

Chapter 2

Introduction to State Space Theory

State space theory deals with dynamical models describing both the internal dynamics of a given physical process and the interaction of the process with the outside world. In this chapter we introduce the general notion of a *dynamical system* and set the basis for the study of various important system classes.

We emphasize that for us a *dynamical system is a mathematical model* and hence should be carefully distinguished from the physical process for which it is a model. Dynamical systems of different types may be used as models of one and the same physical process. Nevertheless it will sometimes be convenient to use the word “system” for the real physical process described by the dynamical model and in this case we shall add the epithet “real” or “physical” whenever this is necessary for a clear distinction.

In Section 2.1 we begin with a description of the components which constitute the mathematical concept of a dynamical system and then give a very general definition. This definition incorporates the basic common structure of most dynamic state space models in current use and in particular comprises all the state space models described in Chapter 1. Its scope will be further illustrated by subsequent sections of this chapter. In the second and third section we focus on the class of *linear* systems and discuss in some detail the dynamics of linear models described by differential or difference equations with constant coefficients. The study of these models represents the core of dynamical systems theory and has strongly influenced the development of other branches. Section 2.2 is concerned with their free motions and Section 2.3 with their forced motions. We also describe some elements of input-output theory and explain the relationship between their representations in time and frequency domain. In Section 2.4 we introduce structure preserving mappings (“morphisms”) between linear systems. We show how new systems of this class can be obtained via standard constructions and describe various interconnection schemes for building complex systems. Finally in the last section we analyze the problem of converting continuous time signals and systems into discrete time versions and vice versa. This is a problem of increasing importance due to the replacement of analog devices by digital ones in the control and measurement of processes which evolve continuously in time. Numerical Analysis offers many techniques for the discretization of

differential equations. We will describe some basic numerical schemes and indicate the difficulties which can occur in their use for approximating differentiable dynamical systems by discrete ones.

2.1 Dynamical Systems

In this section we introduce the general mathematical concept of a dynamical system in state space. This concept has evolved as a unification of a variety of notions which have been used in, for example, the classical theory of differentiable dynamical systems, circuit theory and automata theory. We will illustrate the scope of the general definition by different examples taken from these fields. In order to obtain additional structure we also introduce some basic properties which lead to a broad classification of dynamical systems. Since the section has mainly conceptual objectives the presentation is descriptive and contains just a few mathematical results.

2.1.1 The General Concept of a Dynamical System

Before presenting the formal definition we consider the main terms and relations which need to be specified in order to define a dynamical system.

Time domain. A dynamical system evolves in time and so the variables which describe the behaviour of the system are functions of time. With every dynamical system there is an associated *time domain* $T \subset \mathbb{R}$ which contains all the times t at which the system variables may be evaluated. The time domain may be continuous, i.e. an interval as in Example 1.1.1 where $T = [0, \infty)$ or discrete, i.e. T consists of isolated points in \mathbb{R} e.g. $T = \mathbb{Z}$ or $T = \mathbb{N}$, see Example 1.2.1. For notational convenience we will write $[t_0, t_1)$ rather than $T \cap [t_0, t_1)$ in order to denote the interval $\{t \in T; t_0 \leq t < t_1\}$ in T whenever the underlying time domain is clear.

External variables. These are the variables which describe the interactions of the system with the exterior world. Since a complete description of all the interactions is never possible, the modeller must select a set of variables which are thought to be the most important for the problem in hand. In Example 1.1.1 ecological factors such as pollution may well affect the population dynamics but have not been taken into account in the model.

It is usual to divide the external variables into a family $u = (u_i)$ of *inputs* and a family $y = (y_i)$ of *outputs*. By “inputs” we mean those variables which model the influence of the exterior world on the physical system. These can be of different types — either *controlled* inputs or *uncontrolled* inputs (for instance, disturbances). By “outputs” we mean those variables with which the system acts on the exterior world. Sometimes the outputs are divided into two (not necessarily mutually disjoint) sets of variables. Those which are actually measured will be called *measurements* and those which must be controlled in order to meet specified requirements will be called *regulated*. In certain contexts it is important to distinguish between modelled inputs and outputs and the actual inputs and outputs of the physical system. In this book

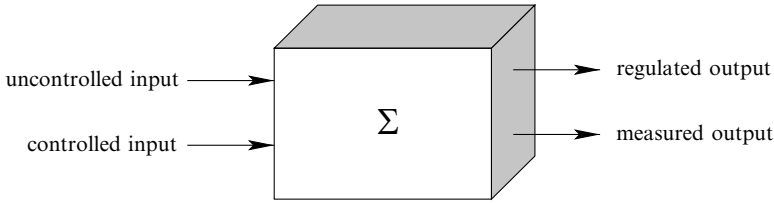


Figure 2.1.1: External variables

the external variables are to be understood as variables of the model and not as quantities of the underlying physical system.

It should be noted that it may not be a priori clear which external variables are to be considered as inputs and which as outputs. For instance, in the electrical circuit problem of Example 1.4.6 it is not obvious that the current should be taken as input and the voltage as output, or vice-versa. A general definition of dynamical systems which does not classify a priori the external variables into inputs and outputs has been developed by *J.C. Willems* (see *Notes and References*). We will not pursue this “behavioural approach” here, but presuppose that a distinction between inputs and outputs has already been made.

A dynamical system must specify the set U of input values (input alphabet) and the set Y of output values (output alphabet), for instance in Example 1.3.4 $U = \mathbb{R}$ and $Y = \mathbb{R}^2$ and in Example 1.1.1 $U = [0, \infty)$, $Y = [0, \infty)$. Throughout the text we assume that the set U of admissible input values does not change with time and does not depend on the values of any other system variables.

Let U^T denote the set of all functions $u(\cdot) : T \rightarrow U$. In general it is not possible to admit arbitrary functions $u(\cdot) \in U^T$ as input signals. For instance, in the controlled differential equations of Example 1.3.2 it would not be possible to allow for non-measurable controls since the equations could not be integrated. Therefore, in addition to the set of input values, we must specify a set $\mathcal{U} \subset U^T$ of admissible input functions. By an appropriate choice of \mathcal{U} , measurability or smoothness properties of the control functions can be imposed as well as time-varying constraints on the control values. Whenever there is a risk of confusion we distinguish in our notation between input *values* $u \in U$ and input *functions* $u(\cdot) \in \mathcal{U}$.

We do not include a space \mathcal{Y} of admissible output functions in our general definition since this space is only occasionally needed in the context of state space theory. However, when we consider input-output systems the space \mathcal{Y} of output signals will become important.

Internal state. The notion of state plays a central role in the definition of a dynamical system. Unlike external variables, the *internal* or *state* variables describe processes in the interior of the system. Not every set of internal variables of a system can be accepted as a state vector. Three basic conditions are required.

- (I) The present state and the chosen control function together determine the future states of the system. More precisely, given the state $x(t_0) = x^0$ of the system at some time $t_0 \in T$ and a control $u(\cdot) \in \mathcal{U}$, the evolution of the system’s state $x(t)$ is uniquely determined for all t in a suitable time interval

$T_{t_0, x^0, u(\cdot)}$ of T starting at t_0 . $T_{t_0, x^0, u(\cdot)}$ may be considered as the “life span” of the trajectory $x(\cdot)$ starting at x^0 at time t_0 under the control $u(\cdot)$.

- (II) Given $x(t_0) = x^0$ at some time $t_0 \in T$, the state $x(t)$ at any later time $t \in T$, $t \geq t_0$ only depends on the input values $u(s)$ for $s \in [t_0, t]$. Thus, at time t , the present state $x(t)$ is not influenced by the present and future values $u(s)$, $s \geq t$ of the control. Moreover, knowledge of the state $x(t_0)$ at some time $t_0 < t$ supersedes the information about all previous input and state values.
- (III) The output value at time t is completely determined by the simultaneous input and state values $u(t)$ and $x(t)$. In other words, the past inputs act on the present output only via accumulated effects on the system’s present state.

These requirements ensure that the principle of causality is built into the concept of state. If we regard the output $y(t)$ as the “effect” of past and present “causes” (= inputs), then $u(t)$ represents the instantaneous cause and the state $x(t)$ incorporates the totality of past causes.

The choice of an adequate state vector is usually a much more difficult problem than the specification of the external variables. There are no general prescriptions. However, in physical systems state variables are often associated with the important energy stores of the system. For example, in mechanical systems the position and velocity of each mass point, or of each rigid body, are possible internal variables which together represent the state of the system at a given time. Similarly in electrical *LRC* circuits the charge on each capacitor and the current through each inductor may be chosen as the components of the state vector. Again, depending essentially on the objectives of the modeller the system may be roughly characterized by a few aggregated internal variables or may be more closely modelled by using a state vector with a large number or even infinitely many components. Since the state mediates the influence of past inputs on the output a rough characterization will in general yield only a rough approximation of the input-output behaviour of the physical plant.

