mathcal{P}(t) > 0$ *for some* $t > 0$.
4. *No left eigenvector* $\mathbf{v}$ *of* $\mathbf{A}$ *is in the left kernel of* $\mathbf{B}$: $\mathbf{v}^*\mathbf{A} = \lambda \mathbf{v}^* \;\; \Rightarrow \;\; \mathbf{v}^*\mathbf{B} \neq 0$.
5. $\operatorname{rank}\, (\mu \mathbf{I}_n - \mathbf{A}, \quad -\mathbf{B}) = n$ *for all* $\mu \in \mathbb{C}$
6. *The polynomial matrices* $s\mathbf{I} - \mathbf{A}$ *and* $\mathbf{B}$ *are left coprime.*

The fourth and fifth conditions in the theorem are known as the Popov–Belevich–Hautus (PBH) tests for reachability. The last condition of the theorem is given for completeness; the concept of *left coprimeness* of polynomial matrices is not used further in this book; for a definition, see [123].

Proof. The equivalence of the first three statements has already been proved. The equivalence between conditions 4 and 5 is straightforward, and 6 can be considered as a different way of stating 5. We will prove the equivalence between conditions 1 and 4.

If there exists some nonzero $\mathbf{v}$ for which $\mathbf{v}^*\mathbf{A} = \lambda \mathbf{v}^*$ and $\mathbf{v}^*\mathbf{B} = \mathbf{0}$, clearly $\mathbf{v}^*\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathbf{0}$; this implies the lack of reachability of $(\mathbf{A}, \mathbf{B})$. Conversely, let $(\mathbf{A}, \mathbf{B})$ be unreachable; there exists a basis in the state space such that $\mathbf{A}$ and $\mathbf{B}$ have the form given by (4.35).

Let $\mathbf{v}_2 \neq \mathbf{0}$ be a left eigenvector of $\mathbf{A}_{\bar{r}}$. Then $\mathbf{v} = (\mathbf{0} \;\; \mathbf{v}_2^*)^*$ (where $\mathbf{0}$ is the zero vector of appropriate dimension) is a left eigenvector of $\mathbf{A}$ that is also in the left kernel of $\mathbf{B}$. This concludes the proof. $\square$

After introducing the reachability property, we introduce a weaker concept, which is sufficient for many problems of interest. Recall the canonical decomposition (4.35) of a pair $(\mathbf{A}, \mathbf{B})$.

Definition 4.16. *The pair* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ *is* stabilizable *if* $\mathbf{A}_{\bar{r}}$ *is stable, i.e., all its eigenvalues either have negative real parts or are inside the unit disk, depending on whether we are dealing with continuous- or discrete-time systems.*

Remark 4.2.2. Reachability is a *generic* property. This means, intuitively, that almost every $n \times n$, $n \times m$ pair of matrices $\mathbf{A}$, $\mathbf{B}$ satisfies

$$\operatorname{rank} \mathcal{R}_n(\mathbf{A}, \mathbf{B}) = n.$$

Put differently, in the space of all $n \times n$, $n \times m$ pairs of matrices, the unreachable pairs form a hypersurface (of “measure” zero).

A concept that is closely related to reachability is that of *controllability*. Here, instead of driving the zero state to a desired state, a given nonzero state is steered to the zero state. More precisely, we have the next definition.

Definition 4.17. *Given* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$*, a (nonzero) state* $\bar{\mathbf{x}} \in \mathbb{X}$ *is* controllable *to the zero state if there exist an input function* $\bar{\mathbf{u}}(t)$ *and a time* $\bar{T} < \infty$*, such that*

$$\phi(\bar{\mathbf{u}}; \bar{\mathbf{x}}; \bar{T}) = \mathbf{0}.$$

The controllable subspace $\mathbb{X}^{\mathrm{contr}}$ *of* Σ *is the set of all controllable states. The system* Σ *is* (completely) controllable *if* $\mathbb{X}^{\mathrm{contr}} = \mathbb{X}$.

The next theorem shows that for continuous-time systems the concepts of reachability and controllability are equivalent, while for discrete-time systems the latter is weaker. For this reason, only the notion of reachability is used in what follows.

Theorem 4.18. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. **(a)** *For continuous-time systems,* $\mathbb{X}^{\mathrm{contr}} = \mathbb{X}^{\mathrm{reach}}$. **(b)** *For discrete-time systems,* $\mathbb{X}^{\mathrm{reach}} \subset \mathbb{X}^{\mathrm{contr}}$*; in particular,* $\mathbb{X}^{\mathrm{contr}} = \mathbb{X}^{\mathrm{reach}} + \ker \mathbf{A}^n$.

Proof. **(a)** By definition, $\mathbf{x} \in \mathbb{X}^{\mathrm{contr}}$ implies $\phi(\mathbf{u}; \mathbf{x}; T) = \mathbf{0}$ for some $\mathbf{u}$; this implies

$$e^{\mathbf{A}T}\mathbf{x} = -\int_0^T e^{\mathbf{A}(T-\tau)}\mathbf{B}\mathbf{u}(\tau)\, d\tau \in \mathbb{X}^{\mathrm{reach}}.$$

Thus, $\mathbf{x} \in e^{-\mathbf{A}T}\mathbb{X}^{\mathrm{reach}} \subset \mathbb{X}^{\mathrm{reach}}$; the latter inclusion follows because by Corollary 4.8, $\mathbb{X}^{\mathrm{reach}}$ is $\mathbf{A}$-invariant. Thus, $\mathbb{X}^{\mathrm{contr}} \subset \mathbb{X}^{\mathrm{reach}}$. Conversely, let $\mathbf{x} \in \mathbb{X}^{\mathrm{reach}}$; there exist $\mathbf{u}$ and T such that

$-\mathbf{x} = \phi(\mathbf{u}; \mathbf{0}; T)$. It follows that $-e^{\mathbf{A}T}\mathbf{x} \in \mathbb{X}^{\text{reach}}$, which in turn implies $\phi(\mathbf{u}; \mathbf{x}; T) = \mathbf{0}$. Thus, $\mathbf{x} \in \mathbb{X}^{\text{contr}}$, i.e., $\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{contr}}$. This completes the proof of (a).

(b) Let $\mathbf{x} \in \mathbb{X}^{\text{reach}}$. Then, $-\mathbf{x} = \phi(\mathbf{u}; \mathbf{0}; T)$. Since $\mathbb{X}^{\text{reach}}$ is $\mathbf{A}$-invariant, $\mathbf{A}^T\mathbf{x} \in \mathbb{X}^{\text{reach}}$. But $\mathbf{x}$ satisfies $\phi(\mathbf{u}, \mathbf{x}, T) = \mathbf{0}$. Thus, $\mathbf{x} \in \mathbb{X}^{\text{contr}}$. The converse does not hold true in general, as $\mathbf{A}$ may be singular. □

Remark 4.2.3. From the above results it follows that for any two states $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^{\text{reach}}$ there exist $\mathbf{u}_{12}$, T_{12} such that $\mathbf{x}_1 = \phi(\mathbf{u}_{12}; \mathbf{x}_2; T_{12})$. To see this, note that since $\mathbf{x}_2$ is reachable it is also controllable; thus there exist $\mathbf{u}_2$, T_2 such that $\phi(\mathbf{u}_2; \mathbf{x}_2; T_2) = \mathbf{0}$. Finally, the reachability of $\mathbf{x}_1$ implies the existence of $\mathbf{u}_1$, T_1 such that $\mathbf{x}_1 = \phi(\mathbf{u}_1; \mathbf{0}; T_1)$. The function $\mathbf{u}_{12}$ is then the concatenation of $\mathbf{u}_2$ with $\mathbf{u}_1$, while $T_{12} = T_1 + T_2$. In general, if $\mathbf{x}_1, \mathbf{x}_2$ are not reachable, there is a trajectory passing through the two points if and only if

$$\mathbf{x}_2 - \mathbf{f}(\mathbf{A}, T)\mathbf{x}_1 \in \mathbb{X}^{\text{reach}} \quad \text{for some} \quad T,$$

where $\mathbf{f}(\mathbf{A}, T) = e^{\mathbf{A}T}$ for continuous-time systems and $\mathbf{f}(\mathbf{A}, T) = \mathbf{A}^T$ for discrete-time systems. This shows that if we start from a reachable state $\mathbf{x}_1 \neq 0$, the states that can be attained are also within the reachable subspace.

Distance to reachability/controllability

Following the considerations in section 3.3.3, the numerical computation of rank is an *ill-posed* problem. Therefore, the same holds for the numerical determination of reachability (controllability) of a given pair $(\mathbf{A}, \mathbf{B})$. One could consider instead the *numerical rank* of the reachability matrix $\mathcal{R}(\mathbf{A}, \mathbf{B})$ or of the reachability gramian $\mathcal{P}(T)$ or, if the system is stable, of the infinite gramian $\mathcal{P}$.

A measure of reachability that is well-posed is the *distance* of the pair to the set of unreachable/uncontrollable ones, denoted by $\delta_r(\mathbf{A}, \mathbf{B})$. Following part 5 of Theorem 4.15, this distance is defined as follows:

$$\delta_r(\mathbf{A}, \mathbf{B}) = \inf_{\mu \in \mathbb{C}} \sigma_{\min}[\mu\mathbf{I}_n - \mathbf{A}, \ \ \mathbf{B}]. \tag{4.36}$$

In other words, this is the infimum over all complex μ of the smallest singular value of $[\mu\mathbf{I}_n - \mathbf{A}, \ \mathbf{B}]$.

4.2.2 The state observation problem

To be able to modify the dynamical behavior of a system, very often the state $\mathbf{x}$ needs to be available. Typically, however, the state variables are inaccessible and only certain linear combinations $\mathbf{y}$ thereof, given by the output equations (4.12), are known. Thus we need to discuss the problem of reconstructing the state $\mathbf{x}(T)$ from observations $\mathbf{y}(\tau)$, where τ is in some appropriate interval. If $\tau \in [T, T + t]$, we have the *state observation problem*, while if $\tau \in [T - t, T]$, we have the *state reconstruction problem*.

The observation problem is discussed first. Without loss of generality, we assume that $T = 0$. Recall (4.17), (4.18), and (4.19). Since the input $\mathbf{u}$ is known, the latter two terms in (4.19) are also known for $t \geq 0$. Therefore, in determining $\mathbf{x}(0)$ we may assume without loss of generality that $\mathbf{u}(\cdot) = 0$. Thus, the observation problem reduces to the following: given $\mathbf{C}\phi(\mathbf{0}; \mathbf{x}(0); t)$ for $t \geq 0$, find $\mathbf{x}(0)$. Since $\mathbf{B}$ and $\mathbf{D}$ are irrelevant, for this subsection,

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{C} \in \mathbb{R}^{p\times n}.$$

Definition 4.19. *A state $\bar{x} \in \mathbb{X}$ is* unobservable *if* $\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{0}; \bar{\mathbf{x}}; t) = 0$ *for all $t \geq 0$, i.e., if $\bar{\mathbf{x}}$ is indistinguishable from the zero state for all $t \geq 0$. The* unobservable subspace $\mathbb{X}^{\text{unobs}}$ *of $\mathbb{X}$ is the set of all unobservable states of $\mathbf{\Sigma}$. Σ is (completely)* observable *if* $\mathbb{X}^{\text{unobs}} = \{\mathbf{0}\}$. *The* observability matrix *of $\mathbf{\Sigma}$ is*

$$\mathcal{O}(\mathbf{C}, \mathbf{A}) = (\mathbf{C}^* \ \ \mathbf{A}^*\mathbf{C}^* \ \ (\mathbf{A}^*)^2\mathbf{C}^* \ \ \cdots\ )^*. \tag{4.37}$$

Again by the Cayley–Hamilton theorem, the kernel of $\mathcal{O}(\mathbf{C}, \mathbf{A})$ is determined by the first n terms, i.e., $\mathbf{CA}^{i-1}$, $i = 1, \ldots, n$. Therefore, for computational purposes, the finite version

$$\mathcal{O}_n(\mathbf{C}, \mathbf{A}) = (\mathbf{C}^* \ \ \mathbf{A}^*\mathbf{C}^* \ \ \cdots \ \ (\mathbf{A}^*)^{n-1}\mathbf{C}^*)^* \tag{4.38}$$

of the observability matrix is used. We are now ready to state the main theorem.

Theorem 4.20. *Given* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$ *for both $t \in \mathbb{Z}$ and $t \in \mathbb{R}$, $\mathbb{X}^{\text{unobs}}$ is a linear subspace of $\mathbb{X}$ given by*

$$\mathbb{X}^{\text{unobs}} = \ker \mathcal{O}(\mathbf{C}, \mathbf{A}) = \{\mathbf{x} \in \mathbb{X} :\ \mathbf{CA}^{i-1}\mathbf{x} = \mathbf{0},\ i > 0\}. \tag{4.39}$$

An immediate consequence of the above formula is the following corollary.

Corollary 4.21. **(a)** *The unobservable subspace $\mathbb{X}^{\text{unobs}}$ is $\mathbf{A}$-invariant.* **(b)** $\mathbf{\Sigma}$ *is observable if and only if* $\operatorname{rank} \mathcal{O}(\mathbf{C}, \mathbf{A}) = n$. **(c)** *Observability is basis independent.*

Remark 4.2.4. Given $\mathbf{y}(t)$, $t \geq 0$, let $\mathbf{Y}_0$ denote the following $np \times 1$ vector:

$$\begin{aligned} \mathbf{Y}_0 &= (\mathbf{y}^*(0) \ \ D\mathbf{y}^*(0) \ \ \cdots \ \ D^{n-1}\mathbf{y}^*(0))^* && \text{for continuous-time systems,} \\ \mathbf{Y}_0 &= (\mathbf{y}^*(0) \ \ \mathbf{y}^*(1) \ \ \cdots \ \ \mathbf{y}^*(n-1))^* && \text{for discrete-time systems,} \end{aligned}$$

where $D = \frac{d}{dt}$. The observation problem reduces to the solution of the set of linear equations,

$$\mathcal{O}_n(\mathbf{C}, \mathbf{A})\mathbf{x}(0) = \mathbf{Y}_0.$$

This set of equations is solvable for all initial conditions $\mathbf{x}(0)$, i.e., it has a *unique* solution if and only if $\mathbf{\Sigma}$ is observable. Otherwise, $\mathbf{x}(0)$ can be determined only modulo $\mathbb{X}^{\text{unobs}}$, i.e., up to an arbitrary linear combination of unobservable states.

Proof. Next, we give the proof of Theorem 4.20. Let $\mathbf{x}_1, \mathbf{x}_2$ be unobservable states. Then

$$\mathbf{C}\phi(\mathbf{0}; \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2; t) = \alpha_1\mathbf{C}\phi(\mathbf{0}; \mathbf{x}_1; t) + \alpha_2\mathbf{C}\phi(\mathbf{0}; \mathbf{x}_2; t) = \alpha_1\mathbf{y}_1(t) + \alpha_2\mathbf{y}_2(t) = 0$$

for all constants α_1, α_2 and $t \geq 0$. This proves the linearity of the unobservable subspace.

For continuous-time systems, by definition, $\mathbf{x}$ is unobservable if $\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x} = \mathbf{0}$, $t \geq 0$. Since $\mathbf{C}e^{\mathbf{A}t}$ is analytic, it is completely determined by all its derivatives at $t = 0$. This implies (4.39). For discrete-time systems, formula (4.39) follows from the fact that the unobservability of $\mathbf{x}$ is equivalent to $\mathbf{y}(i) = \mathbf{C}\mathbf{A}^i\mathbf{x} = \mathbf{0}$, $i \geq 0$. □

Definition 4.22. *Let* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$. *The finite* observability gramians *at time* $t < \infty$ *are*

$$\mathcal{Q}(t) = \int_0^t e^{\mathbf{A}^*\tau}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}\tau}\,d\tau, \qquad t \in \mathbb{R}_+, \tag{4.40}$$

$$\mathcal{Q}(t) = \mathcal{O}_t^*(\mathbf{C}, \mathbf{A})\mathcal{O}_t(\mathbf{C}, \mathbf{A}), \qquad t \in \mathbb{Z}_+. \tag{4.41}$$

It readily follows that $\ker \mathcal{Q}(t) = \ker \mathcal{O}(\mathbf{C}, \mathbf{A})$. As in the case of reachability, this relationship holds for continuous-time systems for $t > 0$ and for discrete-time systems, at least for $t \geq n$. The *energy* of the output function $\mathbf{y}$ at time T caused by the initial state $\mathbf{x}$ is denoted by $\parallel y \parallel$. In terms of the observability gramian, this energy can be expressed as

$$\parallel \mathbf{y} \parallel^2 = \mathbf{x}^*\mathcal{Q}(T)\mathbf{x}. \tag{4.42}$$

Remark 4.2.5. For completeness, we now briefly turn our attention to the reconstructibility problem. A state $\bar{\mathbf{x}} \in \mathbb{X}$ is *unreconstructible* if $\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{0}; \bar{\mathbf{x}}; t) = \mathbf{0}$ for all $t \leq 0$, i.e., if $\bar{\mathbf{x}}$ is indistinguishable from the zero state for all $t \leq 0$. The *unreconstructible subspace* $\mathbb{X}^{\text{unrecon}}$ of $\mathbb{X}$ is the set of all unreconstructible states of $\mathbf{\Sigma}$. $\mathbf{\Sigma}$ is (completely) *reconstructible* if $\mathbb{X}^{\text{unrec}} = \{\mathbf{0}\}$.

Given is the pair $(\mathbf{C}, \mathbf{A})$. For continuous-time systems $\mathbb{X}^{\text{unrec}} = \mathbb{X}^{\text{unobs}}$. For discrete-time systems, $\mathbb{X}^{\text{unrec}} \supset \mathbb{X}^{\text{unobs}}$, in particular, $\mathbb{X}^{\text{unobs}} = \mathbb{X}^{\text{unrec}} \cap \operatorname{im} \mathbf{A}^n$. This shows that while for continuous-time systems the concepts of observability and reconstructibility are equivalent, for discrete-time systems the latter is weaker. For this reason, only the concept of observability is used here.

4.2.3 The duality principle in linear systems

Let $\mathbf{A}^*$, $\mathbf{B}^*$, $\mathbf{C}^*$, $\mathbf{D}^*$, be the dual maps of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, respectively. The *dual* system $\mathbf{\Sigma}^*$ of $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ is formally defined as

$$\mathbf{\Sigma}^* = \left(\begin{array}{c|c} -\mathbf{A}^* & -\mathbf{C}^* \\ \hline \mathbf{B}^* & \mathbf{D}^* \end{array}\right) \in \mathbb{R}^{(n+m)\times(n+p)},$$

i.e., the input map is given by $-\mathbf{C}^*$, the output map by $\mathbf{B}^*$, and the dynamics by $-\mathbf{A}^*$. The matrix representations of $\mathbf{A}^*$, $\mathbf{C}^*$, $\mathbf{B}^*$, $\mathbf{D}^*$ are the complex conjugate transposes of $\mathbf{A}$, $\mathbf{C}$, $\mathbf{B}$, $\mathbf{D}$, respectively, computed in appropriate dual bases. One may think of the dual system $\mathbf{\Sigma}^*$ as the system $\mathbf{\Sigma}$ but with the role of the inputs and outputs interchanged, or with the flow of causality reversed and time running backward. In section 5.2, it is shown that the dual system is also the adjoint system defined by (5.15), with respect to the standard inner product. The main result is the *duality principle*.

Theorem 4.23. *The orthogonal complement of the reachable subspace of* $\mathbf{\Sigma}$ *is equal to the unobservable subspace of its dual* $\mathbf{\Sigma}^*$*:* $(\mathbb{X}_{\Sigma}^{\text{reach}})^{\perp} = \mathbb{X}_{\Sigma^*}^{\text{unobs}}$. *The system* $\mathbf{\Sigma}$ *is reachable if and only if its dual* $\mathbf{\Sigma}^*$ *is observable.*

Proof. The result follows immediately from formulas (4.25) and (4.37), on recalling that for a linear map $\mathbf{M}$ there holds $(\operatorname{im}\mathbf{M})^{\perp} = \ker \mathbf{M}^*$. □

In a similar way, one can prove that controllability and reconstructibility are dual concepts. Since $(\mathbf{A}, \mathbf{B})$ is reachable if and only if $(\mathbf{B}^*, \mathbf{A}^*)$ is observable, we obtain the following results. Their proof follows by duality from the corresponding results for reachability and is omitted.

Lemma 4.24. Observable canonical decomposition. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}$, $\mathbf{C}$ *have the following matrix representations:*

$$\left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cc|c} \mathbf{A}_{\bar o} & \mathbf{A}_{\bar o o} & \\ \mathbf{0} & \mathbf{A}_o & \\ \hline \mathbf{0} & \mathbf{C}_o & \end{array}\right),$$

where $\mathbf{\Sigma}_o = \left(\begin{array}{c|c} \mathbf{A}_o & \\ \hline \mathbf{C}_o & \end{array}\right)$ *is observable.*

The reachable and observable canonical decompositions given in Lemmas 4.14 and 4.24 can be combined to obtain the following decomposition of the triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$.