The state variables need not represent physical quantities of the system. Indeed, from an information processing point of view the system’s state may be regarded as a kind of continually updated memory or information storage. In this respect, the set X of possible states of the system can be substituted by any other set \tilde{X} which is in one-to-one correspondence with X and hence can carry the same amount of information. So there is more scope for the definition of the state than for the definition of the external variables which usually refer to measurable or physically meaningful quantities. The arbitrariness can be reduced by requiring that the state of the system represents the *minimal* amount of information needed to describe the effect of past history on the future development of the system.

Conditions (I), (II) and (III) lead to the introduction of two maps which must be specified in the definition of every dynamical system.

State transition map. According to (I) and (II), the evolution of the state of a system (*trajectory*) can be described by a map φ called the *state transition map* as follows

$$x(t) = \varphi(t; t_0, x^0, u(\cdot)), \quad t \in T_{t_0, x^0, u(\cdot)}. \quad (1)$$

Actually, $\varphi(t; t_0, x^0, u(\cdot))$ only depends upon the restriction $u(\cdot)|[t_0, t]$ because of (II). In most applications this map is implicitly defined by the equations of motion of the system. If these are differential or difference equations in $x(\cdot)$ as in Example 1.3.4 (1.3.25) and Example 1.5.6 (1.5.1) an initial value problem with $x(t_0) = x^0$ must be solved for a given control function $u(\cdot)$ in order to obtain $\varphi(t; t_0, x^0, u(\cdot)) = x(t)$, $t \in T_{t_0, x^0, u(\cdot)}$.

Output map. By requirement (III), the output of the system at time t is completely determined by the state and input values at time t ,

$$y(t) = \eta(t, x(t), u(t)). \quad (2)$$

η is called the *output map*.

Differential equations can be solved forwards and backwards in time. Hence, if the state transition map is defined by a differential equation, the present state $x(t_0) = x^0$ has a life span $T_{t_0, x^0, u(\cdot)}$ which encompasses both past and future moments of time and the state trajectory $x(t) = \varphi(t; t_0, x^0, u(\cdot))$ is defined for $t < t_0$ as well. The following definition allows for this possibility.

Definition 2.1.1 (Dynamical system). A structure $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is said to be a *dynamical system* or *state space system* with *time domain* T , *input value space* U , *input function space* \mathcal{U} , *state space* X , *output value space* Y , *state transition map* φ and *output map* η , if T, U, \mathcal{U}, X, Y are non void sets, $T \subset \mathbb{R}$, $\mathcal{U} \subset U^T$, and $\eta : T \times X \times U \rightarrow Y$, $\varphi : \mathcal{D}_\varphi \rightarrow X$ (where $\mathcal{D}_\varphi \subset T^2 \times X \times \mathcal{U}$) are functions such that the following axioms hold.

Interval Axiom: For every $t_0 \in T$, $x^0 \in X$, $u(\cdot) \in \mathcal{U}$ the *life span* of $\varphi(\cdot; t_0, x^0, u(\cdot))$

$$T_{t_0, x^0, u(\cdot)} = \{t \in T; (t; t_0, x^0, u(\cdot)) \in \mathcal{D}_\varphi\} \quad (3)$$

is an interval in T containing t_0 .

Consistency Axiom: For every $t_0 \in T$, $x^0 \in X$, $u(\cdot) \in \mathcal{U}$

$$\varphi(t_0; t_0, x^0, u(\cdot)) = x^0. \quad (4)$$

Causality Axiom: For all $t_0 \in T$, $x^0 \in X$, $u(\cdot), v(\cdot) \in \mathcal{U}$, $t_1 \in T_{t_0, x^0, u(\cdot)} \cap T_{t_0, x^0, v(\cdot)}$

$$(\forall t \in [t_0, t_1) : u(t) = v(t)) \Rightarrow \varphi(t_1; t_0, x^0, u(\cdot)) = \varphi(t_1; t_0, x^0, v(\cdot)). \quad (5)$$

Cocycle property: If $t_1 \in T_{t_0, x^0, u(\cdot)}$ and $x^1 = \varphi(t_1; t_0, x^0, u(\cdot))$ for some $t_0 \in T$, $x^0 \in X$, $u(\cdot) \in \mathcal{U}$ then $T_{t_1, x^1, u(\cdot)} \subset T_{t_0, x^0, u(\cdot)}$ and

$$\varphi(t; t_0, x^0, u(\cdot)) = \varphi(t; t_1, x^1, u(\cdot)), \quad t \in T_{t_1, x^1, u(\cdot)}. \quad (6)$$

The product space $T \times X$ is sometimes called the *event space* of Σ . We shall say that a control $u(\cdot) \in \mathcal{U}$ *transfers* an event (t_0, x^0) to (t_1, x^1) (notation: $(t_0, x^0) \xrightarrow{u(\cdot)} (t_1, x^1)$) if $x^1 = \varphi(t_1; t_0, x^0, u(\cdot))$. Although this expression intuitively only makes sense if $t_1 \geq t_0$ it is convenient to use it also if $t_1 < t_0$. The cocycle property says that if

a control $u(\cdot)$ transfers the event (t_0, x^0) to the event (t_1, x^1) and (t_1, x^1) to (t_2, x^2) then it also transfers (t_0, x^0) to (t_2, x^2) . Without this assumption it would be impossible to interpret $\varphi(t; t_0, x^0, u(\cdot))$ as the state of Σ at time t when Σ is initialized at (t_0, x^0) and controlled by $u(\cdot)$. The axiom of consistency then implies that the argument x^0 of φ is in fact the initial state $x(t_0)$ of the system.

The interval axiom, the axiom of consistency and the axiom of causality together guarantee that the state of the system satisfies requirements (I) and (II). Requirement (III) is automatically satisfied if we interpret $y(t) = \eta(t, x, u)$ to be the output of Σ at time t when x is its state and u the instantaneous input value at time t .

For any $t_0 \in T$, $x^0 \in X$, $u(\cdot) \in \mathcal{U}$ the function

$$t \mapsto x(t) = \varphi(t; t_0, x^0, u(\cdot)), \quad t \in T_{t_0, x^0, u(\cdot)}$$

describes the evolution of the system's state and is called the (*state*) *trajectory* of Σ determined by the initial condition $x(t_0) = x^0$ and the control function $u(\cdot)$. Its domain of definition, $T_{t_0, x^0, u(\cdot)}$, is the *life span* of the trajectory. Its image $\{\varphi(t; t_0, x^0, u(\cdot)); t \in T_{t_0, x^0, u(\cdot)}\}$ is said to be an *orbit* of Σ . The corresponding *output trajectory* or *output signal* is

$$y(\cdot) = y(\cdot; t_0, x^0, u(\cdot)) : t \mapsto y(t) = \eta(t, x(t), u(t)), \quad t \in T_{t_0, x^0, u(\cdot)}. \quad (7)$$

Definition 2.1.1 allows for the possibility that the state trajectory of a system starting at $x(t_0) = x^0$ under the control $u(\cdot) \in \mathcal{U}$ does not exist for all future times $t \geq t_0$. This may reflect a situation where the system “blows up” or the trajectory “leaves the state space” X under the influence of the control $u(\cdot)$. As an extreme case, Definition 2.1.1 allows for the possibility that $\varphi(t; t_0, x^0, u(\cdot))$ is not defined for any $t > t_0$ and we will express this by saying that the control $u(\cdot)$ is *not applicable* to Σ initialized at (t_0, x^0) .

Remark 2.1.2. For some dynamical systems control aspects do not play a role. This can be expressed in the framework of Definition 2.1.1 by choosing for the input space U a singleton $\{u^*\}$ and for \mathcal{U} the singleton which only consists of the constant input function $u(t) = u^*$, $t \in T$. Such a system will be called *uncontrolled* or *free*. In order to avoid dependency on the specific singleton it is convenient to use the standard singleton $\{\emptyset\}$ for U . In other situations measurement aspects may not be important. This can be expressed in the framework of Definition 2.1.1 by choosing for the output space the standard singleton so that there is only one constant output signal. Such a dynamic model will be called a *system without outputs*. \square

Definition 2.1.3. A dynamical system Σ is said to be *complete* if, for all $(t_0, x^0, u(\cdot)) \in T \times X \times \mathcal{U}$,

$$T_{t_0, x^0, u(\cdot)} \supset T_{t_0} = \{t \in T; t \geq t_0\}.$$

Thus Σ is complete if and only if $\mathcal{D}_\varphi \supset T_{\geq}^2 \times X \times \mathcal{U}$ where $T_{\geq}^2 = \{(t, t_0) \in T^2; t \geq t_0\}$. Now suppose that Σ is complete and the system is initialized at (t_0, x^0) , i.e. the initial state $x(t_0) = x^0$ is fixed. Then the output signal (7) is defined on T_{t_0} and the restriction $y(\cdot)|_{T_{t_0}}$ of $y(\cdot) = y(\cdot; t_0, x^0, u(\cdot))$ only depends upon the restriction $v(\cdot) = u(\cdot)|_{T_{t_0}} \in \mathcal{U}_{t_0} = \{u(\cdot)|_{T_{t_0}}; u(\cdot) \in \mathcal{U}\}$ by the causality axiom. By a slight abuse of notation we may therefore write $y(\cdot; t_0, x^0, u(\cdot)|_{T_{t_0}})$ instead of $y(\cdot; t_0, x^0, u(\cdot))|_{T_{t_0}}$. The input-output behaviour of Σ is then described by the following operator.

Definition 2.1.4. Given a complete system Σ and $(t_0, x^0) \in T \times X$ the *input-output operator* of Σ initialized at (t_0, x^0) is defined by

$$G_{t_0, x^0} : \mathcal{U}_{t_0} \rightarrow Y^{T_{t_0}}, \quad v(\cdot) \mapsto y(\cdot; t_0, x^0, v(\cdot)). \quad (8)$$

A complete dynamical system is called reversible if it is also a dynamical system for reverse time.

Definition 2.1.5. A complete dynamical system Σ is said to be *reversible* if

$$\mathcal{D}_\varphi = T^2 \times X \times \mathcal{U},$$

i.e. $T_{t_0, x^0, u(\cdot)} = T$ for all $(t_0, x^0, u(\cdot)) \in T \times X \times \mathcal{U}$.

Hence all state trajectories of a reversible system are defined on the whole time domain T . Given any event (t_1, x^1) and any $t_0 \in T$, $t_0 < t_1$, $u(\cdot) \in \mathcal{U}$, there exists a unique $x^0 \in X$ such that $u(\cdot)$ transfers (t_0, x^0) into (t_1, x^1) . In fact this state is given by $x^0 = \varphi(t_0; t_1, x^1, u(\cdot))$. It is the only state with this property since, for every other $\hat{x}^0 \in X$ satisfying $(t_0, \hat{x}^0) \xrightarrow{u(\cdot)} (t_1, x^1)$ it follows from the cocycle property and $(t_1, x^1) \xrightarrow{u(\cdot)} (t_0, x^0)$ that $(t_0, \hat{x}^0) \xrightarrow{u(\cdot)} (t_0, x^0)$, hence $\hat{x}^0 = x^0$ by the consistency axiom. Definition 2.1.1 of a dynamical system is far too general a definition on which to build a substantial mathematical theory. However we feel that it is useful

- for showing the unity of similar developments in different fields,
- for establishing bridges for the transfer of ideas from one area of application to another,
- for recognizing more clearly the additional structures of the objects in a particular field.

We will illustrate the definition with a simple example of a digital system (see Example 1.5.6). Digital systems have only finitely many states and are automata in the following sense.

Definition 2.1.6 (Automaton). A five tuple $\mathcal{A} = (U, X, Y, \psi, \eta)$ where U, X, Y are non-void sets and $\psi : X \times U \rightarrow X$, $\eta : X \times U \rightarrow Y$ are maps, is called an *automaton* with *input space* U , *state space* X , *output space* Y , *next-state function* ψ and *output function* η .

The dynamics of an automaton are described by the following state and output equations

$$\begin{aligned} x(t+1) &= \psi(x(t), u(t)), & t \in \mathbb{N} \\ y(t) &= \eta(x(t), u(t)) \end{aligned} \quad (9)$$

It follows that any automaton can be viewed as a dynamical system by setting $T = \mathbb{N}$, $\mathcal{U} = U^{\mathbb{N}}$ and defining $\varphi : T_{\geq}^2 \times X \times \mathcal{U} \rightarrow X$ recursively by

$$\begin{aligned} \varphi(t_0 + k + 1; t_0, x^0, u(\cdot)) &= \psi(\varphi(t_0 + k; t_0, x^0, u(\cdot)), u(t_0 + k)), & k \in \mathbb{N} \\ \varphi(t_0; t_0, x^0, u(\cdot)) &= x^0. \end{aligned}$$

Example 2.1.7 (Switching networks). A (binary) switching network is an automaton whose input, state and output variables admit only two different values (symbolized by 0 and 1), so

$$U = \mathbb{Z}_2^m, \quad X = \mathbb{Z}_2^n, \quad Y = \mathbb{Z}_2^p$$

where $\mathbb{Z}_2 = \mathbb{Z}/2$ is the binary field (see Section 1.5). Physically, these two values may, for example, be realized by two different voltage levels. If $n = 0$ (so that the switching network has the trivial state space $\{0\}$ and trivial state transition map $\varphi \equiv 0$) the output at time t is completely determined by the input in time t , $y(t) = \eta(0, u(t))$. Dynamical systems with this property are called *memoryless*. They represent physical devices which directly transform inputs into outputs without intermediate storage of energy or information. Simple examples of memoryless switching networks are the logic gates described in Chapter 1. Their output map is given by truth tables.

Switching networks with memory are called *sequential* because a sequence of inputs must be specified in order to determine the output. The basic memory elements used in sequential networks are flip-flops. The “ J - K flip-flop” described in Example 1.5.6 has input space $U = \mathbb{Z}_2^2$, output space $Y = \mathbb{Z}_2^2$, state space $X = \mathbb{Z}_2$, output function $y = [x, 1 - x]^\top$ and next state function

$$\psi(x, u) = x(1 - u_2) + (1 - x)u_1 \quad x \in \mathbb{Z}_2, u \in \mathbb{Z}_2^2.$$

In large sequential networks it is common to synchronize the operation of all the flip-flops by a common clock or pulse generator emitting pulses at each time $k\tau$, $k \in \mathbb{N}$ where $\tau > 0$ is fixed. A synchronized sequential circuit changes state only after the occurrence of a clock pulse and the inputs and states of each of the flip-flops are not allowed to change at other times. It is natural to choose $T = \mathbb{Z}\tau$ as the time domain of such a system.

An important and simple example of a sequential network containing several flip-flops is the shift register. This is used in many digital systems to store and shift binary numbers arriving from a serial source. Figure 2.1.2 illustrates a four bit right shift regis-

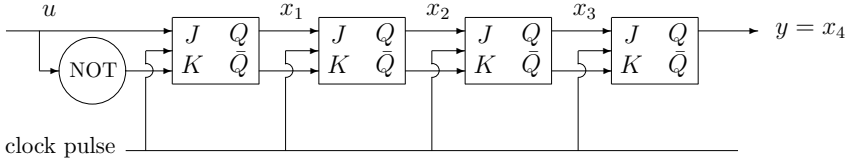


Figure 2.1.2: Shift register

ter constructed from clocked $J - K$ flip-flops. If the initial contents of the register is $x(0) = [x_1, x_2, x_3, x_4]^\top = [1, 0, 1, 1]^\top$ and the input sequence is $u(0) = 1$, $u(\tau) = 0$, $u(2\tau) = 0$, $u(3\tau) = 1$, then its successive states are $x(\tau) = [1, 1, 0, 1]^\top$, $x(2\tau) = [0, 1, 1, 0]^\top$, $x(3\tau) = [0, 0, 1, 1]^\top$, $x(4\tau) = [1, 0, 0, 1]^\top$. The bit shifted out of the right hand end is lost. We can construct a dynamical system modelling the register by setting $U = \mathbb{Z}_2$, $\mathcal{U} = U^T$, $X = \mathbb{Z}_2^4$, $Y = \mathbb{Z}_2$,

$$\begin{aligned} x_1((k+1)\tau) &= u(k\tau), & x_2((k+1)\tau) &= x_1(k\tau) \\ x_3((k+1)\tau) &= x_2(k\tau), & x_4((k+1)\tau) &= x_3(k\tau) \end{aligned}$$

and choosing $y(k\tau) = x_4(k\tau)$. Obviously the output at time $k\tau$, $k \geq 4$ is equal to the delayed input $u((k-4)\tau)$ and the state vector $x(k\tau)$ stores exactly the four preceding input values $u((k-i)\tau)$, $i = 1, 2, 3, 4$. Thus the shift register is an example which highlights the interpretation of the state as a continually updated memory of the system. \square

Traditionally the concept of a dynamical system was more or less synonymous with “a system described by differential equations”. The classical theory of dynamical systems was motivated by problems in mechanics particularly celestial mechanics. Then it was natural to assume that the external forces were given and not subject to human manipulation. This explains why input and output aspects are absent in the classical view of a dynamical system. The following concept of a *differentiable flow* can be regarded as the classical equivalent of a reversible dynamical system, in fact together with the concept of an automaton it motivated the more general Definition 2.1.1.

Definition 2.1.8 (Differentiable flow). A triple (T, X, ψ) is called a *differentiable flow* or *dynamical system in the classical sense* if $T \subset \mathbb{R}$ is an open interval, X an open subset of \mathbb{K}^n , $\mathbb{K} = \mathbb{R}$ or \mathbb{C} (or, more generally a differentiable manifold) and ψ is a continuously differentiable map from $T^2 \times X$ into X , such that

$$\begin{aligned}\psi(t; t, x) &= x, & t \in T, x \in X \\ \psi(t; t_1, \psi(t_1; t_0, x)) &= \psi(t; t_0, x), & t_0, t_1 \in T, x \in X.\end{aligned}$$

Local differentiable flows which avoid the completeness assumption are defined similarly by introducing initial time and state depending life spans T_{t_0, x^0} of the trajectories $\psi(\cdot; t_0, x^0)$. More general local flows will be considered later in the context of stability theory, see Chapter 3.