Lemma 4.25. Reachable-observable canonical decomposition. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}$*,* $\mathbf{B}$*, and* $\mathbf{C}$ *have the following matrix representations:*

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccc|c} \mathbf{A}_{r\bar o} & \mathbf{A}_{12} & \mathbf{A}_{13} & \mathbf{A}_{14} & \mathbf{B}_{r\bar o} \\ \mathbf{0} & \mathbf{A}_{ro} & \mathbf{0} & \mathbf{A}_{24} & \mathbf{B}_{ro} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{\bar r\bar o} & \mathbf{A}_{34} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{\bar r o} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{C}_{ro} & \mathbf{0} & \mathbf{C}_{\bar r o} & \end{array}\right),$$

where the triple $\mathbf{\Sigma}_{ro} = \left(\begin{array}{c|c} \mathbf{A}_{ro} & \mathbf{B}_{ro} \\ \hline \mathbf{C}_{ro} & \end{array}\right)$ *is both reachable and observable.*

The dual of stabilizability is detectability. $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$ is *detectable* if $\mathbf{A}_{\bar o}$ in the observable canonical decomposition is stable, i.e., has eigenvalues either in the left half of the complex plane or inside the unit disk, depending on whether we are dealing with a continuous- or a discrete-time system.

We conclude this subsection by stating the dual to Theorem 4.15.

Theorem 4.26. Observability conditions. *The following are equivalent:*

1. *The pair* $(\mathbf{C}, \mathbf{A})$*,* $\mathbf{C} \in \mathbb{R}^{p\times n}$*,* $\mathbf{A} \in \mathbb{R}^{n\times n}$*, is observable.*
2. *The rank of the observability matrix is full:* $\operatorname{rank} \mathcal{O}(\mathbf{C}, \mathbf{A}) = n$.
3. *The observability gramian is positive definite:* $\mathcal{Q}(t) > 0$ *for some* $t > 0$.
4. *No right eigenvector* $\mathbf{v}$ *of* $\mathbf{A}$ *is in the right kernel of* $\mathbf{C}$: $\mathbf{Av} = \lambda\mathbf{v} \;\Rightarrow\; \mathbf{Cv} \neq \mathbf{0}$.
5. $\operatorname{rank}\,(\mu\mathbf{I}_n - \mathbf{A}^*, \quad \mathbf{C}^*) = n$ *for all* $\mu \in \mathbb{C}$.
6. *The polynomial matrices* $s\mathbf{I} - \mathbf{A}$ *and* $\mathbf{C}$ *are right coprime.*

Again, by section 3.3.3, the numerical determination of the observability of a given pair $(\mathbf{C}, \mathbf{A})$ is an ill-posed problem. Therefore, as in the reachability case the *distance* of the pair to the set of unobservables, denoted by $\delta_o(\mathbf{C}, \mathbf{A})$, will be used instead. This distance is defined by means of the distance to reachability of the dual pair $(\mathbf{A}^*, \mathbf{C}^*)$, which from (4.36) is

$$\delta_o(\mathbf{C}, \mathbf{A}) = \delta_r(\mathbf{A}^*, \mathbf{C}^*) = \inf_{\mu\in\mathbb{C}} \sigma_{\min}[\mu\mathbf{I}_n - \mathbf{A}^*, \quad \mathbf{C}^*] = \inf_{\mu\in\mathbb{C}} \sigma_{\min}\begin{bmatrix} \mu\mathbf{I}_n - \mathbf{A} \\ \mathbf{C} \end{bmatrix}.$$

4.3 The infinite gramians

Consider a continuous-time linear system $\mathbf{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ that is *stable*, i.e., all eigenvalues of $\mathbf{A}$ have *negative real part*. In this case, both (4.28) and (4.40) are defined for $t = \infty$:

$$\boxed{\mathcal{P} = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{BB}^* e^{\mathbf{A}^*\tau}\, d\tau,} \tag{4.43}$$

$$\boxed{\mathcal{Q} = \int_0^\infty e^{\mathbf{A}^*\tau}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}\tau}\, d\tau.} \tag{4.44}$$

$\mathcal{P}$ and $\mathcal{Q}$ are the *infinite reachability* and *infinite observability gramians* associated with $\mathbf{\Sigma}_c$. These gramians satisfy the following linear matrix equations, called Lyapunov equations.

Proposition 4.27. *Given the stable, continuous-time system* $\mathbf{\Sigma}_c$ *as above, the associated infinite reachability gramian* $\mathcal{P}$ *satisfies the continuous-time Lyapunov equation*

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{BB}^* = \mathbf{0}, \tag{4.45}$$

while the associated infinite observability gramian satisfies

$$\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}. \tag{4.46}$$

Proof. It is readily checked that due to stability,

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = \int_0^\infty \left[\mathbf{A}e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau} + e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau}\mathbf{A}^*\right] d\tau$$
$$= \int_0^\infty d(e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau}) = -\mathbf{B}\mathbf{B}^*.$$

This proves (4.45); (4.46) is proved similarly. □

The matrices $\mathcal{P}$ and $\mathcal{Q}$ are indeed gramians in the following sense. Recall that the impulse response of a continuous-time system $\mathbf{\Sigma}_c$ is $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$, $t > 0$. Now, consider the following two maps:

input-to-state map $\xi(t) = e^{\mathbf{A}t}\mathbf{B}$ and

state-to-output map $\eta(t) = \mathbf{C}e^{\mathbf{A}t}$.

If the input to the system is the impulse $\delta(t)$, the resulting state is $\xi(t)$; moreover, if the initial condition of the system is $\mathbf{x}(0)$, in the absence of a forcing function $\mathbf{u}$, the resulting output is $\mathbf{y}(t) = \eta(t)\mathbf{x}(0)$. The *gramians* corresponding to $\xi(t)$ and $\eta(t)$ for time running from 0 to T are

$$\mathcal{P} = \int_0^T \xi(t)\xi(t)^*\,dt = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*t}\,dt$$

and

$$\mathcal{Q} = \int_0^\infty \eta(t)^*\eta(t)\,dt = \int_0^T e^{\mathbf{A}^*t}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}t}\,dt.$$

These are the expressions that we have encountered earlier as (finite) gramians.

Similarly, if the discrete-time system $\mathbf{\Sigma}_d = \left(\begin{array}{c|c}\mathbf{A} & \mathbf{B}\\ \hline \mathbf{C} & \mathbf{D}\end{array}\right)$ is *stable*, i.e., all eigenvalues of $\mathbf{A}$ are inside the unit disk, the gramians (4.29) as well as (4.41) are defined for $t = \infty$:

$$\mathcal{P} = \mathcal{R}\mathcal{R}^* = \sum_{k=0}^{\infty}\mathbf{A}^k\mathbf{B}\mathbf{B}^*(\mathbf{A}^*)^k, \tag{4.47}$$

$$\mathcal{Q} = \mathcal{O}^*\mathcal{O} = \sum_{k=0}^{\infty}(\mathbf{A}^*)^k\mathbf{C}^*\mathbf{C}\mathbf{A}^k. \tag{4.48}$$

Notice that $\mathcal{P}$ can be written as $\mathcal{P} = \mathbf{B}\mathbf{B}^* + \mathbf{A}\mathcal{P}\mathbf{A}^*$; moreover, $\mathcal{Q} = \mathbf{C}^*\mathbf{C} + \mathbf{A}^*\mathcal{Q}\mathbf{A}$. These are the so-called discrete-time Lyapunov or Stein equations.

Proposition 4.28. *Given the stable, discrete-time system $\mathbf{\Sigma}_d$ as above, the associated infinite reachability gramian $\mathcal{P}$ satisfies the discrete-time Lyapunov equation*

$$\mathbf{A}\mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathcal{P}, \tag{4.49}$$

while the associated infinite observability gramian $\mathcal{Q}$ satisfies

$$\mathbf{A}^*\mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathcal{Q}. \tag{4.50}$$

The infinite gramians in the frequency domain

The infinite gramians can also be expressed in the frequency domain. In particular, applying Plancherel's theorem[2] to (4.43) we obtain

$$\boxed{\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\,d\omega;} \tag{4.51}$$

similarly, (4.44) yields

$$\boxed{\mathcal{Q} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(i\omega\mathbf{I} - \mathbf{A})^{-1}\,d\omega.} \tag{4.52}$$

In the discrete-time case the infinite gramians defined by (4.47) and (4.48) can be expressed as

$$\mathcal{P} = \frac{1}{2\pi}\int_{0}^{2\pi} (e^{i\theta}\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(e^{-i\theta}\mathbf{I} - \mathbf{A}^*)^{-1}\,d\theta, \tag{4.53}$$

$$\mathcal{Q} = \frac{1}{2\pi}\int_{0}^{2\pi} (e^{-i\theta}\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(e^{i\theta}\mathbf{I} - \mathbf{A})^{-1}\,d\theta. \tag{4.54}$$

These expressions will be useful in model reduction methods involving *frequency weighting* (section 7.6).

4.3.1 The energy associated with reaching/observing a state

An important consideration in model reduction is the ability to classify states according to their *degree of reachability* or their *degree of observability*. Recall (4.33), (4.34), and (4.42), valid for both discrete- and continuous-time systems. From the definition of the gramians, it follows that

$$\mathcal{P}(t_2) \geq \mathcal{P}(t_1),\ \ \mathcal{Q}(t_2) \geq \mathcal{Q}(t_1), \qquad t_2 \geq t_1,$$

irrespective of whether we are dealing with discrete- or continuous-time systems. Hence from (4.34) it follows that the minimal energy for the transfer from state $\mathbf{0}$ to $\mathbf{x}_r$ is obtained as $\bar{T} \to \infty$; hence, assuming stability and (complete) reachability, the gramian is positive definite, and this minimal energy is

$$\mathbf{x}_r^*\mathcal{P}^{-1}\mathbf{x}_r. \tag{4.55}$$

Similarly, the largest observation energy produced by the state $\mathbf{x}_o$ is also obtained for an infinite observation interval and is equal to

$$\mathbf{x}_o^*\mathcal{Q}\mathbf{x}_o. \tag{4.56}$$

[2] In the theory of Fourier transform, Plancherel's theorem states that the inner product of two (matrix-valued) functions in the time domain and in the frequency domain is (up to a constant) the same. In continuous time we have $2\pi\int_{-\infty}^{\infty}\mathbf{g}^*(t)\mathbf{f}(t)\,dt = \int_{-\infty}^{\infty}\mathbf{F}^*(-i\omega)\mathbf{G}(i\omega)\,d\omega$, while in discrete-time there holds $2\pi\sum_{-\infty}^{\infty}\mathbf{g}^*(t)\mathbf{f}(t) = \int_0^{2\pi}\mathbf{F}^*(e^{-i\theta})\mathbf{G}(e^{i\theta})\,d\theta$.

We summarize these results as follows.

Lemma 4.29. *Let $\mathcal{P}$ and $\mathcal{Q}$ denote the infinite gramians of a stable linear system $\boldsymbol{\Sigma}$.*

(a) *The minimal energy required to steer the state of the system from* $\mathbf{0}$ *to* $\mathbf{x}_r$ *is given by* (4.55).

(b) *The maximal energy produced by observing the output of the system whose initial state is* $\mathbf{x}_o$ *is given by* (4.56).

This lemma provides a way to determine the *degree of reachability* or the *degree of observability* of the states of Σ. The states that are the most difficult, i.e., require the most energy to reach, are (have a significant component) in the span of those eigenvectors of $\mathcal{P}$ which correspond to small eigenvalues. Furthermore, the states that are difficult to observe, i.e., produce small observation energy, are (have a significant component) in the span of those eigenvectors of $\mathcal{Q}$ which correspond to small eigenvalues.

The above conclusion is at the heart of the concept of balancing, discussed in Chapter 7. Recall the definition of equivalent systems (4.24). Under equivalence, the gramians are transformed as follows:

$$\widetilde{\mathcal{P}} = \mathbf{T}\mathcal{P}\mathbf{T}^*,\ \ \widetilde{\mathcal{Q}} = \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} \ \ \Rightarrow\ \ \widetilde{\mathcal{P}}\widetilde{\mathcal{Q}} = \mathbf{T}\,(\mathcal{P}\mathcal{Q})\,\mathbf{T}^{-1}. \tag{4.57}$$

Therefore, the product of the two gramians of equivalent systems is related by similarity transformation, and hence has the same eigenvalues. Quantities that are invariant under state-space transformation are called *input-output invariants* of the associated system Σ.

Proposition 4.30. *The eigenvalues of the product of the reachability and of the observability gramians are input-output invariants.*

Remark 4.3.1. As discussed in Lemma 5.8 and (5.24), the eigenvalues of $\mathcal{P}\mathcal{Q}$ are important invariants called *Hankel singular values* of the system. They turn out to be equal to the singular values of the *Hankel operator* introduced in section 5.1.

Remark 4.3.2. *A formula for the reachability gramian.* Given a continuous-time system described by the pair $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, the reachability gramian is defined by (4.28). If the eigenvalues of $\mathbf{A}$ are assumed to be distinct, $\mathbf{A}$ is diagonalizable. Let the EVD be

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1},\ \ \text{where}\ \ \mathbf{V} = [\mathbf{v}_1\ \ \mathbf{v}_2\ \ \cdots\ \ \mathbf{v}_n],\ \ \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n);$$

$\mathbf{v}_i$ denotes the eigenvector corresponding to the eigenvalue λ_i. Notice that if the ith eigenvalue is complex, the corresponding eigenvector is also complex. Let $\mathbf{W} = \mathbf{V}^{-1}\mathbf{B} \in \mathbb{C}^{n\times m}$, and denote by $\mathbf{W}_i \in \mathbb{C}^{1\times m}$ the ith row of $\mathbf{W}$. With the notation introduced above, the following formula holds:

$$\mathcal{P}(T) = \mathbf{V}\mathcal{R}(T)\mathbf{V}^*,\ \ \text{where}\ \ [\mathcal{R}(T)]_{ij} = \frac{-\mathbf{W}_i\mathbf{W}_j^*}{\lambda_i + \lambda_j^*}\left(1 - \exp\left[(\lambda_i + \lambda_j^*)T\right]\right) \in \mathbb{C}.$$

Furthermore, if $\lambda_i + \lambda_j^* = 0$, $[\mathcal{R}(T)]_{ij} = (\mathbf{W}_i\mathbf{W}_j^*)\,T$. If in addition $\mathbf{A}$ is stable, the infinite gramian (4.43) is given by $\mathcal{P} = \mathbf{V}\mathcal{R}\mathbf{V}^*$, where $\mathcal{R}_{ij} = \frac{-\mathbf{W}_i\mathbf{W}_j^*}{\lambda_i+\lambda_j^*}$. This formula accomplishes both the computation of the exponential and the integration implicitly, in terms of the EVD of $\mathbf{A}$.

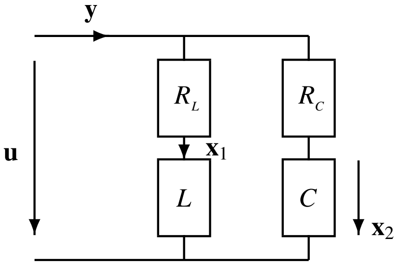

Figure 4.2. *Parallel connection of a capacitor and an inductor.*

Example 4.31. Consider the electric circuit consisting of the parallel connection of two branches, as shown in Figure 4.2. The input is the voltage $\mathbf{u}$ applied to this parallel connection, while the output is the current $\mathbf{y}$; we choose as states the current through the inductor $\mathbf{x}_1$, and the voltage across the capacitor $\mathbf{x}_2$.

The state equations are $L\dot{\mathbf{x}}_1 = -R_L\mathbf{x}_1 + \mathbf{u}$, $CR_C\dot{\mathbf{x}}_2 = -\mathbf{x}_2 + \mathbf{u}$, while the output equation is $R_C\mathbf{y} = R_C\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{u}$. Thus

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cc|c} -\tau_L & 0 & \frac{1}{R_L}\tau_L \\ 0 & -\tau_C & \tau_C \\ \hline 1 & -\frac{1}{R_C} & \frac{1}{R_C} \end{array}\right], \tag{4.58}$$

$$\text{where } \tau_L = \frac{R_L}{L},\ \tau_C = \frac{1}{R_C C},$$

are the time constants of the two branches of the circuit. Therefore, the impulse response is

$$\mathbf{h}(t) = \frac{\tau_L}{R_L}e^{-\tau_L t} - \frac{\tau_C}{R_C}e^{-\tau_C t} + \frac{1}{R_C}\delta(t), \qquad t \geq 0.$$

It readily follows that this system is reachable and observable if the two time constants are different ($\tau_L \neq \tau_C$).

Assuming that the values of these elements are $L = 1$, $R_L = 1$, $C = 1$, $R_C = \frac{1}{2}$:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \Rightarrow e^{\mathbf{A}\bar{t}}\mathbf{B} = \begin{bmatrix} e^{-\bar{t}} \\ 2e^{-2\bar{t}} \end{bmatrix}.$$

Reachability in this case inquires about the existence of an input voltage $\mathbf{u}$ which will steer the state of the system to some desired $\tilde{\mathbf{x}}$, at a given time $T > 0$. In this case, since the system is reachable (for positive values of the parameters), any state can be reached. We choose $\mathbf{x}^1 = [1\ \ 0]^*$, $\mathbf{x}^2 = [0\ \ 1]^*$. The gramian $\mathcal{P}(T)$ and the infinite gramian $\mathcal{P}$ are

$$\mathcal{P}(T) = \begin{bmatrix} -\frac{1}{2}e^{-2T} + \frac{1}{2} & -\frac{2}{3}e^{-3T} + \frac{2}{3} \\ -\frac{2}{3}e^{-3T} + \frac{2}{3} & -e^{-4T} + 1 \end{bmatrix},\ \mathcal{P} = \lim_{T\to\infty}\mathcal{P}(T) = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{bmatrix}.$$

The corresponding inputs valid for $\bar{t}$ between 0 and T are

$$\mathbf{u}_T^1(\bar{t}) = -\frac{6e^{-\bar{t}}\left(3e^{-4T} - 3 - 4e^{-\bar{t}-3T} + 4e^{-\bar{t}}\right)}{e^{-6T} - 9e^{-2T} - 9e^{-4T} + 1 + 16e^{-3T}},$$

$$\Rightarrow\ \mathbf{u}_\infty^1(\bar{t}) = \lim_{T\to\infty} u_{1,T}(t) = 18e^{-\bar{t}} - 24e^{-2\bar{t}}.$$

$$\mathbf{u}_T^2(\bar{t}) = \frac{6e^{-\bar{t}}\left(2e^{-3T} - 2 - 3e^{-\bar{t}-2T} + 3e^{-\bar{t}}\right)}{e^{-6T} - 9e^{-2T} - 9e^{-4T} + 1 + 16e^{-3T}},$$

$$\Rightarrow\ \mathbf{u}_\infty^2(\bar{t}) = \lim_{T\to\infty} u_{2,T}(t) = -12e^{-\bar{t}} + 18e^{-2\bar{t}}.$$

In the above expressions, $\bar{t} = T - t$, where t is the time; in the upper plot of Figure 4.3, the time axis is $\bar{t}$, i.e., time runs backward from T to 0; in the lower plot, the time axis is t, running from 0 to T. Both plots show the minimum energy inputs required to steer the system to $\mathbf{x}_1$ for $T = 1,\ 2,\ 10$ units of time. Notice that for $T = 10$ the input function is zero for most of the interval, starting with $t = 0$; consequently, for $T \to \infty$, the activity occurs close to $T = \infty$ and the input function can thus be plotted only in the $\bar{t}$ axis. If the system is stable, i.e., $\mathcal{R}e(\lambda_i(\mathbf{A})) < 0$, the reachability gramian is defined for $T = \infty$, and it satisfies (4.45). Hence, the infinite gramian can be computed as the solution to this linear matrix equation; explicit calculation of the matrix exponentials, multiplication, and subsequent integration are not required. In MATLAB, if in addition the pair $(\mathbf{A}, \mathbf{B})$ is reachable, we have

$$\texttt{P} = \texttt{lyap}\,(\texttt{A}, \texttt{B} * \texttt{B}').$$

For the matrices defined earlier, using the `lyap` command in the format `short e`, we get

$$\mathcal{P} = \begin{bmatrix} 0.50000 & 0.66666 \\ 0.66666 & 1.00000 \end{bmatrix}.$$

We conclude this example with the computation of the reachability gramian in the frequency domain:

$$\begin{aligned}
\mathcal{P} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I}_2 - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I}_2 - \mathbf{A}^*)^{-1}\,d\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \begin{bmatrix} \frac{1}{\omega^2+1} & \frac{1}{\omega^2+i\omega+2} \\ \frac{1}{\omega^2-i\omega+2} & \frac{4}{\omega^2+4} \end{bmatrix} d\omega \\
&= \frac{1}{2\pi}\begin{bmatrix} 2\arctan\omega & \frac{4}{3}\arctan\frac{\omega}{2} + \frac{4}{3}\arctan\omega \\ \frac{4}{3}\arctan\frac{\omega}{2} + \frac{4}{3}\arctan\omega & 4\arctan\frac{\omega}{2} \end{bmatrix}_{\omega=\infty} \\
&= \frac{1}{2\pi}\begin{bmatrix} \pi & \frac{4\pi}{3} \\ \frac{4\pi}{3} & 2\pi \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{bmatrix}
\end{aligned}$$

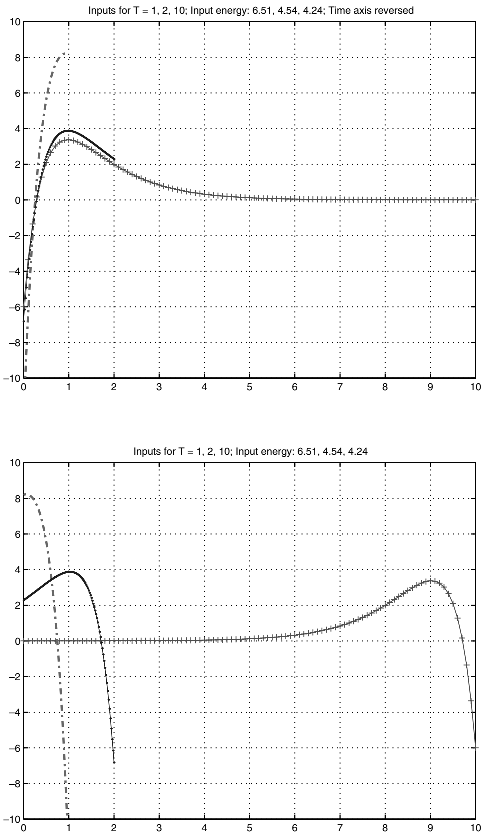

Figure 4.3. *Electric circuit example. Minimum energy inputs steering the system to* $\mathbf{x}_1 = [1\ \ 0]^*$ *for* $T = 1, 2, 10$. *Top plot: time axis running backward; bottom plot: time axis running forward.*

Example 4.32. A second simple example is the following:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix},\ \mathbf{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ \Rightarrow\ e^{\mathbf{A}t} = \begin{pmatrix} -e^{-2t} + 2e^{-t} & e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} & 2e^{-2t} - e^{-t} \end{pmatrix}.$$

This implies

$$\mathcal{P}(T) = \begin{pmatrix} -\frac{1}{2}e^{-2T} + \frac{2}{3}e^{-3T} - \frac{1}{4}e^{-4T} + \frac{1}{12} & -e^{-3T} + \frac{1}{2}e^{-2T} + \frac{1}{2}e^{-4T} \\ -e^{-3T} + \frac{1}{2}e^{-2T} + \frac{1}{2}e^{-4T} & -e^{-4T} + \frac{4}{3}e^{-3T} - \frac{1}{2}e^{-2T} + \frac{1}{6} \end{pmatrix}.$$

And finally, the infinite gramian is

$$\mathtt{P} = \mathtt{lyap}(\mathtt{A}, \mathtt{B} * \mathtt{B}') = \begin{pmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{6} \end{pmatrix}.$$

In the frequency domain

$$\begin{aligned}
\mathcal{P} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega\mathbf{I}_2-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I}_2-\mathbf{A}^*)^{-1}\,d\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\begin{array}{cc} \frac{1}{\omega^4+5\omega^2+4} & \frac{-i\omega}{\omega^4+5\omega^2+4} \\ \frac{i\omega}{\omega^4+5\omega^2+4} & \frac{\omega^2}{\omega^4+5\omega^2+4} \end{array}\right] d\omega \\
&= \frac{1}{2\pi}\left[\begin{array}{cc} \frac{2}{3}\arctan\omega-\frac{1}{3}\arctan\frac{\omega}{2} & 0 \\ 0 & \frac{4}{3}\arctan\frac{\omega}{2}-\frac{2}{3}\arctan\omega \end{array}\right]_{\omega=\infty} \\
&= \frac{1}{2\pi}\left[\begin{array}{cc} \frac{\pi}{6} & 0 \\ 0 & \frac{\pi}{3} \end{array}\right] = \left[\begin{array}{cc} \frac{1}{12} & 0 \\ 0 & \frac{1}{6} \end{array}\right].
\end{aligned}$$

Example 4.33. We will now compute the gramian of a simple discrete-time, second-order system,

$$\mathbf{A}=\begin{pmatrix} 0 & 1 \\ 0 & \frac{1}{2} \end{pmatrix},\ \mathbf{B}=\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

To compute the reachability gramian in the time domain, we make use of (4.47):

$$\mathcal{P}=\mathcal{R}\mathcal{R}^*=\sum_{k=0}^{\infty}\mathbf{A}^k\mathbf{B}\mathbf{B}^*(\mathbf{A}^*)^k=\left[\begin{array}{cc} 1+\frac{1}{4}+\frac{1}{16}+\cdots & \frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\cdots \\ \frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\cdots & 1+\frac{1}{4}+\frac{1}{16}+\cdots \end{array}\right]=\frac{2}{3}\left[\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array}\right].$$

For the frequency domain computation, we make use of formula (4.53):

$$\begin{aligned}
\mathcal{P} &= \frac{1}{2\pi}\int_0^{2\pi}\left[\begin{array}{cc} \frac{4e^{-i\theta}}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} & \frac{4}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} \\ \frac{4}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} & \frac{4e^{-i\theta}}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} \end{array}\right] d\theta \\
&= \frac{1}{2\pi}\left[\begin{array}{cc} \frac{4i}{3}\left[\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] & \frac{2i}{3}\left[4\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] \\ \frac{2i}{3}\left[4\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] & \frac{4i}{3}\left[\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] \end{array}\right]_{\theta=0}^{\theta=2\pi} \\
&= \frac{2}{3}\left[\begin{array}{cc} 2 & 1 \\ 1 & 2 \end{array}\right].
\end{aligned}$$

4.3.2 The cross gramian

In addition to the reachability and observability gramians, a third one, the *cross gramian*, is used. This concept is first defined for *discrete-time* systems $\mathbf{\Sigma}=\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$. Given the (infinite) reachability matrix $\mathcal{R}(\mathbf{A},\mathbf{B})$ (4.25) and the observability matrix $\mathcal{O}(\mathbf{C},\mathbf{A})$ (4.37), the *cross gramian* is the $n\times n$ matrix defined by $\mathcal{X}=\mathcal{R}\mathcal{O}$. Thus, summarizing, the three (infinite) gramians of Σ are

$$\mathcal{P}=\mathcal{R}\mathcal{R}^*,\ \mathcal{Q}=\mathcal{O}^*\mathcal{O},\ \mathcal{X}=\mathcal{R}\mathcal{O}\in\mathbb{R}^{n\times n}.$$

Notice that these gramians are the three finite matrices that can be formed from the reachability matrix (which has infinitely many columns) and the observability matrix (which has infinitely many rows).

These first two gramians satisfy the Stein equations (4.49) and (4.50). The cross gramian satisfies a Stein equation as well, but its form depends on the number of inputs and outputs m, p of $\mathbf{\Sigma}$. If $m = p$, a moment's reflection shows that $\mathcal{X}$ satisfies the following Sylvester equation:

$$\mathbf{A}\mathcal{X}\mathbf{A} + \mathbf{B}\mathbf{C} = \mathcal{X}.$$

If $m = 2p$, let $\mathbf{B} = [\mathbf{B}_1\ \ \mathbf{B}_2]$, $\mathbf{B}_i \in \mathbb{R}^{n\times p}$, $i = 1, 2$; it can be verified that

$$\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2, \quad \text{where } \mathcal{X}_1 = \mathcal{R}(\mathbf{B}_1, \mathbf{A})\mathcal{O}(\mathbf{C}, \mathbf{A}^2) \text{ and } \mathcal{X}_2 = \mathcal{R}(\mathbf{B}_2, \mathbf{A})\mathcal{O}(\mathbf{C}\mathbf{A}, \mathbf{A}^2).$$

Combining these expressions we obtain the Stein equation satisfied by $\mathcal{X}$:

$$\mathbf{A}\mathcal{X}\mathbf{A}^2 + \mathbf{B}\begin{pmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \end{pmatrix} = \mathcal{X}.$$

For general m and p, the Stein equation involves $\mathbf{A}^r$, where r is the least common multiple of m, p.

As in the discrete-time case, if the number of inputs of the stable continuous-time system $\mathbf{\Sigma}$ is equal to the number of outputs $m = p$, the **cross gramian** $\mathcal{X}$ is defined as the solution to the Sylvester equation

$$\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{B}\mathbf{C} = \mathbf{0}. \tag{4.59}$$

Similarly to (4.43) and (4.44) in Proposition 4.27, it can readily be shown that $\mathcal{X}$ can be expressed as

$$\mathcal{X} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{C}e^{\mathbf{A}t}\,dt. \tag{4.60}$$

All three gramians are related to the eigenvalues and singular values of the *Hankel operator*, which will be introduced later. Under a state-space transformation $\mathbf{T}$, the three gramians are transformed to $\mathbf{T}\mathcal{P}\mathbf{T}^*$, $\mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1}$, $\mathbf{T}\mathcal{X}\mathbf{T}^{-1}$, respectively. Therefore, while the eigenvalues of the reachability and observability gramians are *not* input-output invariants, both the product $\mathcal{P}\mathcal{Q}$ and $\mathcal{X}$ are transformed by similarity. Their eigenvalues are input-output invariants for the associated $\mathbf{\Sigma}$, both for discrete- and continuous-time systems. As will be shown in section 5.4, the eigenvalues and the singular values of the *Hankel operator* $\mathcal{H}$ associated with Σ are given by these eigenvalues, namely,

$$\boxed{\lambda_i(\mathcal{H}) = \lambda_i(\mathcal{X}), \quad \sigma_i(\mathcal{H}) = \sqrt{\lambda_i(\mathcal{P}\mathcal{Q})}\,.}$$

The cross gramian for SISO systems was introduced in [113].

Example 4.34. Consider the circuit shown in Figure 4.2. The system matrices are given by (4.58). Thus the reachability, observability, and cross gramians are:

$\mathcal{P} =$	$\mathcal{Q} =$	$\mathcal{X} =$
$\tau_L\tau_C\begin{bmatrix} \frac{1}{R_L^2}\frac{1}{2\tau_C} & \frac{1}{R_L}\frac{1}{\tau_L+\tau_C} \\ \frac{1}{R_L}\frac{1}{\tau_L+\tau_C} & \frac{1}{2\tau_L} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2\tau_L} & -\frac{1}{R_C}\frac{1}{\tau_L+\tau_C} \\ -\frac{1}{R_C}\frac{1}{\tau_L+\tau_C} & \frac{1}{R_C^2}\frac{1}{2\tau_C} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2R_L} & -\frac{1}{R_L R_C}\frac{\tau_L}{\tau_L+\tau_C} \\ \frac{\tau_C}{\tau_L+\tau_C} & -\frac{1}{2R_C} \end{bmatrix}$

.

Notice that $\mathcal{P}$ and $\mathcal{Q}$ become semidefinite if the two time constants are equal, $\tau_L = \tau_R$; if in addition $R_L = R_C$, the product $\mathcal{PQ} = \mathbf{0}$, which is reflected in the fact that $\mathcal{X}$ in this case has two zero eigenvalues.

4.3.3 A transformation between continuous- and discrete-time systems

Often it is advantageous to transform a given problem in a way that its solution becomes easier, either theoretically or computationally. A transformation that is of interest in the present context is the *bilinear transformation*. We will mention here some cases in which this transformation is important.

The theory of optimal approximation in the Hankel-norm discussed in Chapter 8 is easier to formulate for discrete-time systems, while it is easier to solve for continuous-time systems. Thus given a discrete-time system, the bilinear transformation is used to obtain the solution in continuous time and then transform back. (See Example 8.9.) Second, as stated in the next proposition, the gramians remain invariant under the bilinear transformation. In section 12.2, this fact is used to iteratively solve a continuous-time Lyapunov equation in discrete time, that is, by solving the corresponding Stein equation.

The bilinear transformation is defined by $z = \frac{1+s}{1-s}$ and maps the open left half of the complex plane onto the inside of the unit disc and the imaginary axis onto the unit circle. In particular, the transfer function $\mathbf{H}_c(s)$ of a continuous-time system is obtained from that of the discrete-time transfer function $\mathbf{H}_d(z)$ as follows:

$$\mathbf{H}_c(s) = \mathbf{H}_d\left(\frac{1+s}{1-s}\right).$$

Consequently, the matrices

$$\boldsymbol{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \quad \boldsymbol{\Sigma}_d = \left(\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \mathbf{J} \end{array}\right)$$

of these two systems are related as given in the following table:

Continuous time		**Discrete time**
$\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ \mathbf{D}$	$z = \frac{1+s}{1-s}$	$\begin{cases} \mathbf{F} = (\mathbf{I}+\mathbf{A})(\mathbf{I}-\mathbf{A})^{-1} \\ \mathbf{G} = \sqrt{2}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} \\ \mathbf{H} = \sqrt{2}\mathbf{C}(\mathbf{I}-\mathbf{A})^{-1} \\ \mathbf{J} = \mathbf{D}+\mathbf{C}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} \end{cases}$
$\left.\begin{array}{l} \mathbf{A} = (\mathbf{F}+\mathbf{I})^{-1}(\mathbf{F}-\mathbf{I}) \\ \mathbf{B} = \sqrt{2}(\mathbf{F}+\mathbf{I})^{-1}\mathbf{G} \\ \mathbf{C} = \sqrt{2}\mathbf{H}(\mathbf{F}+\mathbf{I})^{-1} \\ \mathbf{D} = \mathbf{J}-\mathbf{H}(\mathbf{F}+\mathbf{I})^{-1}\mathbf{G} \end{array}\right\}$	$s = \frac{z-1}{z+1}$	$\mathbf{F},\ \mathbf{G},\ \mathbf{H},\ \mathbf{J}$

.

Proposition 4.35. *Given the stable continuous-time system* $\mathbf{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ *with infinite gramians* $\mathcal{P}_c$, $\mathcal{Q}_c$, *let* $\mathbf{\Sigma}_d = \left(\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \mathbf{J} \end{array}\right)$, *with infinite gramians* $\mathcal{P}_d$, $\mathcal{Q}_d$, *be the stable discrete-time system obtained by means of the bilinear transformation given above. It follows that this bilinear transformation preserves the gramians:*

$$\mathcal{P}_c = \mathcal{P}_d \text{ and } \mathcal{Q}_c = \mathcal{Q}_d.$$

Consequently, the Hankel-norm of Σ_c *and* Σ_d, *defined by* (5.7), *is the same. Furthermore, the transformation preserves the infinity norms as defined by* (5.8) *and* (5.9).

The above result implies that the bilinear transformation between discrete- and continuous-time systems *preserves* balancing; this concept is discussed in section 7.

4.4 The realization problem

In the preceding sections, we presented two ways of describing linear systems: the internal and the external. The former makes use of the inputs **u**, states **x**, and outputs **y**. The latter makes use *only* of the inputs **u** and the outputs **y**. The question thus arises as to the relationship between these two descriptions.

In one direction, this problem is trivial. Given the internal description $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ of a system, the external description is readily derived. The transfer function of the system is given by (4.22)

$$\mathbf{H}_{\mathbf{\Sigma}}(\xi) = \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B},$$

while from (4.23), the Markov parameters are given by

$$\mathbf{h}_0 = \mathbf{D},\ \mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B} \in \mathbb{R}^{p\times m}, \qquad k = 1, 2, \ldots. \tag{4.61}$$

The converse problem, i.e., given the external description, derive the internal one, is far from trivial. This is the realization problem: given the external description of a linear system, construct an internal or state variable description. In other words, given the impulse response **h** or, equivalently, the transfer function **H**, or the Markov parameters $\mathbf{h}_k$ of a system, construct $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ such that (4.61) holds. It readily follows without computation that $\mathbf{D} = \mathbf{h}_0$. Hence the following problem results.

Definition 4.36. *Given the sequence of* $p \times m$ *matrices* $\mathbf{h}_k$, $k > 0$, *the* realization problem *consists of finding a positive integer* n *and constant matrices* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *such that*

$$\mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}, \qquad \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ \ k = 1, 2, \ldots. \tag{4.62}$$

The triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is then called a* realization *of the sequence* $\mathbf{h}_k$, *and the latter is called a* realizable *sequence.* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is a* minimal *realization if among all realizations of the sequence, its dimension is the smallest possible.*

The realization problem is sometimes referred to as the problem of construction of state for linear systems described by convolution relationships.

Remark 4.4.1. *Realization* was formally introduced in the 1960s (see Kalman, Falb, and Arbib [192]), and eventually two approaches crystallized: the state-space and the polynomial. (See Fuhrmann [121] for an overview of the interplay between these two approaches in linear system theory.) The state-space method uses the Hankel matrix as a main tool and will be presented next. The polynomial approach has the Euclidean division algorithm as a focal point; see, e.g., Kalman [193], Fuhrmann [123], Antoulas [9], and van Barel and Bultheel [329]. Actually, Antoulas [9] presents the complete theory of recursive realization for multi-input, multi-output systems.

Example 4.37. Consider the following (scalar) sequences:

$$\boldsymbol{\Sigma}_1 = \{1, 1, 1, 1, 1, 1, 1, 1, 1, \ldots\},$$

$$\boldsymbol{\Sigma}_2 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, \ldots\} \text{ natural numbers},$$

$$\boldsymbol{\Sigma}_3 = \{1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots\} \text{ Fibonacci numbers},$$

$$\boldsymbol{\Sigma}_4 = \{1, 2, 3, 5, 7, 11, 13, 17, 19, \ldots\} \text{ primes},$$

$$\boldsymbol{\Sigma}_5 = \left\{\frac{1}{1!}, \frac{1}{2!}, \frac{1}{3!}, \frac{1}{4!}, \frac{1}{5!}, \frac{1}{6!}, \ldots\right\} \text{ inverse factorials}.$$

It is assumed that for all sequences, $\mathbf{h}_0 = \mathbf{D} = 0$. Which sequences are realizable? This question will be answered in the example of section 4.43.

Problems. The following problems arise:

(a) Existence: given a sequence $\mathbf{h}_k$, $k > 0$, determine whether there exist a positive integer n and a triple of matrices $\mathbf{C}$, $\mathbf{A}$, $\mathbf{B}$ such that (4.62) holds.

(b) Uniqueness: in case such an integer and triple exist, are they unique in some sense?

(c) Construction: in case of existence, find n and give an algorithm to construct such a triple.

The main tool for answering the above questions is the matrix $\mathcal{H}$ of Markov parameters:

$$\mathcal{H} = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_k & \mathbf{h}_{k+1} & \cdots \\ \mathbf{h}_2 & \mathbf{h}_3 & \cdots & \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \cdots \\ \vdots & \vdots & \iddots & \vdots & \vdots & \iddots \\ \mathbf{h}_k & \mathbf{h}_{k+1} & \cdots & \mathbf{h}_{2k-1} & \mathbf{h}_{2k} & \cdots \\ \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \cdots & \mathbf{h}_{2k} & \mathbf{h}_{2k+1} & \cdots \\ \vdots & \vdots & \iddots & \vdots & \vdots & \end{pmatrix}. \tag{4.63}$$

This is the Hankel matrix; it has infinitely many rows, infinitely many columns, and block Hankel structure, i.e., $(\mathcal{H})_{i,j} = \mathbf{h}_{i+j-1}$, for $i, j > 0$. We start by listing conditions related to the realization problem.

Lemma 4.38. *The following statements are equivalent:*

(a) *The sequence* $\mathbf{h}_k$, $k > 0$, *is* realizable.

(b) *The formal power series* $\sum_{k>0} \mathbf{h}_k s^{-k}$ *is* rational.

(c) *The sequence* $\mathbf{h}_k$, $k > 0$, *satisfies a* recursion *with constant coefficients, i.e., there exist a positive integer* r *and constants* α_i, $0 \leq i < r$, *such that*

$$\mathbf{h}_{r+k} = -\alpha_0 \mathbf{h}_k - \alpha_1 \mathbf{h}_{k+1} - \alpha_2 \mathbf{h}_{k+2} - \cdots - \alpha_{r-2} \mathbf{h}_{r+k-2} - \alpha_{r-1} \mathbf{h}_{r+k-1}, \quad k > 0. \tag{4.64}$$

(d) *The rank of* $\mathcal{H}$ *is* finite.