Local differentiable flows are usually generated via the solution of differential equations. Consider

$$\dot{x}(t) = g(t, x(t)) \tag{10}$$

where $g : T \times X \rightarrow \mathbb{K}^n$ is continuous, with T an open interval and X an open subset of \mathbb{K}^n . We say that $x(\cdot)$ is a solution of (10) on an open interval $I \subset T$ if $x(\cdot)$ is continuously differentiable on I , $(t, x(t)) \in T \times X$ for all $t \in I$ and $x(\cdot)$ satisfies (10) on I . We have the following theorem, see *Notes and References*.

Theorem 2.1.9. *Let $T \subset \mathbb{R}$ be an open interval, X an open subset of \mathbb{K}^n and suppose that $g : T \times X \rightarrow \mathbb{K}^n$ is continuous and continuously differentiable with respect to x on $T \times X$. Then for any $(t_0, x_0) \in T \times X$, there exists a unique solution, $x(\cdot) = \psi(\cdot; t_0, x_0)$ of (10) on some maximal open interval $T_{t_0, x^0} \subset T$ containing t_0 such that $x(t_0) = x^0$. Moreover the set*

$$\mathcal{D}_\psi = \{(t, t_0, x_0); t \in T_{t_0, x^0}, (t_0, x_0) \in T \times X\},$$

is open in $T^2 \times X$ and $\psi : \mathcal{D}_\psi \rightarrow \mathbb{K}^n$ is continuously differentiable (ψ is said to be the general solution of (10)).

We see that under the conditions of the above theorem the differential equation (10) generates a *local* differentiable flow, (T, X, ψ) . It will be shown later that if $X = \mathbb{K}^n$ and g is linearly bounded as in (22), then $T_{t_0, x^0} = T$ and hence in this case (T, X, ψ) is a differentiable flow in the sense of Definition 2.1.1.

To subsume a flow (T, X, ψ) under the general definition of a dynamical system we have to endow it with trivial inputs and outputs as described in Remark 2.1.2. Comparing Definitions 2.1.8 and 2.1.1 (under the completeness assumption) we note the following differences:

- only the evolution of the *state* is described,
- a smoothness condition is imposed on the state transition map,
- reversibility is built into the definition of a differentiable flow.

The following example illustrates the concept of a differential flow.

Example 2.1.10 (Pendulum). In Example 1.3.3 we saw that the equation of motion of a simple swinging pendulum of length l and mass m suspended from a fixed point is

$$ml^2\ddot{\theta} = -mgl \sin \theta \quad (11)$$

where g is the gravitational constant. Let $x = [x_1, x_2]^\top = [\theta, \dot{\theta}]^\top$, then

$$\dot{x}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ -gl^{-1} \sin x_1(t) \end{bmatrix} =: g(t, x(t)).$$

Suppose $T = \mathbb{R}$, $X = \mathbb{R}^2$, then it is easy to see that $g(\cdot, \cdot)$ satisfies the conditions of Theorem 2.1.9 and is linearly bounded. Hence there exists a unique solution $x(t) = \psi(t; t_0, x^0)$ on T satisfying $x(t_0) = x^0$, and (T, X, ψ) is a differentiable flow.

To obtain a graphical representation of the flow ψ the corresponding orbits $\{\psi(t; t_0, x^0); t \in T\}$ are provided with an orientation indicating the direction of motion as time increases. For a given $t_0 \in T$ the collection of oriented orbits corresponding to various initial conditions $x(t_0) = x^0$ in a given region of the state space form a so-called *phase-portrait* of the flow at time t_0 , see Figure 2.1.3. The different character of these trajectories correspond

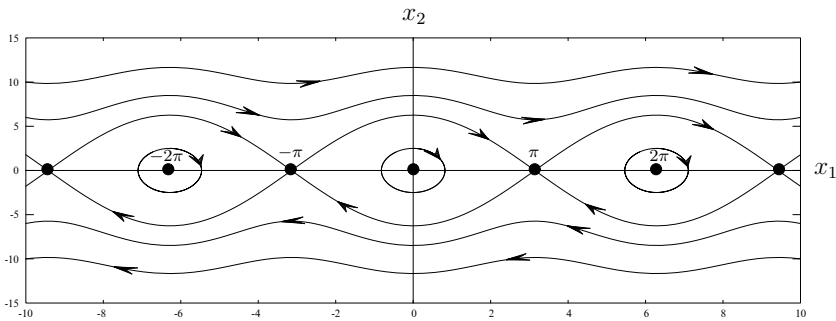


Figure 2.1.3: Phase portrait for the pendulum

to different motions of the pendulum. For example the pendulum stays at rest if it starts at $x^0 = [0, 0]^\top$ (vertically downwards with zero velocity) or at $x^0 = [\pi, 0]^\top$ (vertically upwards with zero velocity). It swings periodically backwards and forwards if it starts at $x^0 = [\theta_0, 0]^\top$ where $0 < \theta_0 < 2\pi$, $\theta_0 \neq \pi$ and rotates continuously around 0 if it starts at $x^0 = [0, \omega_0]^\top$ where $|\omega_0|$ is large enough. \square

Remark 2.1.11. The equations of motion of a physical system are often described by higher order differential equations. In these cases a state vector must be found which enables the equations of motion to be transformed into an equivalent system of the form (10), see Ex. 8 and Ex. 9. \square

Both examples considered in this subsection are complete systems. An example of a differentiable system which is not complete will be given in Example 2.1.16.

2.1.2 Differentiable Dynamical Systems

Let us now consider differentiable systems which are controlled and measured. Since we do not intend to develop a systematic theory of nonlinear control systems in this book, we will only deal with differentiable systems on open subsets $X \subset \mathbb{K}^n$ and not on general differential manifolds. In the following we suppose that every space \mathbb{K}^ℓ , $\ell \in \mathbb{N}^*$ is provided with an arbitrary norm denoted by $\|\cdot\|$.

Definition 2.1.12 (Differentiable system). A dynamical system $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is called *differentiable* if the following conditions are satisfied.

- (i) $T \subset \mathbb{R}$ is an open interval.
- (ii) U, Y are subsets of \mathbb{K}^m and \mathbb{K}^p , X is an open subset of \mathbb{K}^n .
- (iii) There exists a function $f : T \times X \times U \rightarrow \mathbb{K}^n$ such that for all $t_0 \in T$, $x^0 \in X$, $u(\cdot) \in \mathcal{U}$ the initial value problem

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), & t \geq t_0, t \in T \\ x(t_0) &= x^0 \end{aligned} \quad (12)$$

has a unique solution $x(\cdot)$ on a maximal open time interval I satisfying $I = T_{t_0, x^0, u(\cdot)}$ and $x(t) = \varphi(t; t_0, x^0, u(\cdot))$, $t \in I$.

- (iv) $\eta : T \times X \times U \rightarrow Y$ is continuous.

Remark 2.1.13. A continuous time system $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ whose time interval $T \subset \mathbb{R}$ is not open will be called *differentiable* if it is obtained by restriction of the time domain from a differentiable system in the sense of the previous definition. \square

Some remarks concerning the choice of \mathcal{U} and the underlying solution concept for (12) are in order. Often it is necessary to consider jumps in the input functions. For example, if $u(\cdot)$ is a set point control a switch from one set point $u(t) = u_1$, $t \leq t_1$ to another $u(t) = u_2$, $t > t_1$ should be allowed. This leads to choices for \mathcal{U} as the space of piecewise constant functions from T to U or the space of piecewise continuous functions $PC(T; U)$. Sometimes it will be necessary to extend the set of input signals to arbitrary Lebesgue measurable functions u which are *locally integrable* on T (i.e. $\int_a^b \|u(t)\| dt < \infty$ for all $a, b \in T$, $a < b$). Then $f(t, x, u(t))$ will not, in general, depend continuously on t for each fixed x , hence the solution concept used in Theorem 2.1.9 is not applicable. Instead we will call $x(\cdot) : I \rightarrow X$ a *solution of* (12) on an interval $I \subset T$ if it is *absolutely continuous* and satisfies (12) “almost everywhere” on I (that is “except on a set of Lebesgue measure zero”). Here “absolutely continuous” means that $x(\cdot)$ is continuous, differentiable almost everywhere (a.e.) with locally integrable derivative and can be reconstructed from its derivative by integration (see Definition A.3.12)

$$\int_{t_0}^t \dot{x}(s) ds = x(t) - x(t_0), \quad t_0, t \in T, t \geq t_0.$$

For later use we formulate two basic results concerning the existence and uniqueness of solutions of differential equations with measurable RHS¹

$$\dot{x}(t) = g(t, x(t)) \quad (13)$$

¹RHS: right hand side, LHS: left hand side

where $g : T \times X \rightarrow \mathbb{K}^n$, $T \subset \mathbb{R}$ an interval and X an open subset of \mathbb{K}^n . We say that $g : T \times X \rightarrow \mathbb{K}^n$ satisfies the *Carathéodory conditions* if

(Car 1) $g(\cdot, x) : T \rightarrow \mathbb{K}^n$ is measurable for each fixed $x \in X$;

(Car 2) $g(t, \cdot) : X \rightarrow \mathbb{K}^n$ is continuous for each fixed $t \in T$;