Proof. **(a) ⇒ (b).** Realizability implies (4.62). Hence

$$\sum_{k>0} \mathbf{h}_k s^{-k} = \sum_{k>0} \mathbf{C}\mathbf{A}^{k-1}\mathbf{B} s^{-k} = \mathbf{C}\left(\sum_{k>0} \mathbf{A}^{k-1} s^{-k}\right)\mathbf{B} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}.$$

This proves (b). Notice that the quantity in parentheses is a *formal* power series and convergence is not an issue.

(b) ⇒ (c). Let $\det(s\mathbf{I} - \mathbf{A}) = \alpha_0 + \alpha_1 s + \cdots + \alpha_{r-1} s^{r-1} + s^r = \chi_{\mathbf{A}}(s)$. The previous relationship implies

$$\chi_{\mathbf{A}}(s)\left(\sum_{k>0} \mathbf{h}_k s^{-k}\right) = \mathbf{C}\left[\operatorname{adj}(s\mathbf{I} - \mathbf{A})\right]\mathbf{B},$$

where adj $(\mathbf{M})$ denotes the *matrix adjoint* of the $\mathbf{M}$, i.e., the matrix of *cofactors*. (For the definition and properties of the cofactors and of the adjoint, see Chapter 6 of Meyer's book [238].) On the left-hand side are terms having both positive and negative powers of s, while on the right-hand side are only terms having positive powers of s. Hence the coefficients of the negative powers of s on the left-hand side must be identically zero; this implies precisely (4.64).

(c) ⇒ (d). Relationships (4.64) imply that the $(r + 1)$st block column of $\mathcal{H}$ is a linear combination of the previous r block columns. Furthermore, because of the block Hankel structure, every block column of $\mathcal{H}$ is a subcolumn of the previous one; this implies that all block columns after the rth are linearly dependent on the first r, which in turn implies the finiteness of the rank of $\mathcal{H}$. □

The following lemma describes a fundamental property of $\mathcal{H}$; it also provides a direct proof of the implication (a) ⇒ (d).

Lemma 4.39. Factorization of $\mathcal{H}$. *If the sequence of Markov parameters is* realizable *by means of the triple* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$*,* $\mathcal{H}$ *can be factored,*

$$\mathcal{H} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}). \tag{4.65}$$

Consequently, if the sequence of Markov parameters is realizable, the rank of $\mathcal{H}$ *is finite.*

Proof. If the sequence $\{\mathbf{h}_n,\ n = 1, 2, \ldots\}$ is realizable, the relationships $\mathbf{h}_n = \mathbf{C}\mathbf{A}^{n-1}\mathbf{B}$ hold. Hence,

$$\mathcal{H} = \begin{pmatrix} \mathbf{CB} & \mathbf{CAB} & \cdots \\ \mathbf{CAB} & \mathbf{CA}^2\mathbf{B} & \cdots \\ \vdots & \vdots & \\ & & \end{pmatrix} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}).$$

It follows that rank $\mathcal{H} \leq \max\{\text{rank } \mathcal{O}, \text{rank } \mathcal{R}\} \leq \dim(\mathbf{A})$. □

To discuss the uniqueness issue of realizations, we need to recall the concept of equivalent systems defined by (4.24). In particular, Proposition 4.4 asserts that equivalent triples $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ have the same Markov parameters. Hence the best one can hope for the uniqueness question is that realizations be equivalent. Indeed, as shown in the next section, this holds for realizations with the smallest possible dimension.

4.4.1 The solution of the realization problem

We are now ready to answer the three questions posed at the beginning of this section. In the process we prove the implication (d) ⇒ (a) and hence the equivalence of the statements in Lemma 4.38.

Theorem 4.40. Main Result.

(1) *The sequence* $\mathbf{h}_k$*,* $k > 0$*, is realizable if and only if* rank $\mathcal{H} = n < \infty$.

(2) *The state-space dimension of any solution is at least n. All realizations that are minimal are both reachable and observable. Conversely, every realization that is reachable and observable is minimal.*

(3) *All minimal realizations are equivalent.*

Lemma 4.39 proves part (1) of the main theorem in one direction. To prove (1) in the other direction we will actually construct a realization assuming that the rank of $\mathcal{H}$ is finite. For this we need to define the shift σ. It acts on the columns on the Hankel matrix; if $(\mathcal{H})_k$ denotes the kth column of $\mathcal{H}$, $\sigma(\mathcal{H})_k = (\mathcal{H})_{k+m}$; in other words, σ is a shift by m columns. The shift applied to a submatrix of $\mathcal{H}$ consisting of several columns is applied to each column separately.

Lemma 4.41. Silverman realization algorithm. *Let* rank $\mathcal{H} = n$. *Find an* $n \times n$ *submatrix* Φ *of* $\mathcal{H}$ *that has full rank. Construct the following matrices:*

(i) $\sigma\Phi \in \mathbb{R}^{n\times n}$ *is the submatrix of* $\mathcal{H}$ *having the rows with the same index as those of* Φ *and the columns obtained by shifting each individual column of* Φ *by one block column (i.e., m columns).*

(ii) $\Gamma \in \mathbb{R}^{n\times m}$ *is composed of the same rows as* Φ*; its columns are the first m columns of* $\mathcal{H}$.

(iii) $\Lambda \in \mathbb{R}^{p\times n}$ *is composed of the same columns as* Φ*; its rows are the first p rows of* $\mathcal{H}$.

The triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$*, where* $\mathbf{C} = \Lambda$, $\mathbf{A} = \Phi^{-1}\sigma\Phi$*, and* $\mathbf{B} = \Phi^{-1}\Gamma$*, is a realization of dimension n of the given sequence of Markov parameters.*

Proof. By assumption there exist $n = \text{rank}\mathcal{H}$ columns of $\mathcal{H}$ that span its column space. Denote these columns by Φ_∞; note that the columns making up Φ_∞ need not be consecutive columns of $\mathcal{H}$. We denote by σ the column right-shift operator. Let $\sigma\Phi_\infty$ denote the n columns of $\mathcal{H}$ obtained by shifting those of Φ_∞ by one block column, i.e., by m individual columns; let Γ_∞ denote the first m columns of $\mathcal{H}$. Since the columns of Φ_∞ form a basis for the space spanned by the columns of $\mathcal{H}$, there exist *unique* matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$ such that

$$\sigma\Phi_\infty = \Phi_\infty\mathbf{A}, \tag{4.66}$$

$$\Gamma_\infty = \Phi_\infty\mathbf{B}. \tag{4.67}$$

Finally, define $\mathbf{C}$ as the first block row, i.e., the first p individual rows, of Φ_∞:

$$\mathbf{C} = (\Phi_\infty)_1. \tag{4.68}$$

For this proof $(\mathbf{M})_k$, $k > 0$, denotes the kth block row of the matrix $\mathbf{M}$. Recall that the first block element of Γ_∞ is $\mathbf{h}_1$, i.e., using our notation $(\Gamma_\infty)_1 = \mathbf{h}_1$. Thus (4.67), together with (4.68), implies

$$\mathbf{h}_1 = (\Gamma_\infty)_1 = (\Phi_\infty\mathbf{B})_1 = (\Phi_\infty)_1\mathbf{B} = \mathbf{CB}.$$

For the next Markov parameter, notice that

$$\mathbf{h}_2 = (\sigma\Gamma_\infty)_1 = (\Gamma_\infty)_2.$$

Thus making use of (4.66), we have

$$h_2 = (\sigma\Gamma_\infty)_1 = (\sigma\Phi_\infty\mathbf{B})_1 = (\Phi_\infty\mathbf{AB})_1 = (\Phi_\infty)_1\mathbf{AB} = \mathbf{CAB}.$$

For the kth Markov parameter, combining (4.67), (4.66), and (4.68), we obtain

$$\mathbf{h}_k = (\sigma^{k-1}\Gamma_\infty)_1 = (\sigma^{k-1}\Phi_\infty\mathbf{B})_1 = (\Phi_\infty\mathbf{A}^{k-1}\mathbf{B})_1 = (\Phi_\infty)_1\mathbf{A}^{k-1}\mathbf{B} = \mathbf{CA}^{k-1}\mathbf{B}.$$

Thus $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ is indeed a realization of dimension n. □

The state dimension of a realization cannot be less than n; indeed, if such a realization exists, the rank of $\mathcal{H}$ will be less than n, which is a contradiction to the assumption that the rank of $\mathcal{H}$ is equal to n. Thus a realization of Σ whose dimension equals rank $\mathcal{H}$ is called a *minimal realization*; notice that the Silverman algorithm constructs minimal realizations. In this context the following holds true.

Lemma 4.42. *A realization of* $\boldsymbol{\Sigma}$ *is minimal if and only if it is reachable and observable.*

Proof. Let $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ be some realization of $\mathbf{h}_n$, $n > 0$. Since $\mathcal{H} = \mathcal{O}\mathcal{R}$,

$$\text{rank}\mathcal{H} \leq \min\{\text{rank}\mathcal{O}, \text{rank}\mathcal{R}\} \leq \dim(\mathbf{A}).$$

Let $(\hat{\mathbf{C}}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ be a reachable and observable realization. Since $\mathcal{H} = \hat{\mathcal{O}}\hat{\mathcal{R}}$, and each of the matrices $\hat{\mathcal{O}}$, $\hat{\mathcal{R}}$ contains a nonsingular matrix of size $\hat{\mathbf{A}}$, we conclude that $\dim(\hat{\mathbf{A}}) \leq \text{rank}\mathcal{H}$, which concludes the proof. □

We are now left with the proof of part (3) of the main theorem, namely, that minimal realizations are equivalent. We provide the proof only for a special case; the proof of the general case follows along similar lines.

Outline of proof. *SISO case (i.e.,* $p = m = 1$*).* Let $(\mathbf{C}_i, \mathbf{A}_i, \mathbf{B}_i)$, $i = 1, 2$, be minimal realizations of Σ. We will show the existence of a transformation $\mathbf{T}$, $\det \mathbf{T} \neq 0$ such that (4.24) holds. From Lemma 4.39 we conclude that

$$\mathcal{H}_{n,n} = \mathcal{O}_n^1 \mathcal{R}_n^1 = \mathcal{O}_n^2 \mathcal{R}_n^2, \tag{4.69}$$

where the superscript is used to distinguish between the two different realizations. Furthermore, the same lemma also implies

$$\mathcal{H}_{n,n+1} = \mathcal{O}_n^1 [\mathbf{B}_1 \;\; \mathbf{A}_1 \mathcal{R}_n^1] = \mathcal{O}_n^2 [\mathbf{B}_2 \;\; \mathbf{A}_2 \mathcal{R}_n^2],$$

which in turn yields

$$\mathcal{O}_n^1 \mathbf{A}_1 \mathcal{R}_n^1 = \mathcal{O}_n^2 \mathbf{A}_2 \mathcal{R}_n^2. \tag{4.70}$$

Because of minimality, the following determinants are nonzero: $\det \mathcal{O}_n^i \neq 0$, $\det \mathcal{R}_n^i \neq 0$, $i = 1, 2$. We now define

$$\mathbf{T} = (\mathcal{O}_n^1)^{-1} \mathcal{O}_n^2 = \mathcal{R}_n^1 (\mathcal{R}_n^2)^{-1}.$$

Equation (4.69) implies $\mathbf{C}_1 = \mathbf{C}_2 \mathbf{T}^{-1}$ and $\mathbf{B}_1 = \mathbf{T}\mathbf{B}_2$, while (4.70) implies $\mathbf{A}_1 = \mathbf{T}\mathbf{A}_2\mathbf{T}^{-1}$. □

Example 4.43. We now investigate the realization problem for the *Fibonacci sequence* given in Example 4.37,

$$\Sigma_3 = \{1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots\},$$

which is constructed according to the rule $\mathbf{h}_1 = 1$, $\mathbf{h}_2 = 2$, and

$$\mathbf{h}_{k+2} = \mathbf{h}_{k+1} + \mathbf{h}_k, \qquad k > 0.$$

The Hankel matrix (4.63) becomes

$$\mathcal{H} = \begin{pmatrix} 1 & 2 & 3 & 5 & 8 & 13 & \cdots \\ 2 & 3 & 5 & 8 & 13 & 21 & \cdots \\ 3 & 5 & 8 & 13 & 21 & 34 & \cdots \\ 5 & 8 & 13 & 21 & 34 & 55 & \cdots \\ 8 & 13 & 21 & 34 & 55 & 89 & \cdots \\ 13 & 21 & 34 & 55 & 89 & 144 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

It readily follows from the law of construction of the sequence that the rank of the Hankel matrix is two. Φ is chosen so that it contains rows 2 and 4 and columns 2 and 5 of $\mathcal{H}$:

$$\Phi = \begin{pmatrix} 3 & 13 \\ 8 & 34 \end{pmatrix} \Rightarrow \Phi^{-1} = \begin{pmatrix} -17 & \frac{13}{2} \\ 4 & -\frac{3}{2} \end{pmatrix}.$$

The remaining matrices are now

$$\sigma\Phi = \begin{pmatrix} 5 & 21 \\ 13 & 55 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \quad \text{and} \quad \Lambda = (2 \quad 8).$$

It follows that

$$\mathbf{A} = \Phi^{-1}\sigma\Phi = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix}, \quad \mathbf{B} = \Phi^{-1}\Gamma = \begin{pmatrix} -3/2 \\ 1/2 \end{pmatrix}, \quad \text{and} \quad \mathbf{C} = (2 \quad 8).$$

Furthermore,

$$\mathbf{H}_3(s) = \sum_{k>0} \mathbf{h}_k s^{-k} = \frac{s+1}{s^2 - s - 1}.$$

Concerning the remaining four sequences of Example 4.37, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are realizable, while the last two, namely, $\mathbf{\Sigma}_4$ and $\mathbf{\Sigma}_5$, are not realizable. In particular,

$$\mathbf{\Sigma}_1 = \left(\begin{array}{c|c} 1 & 1 \\ \hline 1 & 0 \end{array}\right), \quad \mathbf{H}_1(s) = \frac{1}{s-1}, \quad \text{and} \quad \mathbf{\Sigma}_2 = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ -1 & 2 & 1 \\ \hline 0 & 1 & 0 \end{array}\right], \quad \mathbf{H}_2(s) = \frac{s}{(s-1)^2}.$$

In the last case, $\mathbf{H}_5(s) = e^{s^{-1}} - 1$, which is not rational and hence has no finite-dimensional realization. The fact that $\mathbf{\Sigma}_5$ is not realizable follows also from the fact that the determinant of the associated Hankel matrix $\mathcal{H}_{i,j} = \frac{1}{i+j-1}$, of size n, also known as the *Hankel determinant*, is nonzero; it has been shown in [5], namely, that

$$\det \mathcal{H}_n = \prod_{i=2}^{n} \prod_{j=2}^{i} \frac{(j-1)^2}{(2j-1)(2j-2)^2(2j-3)},$$

which implies that the Hankel determinant for $n = 4, 5, 6$ is of the order 10^{-7}, 10^{-17}, and 10^{-43}, respectively.

4.4.2 Realization of proper rational matrix functions

Given is a $p \times m$ matrix $\mathbf{H}(s)$ with proper rational entries, i.e., entries whose numerator degree is no larger than the denominator degree. Consider first the scalar case, i.e., $p = m = 1$. We can write

$$\mathbf{H}(s) = \mathbf{D} + \frac{\mathbf{p}(s)}{\mathbf{q}(s)},$$

where $\mathbf{D}$ is a constant in $\mathbb{R}$ and $\mathbf{p}$, $\mathbf{q}$ are polynomials in s,

$$\begin{aligned} \mathbf{p}(s) &= p_0 + p_1 s + \cdots + p_{\nu-1}s^{\nu-1}, \qquad && p_i \in \mathbb{R}, \\ \mathbf{q}(s) &= q_0 + q_1 s + \cdots + q_{\nu-1}s^{\nu-1} + s^{\nu}, \qquad && q_i \in \mathbb{R}. \end{aligned}$$

In terms of these coefficients p_i and q_i, we can write down a realization of $\mathbf{H}(s)$ as follows:

$$\mathbf{\Sigma_H} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{ccccc|c} 0 & 1 & 0 & & 0 & 0 \\ 0 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & & 1 & 0 \\ -q_0 & -q_1 & -q_2 & \cdots & -q_{\nu-1} & 1 \\ \hline p_0 & p_1 & p_2 & \cdots & p_{\nu-1} & \mathbf{D} \end{array}\right) \in \mathbb{R}^{(\nu+1)\times(\nu+1)}. \quad (4.71)$$

It can be shown that $\mathbf{\Sigma_H}$ is indeed a realization of $\mathbf{H}$, i.e.,

$$\mathbf{H}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I}_\nu - \mathbf{A})^{-1}\mathbf{B}.$$

This realization is reachable but not necessarily observable; this means that the rank of the associated Hankel matrix is at most ν. The realization is in addition observable if the polynomials $\mathbf{p}$ and $\mathbf{q}$ are coprime. Thus (4.71) is *minimal* if $\mathbf{p}$ and $\mathbf{q}$ are coprime. In this case the rank of the associated Hankel matrix $\mathcal{H}$ is precisely ν.

In the general case, we can write

$$\mathbf{H}(s) = \mathbf{D} + \frac{1}{\mathbf{q}(s)}\mathbf{P}(s),$$

where $\mathbf{q}$ is a scalar polynomial that is the least common multiple of the denominators of the entries of $\mathbf{H}$ and $\mathbf{P}$ is a polynomial matrix of size $p \times m$:

$$\begin{aligned} \mathbf{P}(s) &= \mathbf{P}_0 + \mathbf{P}_1 s + \cdots + \mathbf{P}_{\nu-1}s^{\nu-1}, \qquad && \mathbf{P}_i \in \mathbb{R}^{p\times m}, \\ \mathbf{q}(s) &= q_0 + q_1 s + \cdots + q_{\nu-1}s^{\nu-1} + s^{\nu}, \qquad && q_i \in \mathbb{R}. \end{aligned}$$

The construction given above provides a realization:

$$\mathbf{\Sigma_H} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{ccccc|c} \mathbf{0}_m & \mathbf{I}_m & \mathbf{0}_m & \cdots & \mathbf{0}_m & \mathbf{0}_m \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{I}_m & \cdots & \mathbf{0}_m & \mathbf{0}_m \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{0}_m & & \mathbf{I}_m & \mathbf{0}_m \\ -q_0\mathbf{I}_m & -q_1\mathbf{I}_m & -q_2\mathbf{I}_m & \cdots & -q_{\nu-1}\mathbf{I}_m & \mathbf{I}_m \\ \hline \mathbf{P}_0 & \mathbf{P}_1 & \mathbf{P}_2 & \cdots & \mathbf{P}_{\nu-1} & \mathbf{D} \end{array}\right),$$

where $\mathbf{0}_m$ is a square zero matrix of size m, $\mathbf{I}_m$ is the identity matrix of the same size, and, consequently, the state of this realization has size νm. Unlike the scalar case, however, the realization $\Sigma_{\mathbf{H}}$ need not be minimal. One way to obtain a minimal realization is by applying the reachable observable-canonical decomposition given in Lemma 4.25. An alternative way is to apply the Silverman algorithm; in this case, $\mathbf{H}$ has to be expanded into a formal power series,

$$\mathbf{H}(s) = \mathbf{h}_0 + \mathbf{h}_1 s^{-1} + \mathbf{h}_2 s^{-2} + \cdots + \mathbf{h}_t s^{-t} + \cdots .$$

The Markov parameters can be computed using the following relationship. Given the polynomial q as above, let

$$q^{(k)}(s) = s^{\nu-k} + q_{n-1}s^{\nu-k-1} + \cdots + q_{k+1}s + q_k, \qquad k = 1, \ldots, \nu, \tag{4.72}$$

denote its ν *pseudo-derivative* polynomials. It follows that the numerator polynomial $\mathbf{P}(s)$ is related to the Markov parameters $\mathbf{h}_k$ and the denominator polynomial q as follows:

$$\mathbf{P}(s) = \mathbf{h}_1 q^{(1)}(s) + \mathbf{h}_2 q^{(2)}(s) + \cdots + \mathbf{h}_{\nu-1} q^{(\nu-1)}(s) + \mathbf{h}_\nu q^{(\nu)}(s). \tag{4.73}$$

This can be verified by direct calculation. Alternatively, assume that $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$, and let $q(s)$ denote the characteristic polynomial of $\mathbf{A}$. Then

$$\text{adj}\,(s\mathbf{I} - \mathbf{A}) = q^{(\nu)}(s)\mathbf{A}^{\nu-1} + q^{(\nu-1)}(s)\mathbf{A}^{\nu-2} + \cdots q^{(2)}(s)\mathbf{A}^1 + q^{(1)}(s)\mathbf{I}. \tag{4.74}$$

The result (4.73) follows by noting that $\mathbf{P}(s) = \mathbf{C}\,\text{adj}\,(s\mathbf{I} - \mathbf{A})\,\mathbf{B}$.

Since $\mathbf{H}$ is rational, the rank of the ensuing *Hankel* matrix associated with the sequence of Markov parameters $\mathbf{h}_k$, $k > 0$, is guaranteed to have *finite rank*. In particular, the following upper bound holds:

$$\text{rank}\mathcal{H} \leq \min\{\nu m, \nu p\}.$$

An important attribute of a rational matrix function is its *McMillan degree*. For proper rational matrix functions $\mathbf{H}$, the McMillan degree turns out to equal the rank of the associated Hankel matrix $\mathcal{H}$; in other words, the McMillan degree in this case is equal to the dimension of any minimal realization of $\mathbf{H}$.

4.4.3 Symmetric systems and symmetric realizations

A system $\mathbf{\Sigma}$ is called symmetric if its Markov parameters are symmetric, $\mathbf{h}_k = \mathbf{h}_k^*$, $k \geq 0$.[3] In other words, $\mathbf{\Sigma}$ is symmetric if $\mathbf{h}_0 = \mathbf{h}_0^*$ and the associated Hankel matrix (4.63) is symmetric, $\mathcal{H} = \mathcal{H}^*$.

Definition 4.44. *A realization is called* symmetric *if* $\mathbf{D} = \mathbf{D}^*$ *and there exists a symmetric matrix* $\Psi = \Psi^*$ *such that*

$$\mathbf{A}\Psi = \Psi\mathbf{A}^*, \;\; \mathbf{B} = \Psi\mathbf{C}^*. \tag{4.75}$$

[3] Recall that if a matrix is real, the superscript $(\cdot)^*$ denotes simple transposition.

It follows that every symmetric system has a symmetric realization.

Lemma 4.45. *A reachable and observable system $\mathbf{\Sigma}$ is symmetric if and only if it possesses a symmetric realization.*

Proof. A moment's reflection shows that if $\mathbf{\Sigma}$ has a symmetric realization, it is symmetric. Conversely, let the system be symmetric; this together with the factorization (4.65) implies

$$\mathcal{H} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{H}^*.$$

Thus, since the column span of $\mathcal{H}$ and $\mathcal{H}^*$ are the same, there exists a matrix $\Psi \in \mathbb{R}^{n\times n}$ such that $\mathcal{O}(\mathbf{C}, \mathbf{A})\Psi = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)$; hence $\mathbf{C}\Psi = \mathbf{B}^*$. Furthermore, $\mathcal{O}(\mathbf{C}, \mathbf{A})\mathbf{A}\Psi = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathbf{A}^* = \mathcal{O}(\mathbf{C}, \mathbf{A})\Psi\mathbf{A}^*$. Since $\mathcal{O}$ has full column rank, the equality $\mathbf{A}\Psi = \Psi\mathbf{A}^*$ follows. It remains to show that Ψ is symmetric. Notice that $\mathcal{O}(\mathbf{C}, \mathbf{A})\Psi\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{H}$ is symmetric. Again, since $\mathcal{O}$ has full column rank, it has an $n \times n$ nonsingular submatrix, composed of the rows with index $I = \{i_1, \dots, i_n\}$, which we denote by $\mathcal{O}_I$; thus $\mathcal{O}_I\Psi\mathcal{O}_I^* = \mathcal{H}_{I,I}$, where the latter is the submatrix of the Hankel matrix composed of those rows and columns indexed by I. Thus $\Psi = [\mathcal{O}_I]^{-1}\mathcal{H}_{I,I}\left[\mathcal{O}_I^*\right]^{-1}$. Since $\mathcal{H}_{I,I}$ is symmetric, this proves the symmetry of Ψ. The proof is thus complete. □

4.4.4 The partial realization problem

This problem was studied in [193]. Recursive solutions were provided in [12] and [11]. Recall section 4.4 and in particular Definition 4.36. The realization problem with partial data is defined as follows.

Definition 4.46. *Given the finite sequence of $p \times m$ matrices $\mathbf{h}_k$, $k = 1, \dots, r$, the* partial realization problem *consists of finding a positive integer n and constant matrices* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *such that*

$$\mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B};\ \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m}, \qquad k = 1, 2, \dots, N.$$

The triple $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ *is then called a* partial realization *of the sequence* $\mathbf{h}_k$.

Because of Lemma 4.38, a finite sequence of matrices is always realizable. As a consequence, the set of problems arising consists of

(a) *minimality*: given the sequence $\mathbf{h}_k$, $k = 1, \dots, r$, find the smallest positive integer n for which the partial realization problem is solvable.

(b) *parametrization of solutions*: parametrize all minimal and other solutions.

(c) *recursive construction*: recursive construction of solutions.

Similarly to the realization problem, the partial realization problem can be studied by means of the partially defined Hankel matrix:

$$\mathcal{H}_r = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & & \mathbf{h}_r \\ \mathbf{h}_2 & & \cdots & \mathbf{h}_r & ? \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ & \mathbf{h}_r & \cdots & ? & ? \\ \mathbf{h}_r & ? & \cdots & ? & ? \end{pmatrix} \in \mathbb{R}^{rp\times rm},$$

where ? denote unknown matrices defining the continuation of the given finite sequence $\mathbf{h}_k$, $k = 1, \dots, N$.

The *rank* of the partially defined Hankel matrix $\mathcal{H}_k$ is defined as the size of the largest nonsingular submatrix of $\mathcal{H}_k$, independently of the unknown parameters "?". It then follows that the dimension of any partial realization $\mathbf{\Sigma}$ satisfies

$$\dim \mathbf{\Sigma} \geq \operatorname{rank} \mathcal{H}_k = n.$$

Furthermore, there always exists a partial realization of dimension n, which is a minimal partial realization. Once the rank of $\mathcal{H}_k$ is determined, Silverman's algorithm (see Lemma 4.41) can be used to construct such a realization. We illustrate this procedure by means of a simple example.

Example 4.47. Consider the scalar (i.e., $m = p = 1$) sequence $\mathbf{\Sigma} = (1, 1, 1, 2)$; the corresponding Hankel matrix $\mathcal{H}_4$ and its first three submatrices are

$$\mathcal{H}_4 = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & a \\ 1 & 2 & a & b \\ 2 & a & b & c \end{pmatrix}, \ \mathcal{H}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & a \end{pmatrix}, \ \mathcal{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \ \mathcal{H}_1 = (1),$$

where a, b, c denote the unknown continuation of the original sequence. The determinants of these matrices are

$$\det \mathcal{H}_4 = -a^3 + 4a^2 - 8a + 8 + 2ab - 3b - c, \ \det \mathcal{H}_3 = -1, \ \det \mathcal{H}_2 = 0, \ \det \mathcal{H}_1 = 1.$$

It follows that

$$\operatorname{rank} \mathcal{H}_4 = 3.$$

By Lemma 4.41, we choose $\Phi = \mathcal{H}_3$, $\Gamma = \Lambda^* = (1\ \ 1\ \ 1)^*$, which implies

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & a^2 - 4a + 8 - b \\ 1 & 0 & -a^2 + 3a - 4 + b \\ 0 & 1 & a - 2 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{C} = (1\ \ 1\ \ 1).$$

Hence, there are multiple minimal partial realizations of $\mathbf{\Sigma}$. Indeed, the above expressions provide a parametrization of all minimal solutions; the parameters are $a, b \in \mathbb{R}$. Finally, we note that the value of c is uniquely determined by a, b. In this case, for realizations of minimal degree 3, we must have $c = -a^3 + 5a^2 - 12a + 2ab - 4b + 16$.

4.5 The rational interpolation problem*

The realization problem aims at constructing a linear system in internal (state-space) form, from some or all Markov parameters $\mathbf{h}_k$. As already noted, the Markov parameters constitute information about the transfer function $\mathbf{H}(s)$ obtained by expanding it around infinity $\mathbf{H}(s) = \sum_{k\geq 0} \mathbf{h}_k s^{-k}$.

In this section we will address the problem of constructing an external or internal description of a system based on finitely many samples of the transfer function taken at finite points in the complex plane. This problem is known as *rational interpolation*. For simplicity we will assume that the systems in question have a single input and a single output and thus their transfer function is scalar rational. We distinguish between two approaches.

The first approach, presented in subsection 4.5.2, has as a main tool the so-called Löwner matrix. The Löwner matrix encodes the information about the admissible complexity of the solutions as a simple function of its rank. The computation of the solutions can be carried out both in the *external* (transfer function) and in the *internal* (state-space) frameworks. This approach to rational interpolation leads to a generalization of the classical, system theoretic, concept of *realization* of linear dynamical systems.

The second approach, presented in subsection 4.5.3, is known as the *generating system approach* and involves the construction of a polynomial or rational matrix, such that any polynomial or rational combination of its rows yields a solution of the problem at hand. This construction has a system theoretic interpretation as a *cascade interconnection* of two systems, one of which can be chosen freely. This method leads to *recursive solutions* in a natural way by expanding, namely, the cascade interconnection just mentioned. The solutions can be classified according to properties like *complexity*, *norm boundedness*, or *positive realness*.

4.5.1 Problem formulation*

Consider the array of pairs of points

$$\mathbb{P} = \{(s_i;\phi_{ij}) : \ j = 0, 1, \dots, \ell_i - 1; \ i = 1, \dots, k, \ s_i \neq s_j, \ i \neq j\}, \tag{4.76}$$

where we will assume that there are N given pieces of data, that is, $\sum_{i=1}^{k} \ell_i = N$. We are looking for *all* rational functions,

$$\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}, \ \ \gcd(\mathbf{n}, \mathbf{d}) = 1, \tag{4.77}$$

where $\mathbf{n}(s)$, $\mathbf{d}(s)$ are *coprime* polynomials, that is, their *greatest common divisor* is a (nonzero) constant, which *interpolate* the points of the array $\mathbb{P}$, i.e.,

$$\left.\frac{d^j\phi(s)}{ds^j}\right|_{s=s_i} = \phi_{ij}, \qquad j = 0, 1, \dots, \ell_i - 1, \ i = 1, \dots, k. \tag{4.78}$$

In other words, the jth derivative of $\phi(s)$ evaluated at $s = s_i$ is equal to ϕ_{ij}. We distinguish two special cases: (a) the distinct point interpolation problem,

$$\mathbb{P} = \{(s_i;\phi_i) : \ i = 1, \dots, N, \ s_i \neq s_j, \ i \neq j\}, \tag{4.79}$$

and (b) the single multiple point interpolation problem,

$$\mathbb{P} = \{(s_0;\, \phi_{0j}) :\ j = 0, 1, \ldots, N-1\}, \tag{4.80}$$

where the value of the function and derivatives thereof are provided only at $s = s_0$.

Solution of the unconstrained problem

The *Lagrange interpolating polynomial* associated with $\mathbb{P}$ is the unique polynomial of degree less than N which interpolates the points of this array. In the distinct point case (4.79), it is

$$\ell(s) = \sum_{j=1}^{N} \phi_j \prod_{i \neq j} \frac{s - s_i}{s_j - s_i}. \tag{4.81}$$

A similar formula holds for the general case. A parametrization of *all* solutions to (4.77), (4.78) can be given in terms of $\ell(s)$ as follows:

$$\phi(s) = \ell(s) + \mathbf{r}(s)\Pi_{i=1}^{N}(s - s_i), \tag{4.82}$$

where the parameter $\mathbf{r}(s)$ is an arbitrary rational function with no poles at the s_i. Most often, however, one is interested in parametrizing all solutions to the interpolation problem (4.78) which satisfy additional constraints. In such cases, this formula, although general, provides little insight.

Constrained interpolation problems

The first parameter of interest is the *complexity* or *degree* of rational interpolants (4.77). It is defined as

$$\deg \phi = \max\{\deg \mathbf{n},\ \deg \mathbf{d}\}$$

and is sometimes referred to as the *McMillan degree* of the rational function ϕ. The following problems arise.

Problem (A): ***Parametrization of interpolants by complexity***.

(a) Find the *admissible* degrees of complexity, i.e., those positive integers π for which there exist solutions $\phi(s)$ to the interpolation problem (4.77), (4.78), with $\deg \phi = \pi$.

(b) Given an admissible degree π, construct *all* corresponding solutions.

Another constraint of interest is *bounded realness*, that is, finding interpolants which have poles in the left half of the complex plane (called *stable*) and whose magnitude on the imaginary axis is less than some given positive number μ.

Problem (B): Nevanlinna–Pick. ***Parametrization of interpolants by norm.***

(a) Do there exist bounded real interpolating functions?

(b) If so, what is the minimum norm and how can such interpolating functions be constructed?

A third constraint of interest is *positive realness* of interpolants. A function $\phi : \mathbb{C} \rightarrow \mathbb{C}$ is positive real (p.r.) if it maps the closed right half of the complex plane onto itself:

$$s \in \mathbb{C} : \mathcal{R}e(s) \geq 0 \mapsto \phi(s) : \mathcal{R}e(\phi(s)) \geq 0 \ \text{ for } \ s \ \text{ not a pole of } \ \phi.$$

Thus given the array of points $\mathbb{P}$, the following problem arises.

Problem (C): ***Parametrization of positive real interpolants.***

(a) Does there exist a p.r. interpolating function?

(b) If so, give a procedure to construct such interpolating functions.

In sections 4.5.2 and 4.5.3, we will investigate the constrained interpolation Problem (A) for the special case of *distinct* interpolating points. At the end of section 4.5.3, a short summary of the solution of the other two, that is, Problems (B) and (C), will be provided.

4.5.2 The Löwner matrix approach to rational interpolation*

The idea behind this approach to rational interpolation is to use a formula similar to (4.81) which would be valid for rational functions. Before introducing this formula, we partition the array $\mathbb{P}$ given by (4.79) into two disjoint subarrays $\mathbb{J}$ and $\mathbb{I}$ as follows:

$$\mathbb{J} = \{(s_i, \phi_i) : \ i = 1, \ldots, r\}, \ \ \mathbb{I} = \{(\hat{s}_i, \hat{\phi}_i) : \ i = 1, \ldots, p\},$$

where for simplicity of notation some of the points have being redefined as follows: $\hat{s}_i = s_{r+i}, \hat{\phi}_i = \phi_{r+i}, i = 1, \ldots, p, p+r = N$. Consider $\phi(s)$ defined by the following equation:

$$\sum_{i=1}^{r} \gamma_i \frac{\phi(s) - \phi_i}{s - s_i} = 0, \qquad \gamma_i \neq 0, \ i = 1, \ldots, r, \ r \leq N.$$

Solving for $\phi(s)$ we obtain

$$\phi(s) = \frac{\sum_{j=1}^{r} \phi_j \gamma_j \prod_{i \neq j}(s - s_i)}{\sum_{j=1}^{r} \gamma_j \prod_{i \neq j}(s - s_i)}, \qquad \gamma_j \neq 0. \tag{4.83}$$

Clearly, the above formula, which can be regarded as the rational equivalent of the Lagrange formula, interpolates the first r points of the array $\mathbb{P}$, i.e., the points of the array $\mathbb{J}$. For $\phi(s)$ to interpolate the points of the array $\mathbb{I}$, the coefficients γ_i have to satisfy the following equation:

$$\boldsymbol{L}\gamma = \boldsymbol{0},$$

where

$$\boldsymbol{L} = \begin{bmatrix} \frac{\hat{\phi}_1 - \phi_1}{\hat{s}_1 - s_1} & \cdots & \frac{\hat{\phi}_1 - \phi_r}{\hat{s}_1 - s_r} \\ \vdots & & \vdots \\ \frac{\hat{\phi}_p - \phi_1}{\hat{s}_p - s_1} & \cdots & \frac{\hat{\phi}_p - \phi_r}{\hat{s}_p - s_r} \end{bmatrix} \in \mathbb{R}^{p \times r}, \ \ \gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_r \end{bmatrix} \in \mathbb{R}^r. \tag{4.84}$$

$\boldsymbol{L}$ is called the *Löwner matrix*, defined by means of the *row array* $\mathbb{I}$ and the *column array* $\mathbb{J}$. As it turns out, $\boldsymbol{L}$ is the main tool of this approach to the rational interpolation problem.

Remark 4.5.1. As shown by Antoulas and Anderson [25], the (generalized) Löwner matrix associated with the array $\mathbb{P}$ consisting of *one multiple point* (4.80) has *Hankel* structure. In particular,

$$\boldsymbol{L} = \begin{bmatrix} \frac{\phi^{(1)}(s_0)}{1!} & \frac{\phi^{(2)}(s_0)}{2!} & \frac{\phi^{(3)}(s_0)}{3!} & \cdots & \frac{\phi^{(k)}(s_0)}{k!} \\ \frac{\phi^{(2)}(s_0)}{2!} & \frac{\phi^{(3)}(s_0)}{3!} & \frac{\phi^{(4)}(s_0)}{4!} & \cdots & \frac{\phi^{(k+1)}(s_0)}{(k+1)!} \\ \frac{\phi^{(3)}(s_0)}{3!} & \frac{\phi^{(4)}(s_0)}{4!} & \frac{\phi^{(5)}(s_0)}{5!} & \cdots & \frac{\phi^{(k+2)}(s_0)}{(k+2)!} \\ \vdots & \vdots & \vdots & \iddots & \vdots \\ \frac{\phi^{(k)}(s_0)}{k!} & \frac{\phi^{(k+1)}(s_0)}{(k+1)!} & \frac{\phi^{(k+2)}(s_0)}{(k+2)!} & \cdots & \frac{\phi^{(2k)}(s_0)}{(2k)!} \end{bmatrix}.$$

This shows that the Löwner matrix is the right tool for generalizing realization theory to rational interpolation.

From rational function to Löwner matrix

The key result in connection with the Löwner matrix is the following.

Lemma 4.48. *Consider the array of points* $\mathbb{P}$ *defined by* (4.76), *consisting of samples taken from a given rational function* $\phi(s)$*, together with a partition into subarrays* $\mathbb{J}$, $\mathbb{I}$ *as defined in the beginning of the subsection. Let* $\boldsymbol{L}$ *be any* $p \times r$ *Löwner matrix with* $p, r \geq \deg \phi$. *It follows that* $\operatorname{rank} \boldsymbol{L} = \deg \phi$.

Corollary 4.49. *Under the assumptions of the lemma, any square Löwner submatrix of* $\boldsymbol{L}$ *of size* $\deg \phi$ *is nonsingular.*

In what follows, given $\mathbf{A} \in \mathbb{R}^{\pi \times \pi}$, $\mathbf{b}, \mathbf{c}^* \in \mathbb{R}^{\pi}$, the following matrices will be of interest:

$$\mathcal{R}_r = [(s_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \ \cdots \ (s_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}] \in \mathbb{R}^{\pi \times r}, \tag{4.85}$$

$$\mathcal{O}_p = [(\hat{s}_1\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{c}^* \ \cdots \ (\hat{s}_p\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{c}^*]^* \in \mathbb{R}^{p \times \pi}. \tag{4.86}$$

As will be shown subsequently in (4.87), the Löwner matrix factors in a product of $\mathcal{O}_p$ times $\mathcal{R}_r$. Therefore, in analogy with the realization problem (where the Hankel matrix factors in a product of an observability times a reachability matrix), we will call $\mathcal{O}_p$ the *generalized observability matrix* and $\mathcal{R}_r$ the *generalized reachability matrix* associated with the underlying interpolation problem.

Proposition 4.50. *Let* $(\mathbf{A}, \mathbf{b})$ *be a reachable pair, where* $\mathbf{A}$ *is a square matrix and* $\mathbf{b}$ *is a vector; in addition, let* s_i*,* $i = 1, \ldots, r$*, be scalars that are not eigenvalues of* $\mathbf{A}$*. It follows that the generalized reachability matrix defined by* (4.85) *has rank equal to the size of* $\mathbf{A}$*, provided that* $r \geq \operatorname{size}(\mathbf{A})$.

For a proof of this proposition see Antoulas and Anderson [24]. Based on this proof, we can now provide a proof of Lemma 4.48.

Proof. We distinguish two cases. **(a)** $\phi(s)$ is *proper rational.* According to the results in section 4.4.2, there exists a minimal quadruple $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$ of dimension π such that

$$\phi(s) = d + \mathbf{c}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

This expression implies

$$\begin{aligned}\hat{\phi}_i - \phi_j &= \mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}[(s_j\mathbf{I} - \mathbf{A}) - (\hat{s}_i\mathbf{I} - \mathbf{A})](s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\\ &= \mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}[s_j - \hat{s}_i](s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b},\end{aligned}$$

and hence

$$[\boldsymbol{L}]_{i,j} = \frac{\hat{\phi}_i - \phi_j}{\hat{s}_i - s_j} = -\mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}(s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

Consequently, $\boldsymbol{L}$ can be factorized as follows:

$$\boldsymbol{L} = -\mathcal{O}_p\mathcal{R}_r, \tag{4.87}$$

where $\mathcal{R}_r$ and $\mathcal{O}_p$ are the generalized reachability, observability matrices defined by (4.85), (4.86), respectively. Because of the proposition given above, the rank of both $\mathcal{O}_p$ and $\mathcal{R}_r$ is π. This implies that the rank of their product $\boldsymbol{L}$ is also π. This completes the proof when $\phi(s)$ is proper.