(Car 3) $\|g(\cdot, \tilde{x})\|$ is locally integrable on T for some $\tilde{x} \in X$;

(Car 4) for each compact set $C = I \times K \subset T \times X$ there exists an integrable function $L_C(\cdot) : I \rightarrow \mathbb{R}_+$ such that

$$\|g(t, x) - g(t, y)\| \leq L_C(t)\|x - y\|, \quad (t, x), (t, y) \in C. \quad (14)$$

Recall that in any metric space (X, d) the *distance* between a point $x \in X$ and a subset $S \subset X$ is defined by

$$\text{dist}(x, S) = \inf\{d(x, y); y \in S\}. \quad (15)$$

Theorem 2.1.14 (Carathéodory). *If T is an open interval, X is an open subset of \mathbb{K}^n and $g : T \times X \rightarrow \mathbb{K}^n$ satisfies the Carathéodory conditions on $T \times X$, then for any $(t_0, x^0) \in T \times X$ there exists a unique solution $x(\cdot) = \psi(\cdot; t_0, x_0)$ of (13) on some maximal open interval $T_{t_0, x^0} \subset T$ containing t_0 , such that $x(t_0) = x^0$. Moreover*

(i) *if $t_+(t_0, x^0) := \sup T_{t_0, x^0} < \sup T$ then $x(t)$ is unbounded as $t \nearrow t_+(t_0, x^0)$ or the boundary ∂X of X is not empty and $\text{dist}(x(t), \partial X) \rightarrow 0$ as $t \nearrow t_+(t_0, x^0)$. An analogous statement holds for $t \searrow t_-(t_0, x^0) := \inf T_{t_0, x^0}$ if $t_-(t_0, x^0) > \inf T$.*

(ii) *If \mathcal{D}_ψ is the domain of definition of the general solution ψ ,*

$$\mathcal{D}_\psi = \{(t, t_0, x_0); t \in T_{t_0, x^0}, (t_0, x_0) \in T \times X\},$$

then \mathcal{D}_ψ is open in $T^2 \times \mathbb{K}^n$ and $\psi : \mathcal{D}_\psi \rightarrow \mathbb{K}^n$ is continuous.

If $t_+ = t_+(t_0, x^0) < \sup T$ as in (i) then t_+ is called a *finite escape time* of the solution $\psi(\cdot; t_0, x_0)$. If additionally $\psi(\cdot; t_0, x_0)$ is unbounded on $[t_0, t_+)$ we say that it “blows up” or “explodes” in finite time.

In order to establish that differentiable equations of the form (12) define a differentiable dynamical system one must verify that $g(t, x) = f(t, x, u(t))$ satisfies the Carathéodory conditions for all $u(\cdot) \in \mathcal{U}$. The following corollary gives a sufficient condition.

Corollary 2.1.15. *Suppose T, U, \mathcal{U}, X, Y are sets as in Definition 2.1.12, $\eta : T \times X \times U \rightarrow Y$ is continuous and $f : T \times X \times U \rightarrow \mathbb{K}^n$ is jointly measurable in $(t, u) \in T \times U$ for every $x \in X$ and continuous in $x \in X$ for each fixed $(t, u) \in T \times U$. If $\mathcal{U} \subset U^T$ consists of locally L^p -integrable functions ($1 \leq p < \infty$) on T and for each compact set $C = I \times K \subset T \times X$ there exist constants m_C, l_C such that*

$$\|f(t, x, u)\| \leq m_C(\|u\|^p + 1), \quad t \in I, \quad u \in U \text{ for some } x \in X, \quad (16)$$

$$\|f(t, x, u) - f(t, y, u)\| \leq l_C(\|u\|^p + 1)\|x - y\|, \quad (t, x), (t, y) \in C, \quad u \in U, \quad (17)$$

then the initial value problem

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \quad t \in T \\ x(t_0) &= x^0 \end{aligned}$$

has a unique solution $x(\cdot) = x(\cdot; t_0, x^0, u(\cdot))$ on a maximal interval of existence $T_{t_0, x^0, u(\cdot)}$ for all $(t_0, x^0, u(\cdot)) \in T \times X \times \mathcal{U}$. Moreover, if we define the state transition map $\varphi: \mathcal{D}_\varphi \rightarrow X$ by

$$\varphi(t; t_0, x^0, u(\cdot)) = x(t; t_0, x^0, u(\cdot)), \quad \mathcal{D}_\varphi = \{(t; t_0, x^0, u(\cdot)) \in T^2 \times X \times \mathcal{U}; t \in T_{t_0, x^0, u(\cdot)}\}$$

then $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is a differentiable dynamical system.

In general, differentiable systems are not complete. The following example illustrates this fact and shows that the maximal intervals of existence $T_{t_0, x^0, u(\cdot)}$ will in general depend on both x^0 and $u(\cdot) \in \mathcal{U}$.

Example 2.1.16 (Exploding solutions). Consider the initial value problem

$$\dot{x}(t) = x(t)^2 + u(t), \quad x(0) = x^0 \quad (18)$$

where $t \in T := \mathbb{R}$, $x^0 \in X := \mathbb{R}$. For the constant control $u(t) \equiv 1$, $t \geq 0$ we obtain the solution

$$x(t) = \tan(t + c(x^0)), \quad t \geq 0, \quad c(x^0) = \arctan x^0 \in (-\pi/2, \pi/2)$$

which “explodes” at the times $t_\pm(x^0) = \pm\pi/2 - c(x^0)$. Hence in this case the interval of existence is $(-\pi/2 - c(x^0), +\pi/2 - c(x^0))$. For the constant control $u(t) \equiv 0$ it is easily seen that $x(t) = x^0/(1 - x^0 t)$ is a solution of (18) on $(1/x^0, \infty)$ if $x^0 < 0$. For $x^0 = 0$ the solution is zero for all $t \in \mathbb{R}$ and for $x^0 > 0$ the interval of existence is $(-\infty, 1/x^0)$. \square

We will now determine conditions under which a differentiable dynamical system with state space $X = \mathbb{K}^n$ is complete. The existence of solutions *in the large* (i.e. for all $\inf T < t < \sup T$) can be derived from Theorem 2.1.14 (i). Indeed, if $X = \mathbb{K}^n$ then $\sup T_{t_0, x^0} < \sup T$ (resp. $\inf T < \inf T_{t_0, x^0}$) can only occur when $x(t)$ is unbounded as $t \rightarrow \sup T_{t_0, x^0}$ (resp. $t \rightarrow \inf T_{t_0, x^0}$). Thus we need criteria to ensure that a given solution will not escape to infinity at some time $\inf T < t_1 < \sup T$. Gronwall’s lemma is fundamental for estimating the growth of solutions of differential equations. We give two versions of the lemma. The first one is important in this chapter, the more standard second version (which cannot be deduced from the first one) will be used in later chapters.

Lemma 2.1.17 (Generalized Gronwall inequality). Suppose that T is an interval, $a \in T$, $\beta(\cdot)$ is a locally integrable non-negative function on T and $\alpha(\cdot)$, $\xi(\cdot)$ are non-negative continuous functions on T such that

$$\xi(t) \leq \alpha(t) + \left| \int_a^t \beta(r) \xi(r) dr \right|, \quad t \in T. \quad (19)$$

Then

$$\xi(t) \leq \alpha(t) + \left| \int_a^t \alpha(r) \beta(r) \exp \left(\left| \int_r^t \beta(s) ds \right| \right) dr \right|, \quad t \in T. \quad (20)$$

Lemma 2.1.18 (Gronwall). *Suppose that T is an interval, $a \in T$, $\alpha \in \mathbb{R}$, $\beta(\cdot)$ is a locally integrable non-negative function on T and $\xi(\cdot)$ is a continuous function on T satisfying*

$$\xi(t) \leq \alpha + \int_a^t \beta(r)\xi(r)dr, \quad t \in T, t \geq a.$$

Then

$$\xi(t) \leq \alpha \exp \left(\int_a^t \beta(s)ds \right), \quad t \in T, t \geq a. \quad (21)$$

Proposition 2.1.19. *Suppose $T \subset \mathbb{R}$ is an open interval, $X \subset \mathbb{K}^n$ is open and $g : T \times X \rightarrow \mathbb{K}^n$ is affinely bounded, that is*

$$\|g(t, x)\| \leq M(t)\|x\| + m(t), \quad (t, x) \in T \times X. \quad (22)$$

where $M(\cdot)$, $m(\cdot)$ are locally integrable non-negative functions on T . Then every solution of (13) is bounded on every finite interval (t_1, t_2) , $t_1, t_2 \in T, t_1 < t_2$ on which it is defined. If moreover $X = \mathbb{K}^n$ then every solution of (13) can be continued to all of T .