(b) $\phi(s)$ is *not proper rational.* In this case, by means of a bilinear transformation

$$s \mapsto \frac{\alpha s + \beta}{s + \gamma}, \qquad \alpha\gamma - \beta \neq 0,$$

for almost all α, β, γ, the rational function

$$\tilde{\phi}(s) = \phi\left(\frac{\alpha s + \beta}{s + \gamma}\right)$$

will be proper. The Löwner matrices $\boldsymbol{L}$, $\bar{\boldsymbol{L}}$ attached to ϕ, $\tilde{\phi}$, respectively, are related as follows:

$$(\alpha\gamma - \beta)\tilde{\boldsymbol{L}} = \operatorname{diag}(\alpha - \hat{s}_i)\,\boldsymbol{L}\,\operatorname{diag}(\alpha - s_i).$$

The parameter α can be chosen so that $\operatorname{diag}(\alpha - \hat{s}_i)$ and $\operatorname{diag}(\alpha - s_j)$ are nonsingular, which implies the desired result. This concludes the proof of the lemma. □

From Löwner matrix to rational function

Given the array of points $\mathbb{P}$ defined by (4.76), we are now ready to tackle the interpolation problem (4.77), (4.78) and, in particular, solve the two problems (a) and (b) of Problem (A). The following definition is needed first.

Definition 4.51. **(a)** *The rank of the array* $\mathbb{P}$ *is*

$$\operatorname{rank}\mathbb{P} = \max_{\boldsymbol{L}} \{\operatorname{rank}\boldsymbol{L}\} = q,$$

where the maximum is taken over all possible Löwner matrices which can be formed from $\mathbb{P}$.

(b) *We will call a Löwner matrix almost square if it has at most one more row than column or vice versa, with the sum of the number of rows and columns being equal to* N.

The following is a consequence of Lemma 4.48.

Proposition 4.52. *The rank of all Löwner matrices having at least* q *rows and* q *columns is equal to* q. *Consequently, almost square Löwner matrices have rank* q.

Let $q = \operatorname{rank}\mathbb{P}$, and assume that $2q < N$. For any Löwner matrix with $\operatorname{rank}\boldsymbol{L} = q$, there exists a column vector $\gamma \neq \mathbf{0}$ of appropriate dimension, say, $r+1$, satisfying

$$\boldsymbol{L}\gamma = \mathbf{0} \;\text{ or }\; \gamma^*\boldsymbol{L} = \mathbf{0}. \tag{4.88}$$

In this case, we can attach to $\boldsymbol{L}$ a rational function denoted by

$$\phi_{\boldsymbol{L}}(s) = \frac{\mathbf{n}_{\boldsymbol{L}}(s)}{\mathbf{d}_{\boldsymbol{L}}(s)} \tag{4.89}$$

using formula (4.83), i.e.,

$$\mathbf{n}_{\boldsymbol{L}}(s) = \sum_{j=1}^{r+1} \gamma_j \phi_j \prod_{i\neq j}(s - s_i), \quad \mathbf{d}_{\boldsymbol{L}}(s) = \sum_{j=1}^{r+1} \gamma_j \prod_{i\neq j}(s - s_i). \tag{4.90}$$

The rational function $\phi_{\boldsymbol{L}}(s)$ just defined has the following properties.

Lemma 4.53. **(a)** $\deg\phi_{\boldsymbol{L}} \leq r \leq q < N$. **(b)** *There is a unique* $\phi_{\boldsymbol{L}}$ *attached to all* $\boldsymbol{L}$ *and* γ *satisfying* (4.88) *as long as* $\operatorname{rank}\boldsymbol{L} = q$. **(c)** *The numerator, denominator polynomials* $\mathbf{n}_{\boldsymbol{L}}$, $\mathbf{d}_{\boldsymbol{L}}$ *have* $q - \deg\phi_{\boldsymbol{L}}$ *common factors of the form* $(s - s_i)$. **(d)** $\phi_{\boldsymbol{L}}$ *interpolates exactly* $N - q + \deg\phi_{\boldsymbol{L}}$ *points of the array* $\mathbb{P}$.

The proof of this result can be found in Antoulas and Anderson [25]. As a consequence of the above lemma and Lemma 4.48, we obtain the next corollary.

Corollary 4.54. $\phi_{\boldsymbol{L}}$ *interpolates all given points if and only if* $\deg\phi_{\boldsymbol{L}} = q$ *if and only if all* $q \times q$ *Löwner matrices which can be formed from the data array* $\mathbb{P}$ *are nonsingular.*

We are now ready to state, from Antoulas and Anderson [25], the main result.

Theorem 4.55. *Given the array of N points* $\mathbb{P}$, *let* $\operatorname{rank}\mathbb{P} = q$.

(a) *If* $2q < N$, *and all square Löwner matrices of size* q *which can be formed from* $\mathbb{P}$ *are nonsingular, there is a unique interpolating function of minimal degree denoted by* $\phi^{min}(s)$, *and* $\deg\phi^{min} = q$.

(b) *Otherwise,* $\phi^{min}(s)$ *is not unique and* $\deg\phi^{min} = N - q$.

The first part of the theorem follows from the previous corollary. The second part can be justified as follows. Part (b) of the proposition above says that as long as $\boldsymbol{L}$ has rank q, there is a unique rational function $\phi_{\boldsymbol{L}}$ attached to it. Consequently, for $\boldsymbol{L}$ to yield a different rational function $\phi_{\boldsymbol{L}}$ defined by (4.89), (4.90), it will have to *lose* rank. This occurs when $\boldsymbol{L}$ has at most $q-1$ rows. In this case, its rank is $q-1$ and there exists a column vector γ such that $\boldsymbol{L}\gamma = \mathbf{0}$. Since $\boldsymbol{L}$ has $N-q+1$ columns, the degree of the attached $\phi_{\boldsymbol{L}}$ will generically (i.e., for almost all γ) be $N-q$. It readily follows that for almost all γ, $\phi_{\boldsymbol{L}}$ will interpolate all the points of the array $\mathbb{P}$. This argument shows that there can *never* exist interpolating functions of degree between q and $N-q$. The admissible degree problem can now be solved in terms of the rank of the array $\mathbb{P}$.

Corollary 4.56. *Under the assumptions of the main theorem, if* $\deg \phi^{min} = q$*, the admissible degrees are* q*, and all integers greater than or equal to* $N-q$*, while if* $\deg \phi^{min} = N-q$*, the admissible degrees are all integers greater than or equal to* $N-q$*.*

Remark 4.5.2. (i) If $2q = N$, the only solution γ of (4.88) is $\gamma = \mathbf{0}$. Hence, $\phi_{\boldsymbol{L}}$, defined by (4.89), (4.90) does not exist, and part (b) of the above theorem applies.

(ii) To distinguish between case (a) and case (b) of this theorem, we need only to check the nonsingularity of $2q+1$ Löwner matrices. Construct from $\mathbb{P}$ *any* Löwner matrix of size $q \times (q+1)$, with row, column sets denoted by $\mathbb{I}_q$, $\mathbb{J}_q$, and call it $\boldsymbol{L}_q$. The Löwner matrix $\boldsymbol{L}_q^*$ of size $(q+1) \times q$ is now constructed. Its row set $\mathbb{I}_q$ contains the points of the row set $\mathbb{I}_q$ together with the last point of the column set $\mathbb{J}_q$; moreover, its column set $\mathbb{J}_q^*$ contains the points of the column set $\mathbb{J}_q$ with the exception of the last one. The $2q+1$ Löwner matrices which need to be checked are the $q \times q$ submatrices of $\boldsymbol{L}_q$ and $\boldsymbol{L}_q^*$.

The construction of interpolating functions

Given an admissible degree, we will discuss in this section the construction of all corresponding interpolating functions. Two construction methods will be presented: the first is based on an external (input-output) framework, while the second is based on a state-space framework.

Given the array $\mathbb{P}$, let π be an admissible degree. For the polynomial construction we need to form from $\mathbb{P}$ *any* Löwner matrix having $\pi+1$ columns,

$$\boldsymbol{L} \in \mathbb{R}^{(N-\pi-1)\times(\pi+1)},$$

and determine a parametrization of all γ such that

$$\boldsymbol{L}\gamma = \mathbf{0}.$$

A parametrization of all interpolating functions of degree π is then

$$\phi_{\boldsymbol{L}}(s) = \frac{\mathbf{n}_{\boldsymbol{L}}(s)}{\mathbf{d}_{\boldsymbol{L}}(s)},$$

where the numerator and denominator polynomials are defined by (4.90). If $\pi \geq N-q$, we have to make sure that there are no common factors between the numerator and the denominator of $\phi_{\boldsymbol{L}}$; this is the case for almost all γ. More precisely, the $2\pi+1-N$ (scalar) parameters which parametrize all γ have to avoid the hypersurfaces defined by the equations

$$\mathbf{d}_{\boldsymbol{L}}(s_i)=0, \qquad i=1,\dots,N.$$

Since we can always make sure that γ depends affinely on these parameters, we are actually dealing with hyperplanes. For details and examples, see Antoulas and Anderson [25].

For use below, notice that $\phi_{\boldsymbol{L}}$ will be proper rational if and only if the leading coefficient of $\mathbf{d}_{\boldsymbol{L}}$ is different from zero; i.e., from the second formula (4.90), we must have

$$\gamma_{\pi_1}+\cdots+\gamma_{\pi_{\pi+1}}\neq 0.$$

For the *state-space* construction of interpolating functions of admissible degree π, we need a Löwner matrix of size $\pi\times(\pi+1)$:

$$\bar{\boldsymbol{L}}\in\mathbb{R}^{\pi\times(\pi+1)}.$$

Thus, in case $\pi\geq N-q$, we need an array $\bar{\mathbb{P}}$ which contains besides the original N points of the array $\mathbb{P}$, another $2\pi+1-N$ points, chosen arbitrarily but subject to the nonsingularity condition given in part (a) of the main theorem (see also the remark at the end of the previous section). Let $\bar{\gamma}\in\mathbb{R}^{\pi+1}$ be such that

$$\bar{\boldsymbol{L}}\bar{\gamma}=\mathbf{0}.$$

If $\bar{\gamma}_{\pi_1}+\cdots+\bar{\gamma}_{\pi_{\pi+1}}\neq 0$, the underlying interpolating function is proper. Otherwise, we need to perform a bilinear transformation which will ensure the properness of the function under construction. (See the proof of Lemma 4.48.) Once the properness condition is guaranteed, the state-space construction proceeds by defining the following two $\pi\times\pi$ matrices:

$$\mathbf{Q}=\bar{\boldsymbol{L}}\begin{bmatrix}1&&&\\&1&&\\&&\ddots&\\&&&1\\-1&-1&\cdots&-1\end{bmatrix},\quad \sigma\mathbf{Q}=\bar{\boldsymbol{L}}\begin{bmatrix}s_1&&&\\&s_2&&\\&&\ddots&\\&&&s\\-s_{\pi+1}&-s_{\pi+1}&\cdots&-s_{\pi+1}\end{bmatrix}\in\mathbb{R}^{(\pi+1)\times\pi},$$

where s_i, $i=1,\dots,\pi+1$, are the points which define the column array of $\bar{\boldsymbol{L}}$. Let the quadruple of constant matrices $(\mathbf{A},\mathbf{b},\mathbf{c},d)$ be defined as follows:

$$\left.\begin{array}{rcl}\mathbf{A}&=&(\sigma\mathbf{Q})\mathbf{Q}^{-1}\\ \mathbf{b}&=&(s_1\mathbf{I}-\mathbf{A})[\bar{\boldsymbol{L}}]_{(:,1)}\\ \mathbf{c}&=&[(s_1\mathbf{I}-\mathbf{A})]_{(1,:)}\\ d&=&y_i-\mathbf{c}(s_i\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\end{array}\right\}\tag{4.91}$$

for any s_i, where $[\mathbf{M}]_{(:,1)}$ denotes the first column of $\mathbf{M}$, while $[\mathbf{M}]_{(1,:)}$ denotes the first row. It can be shown that the above quadruple is a minimal realization of the desired interpolating function $\phi(s)$ of degree π:

$$\phi(s)=d+\mathbf{c}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}.\tag{4.92}$$

The steps involved in proving the above result are as follows. First, because of the properness of the underlying function, the matrix **Q** is nonsingular. Next, we need to show that none of the s_i's are an eigenvalue of **A**, that is, $(s_i\mathbf{I} - \mathbf{A})$ is invertible. Finally, we need to show that the rational function given by (4.92) is indeed an interpolating function of the prescribed degree π. These steps can be found in Anderson and Antoulas [24].

Remark 4.5.3. (i) In the realization problem the *shift* is defined in terms of the associated Hankel matrix, as the operation that assigns to the ith column the $(i + 1)$st column. It follows that **A** is determined by this shift. For the more general interpolation problem, formula (4.91) shows that

$$\mathbf{AQ} = \sigma\mathbf{Q}.$$

If we define the shift operation in this case as assigning to the ith column of the Löwner matrix, s_i times itself, then $\sigma\mathbf{Q}$ is indeed the shifted version of **Q**, and, consequently, **A** is again determined by the shift.

(ii) The theory, presented above, has been worked out for the multiple-point as well as for more general interpolation problems; see [25] and [24].

(iii) It is readily checked that the classical system theoretic problem of realization can be interpreted as a rational interpolation problem where all the data are provided at a single point. Our theory has generalized the theory of *realization* to the theory of *interpolation*.

All missing proofs, as well as other details and examples, can be found in [25] and [24]. Some of the results discussed can also be found in [50].

4.5.3 Generating system approach to rational interpolation*

This method for dealing with the rational interpolation problem is based on the factorization of a rational matrix expressed in terms of the data (4.76). It leads to a parametrization of all interpolants that solve Problems (A), (B), and (C) in section 4.5.1. Through the years, solutions to various special cases of the general rational interpolation problem have been worked out in what amounts to a generating system approach. For example, more than three-quarters of a century ago, Problem (B) was solved using this approach. Actually, the generating system was constructed recursively.

This section will make use of certain elementary concepts and results concerning polynomial and rational matrices, for example, invariant factors of polynomial matrices, left coprime polynomial matrices, row-reduced polynomial matrices, unimodular polynomial matrices, and Bezout equations. See section 6.3 of the book by Kailath [191] for an exposition of these concepts and the underlying theory.

To keep the exposition simple, only the distinct point and the single multiple point interpolation problems will be considered. The tools presented, however, are applicable in the general case.

The data in terms of time series

The interpolation array $\mathbb{P}$ defined by (4.76) can be interpreted in terms of time functions. To the distinct point array $\mathbb{P}$ defined by (4.79) we associate the following exponential time series (i.e., functions of time):

$$\mathbb{D} = \left\{ \mathbf{w}_k(t) = \begin{pmatrix} \mathbf{u}_k(t) \\ -\mathbf{y}_k(t) \end{pmatrix},\ \mathbf{u}_k(t) = e^{s_k t},\ \mathbf{y}_k(t) = \phi_k e^{s_k t},\ t \geq 0,\ k = 1, \ldots, N \right\}. \tag{4.93}$$

We will also consider the (unilateral) Laplace transform of these time series; in particular, we will consider the $2 \times N$ matrix of rational entries whose kth column is the transform of $\mathbf{w}_k(t)$:

$$\mathbf{W}(s) = [\mathbf{W}_1(s) \ \cdots \ \mathbf{W}_N(s)]\,, \ \text{ where } \ \mathbf{W}_k(s) = \frac{1}{s - s_k} \begin{pmatrix} 1 \\ -\phi_k \end{pmatrix}, \qquad k = 1, \ldots, N. \tag{4.94}$$

It is easy to see that $\mathbf{W}(s)$ has a realization,

$$\mathbf{W}(s) = \mathbf{C}(s\mathbf{I}_N - \mathbf{A})^{-1}, \tag{4.95}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -\phi_1 & -\phi_2 & \cdots & -\phi_N \end{bmatrix} \in \mathbb{C}^{2 \times N}, \ \ \mathbf{A} = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_N \end{bmatrix} \in \mathbb{C}^{N \times N}. \tag{4.96}$$

Since the interpolation points are distinct, i.e., $s_i \neq s_j$, $i \neq j$, the pair $(\mathbf{C}, \mathbf{A})$ is observable.

For the single multiple-point interpolation array $\mathbb{P}$ defined by (4.80), a similar construction holds. Let $\mathbf{p}_k$ be the vector-valued polynomial function

$$\mathbf{p}_k(t) = \begin{pmatrix} 1 \\ -\phi_0 \end{pmatrix} \frac{t^{k-1}}{(k-1)!} + \cdots + \begin{pmatrix} 0 \\ -\frac{\phi_{0j}}{j!} \end{pmatrix} \frac{t^{k-j-1}}{(k-j-1)!} + \cdots + \begin{pmatrix} 0 \\ -\frac{\phi_{0,k-1}}{(k-1)!} \end{pmatrix},\ k = 1, \ldots, N.$$

The time series in this case are polynomial-exponential:

$$\mathbb{D} = \left\{ \mathbf{w}_k(t) = \begin{pmatrix} \mathbf{u}_k(t) \\ -\mathbf{y}_k(t) \end{pmatrix} = e^{s_0 t}\, \mathbf{p}_k(t),\ t \geq 0,\ k = 1, \ldots, N \right\}. \tag{4.97}$$

A straightforward calculation yields the following realization for the (unilateral) Laplace transform of this set of time series $\mathbf{W}(s) = [\mathbf{W}_1(s) \cdots \mathbf{W}_N(s)] = \mathbf{C}(s\mathbf{I}_N - \mathbf{A})^{-1}$, where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_0 & -\frac{\phi_{01}}{1!} & -\frac{\phi_{02}}{2!} & \cdots & -\frac{\phi_{0,N-1}}{(N-1)!} \end{bmatrix} \in \mathbb{C}^{2 \times N},$$

$$\mathbf{A} = \begin{bmatrix} s_0 & 1 & & & \\ & s_0 & 1 & & \\ & & \ddots & \ddots & \\ & & & s_0 & 1 \\ & & & & s_0 \end{bmatrix} \in \mathbb{C}^{N \times N}.$$

Again by construction, the pair $(\mathbf{C}, \mathbf{A})$ is observable.

The *realization problem* discussed in section 4.4 can be expressed in terms of the rational interpolation problem. Given the scalar Markov parameters $h_0, h_1, \dots, h_{N-1}$, we seek to determine all rational functions $\phi(s)$ whose behavior at infinity (i.e., formal power series) is

$$\phi(s) = h_0 + h_1 s^{-1} + \cdots + h_{N-1} s^{-N+1} + \cdots.$$

By introducing s^{-1} as the new variable, the behavior at infinity is transformed into the behavior at zero and the Markov parameters become *moments*:

$$\tilde{\phi}(s) = \phi(s^{-1}) = h_0 + h_1 s + \cdots + h_{N-1} s^{N-1} + \cdots.$$

Consequently, $h_k = \frac{1}{k!}\frac{d^k\tilde{\phi}}{dt^k}|_{s=0}$, i.e., the realization problem is equivalent to a rational interpolation problem where all the data are provided at zero. From the above considerations, the corresponding time series $\mathbf{W} = [\mathbf{W}_1 \cdots \mathbf{W}_N]$ can be expressed in terms of (4.95), where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -h_0 & -h_1 & -h_2 & \cdots & -h_{N-1} \end{bmatrix} \in \mathbb{R}^{2\times N}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{N\times N}. \tag{4.98}$$

Interpolation in terms of the time series data

Given the polynomial $\mathbf{r}(s) = \sum_{i=0}^{k} r_i s^i$, $r_i \in \mathbb{R}$, we will denote by $\mathbf{r}(\frac{d}{dt})$ the constant coefficient differential operator

$$\mathbf{r}\left(\frac{d}{dt}\right) = r_0 + r_1\frac{d}{dt} + r_2\frac{d^2}{dt^2} + \cdots + r_k\frac{d^k}{dt^k}.$$

The following is a characterization of rational interpolants in terms of the time series in both the time and the frequency domains.