Proof: Let $x(\cdot)$ be a solution of (13) on $(t_1, t_2) \subset T$, $t_1, t_2 \in T$ and let $t_0 \in (t_1, t_2)$. It suffices to show that $x(\cdot)$ is bounded on $[t_0, t_2]$. The proof for $(t_1, t_0]$ is similar. Now

$$\begin{aligned} \|x(t)\| &\leq \|x(t_0)\| + \int_{t_0}^t \|g(r, x(r))\|dr \\ &\leq \left[\|x(t_0)\| + \int_{t_0}^t m(r)dr \right] + \int_{t_0}^t M(r)\|x(r)\|dr, \quad t_0 \leq t \leq t_2. \end{aligned}$$

Applying the generalized Gronwall inequality with $\alpha(t) = \|x(t_0)\| + \int_{t_0}^t m(r)dr$, $\xi(t) = \|x(t)\|$ and $\beta(t) = M(t)$ we see that $\|x(t)\|$ is bounded on $[t_0, t_2]$.

To conclude the proof, suppose that $X = \mathbb{K}^n$. It suffices to show that every *maximal* solution² $x(\cdot) : (t_-, t_+) \rightarrow X$ of (13) is defined on T . But if $t_+ < \sup T$ then $x(\cdot)$ would be bounded on $[t_0, t_+)$ for any $t_0 \in (t_-, t_+)$ and this would contradict Theorem 2.1.14 (i) since $\partial X = \emptyset$. $t_- = \inf T$ is shown similarly. \square

As a corollary we obtain the following sufficient criterion for the completeness of a differentiable system.

Corollary 2.1.20. *Under the conditions of Corollary 2.1.15 with $X = \mathbb{K}^n$, if for every compact subinterval $I \subset T$ there exist constants C_I and c_I such that*

$$\|f(t, x, u)\| \leq C_I(\|u\|^p + 1)\|x\| + c_I(\|u\|^p + 1), \quad (t, x, u) \in I \times X \times U \quad (23)$$

then the differentiable system $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is complete and reversible.

Now let Σ be a differentiable system as in Definition 2.1.12 and $u(\cdot) \in \mathcal{U}$. The input function $u(\cdot)$ defines at every time $t \in T$ a *vector field* $x \mapsto f(t, x, u(t))$ on X . Of

²A solution of (13) which cannot be continued to a solution of (13) on a larger interval is called *maximal*.

particular importance are those states \bar{x} at which the vector fields $x \mapsto f(t, x, u(t))$ vanish for all times $t \in T$

$$f(t, \bar{x}, u(t)) = 0, \quad t \in T. \quad (24)$$

These states are *singular points* for all the vector fields $x \mapsto f(t, x, u(t))$, $t \in T$. They represent equilibria of the system in the sense that if the state at an arbitrary initial time $t_0 \in T$ is \bar{x} and Σ is controlled by $u(\cdot)$ then it remains in this state for all $t \in T_{t_0}$. The following definition applies to arbitrary dynamical systems.

Definition 2.1.21 (Equilibrium state). Let Σ be a dynamical system and $u(\cdot) \in \mathcal{U}$, then $\bar{x} \in X$ is said to be an *equilibrium state* of Σ under the control $u(\cdot)$ if

$$\varphi(t; t_0, \bar{x}, u(\cdot)) = \bar{x}, \quad t_0, t \in T, \quad t \geq t_0.$$

Systems which arise from technical processes are often designed to operate at a variety of equilibrium states. These different states are obtained by altering the input signal $u(\cdot)$. The next example describes a simple differentiable system in the sense of Definition 2.1.12 which has this property.

Example 2.1.22. Consider a tank of infinite height with constant cross sectional area a to which an incompressible fluid is supplied by a pipe with flow rate $u(t)$. The fluid leaves the tank via an orifice of cross sectional area a_0 (see Figure 2.1.4). Neglecting all inertia

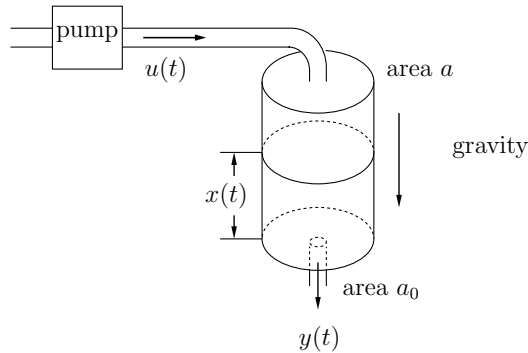


Figure 2.1.4: Fluid level control in a tank

effects of the fluid in the tank, the outlet flow rate y is related to the height x of the liquid level in the tank by the equation

$$y = a_0 \gamma \sqrt{2gx} \quad (25)$$

where γ is a “discharge coefficient” (0.62 for a sharp-edged orifice) and g is the gravitational constant. The principle of conservation of mass yields the following differential equation

$$\dot{x} = -\frac{a_0}{a} \gamma \sqrt{2gx(t)} + \frac{u(t)}{a}. \quad (26)$$

Clearly, (25) and (26) make sense only for $x > 0$. We regard x as the state of our system and choose $X = (0, \infty)$ to be the state space and \mathcal{U} to be the set of all piecewise continuous non-negative functions $u(\cdot) : \mathbb{R} \rightarrow U = \mathbb{R}_+$. Applying Corollary 2.1.15 we see that for each

initial height $x^0 > 0$ of the liquid level in the tank and any control $u(\cdot) \in \mathcal{U}$ (26) admits a unique solution $x(t) > 0$ with $x(0) = x^0$ on some maximal time interval (t_1, t_2) . Thus (25) and (26) define a differentiable dynamical system Σ with the above specification of U, \mathcal{U}, X and $Y = (0, \infty)$. Since the RHS of (26) is affinely bounded $x(t)$ does not explode in finite time by Proposition 2.1.19. But $x(t)$ may leave the state space $X = (0, \infty)$ in finite time (e.g. for $u(\cdot) = 0$) so that Σ is not complete. If, however, $u(t) \geq \varepsilon > 0$ for all t then $x(t)$ cannot tend to 0 in finite time. Hence Σ is complete for all controls which are bounded away from zero.

Now suppose that the control is kept constant $u(t) \equiv \bar{u} > 0$, then for each value of \bar{u} there is exactly one equilibrium state \bar{x} namely

$$\bar{x}(\bar{u}) = \frac{1}{2g} \left(\frac{\bar{u}}{a_0 \gamma} \right)^2.$$

The corresponding equilibrium output value is, as it should be, $\bar{y} = \bar{u}$. □

2.1.3 System Properties

In the previous subsections we introduced some special properties of dynamical systems - complete, reversible, differentiable. We shall now define further classifying properties which will play an important role later.

With respect to the time domain we distinguish between *continuous time* systems where T is a bounded or unbounded interval and *discrete time* systems where T is a discrete subset of \mathbb{R} . Typical discrete time domains are $T = \mathbb{Z}$, $T = \mathbb{N}$ or some corresponding equidistant time sequences $\mathbb{Z}\tau = \{k\tau; k \in \mathbb{Z}\}$, $\mathbb{N}\tau = \{k\tau; k \in \mathbb{N}\}$ where $\tau > 0$. Discrete time counterparts to differentiable systems are systems described by *difference equations*. Unlike differentiable systems they can be defined in a purely set theoretic framework.

Example 2.1.23 (Recursive system). Let U, X, Y be non-empty sets, $T = \mathbb{N}$ or \mathbb{Z} and

$$f: T \times X \times U \rightarrow X, \quad \eta: T \times X \times U \rightarrow Y$$

be two arbitrary mappings. For any $u(\cdot) \in \mathcal{U} = U^T$, $t_0 \in T$, $x^0 \in X$ let $\varphi(t; t_0, x^0, u(\cdot))$, $t \in T$, $t \geq t_0$ be the unique solution of the recursive (or difference) equation

$$x(t+1) = f(t, x(t), u(t)) \tag{27}$$

with initial value $x(t_0) = x^0$. Then $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is a discrete-time dynamical system. Every discrete time system (with the above time domains) can be described in this way (with possible restriction of \mathcal{U}) and, in particular, automata may be regarded as special recursive systems. □

Although the state of a dynamical system evolves in time, the system itself may be *time-invariant* in the sense that the state transition map is invariant with respect to time shifts and the output map does not depend explicitly on time. These systems are more easily analyzed than time-varying ones and so time invariance is often assumed although in reality the system dynamics may change slowly by the effect of growth, ageing, wear and tear, etc.. If $T \subset \mathbb{R}$, U is any non-empty set and $\tau \in \mathbb{R}$ we denote by S_τ the *shift operator* on U^T defined by

$$(S_\tau u)(t) = \begin{cases} u(t - \tau) & \text{if } t - \tau \in T \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

Definition (28) does not make sense if $t - \tau \notin T$ for some $t \in T$ and $0 \notin U$ (for instance, if U is not a subset of a vector space). Whenever we make use of the shift operator S_τ , it is implicitly assumed that either $t - \tau \in T$ for all $t \in T$ or U is a subset of a vector space V and contains the zero element of V . S_τ is called the *right* or *forward* shift if $\tau > 0$ and the *left* or *backward* shift if $\tau < 0$.

Definition 2.1.24 (Time-invariant system). A dynamical system Σ is said to be *time-invariant* if it satisfies the following axioms

- (i) $T \subset \mathbb{R}$ contains 0 and is closed under addition, i.e. $T + T \subset T$.
- (ii) \mathcal{U} is invariant under the right shift, i.e. $S_\tau \mathcal{U} \subset \mathcal{U}$ for all $\tau \in T$, $\tau \geq 0$.
- (iii) For every $t_0, t, \tau \in T$, $t \geq t_0$, $\tau \geq 0$ and every $x^0 \in X$, $u(\cdot) \in \mathcal{U}$

$$\varphi(t + \tau; t_0 + \tau, x^0, S_\tau u(\cdot)) = \varphi(t; t_0, x^0, u(\cdot)).$$
- (iv) The output map η does not depend on time, i.e. $\eta(t, x, u) = \eta(x, u)$, $t \in T$.

From (iii) we see that if $x(\cdot)$ is the state response to $u(\cdot)$ starting at (t_0, x^0) , the state response $\tilde{x}(\cdot)$ to the control $\tilde{u} = S_\tau u$ starting at $(t_0 + \tau, x^0)$ is given by $\tilde{x}(t) = (S_\tau x)(t)$, $t \geq t_0 + \tau$ (see Figure 2.1.5).

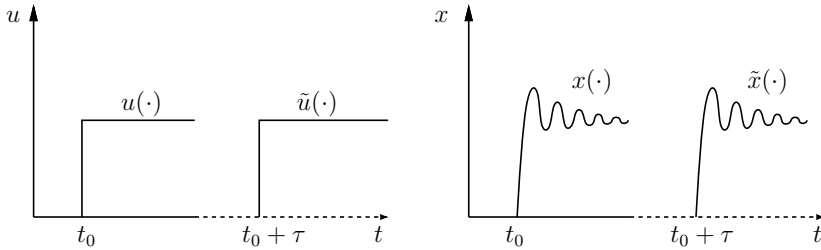


Figure 2.1.5: Time invariance.

The state transition map of a time-invariant system is completely determined by its state transition map at the fixed initial time $t_0 = 0$

$$\varphi(t; x^0, u(\cdot)) := \varphi(t; 0, x^0, u(\cdot)), \quad (t, x^0, u(\cdot)) \in T \times X \times \mathcal{U}.$$

A differentiable or recursive system

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \quad t \in \mathbb{R}_+ & x(t+1) &= f(t, x(t), u(t)), \quad t \in \mathbb{N} \\ y(t) &= \eta(t, x(t), u(t)) & y(t) &= \eta(t, x(t), u(t)) \end{aligned}$$

is time-invariant if f and η do not depend explicitly on time. In particular every automaton is a time-invariant dynamical system.

Sometimes it is mathematically convenient to convert a time-varying differentiable

or difference system into a time-invariant one by introducing time as a new state variable $x_{n+1}(t) = t$. Then the system equations become

$$\begin{aligned} \dot{x}(t) &= f(x_{n+1}(t), x(t), u(t)) & x(t+1) &= f(x_{n+1}(t), x(t), u(t)) \\ \dot{x}_{n+1}(t) &= 1 & x_{n+1}(t+1) &= x_{n+1}(t) + 1 \\ y(t) &= \eta(x_{n+1}(t), x(t), u(t)), & y(t) &= \eta(x_{n+1}(t), x(t), u(t)). \end{aligned}$$

Note that this method increases the dimension of the state space by one.

A system is called *finite*, *finite dimensional* or *infinite dimensional* depending on whether its state space X is a finite set, or a finite or infinite dimensional vector space. The system described in Examples 2.1.7 and 1.5.6 are finite, those in Examples 2.1.10, 2.1.22 are finite dimensional, and the heat equation of Section 1.6 describes an infinite dimensional system. Another infinite dimensional system is presented in the next example which illustrates very clearly the relationship between the state space and the memory of a system.

Example 2.1.25 (Delay system). Consider the system described by

$$\begin{aligned} \dot{x}(t) &= A_0 x(t) + A_1 x(t-h) + Bu(t) \\ y(t) &= Cx(t) \end{aligned} \tag{29}$$

where $A_0, A_1 \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$ and $h > 0$ are given. Here the velocity $\dot{x}(t)$ depends not only on the present state $x(t)$ and control value $u(t)$ but also on the past value $x(t-h)$. Thus the system has a memory of positive length h whereas differentiable systems, in the sense of Definition 2.1.12, have only a memory of infinitesimal duration.

Mathematical models of the above type (both linear and nonlinear) play an important role whenever an action produces an effect with some delay. For example in engineering, feedback control systems sometimes contain long transmission lines which induce non-negligible time lags in the response of the plant regulator. In biology, the growth of a species is influenced by the time lag between birth and procreation. Also in economics, there is a time delay between an investment decision and its effect on productive capacity—the so-called “period of realization” of an investment.

In order to obtain a suitable state space for the system we have to find the amount of initial data required at any time $t_0 \in T = \mathbb{R}$ to determine the future evolution of $x(\cdot)$ on $[t_0, \infty)$. Obviously we need to know the values of $x(s)$ for $t_0 - h \leq s \leq t_0$. In fact we will show that for an arbitrary continuous initial function $z(\cdot) \in X := \mathcal{C}([-h, 0], \mathbb{R}^n)$ and piecewise continuous control $u(\cdot) : [t_0, \infty) \rightarrow \mathbb{R}^m$ there exists a unique continuous function $x(\cdot) : [t_0 - h, \infty) \rightarrow \mathbb{R}^n$ which coincides with $S_{t_0} z(\cdot)$ on $[t_0 - h, t_0]$ and satisfies (29) for $t \geq t_0$. We construct this solution by the *method of steps*. On the interval $[t_0, t_0 + h]$, $x(t)$ is uniquely determined by the variation-of-parameters formula for ordinary differential equations (see Example 2.2.1)

$$x(t) = e^{A_0(t-t_0)} z(0) + \int_{t_0}^t e^{A_0(t-s)} [A_1 z(s-t_0-h) + Bu(s)] ds, \quad t \in [t_0, t_0 + h]. \tag{30}$$

If we set $x(t_0 + s) = z(s)$ for $s \in [-h, 0]$, then obviously $x(\cdot)$ is continuous on $[t_0 - h, t_0 + h]$. Now knowledge of $x(\cdot)$ on $[t_0, t_0 + h]$ enables us via (30) to construct $x(\cdot)$ on $[t_0 + h, t_0 + 2h]$ (replace t_0 by $t_0 + h$, $z(0)$ by $x(t_0 + h)$ and $z(s - t_0 - h)$ by $x(s - t_0 - h)$). Continuing this process we see that there is a unique continuous solution $x(\cdot)$ of (29) on $[t_0, \infty)$ with $x(t_0 + s) = z(s)$ for $s \in [-h, 0]$. Since we need to know the whole function segment

$$x_t : s \mapsto x(t + s) \quad s \in [-h, 0]$$

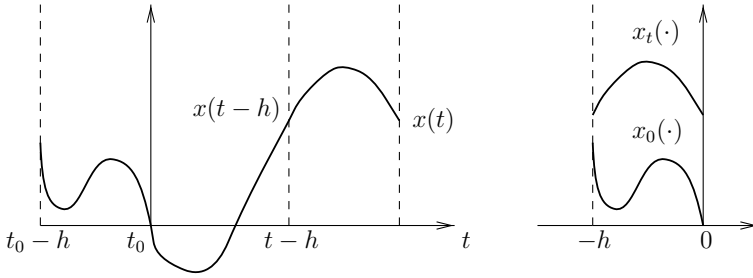


Figure 2.1.6: State of the delay system

in order to determine the system's future evolution under a given control $u(\cdot)$, we regard x_t as the state of our system at time t (see Figure 2.1.6) and take a suitable function space, e.g. $X = \mathcal{C}([-h, 0], \mathbb{R}^n)$, as state space. $x_t(\cdot)$ is simply the trajectory $x(\cdot)$ seen through a window of width h moving with time. The corresponding state transition map φ is given by

$$\varphi(t; t_0, z(\cdot), u(\cdot)) = x_t(\cdot)$$

where $t \in [t_0, \infty)$, $z(\cdot) \in X$ and $x(\cdot)$ is the corresponding solution of (29) with initial state $x_{t_0}(\cdot) = z(\cdot)$. If we apply a time shift $\tau \geq 0$, the solution of (29) on $[t_0 + \tau, \infty)$ with initial state $x_{t_0 + \tau}(\cdot) = z(\cdot)$ and shifted input function $S_\tau u(\cdot)$ will be $S_\tau x(\cdot)$. Hence if the output map is given by

$$\eta(t, z(\cdot), u) = Cz(0), \quad z(\cdot) \in X, \quad u \in \mathbb{R}^m$$

we have an example of a time-invariant infinite dimensional system.

Note that any solution of (29) must be absolutely continuous on its domain of definition. So if $z(\cdot) \in X$ is not absolutely continuous, then there cannot exist a solution of (29) on all of \mathbb{R} which coincides with $z(\cdot)$ on $[-h, 0]$. Hence the system is not reversible. \square

Besides *time-invariance* another system property which will play a central role throughout this book is that of *linearity*.