Proposition 4.57. *With the notation introduced above, consider the rational function* $\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$, *where* $\mathbf{n}$, $\mathbf{d}$ *are coprime. This rational function interpolates the points of the array* $\mathbb{P}$ *defined by* (4.79) *if and only if one of the following equivalent conditions holds:*

$$\left[\mathbf{n}\left(\frac{d}{dt}\right) \quad \mathbf{d}\left(\frac{d}{dt}\right)\right] \mathbf{w}_k(t) = 0, \qquad t \geq 0, \tag{4.99}$$

$$[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}_k(s) = \mathbf{r}_k(s) \tag{4.100}$$

for $k = 1, \dots, N$, *where* $\mathbf{r}_k(s)$ *is a polynomial.*

Equation (4.99) provides a *time domain characterization*, while (4.100) provides a *frequency domain characterization* of rational interpolants.

Proof. We will give the proof only for the distinct point interpolation problem. From (4.99), given the definition of the time series $\mathbf{w}_k$, follows $(\mathbf{n}(s_k) - \phi_k \mathbf{d}(s_k))\, e^{s_k t} = 0$; since this must hold for all $t \geq 0$, the necessity and sufficiency of the interpolation conditions $\frac{\mathbf{n}(s_k)}{\mathbf{d}(s_k)} = \phi_k$ follow for all k. If we take the unilateral Laplace transform of (4.99), we obtain (4.100), and in particular $\frac{\mathbf{n}(s)-\phi_k \mathbf{d}(s)}{s-s_k} = \mathbf{r}_k(s)$, where $\mathbf{r}_k(s)$ is a polynomial resulting from initial conditions of $\mathbf{y}_k$ at $t = 0^-$. Thus the expression on the left-hand side is a polynomial if and only if $\phi(s_k) = \phi_k$ for all k. □

The solution of Problem (A)

From the frequency domain representation (4.95) of the data, we construct the pair of polynomial matrices $\Xi(s)$, $\Theta(s)$ of size $2 \times N$, 2×2, respectively, such that $\det \Theta(s) \neq 0$, and

$$\mathbf{W}(s) = \Theta(s)^{-1}\Xi(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}. \tag{4.101}$$

The above operation consists of computing a *left* polynomial denominator $\Theta(s)$ for the data $\mathbf{W}(s)$. Moreover, the polynomial matrices $\Theta(s)$, $\Xi(s)$ must be *left coprime*, that is, every nonsingular polynomial matrix $\mathbf{L}(s)$ such that $\mathbf{P} = \mathbf{L}\tilde{\mathbf{P}}$ and $\mathbf{Q} = \mathbf{L}\tilde{\mathbf{Q}}$, for appropriate polynomial matrices $\tilde{\mathbf{P}}$, $\tilde{\mathbf{Q}}$, is unimodular, that is, its determinant is a nonzero constant. A consequence of left coprimeness is the existence of polynomial matrices $\mathbf{P}(s)$, $\mathbf{Q}(s)$ of size 2×2, $N \times 2$, respectively, such that the so-called *Bezout equation* is satisfied:

$$\Theta(s)\mathbf{P}(s) + \Xi(s)\mathbf{Q}(s) = \mathbf{I}_2.$$

A Θ constructed this way has the following properties.

Proposition 4.58. *The matrix $\Theta(s)$ constructed above satisfies the following:* **(a)** *its invariant factors are* 1 *and* $\chi(s) = \Pi_i(s - s_i)$*;* **(b)** *its* $(1,2)$, $(2,2)$ *entries* $\theta_{12}(s)$, $\theta_{22}(s)$ *are coprime.*

Proof. Because of the observability of the pair $\mathbf{C}, \mathbf{A}$ and the coprimeness of Ξ, Θ, the polynomial matrices $s\mathbf{I} - \mathbf{A}$ and $\Theta(s)$ have a *single* nonunity invariant factor that is the *same*, namely, the characteristic polynomial of $\mathbf{A}$: $\chi(s) = \Pi_{i=1}^{N}(s - s_i)$. (See, e.g., [8] or Chapter 6 of [191].) Therefore, after a possible normalization by a (nonzero) constant,

$$\det \Theta(s) = \chi(s). \tag{4.102}$$

Let θ_{ij} denote the (i, j)th entry of Θ. The ith column of (4.101) yields the equation

$$(s - s_i)[\Xi(s)]_{(:,i)} = \begin{pmatrix} \theta_{11}(s) & \theta_{12}(s) \\ \theta_{21}(s) & \theta_{22}(s) \end{pmatrix} \begin{pmatrix} 1 \\ -\phi_i \end{pmatrix} = \begin{pmatrix} \theta_{11}(s) - \phi_i\theta_{12}(s) \\ \theta_{21}(s) - \phi_i\theta_{22}(s) \end{pmatrix},$$
$$i = 1, \ldots, N.$$

Evaluating this expression at $s = s_i$, we obtain $\theta_{11}(s_i) = \phi_i\theta_{12}(s_i)$, $\theta_{21}(s_i) = \phi_i\theta_{22}(s_i)$, $i = 1, \ldots, N$. Because of (4.102), if θ_{12}, θ_{22} were not coprime, their greatest common divisor would have to be a product of terms $(s - s_i)$, where the s_i are the interpolation points. Therefore, by the latter equation, all four entries of Θ would have the same common factor. This, however, contradicts the fact that one of the two invariant factors of Θ is equal to 1. The desired coprimeness is thus established. □

Lemma 4.59. *The rational function* $\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$, *with* $\mathbf{n}, \mathbf{d}$ *coprime, is an interpolant for the array* $\mathbb{P}$ *if and only if there exist coprime polynomials* $\mathbf{a}(s)$, $\mathbf{b}(s)$ *such that*

$$[\mathbf{n}(s) \quad \mathbf{d}(s)] = [\mathbf{a}(s) \quad \mathbf{b}(s)]\,\Theta(s) \tag{4.103}$$

and

$$\mathbf{a}(s_i)\theta_{12}(s_i) + \mathbf{b}(s_i)\theta_{22}(s_i) \neq 0, \qquad i = 1, \ldots, N. \tag{4.104}$$

Proof. If the numerator and the denominator of ϕ satisfy (4.103), there holds

$$[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}(s) = [\mathbf{a}(s) \quad \mathbf{b}(s)]\Xi(s).$$

The latter expression is polynomial, and hence by Proposition 4.57 ϕ is an interpolant. Also, $\mathbf{a}$, $\mathbf{b}$ are coprime; otherwise $\mathbf{n}$, $\mathbf{d}$ would not be coprime, which is a contradiction. Finally, we notice that $\mathbf{d} = \mathbf{a}\theta_{12} + \mathbf{b}\theta_{22}$, which implies that conditions (4.104) must be satisfied. This is possible due to part (b) of the proposition above.

Conversely, let ϕ be an interpolant. According to Proposition 4.57, $[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}(s)$ is a polynomial row vector. From the Bezout equation, it follows that $\mathbf{P}(s) + \mathbf{W}(s)\mathbf{Q}(s) = [\Theta(s)]^{-1}$. Multiplying this relationship on the left by the row vector $[\mathbf{n}(s) \quad \mathbf{d}(s)]$, we conclude that $[\mathbf{n}(s) \quad \mathbf{d}(s)][\Theta(s)]^{-1}$ must be a polynomial row vector, i.e., there exist polynomials $\mathbf{a}$, $\mathbf{b}$ such that (4.103) holds. Furthermore, the coprimeness of $\mathbf{n}$, $\mathbf{d}$ implies the coprimeness of $\mathbf{a}$, $\mathbf{b}$. □

As shown above, *all* interpolants $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ can be parametrized by means of (4.103), where the parameter $\Gamma = \frac{\mathbf{a}}{\mathbf{b}}$ is an arbitrary rational function subject to constraints (4.104). We can interpret Θ as a two-port with inputs $\mathbf{u}$, $\hat{\mathbf{u}}$, and outputs $\mathbf{y}$, $\hat{\mathbf{y}}$ (see Figure 4.4):

$$\begin{pmatrix} \hat{\mathbf{y}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix}$$

Γ can be seen as relating $\hat{\mathbf{u}}$ and $\hat{\mathbf{y}}$ in the following manner: $\mathbf{b}(\frac{d}{dt})\hat{\mathbf{u}} = \mathbf{a}(\frac{d}{dt})\hat{\mathbf{y}}$. Then the parametrization of *all* solutions ϕ can be interpreted as a linear system described by the linear, constant coefficient, differential equation $\mathbf{d}(\frac{d}{dt})\mathbf{y} = \mathbf{n}(\frac{d}{dt})\mathbf{u}$. This system, in turn, can be represented by means of a feedback interconnection between Θ and Γ, where Θ is

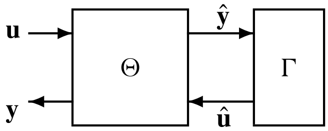

Figure 4.4. *Feedback interpretation of the parametrization of all solutions of the rational interpolation problem.*

fixed and Γ is arbitrary, subject to (4.104). As a consequence of the above interpretation, Θ is called the generating system or generating matrix of the *rational interpolation problem* at hand. Furthermore, (4.103) shows that every interpolant can be expressed in terms of the linear fractional representation,

$$\phi(s) = \frac{\mathbf{a}(s)\theta_{11}(s) + \mathbf{b}(s)\theta_{21}(s)}{\mathbf{a}(s)\theta_{12}(s) + \mathbf{b}(s)\theta_{22}(s)}.$$

Remark 4.5.4. The coprimeness constraint on $\mathbf{n}$, $\mathbf{d}$ has been expressed equivalently as a coprimeness constraint on the parameter polynomials $\mathbf{a}$, $\mathbf{b}$ together with constraints (4.104). The former is a nonlinear constraint in the space of coefficients of $\mathbf{a}$, $\mathbf{b}$; it is automatically satisfied in the case of minimal interpolants discussed below. Constraints (4.104) are linear in the coefficients of $\mathbf{a}$, $\mathbf{b}$. Examples will be worked out later.

To tackle Problem **(A)**, the concept of a *row reduced* generating matrix is needed. Let ν_i be the degree of the ith row of Θ. The row-wise highest coefficient matrix $[\Theta]_{\mathrm{hr}}$ is a 2×2 constant matrix, whose (i, j)th entry is the coefficient of the term s^{ν_i} of the polynomial $\theta_{i,j}(s)$. We will call Θ *row reduced* if $[\Theta]_{\mathrm{hr}}$ is nonsingular. Notice that the row degrees of Θ and the degree of its determinant, which by (4.102) is N, satisfy $\nu_1 + \nu_2 \geq N$, in general. An equivalent characterization of row reducedness is that the row degrees of Θ satisfy $\nu_1 + \nu_2 = N$.

The matrix Θ in (4.101) is unique, up to left multiplication with a unimodular matrix (which is a polynomial matrix with constant nonzero determinant). We use this freedom to transform Θ into *row reduced* form. For simplicity we will use the same symbol, namely, Θ, to denote the row reduced version of this matrix. Let the corresponding row degrees be

$$\kappa_1 = \deg\ (\theta_{11}(s) \quad \theta_{12}(s)) \leq \kappa_2 = \deg\ (\theta_{21}(s) \quad \theta_{22}(s)), \qquad \kappa_1 + \kappa_2 = N.$$

The row degrees of row reduced polynomial matrices have two important properties. First, although the row reduced version of any polynomial matrix is nonunique, the corresponding row degrees are *unique*. Second, because of the so-called predictable-degree property of row reduced polynomial matrices (see, e.g., Chapter 6 of [191]), the degree of $\mathbf{r}(s)\Theta(s)$, with Θ row reduced and $\mathbf{r}$ some polynomial row vector with coprime entries, either can be κ_1 or be greater than or equal to κ_2.

Construction of a row reduced Θ using $\mathcal{O}(\mathbf{C}, \mathbf{A})$. We will now show how a row reduced generating matrix can be constructed directly from the observability matrix $\mathcal{O}(\mathbf{C}, \mathbf{A})$. The procedure involves the determination of two linear dependencies among the rows of this observability matrix, which leads to the two *observability indices* κ_i, $i = 1, 2$, of the pair $(\mathbf{C}, \mathbf{A})$. (For details, see [8] or [191].)

Let $\mathbf{c}_i$ denote the ith row of $\mathbf{C}$, $i = 1, 2$. For simplicity, we will assume that working from top to bottom of the observability matrix, $\mathbf{c}_2\mathbf{A}^{\kappa_1}$ is the first row of $\mathcal{O}$ that is linearly dependent on the preceding ones, i.e., $\mathbf{c}_i\mathbf{A}^j$, $i = 1, 2$, $j \leq \kappa_1$. Then, because of observability, the next row to be linearly dependent on the previous ones will be $\mathbf{c}_1\mathbf{A}^{\kappa_2}$, where $\kappa_1 < \kappa_2$, and $\kappa_1 + \kappa_2 = N$:

$$\begin{aligned}
\mathbf{c}_2\mathbf{A}^{\kappa_1} &= \textstyle\sum_{i=0}^{\kappa_1} \alpha_i \mathbf{c}_1\mathbf{A}^i &+& \textstyle\sum_{j=0}^{\kappa_1-1} \beta_j \mathbf{c}_2\mathbf{A}^j, \\
\mathbf{c}_1\mathbf{A}^{\kappa_2} &= \textstyle\sum_{i=0}^{\kappa_2-1} \gamma_i \mathbf{c}_1\mathbf{A}^i &+& \textstyle\sum_{j=0}^{\kappa_1} \delta_j \mathbf{c}_2\mathbf{A}^j.
\end{aligned}$$

It follows that Θ can be read off of the above relationships:

$$\Theta(s) = \begin{pmatrix} -\sum_{i=0}^{\kappa_1} \alpha_i s^i & s^{\kappa_1} - \sum_{i=0}^{\kappa_1 - 1} \beta_j s^j \\ s^{\kappa_2} - \sum_{i=0}^{\kappa_2 - 1} \gamma_i s^i & -\sum_{i=0}^{\kappa_1} \delta_j s^j \end{pmatrix}. \tag{4.105}$$

Clearly, $\det [\Theta]_{\text{hr}} = -1$, which implies that Θ is row reduced.

Combining the preceding lemma with the above considerations, we obtain the main result, which provides the solution of Problem (A). This result was proved in Antoulas and Willems [15] as well as Antoulas et al. [22].

Theorem 4.60. *Consider* Θ *defined by* (4.101), *which is row reduced, with row degrees* $\kappa_1 \leq \kappa_2$.

(i) *If* $\kappa_1 < \kappa_2$ *and* θ_{11}, θ_{21} *are coprime,*

$$\phi^{min}(s) = \frac{\theta_{11}(s)}{\theta_{12}(s)}, \qquad \delta(\phi^{min}) = \kappa_1,$$

is the unique minimal interpolant. Furthermore, there are no interpolants of complexity between κ_1 *and* κ_2.

(ii) *Otherwise, there is a family of interpolating functions of minimal complexity which can be parametrized as follows:*

$$\phi^{min}(s) = \frac{\theta_{21}(s) + \mathbf{a}(s)\theta_{11}(s)}{\theta_{22}(s) + \mathbf{a}(s)\theta_{12}(s)}, \qquad \delta(\phi^{min}) = \kappa_2 = N - \kappa_1,$$

where the polynomial $\mathbf{a}(s)$ *satisfies*

$$\deg \mathbf{a} = \kappa_2 - \kappa_1, \quad \theta_{22}(s_i) + \mathbf{a}(s_i)\theta_{12}(s_i) \neq 0, \qquad i = 1, \ldots, N.$$

(iii) *In both cases* (i) *and* (ii), *there are families of interpolants* $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ *of every degree* $\kappa \geq \kappa_2$, *satisfying* (4.103), *where* $\deg \mathbf{a} = \kappa - \kappa_1$, $\deg \mathbf{b} = \kappa - \kappa_2$, *and* $\mathbf{a}$, $\mathbf{b}$ *are coprime.*

Corollary 4.61. Proper rational and polynomial interpolants. *The interpolants above are proper rational provided that* $\mathbf{a}$, $\mathbf{b}$ *satisfy* $[\mathbf{b}(s)\theta_{22}(s) + \mathbf{a}(s)\theta_{12}(s)]_h \neq 0$, *where* $[\mathbf{r}]_h$ *is used to denote the coefficient of the highest power of the polynomial* $\mathbf{r}$. *All polynomial interpolants* $\phi(s) = \mathbf{n}(s)$ *are given by* $-\mathbf{n}(s) = \mathbf{b}(s)\theta_{21}(s) + \mathbf{a}(s)\theta_{11}(s)$, *where the polynomials* $\mathbf{a}$, $\mathbf{b}$ *satisfy the Bezout equation* $\mathbf{b}(s)\theta_{22}(s) + \mathbf{a}(\mathbf{s})\theta_{12}(s) = 1$.

One polynomial interpolant is the *Lagrange interpolating polynomial* $\ell(s)$ given in the distinct point case by (4.81). It is also worth mentioning that a generating matrix (which is *not* row reduced) can be written in terms of $\ell(s)$ and $\chi(s)$ given by (4.102):

$$\Theta(s) = \begin{pmatrix} \chi(s) & 0 \\ \ell(s) & 1 \end{pmatrix}.$$

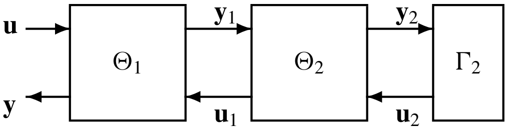

Figure 4.5. *Feedback interpretation of recursive rational interpolation.*

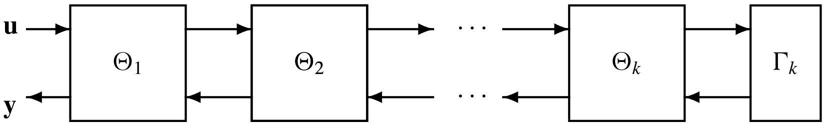

Figure 4.6. *A general cascade decomposition of systems.*

The solution of the unconstrained interpolation problem (4.82) can be obtained in this framework by means of polynomial linear combinations of the rows of this particular generating matrix.

Recursive interpolation

The fact that in the parametrization of all interpolants via the generating system, Γ is arbitrary—except for the avoidance conditions (4.104)—yields, at least conceptually, the solution to the recursive interpolation problem with no additional effort. In particular, if $\mathbb{D} = \mathbb{D}_1 \cup \mathbb{D}_2$, we define Θ_1 as a generating system for the data set $\mathbb{D}_1$ and Θ_2 as a generating system for the modified data set $\Theta_1(\frac{d}{dt})\mathbb{D}_2$. Then the cascade $\Theta = \Theta_2\Theta_1$ of the two generating systems Θ_1 and Θ_2 provides a generating system for $\mathbb{D}$. More generally, Figures 4.5 and 4.6 give a pictorial representation of the solution to the recursive interpolation problem. The *cascade interconnection* is associated with the Euclidean algorithm, the Schur algorithm, the Nevanlinna algorithm, the Berlekamp–Massey algorithm, Darlington synthesis, continued fractions, and Hankel-norm approximation.

Problem (A) was solved recursively in [12]. Earlier, the recursive realization problem with a degree constraint was solved in [9]. The main result of these two papers, which is shown in Figures 4.6 and 4.7, is that a recursive update of the solution of the realization or interpolation problems corresponds to attaching an appropriately defined component to a cascade interconnection of systems.

The scalar case. We conclude this account on recursiveness by making this cascade interconnection explicit for SISO systems. Let the transfer function be $\mathbf{H}(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$, assumed for simplicity, strictly proper. As shown by Kalman [193], recursive realization corresponds to the decomposition of $\mathbf{H}(s)$ is a *continued fraction*:

$$\mathbf{H}(s) = \cfrac{1}{\mathbf{d}_1(s) + \cfrac{1}{\mathbf{d}_2(s) + \cfrac{1}{\mathbf{d}_3(s) + \cfrac{1}{\ddots + \frac{1}{\mathbf{d}_N(s)}}}}}, \tag{4.106}$$

where $\mathbf{d}_i$ are polynomials of degree κ_i and $\sum_i \kappa_i = \deg \mathbf{d} = n$. In the generating system framework, we have

$$\Theta = \prod_{i=1}^{N} \Theta_i, \quad \text{where } \Theta_i = \begin{pmatrix} 0 & 1 \\ 1 & -\mathbf{d}_i \end{pmatrix}.$$

This cascade decomposition can be simplified in this case, as shown in Figure 4.7.

Furthermore, a state-space realization follows from this decomposition. For details, see [146] and [10]. Here we will illustrate the generic case only, that is, the case where $\kappa_i = 1$ for all i, and hence $N = n$. By appropriate scaling, we will assume that the polynomials are monic (coefficient of highest degree is one); (4.106) becomes

$$\mathbf{H}(s) = \cfrac{\beta_1}{s + \alpha_1 + \cfrac{\beta_2}{s + \alpha_2 + \cfrac{\beta_3}{\ddots + \frac{\beta_n}{s + \alpha_n}}}}. \tag{4.107}$$

The following triple is a minimal realization of $\mathbf{H}(s)$, whenever it has a generic decomposition of the form (4.107):

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccccc|c} -\alpha_1 & \beta_2 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -\alpha_2 & \beta_3 & \cdots & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & -\alpha_{n-1} & \beta_n & 0 \\ 0 & 0 & \cdots & 0 & -1 & -\alpha_n & 0 \\ \hline \beta_1 & 0 & \cdots & 0 & 0 & 0 & \end{array}\right). \tag{4.108}$$

Notice that $\mathbf{A}$ is in *tridiagonal* form, while $\mathbf{B}$ and $\mathbf{C}^*$ are multiples of the first canonical unit vector in $\mathbb{R}^n$. To summarize, we have seen important connections between the following topics:

(a) realization/interpolation,

(b) cascade/feedback interconnection,

(c) linear fractions,

(d) continued fractions,

(e) tridiagonal state space realizations.

Thus partial realization consists of truncating the tail of the continued fraction or, equivalently, of the cascade decomposition of the system or of the tridiagonal state space realization. These issues will play a role in Chapter 10, where an iterative method, the so-called Lanczos procedure, of constructing this tridiagonal realization will be of central importance.

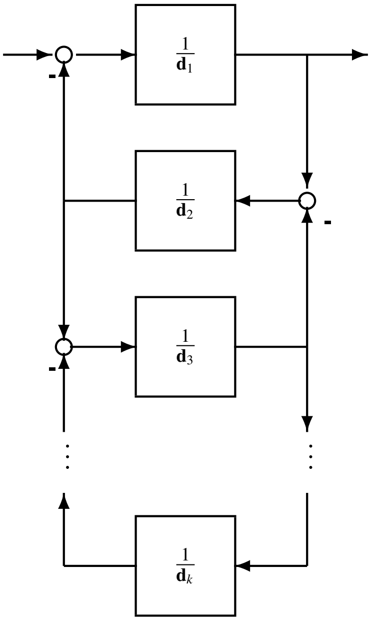

Figure 4.7. *Cascade (feedback) decomposition of scalar systems.*

The solution of Problems (B) and (C)

Given the array $\mathbb{P}$ defined by (4.79)—again we restrict our attention to the distinct point case—together with $\mu > 0$, we wish to find out whether there exist interpolants that are stable (poles in the left half of the complex plane) with magnitude bounded by μ on the imaginary axis. The tool for investigating the existence issue is the *Nevanlinna–Pick* matrix

$$\Pi_\mu = \begin{bmatrix} \frac{\mu^2-\bar{\phi}_1\phi_1}{\bar{s}_1+s_1} & \cdots & \frac{\mu^2-\bar{\phi}_1\phi_N}{\bar{s}_1+s_N} \\ \vdots & \ddots & \vdots \\ \frac{\mu^2-\bar{\phi}_N\phi_1}{\bar{s}_N+s_1} & \cdots & \frac{\mu^2-\bar{\phi}_N\phi_N}{\bar{s}_N+s_N} \end{bmatrix},$$

where $\bar{(\cdot)}$ denotes complex conjugation. A solution exists if and only if this matrix is positive (semi-) definite $\Pi_\mu \geq 0$. Write $\Pi_\mu = \mu^2\Pi_1 - \Pi_2$, where $\Pi_1 > 0$, $\Pi_2 > 0$. Let μ_i^2 be the eigenvalues of $\Pi_1^{-1}\Pi_2$, with μ_1^2 the largest. As long as $\mu > \mu_1$, $\Pi_\mu > 0$, for $\mu = \mu_1$ it becomes semidefinite and for $\mu < \mu_1$ it is indefinite. Thus the smallest norm for which there exist solutions to Problem (B) is the square root of the largest eigenvalue of $\Pi_1^{-1}\Pi_2$.

In [26] it was shown that the Nevanlinna–Pick interpolation problem can be transformed into an interpolation problem *without norm constraint*. This is achieved by adding the so-called mirror image interpolation points to the original data. In terms of trajectories, the mirror image set $\hat{\mathbb{D}}$ of $\mathbb{D}$, defined by (4.93), is

$$\mathbb{D} = \left\{ \begin{pmatrix} 1 \\ -\phi_i \end{pmatrix} e^{s_i t},\ i = 1, \ldots, N \right\} \text{ and } \hat{\mathbb{D}} = \left\{ \begin{pmatrix} -\bar{\phi}_i \\ 1 \end{pmatrix} e^{-\bar{s}_i t},\ i = 1, \ldots, N \right\}.$$

The augmented data set is thus $\mathbb{D}_{aug} = \mathbb{D} \cup \hat{\mathbb{D}}$, and the corresponding pair of matrices is

$$\mathbf{C}_{aug} = (\mathbf{C} \quad \mathbf{J}\bar{\mathbf{C}}), \quad \mathbf{A}_{aug} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\bar{\mathbf{A}} \end{pmatrix}, \quad \text{where } \mathbf{J} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and $\bar{(\cdot)}$ applied to a matrix denotes complex conjugation (without transposition). We now construct left coprime polynomial matrices Θ_{aug}, Ξ_{aug} such that $\mathbf{C}_{aug}(s\mathbf{I} - \mathbf{A}_{aug})^{-1} = [\Theta_{aug}(s)]^{-1}\Xi_{aug}(s)$. The main result is that the generating system for the data set $\mathbb{D}_{aug}$ is the generating system that solves Problem (B), provided that the parameters $\mathbf{a}$, $\mathbf{b}$ are appropriately restricted. For simplicity of notation, let the entries of Θ_{aug} be denoted by θ_{ij}. The following result can be proved using the results in [26].

Theorem 4.62. Classification by norm. *Let Θ_{aug} be as defined above. The interpolants $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ of $\mathbb{D}$ with norm of ϕ less than or equal to μ on the imaginary axis are given, if they exist, by $\phi = \frac{\mathbf{a}\theta_{11}+\mathbf{b}\theta_{21}}{\mathbf{a}\theta_{12}+\mathbf{b}\theta_{22}}$, where the magnitude of $\frac{\mathbf{a}}{\mathbf{b}}$ on the imaginary axis must be at most μ.*

The above result achieves an *algebraization* of the Nevanlinna–Pick interpolation problem. For related treatments of Problem (B), see [271] and [195]. See also the book by Ball, Gohberg, and Rodman [38]. We conclude this part by mentioning that Problem (C), that is, the problem of positive real interpolation, can also be turned into an unconstrained interpolation problem by adding an appropriately defined mirror-image set of points. For details, see [27], [234], [309].

Concluding remarks and generalizations

The problem of *rational interpolation* has a long history. It was only recently recognized, however, as a problem that generalizes the realization problem. One can distinguish two approaches: state-space and polynomial. The generalization of the state-space framework from the realization to the rational interpolation problem is due to Antoulas and Anderson [25], [23], [26] and Anderson and Antoulas [24]. Therein, the Löwner matrix replaces and generalizes the Hankel matrix as the main tool. The generating system or polynomial approach to rational interpolation with the complexity (McMillan degree) as constraint was put forward in Antoulas and Willems [15] and Antoulas et al. [22].

The above results can be generalized considerably. Consider the array consisting of the distinct interpolation data s_i, V_i, Y_i of size 1×1, $r_i \times p$, $r_i \times m$, respectively, satisfying $s_i \neq s_j$, $i \neq j$, and rank $V_i = r_i \leq p$, of $i = 1, \ldots, N$. The *left tangential* or *left directional interpolation problem* consists of finding all $p \times m$ rational matrices $\Phi(s)$ satisfying $V_i\Phi(s_i) = Y_i$, $i = 1, \ldots, N$, keeping track of their complexity, norm boundedness, or positive realness at the same time. In this case, the generating matrix Θ is a square of size $p + m$, and there are $p + m$ (observability) indices that enter the picture. The *right tangential* or *right directional interpolation problem*, as well as the *bitangential* or *bidirectional interpolation problem*, can be defined similarly. The solution of Problem (A) in all its matrix and tangential versions has been given in the generating system framework in [22]. For a general account on the generating system approach to rational interpolation, see [38].

Examples

Example 4.63. Consider the data array containing four pairs:

$$\mathbb{P} = \{(0,0),\ (1,3),\ (2,4),\ (1/2,2)\}.$$

According to (4.96),

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -3 & -4 & -2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 2 & \\ & & & \frac{1}{2} \end{pmatrix}.$$

Following the construction leading to (4.105), we need the observability matrix of the $(\mathbf{C}, \mathbf{A})$ pair,

$$\mathcal{O}_4 = \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 0 & -3 & -4 & -2 \\ \hline 0 & 1 & 2 & \frac{1}{2} \\ 0 & -3 & -8 & -1 \\ \hline 0 & 1 & 4 & \frac{1}{4} \\ 0 & -3 & -16 & -\frac{1}{2} \\ \hline 0 & 1 & 8 & \frac{1}{8} \\ 0 & -3 & -32 & -\frac{1}{4} \end{array}\right).$$

Examining the rows from top to bottom, the first linear dependence occurring in $\mathcal{O}_4$ is that of the fourth row on the previous ones:

$$\mathbf{c}_2\mathbf{A} + 6\mathbf{c}_1\mathbf{A} + \mathbf{c}_2 = 0.$$

It follows that $\kappa_1 = 1$. Therefore, $\kappa_2 = N - \kappa_1 = 3$. This means that the next linear dependence is that of the seventh row on the previous ones:

$$2\mathbf{c}_1\mathbf{A}^3 - 9\mathbf{c}_1\mathbf{A}^2 + 4\mathbf{c}_1\mathbf{A} - 2\mathbf{c}_2\mathbf{A} + \mathbf{c}_2 = 0.$$

According to formula (4.105), a row reduced generating matrix Θ is therefore

$$\Theta(s) = \begin{pmatrix} 6s & s+1 \\ s(s-4)(2s-1) & -(2s-1) \end{pmatrix}.$$

By (i) of the theorem, since $6s$ and $s+1$ are coprime polynomials, there is a unique minimal interpolant with McMillan degree 1, namely,

$$\phi^{min}(s) = \frac{6s}{s+1}.$$

Furthermore, there are no interpolants of degree 2. The next family of interpolants has McMillan degree 3. It can be parametrized in terms of the second-degree polynomial $\mathbf{p}(s) = p_0 + p_1 s + p_2 s^2$, as follows:

$$\phi(s) = \frac{s(s-4)(2s-1) + (p_2 s^2 + p_1 s + p_0)6s}{-(2s-1) + (p_2 s^2 + p_1 s + p_0)(s+1)}.$$

The coefficients of $\mathbf{p}$ must satisfy the constraints (4.104) in $\mathbb{R}^3$, which in this case turn out to be

$$\begin{aligned}&p_0 + 1 \neq 0,\\ &2p_2 + 2p_1 + 2p_0 - 1 \neq 0,\\ &4p_2 + 2p_1 + p_0 - 1 \neq 0,\\ &p_2 + 2p_1 + 4p_0 \neq 0.\end{aligned}$$

By letting $p_2 = p_1 = 0$ and $p_0 = 2$ in the above family of interpolants, we obtain the Lagrange interpolating polynomial for the data array $\mathbb{P}$:

$$\ell(s) = \frac{s}{3}(2s^2 - 9s + 16).$$

The next family of interpolants has McMillan degree 4. It can be parametrized in terms of a third-degree polynomial $\mathbf{p}$ and a first-degree polynomial $\mathbf{q}$, as follows:

$$\phi(s) = \frac{(p_3 s^3 + p_2 s^2 + p_1 s + p_0)6s + (q_1 s + q_0)s(s-4)(2s-1)}{(p_3 s^3 + p_2 s^2 + p_1 s + p_0)(s+1) - (q_1 s + q_0)(2s-1)}.$$

Firstly, $\mathbf{p}$, $\mathbf{q}$ must be coprime, i.e.,

$$p_3 q_0^3 - p_2 q_0^2 q_1 + p_1 q_0 q_1^2 - p_0 q_1^3 \neq 0.$$

Then, constraints (4.104) must also be satisfied, that is, the free parameters must avoid the following hyperplanes in $\mathbb{R}^6$:

$$\begin{aligned}&p_0 + q_0 \neq 0,\\ &2p_3 + 2p_2 + 2p_1 + 2p_0 - q_1 - q_0 \neq 0,\\ &8p_3 + 4p_2 + 2p_1 + p_0 - 2q_1 - q_0 \neq 0,\\ &p_3 + 2p_2 + 4p_1 + 8p_0 \neq 0.\end{aligned}$$

Example 4.64. *Continuation of Example* 4.63*: Recursive construction of interpolants.* The time series associated with $\mathbb{P}$ are

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ -3 \end{pmatrix} e^t, \quad \mathbf{w}_3 = \begin{pmatrix} 1 \\ -4 \end{pmatrix} e^{2t}, \quad \mathbf{w}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} e^{\frac{t}{2}}.$$

Following Figure 4.6, we will now construct the generating systems Θ_i, $i = 1, 2, 3, 4$, satisfying $\Theta = \Theta_4\Theta_3\Theta_2\Theta_1$; according to (4.102), since $\det \Pi_{i=1}^4 \Theta(s) = \Pi_{i=1}^4 (s - s_i)$, there must hold $\det \Theta_i = s - s_i$, $i = 1, 2, 3, 4$. The generating system that annihilates $\mathbf{w}_1$ is thus

$$\Theta_1(s) = \begin{pmatrix} s & 0 \\ 0 & 1 \end{pmatrix},$$

since the first error time series, defined as $\mathbf{e}_1 = \Theta_1(\frac{d}{dt})\mathbf{w}_1$, is zero: $\mathbf{e}_1 = \mathbf{0}$. The second error time series is defined similarly, namely, $\mathbf{e}_2 = \Theta_1(\frac{d}{dt})\mathbf{w}_2$, and we have $\mathbf{e}_2 = \mathbf{w}_2$; thus

$$\Theta_2(s) = \begin{pmatrix} 3 & 1 \\ 0 & s-1 \end{pmatrix}.$$

The first error time series remains zero, $\Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_1 = 0$, and $\Theta_2(\frac{d}{dt})\mathbf{e}_2 = \Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_2 = \mathbf{0}$; the third error time series is $\Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_3 = \left(\begin{smallmatrix} 2 \\ -4 \end{smallmatrix}\right)e^{2t}$, which implies

$$\Theta_3(s) = \begin{pmatrix} s-2 & 0 \\ 2 & 1 \end{pmatrix} \quad \Rightarrow \quad \Theta_3\Theta_2\Theta_1\left(\frac{d}{dt}\right)\mathbf{w}_i = \mathbf{0}, \qquad i = 1, 2, 3,$$

$$\text{while } \mathbf{e}_4 = \Theta_3\Theta_2\Theta_1\left(\frac{d}{dt}\right)\mathbf{w}_4 = \begin{pmatrix} \frac{3}{4} \\ 0 \end{pmatrix} e^{\frac{t}{2}}.$$

Finally

$$\Theta_4(s) = \begin{pmatrix} 2s-1 & 0 \\ 0 & 1 \end{pmatrix} \quad \Rightarrow \quad \Theta(s) = \Theta_4(s)\Theta_3(s)\Theta_2(s)\Theta_1(s)$$
$$= \begin{pmatrix} 3s(s-2)(2s-1) & (s-2)(2s-1) \\ 6s & s+1 \end{pmatrix}.$$

Although this generating matrix is not the same as the one obtained in the previous example, it differs only by left multiplication with a unimodular matrix. In particular, $\mathbf{U}\Theta_4\Theta_3\Theta_2\Theta_1$, where $\mathbf{U}(s) = \left(\begin{smallmatrix} 0 & 1 \\ \frac{1}{3} & \frac{1-2s}{3} \end{smallmatrix}\right)$ is equal to the generating matrix obtained in the previous example.

Example 4.65. *Realization Example* 4.47 *revisited.* Following the construction leading to (4.105), we need the observability matrix of the $(\mathbf{C}, \mathbf{A})$ pair defined by (4.98):

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & -2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The associated observability matrix is $\mathcal{O}_5$:

$$\mathcal{O}_5(\mathbf{C}, \mathbf{A}) = \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & -2 \\ \hline 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The first row that is linearly dependent on the preceding ones is the sixth; the next is the seventh:

$$\mathbf{c}_1\mathbf{A}^2 - \mathbf{c}_2\mathbf{A}^2 + \mathbf{c}_2\mathbf{A} = 0, \quad \mathbf{c}_1\mathbf{A}^3 + \mathbf{c}_1\mathbf{A}^2 - \mathbf{c}_1\mathbf{A} + 2\mathbf{c}_2\mathbf{A} - \mathbf{c}_2 = 0.$$

The corresponding generating system is

$$\Theta(s) = \begin{pmatrix} s^2 & -s^2 + s \\ s^3 + s^2 - s & 2s - 1 \end{pmatrix}.$$

According to the theory, the minimal interpolant has degree 3, and all minimal interpolants form a two-parameter family obtained by multiplying Θ on the left by $[\alpha s + \beta \quad 1]$:

$$\tilde{\phi}^{min}(s) = -\frac{(\alpha s + \beta)s^2 + (s^3 + s^2 - s)}{(\alpha s + \beta)(s^2 - s) - (2s - 1)}, \qquad \alpha,\ \beta \in \mathbb{R}.$$

According to (4.104), the parameters α, β must be such that the denominator of the above expression is nonzero for $s = 0$; therefore, since the value of the denominator for $s = 0$ is 1, these two parameters are *free*. And to obtain matching of the Markov parameters, we replace s by s^{-1}:

$$\phi^{min}(s) = \tilde{\phi}^{min}(s^{-1}) = -\frac{(\alpha + \beta s) + (1 + s - s^2)}{(\alpha + \beta s)(1 - s) - 2s^2 + s^3} = \frac{s^2 - (\beta + 1)s - (\alpha + 1)}{s^3 - (\beta + 2)s^2 + (\beta - \alpha)s + \alpha}.$$

It is readily checked that the power series expansion of ϕ around infinity is

$$\phi^{min}(s) = s^{-1} + s^{-2} + s^{-3} + 2s^{-4} + (\beta + 4)s^{-5} + (\beta^2 + 4\beta + \alpha + 8)s^{-6} + \cdots .$$

4.6 Chapter summary

The purpose of this chapter was to familiarize the reader with fundamental concepts from system theory. Three topics were discussed, namely, the external description, the internal description, and the realization/interpolation problem. The last section on the rational interpolation problem can be omitted on first reading. It forms a natural extension and generalization of the realization problem, which is interesting in its own right. Many aspects discussed in this section, however, will turn out to be important in what follows.

The section on the external description states that a linear time-invariant system is an operator (map) which assigns inputs to outputs. In particular, it is the convolution operator defined in terms of its kernel, which is the impulse response of the system. In the frequency domain, the external description is given in terms of the transfer function, whose series expansion around infinity yields the Markov parameters.

The internal description introduces, besides the external variables (that is, the input and the output), an internal variable called the state. The internal description thus consists of a set of equations describing how the input affects the state and another set describing how the output is obtained from the input and the state. The former set contains differential or difference equations, which are first order in the state (only the first derivative or shift of the state is involved), and it describes the dynamics of the system. The latter set contains algebraic equations (no derivatives or shifts of the state are allowed). The internal description is thus completely determined by means of the quadruple of matrices $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, defined by (4.13).

To get a better understanding of a dynamical system, two *structural concepts* are introduced, namely, those of *reachability* and *observability*. The former is used to investigate the extent to which the states can be influenced by manipulating the input. The latter is used to assess the influence of the state on the output. The *reachability* and *observability* matrices are used to quantitatively answer these questions. Besides these matrices, the so-called gramians can be used as well. Their advantage lies in the fact that they are square symmetric and semidefinite. Furthermore, if the assumption of stability is made, the *infinite gramians* arise, which can be computed by solving appropriately defined linear matrix equations, the so-called Lyapunov equations. These equations play a fundamental role in the computation of approximants to a given system. They will be the subject of detailed study in Chapter 6.

Eliminating the state from the internal description is straightforward and leads to the external description. The converse, however, that is deducing an internal description from the external description is nontrivial. It involves the *construction of state*, and if the data consist of *all* Markov parameters, it is known as the *realization problem*. Conditions for solvability and ways of constructing solutions (if they exist) are discussed in section 4.4. If the data consist of a partial set of Markov parameters, the existence of solutions being no longer an issue, the parametrization of all (minimal complexity) solutions becomes important; this is discussed briefly.

The final section on rational interpolation provides a generalization of the realization results in several ways. First, the input-output data need not be confined to Markov parameters; instead, samples of the transfer function at arbitrary points in the complex plane are allowed. Second, both state-space and polynomial ways of constructing all solutions of a given complexity are discussed. Finally it is pointed out that the machinery which was set up (generating system method) can be used to solve the problems of constructing systems with special properties, like bounded real or positive real transfer function.