Definition 2.1.26 (Linear system). Let \mathbb{K} be an arbitrary field. A dynamical system Σ is said to be \mathbb{K} -linear if

(i) U, \mathcal{U}, X, Y are vector spaces over \mathbb{K} ,

(ii) the maps

$$\varphi(t; t_0, \cdot, \cdot) : X \times \mathcal{U} \rightarrow X \quad \text{and} \quad \eta(t, \cdot, \cdot) : X \times U \rightarrow Y$$

are \mathbb{K} -linear for all $t, t_0 \in T, t \geq t_0$.

Condition (ii) implies that

$$\varphi(t; t_0, 0_X, 0_U) = 0_X, \quad t, t_0 \in T, \quad t \geq t_0$$

where 0_X is the origin in X and 0_U the origin in \mathcal{U} (zero function). This means that 0_X is an equilibrium state of Σ under the control 0_U whenever Σ is linear. Example 2.1.25 is a linear system as is the system described by equations (1.3.29), (1.3.33) in Example 1.3.2.

2.1.4 Linearization

We conclude this section with some remarks on how linear models can be used to approximate the behaviour of a nonlinear differentiable system close to a given trajectory or equilibrium point. Let Σ be a differentiable dynamical system with state equation

$$\dot{x}(t) = f(t, x(t), u(t)), \quad t \in T \quad (31)$$

and output equation

$$y(t) = \eta(t, x(t), u(t)) \quad (32)$$

where $T \subset \mathbb{R}$ is an open interval, $U \subset \mathbb{R}^m$ and $X \subset \mathbb{R}^n$ are open, $Y = \mathbb{R}^p$, $\mathcal{U} = \mathcal{C}(T, U)$. Let $\tilde{x}(\cdot)$ be the trajectory corresponding to a given control $\tilde{u}(\cdot) \in \mathcal{U}$ and initial condition $(t_0, \tilde{x}^0) \in T \times X$, so that

$$\begin{aligned} \dot{\tilde{x}}(t) &= f(t, \tilde{x}(t), \tilde{u}(t)), \quad t \geq t_0, \quad t \in T \\ \tilde{x}(t_0) &= \tilde{x}^0. \end{aligned}$$

We assume that the functions $f : T \times X \times U \rightarrow \mathbb{R}^n$ and $\eta : T \times X \times U \rightarrow Y$ are continuous and continuously differentiable with respect to (x, u) on $T \times X \times U$. Consider the Fréchet derivatives (Jacobians)

$$\begin{aligned} A(t) &= D_x f(t, \tilde{x}(t), \tilde{u}(t)) = \left[\frac{\partial f_i}{\partial x_j}(t, \tilde{x}(t), \tilde{u}(t)) \right]_{n \times n} \\ B(t) &= D_u f(t, \tilde{x}(t), \tilde{u}(t)) = \left[\frac{\partial f_i}{\partial u_k}(t, \tilde{x}(t), \tilde{u}(t)) \right]_{n \times m} \\ C(t) &= D_x \eta(t, \tilde{x}(t), \tilde{u}(t)) = \left[\frac{\partial \eta_i}{\partial x_j}(t, \tilde{x}(t), \tilde{u}(t)) \right]_{p \times n} \\ D(t) &= D_u \eta(t, \tilde{x}(t), \tilde{u}(t)) = \left[\frac{\partial \eta_i}{\partial u_k}(t, \tilde{x}(t), \tilde{u}(t)) \right]_{p \times m}. \end{aligned} \quad (33)$$

The linear differentiable system described by

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \end{aligned} \quad (34)$$

is said to be the *linearization* of (31) and (32) along the pair $(\tilde{x}(\cdot), \tilde{u}(\cdot))$.

Let $\xi^0 \in \mathbb{R}^n$, $u(\cdot) \in \mathcal{U}$ and for all small $\varepsilon > 0$ denote by $x(t, \varepsilon)$ the solution of (31) corresponding to the control $u(t, \varepsilon) = \tilde{u}(t) + \varepsilon u(t)$ and the initial condition $x(t_0, \varepsilon) = \tilde{x}^0 + \varepsilon \xi^0$. It follows from basic results concerning the dependence of solutions on parameters and initial conditions that $x(t, \varepsilon)$ is differentiable with respect to ε at $\varepsilon = 0$ and the derivative $\xi(t) = \frac{\partial x}{\partial \varepsilon}(t, 0)$ satisfies

$$\dot{\xi}(t) = A(t)\xi(t) + B(t)u(t), \quad t \in T, \quad t \geq t_0$$

(see *Notes and References*). Hence, if $\xi(\cdot)$ is a solution of (34) corresponding to a control $u(\cdot)$ and initial state ξ^0 then, for small $\varepsilon > 0$, $\tilde{x}(t) + \varepsilon \xi(t)$ is a first order approximation to the solution of (31) corresponding to the control $\tilde{u}(t) + \varepsilon u(t)$ and

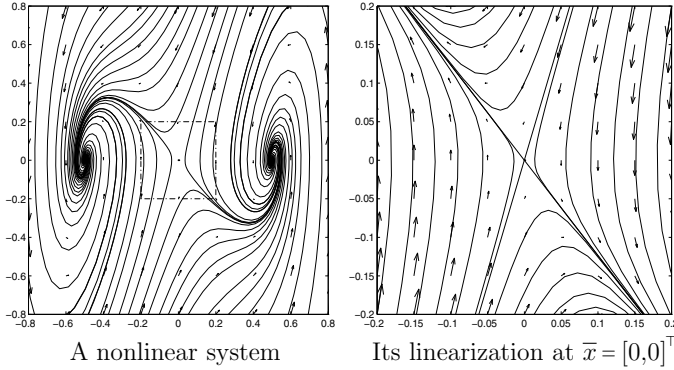


Figure 2.1.7: Phase portraits near an equilibrium point

initial state $\tilde{x}^0 + \varepsilon \xi^0$. Note however, that as $\varepsilon \rightarrow 0$ this approximation is, in general, only uniform in t on compact intervals. Nevertheless, the behaviour of the linear system (34) near the origin gives an approximate picture of the behaviour of the nonlinear system (31), (32) in a sufficiently small neighbourhood of the trajectory $\tilde{x}(t)$. The phase portraits in Figure 2.1.7 illustrate this for a time-invariant free system near an equilibrium solution $\tilde{x}(t) \equiv \bar{x}$ (saddle point). Note that the global properties of the nonlinear and linear systems are quite different.

As (33) shows, the linearized model is, in general, time varying *even if the nonlinear system is time-invariant*, and this is one of the main reasons for the importance of time-varying linear systems in control theory. However, if we linearize a time-invariant system at an equilibrium point corresponding to some constant control the linearized model will again be time-invariant.

Example 2.1.27 (Satellite). The motion of a satellite of mass $m=1$ in a 2-dimensional central gravitational field of the form $k(r) = -\gamma r^{-2}$, $r \neq 0$ can be described by the following equations

$$\ddot{r}(t) = r(t)\dot{\theta}^2(t) - \gamma r^{-2}(t) + u_1(t) \quad (35)$$

$$r(t)\ddot{\theta}(t) = -2\dot{r}(t)\dot{\theta}(t) + u_2(t). \quad (36)$$

Here $r(t)$ is the distance of the satellite from the centre of gravitation at time t , $\dot{\theta}(t)$ is the angular velocity of the radius vector from the centre of gravitation to the satellite at time t , and $u_1(t)$, $u_2(t)$ are radial and tangential thrusts which we take to be control inputs. If $u_1(\cdot) = u_2(\cdot) = 0$, the circular motion $r(t) = 1$, $\theta(t) = \sqrt{\gamma}t$ solves (35). Introducing the state variables $x_1 = r$, $x_2 = \dot{r}$, $x_3 = \theta$, $x_4 = \dot{\theta}$, we see that (35), (36) can be written in the form (31) where the coordinates of $f(t, x, u)$ are given by

$$\begin{aligned} f_1(t, x, u) &= x_2, & f_2(t, x, u) &= x_1 x_4^2 - \gamma x_1^{-2} + u_1 \\ f_3(t, x, u) &= x_4, & f_4(t, x, u) &= -2x_2 x_4 x_1^{-1} + u_2 x_1^{-1}. \end{aligned}$$

If $\tilde{x}(t) = [1, 0, \sqrt{\gamma}t, \sqrt{\gamma}]^T$, $\tilde{u}(t) = [0, 0]^T$ the linearized equation about this trajectory is

$$\dot{x} = Ax + Bu$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \quad (37)$$

and $\omega = +\sqrt{\gamma}$. If the distance $x_1(t)$ and the angle $x_3(t)$ are measured then the (linear) output equation is

$$y = Cx, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (38)$$

□

2.1.5 Exercises

1. (*RC network*) Introduce a suitable state vector and determine the state and the output equations of the electrical circuit represented in Figure 2.1.8. Choose the driving voltage

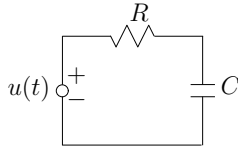


Figure 2.1.8: RC circuit

$u(\cdot)$ (piecewise continuous) as input and the current through the resistor R as output. Specify all the components of a differentiable dynamical system $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ modelling this circuit.

2. (*Tank system*) Determine the equation of motion and the output map of the fluid system shown in Figure 2.1.9. The cross sectional areas of the tanks are $a_1, a_2 > 0$. The

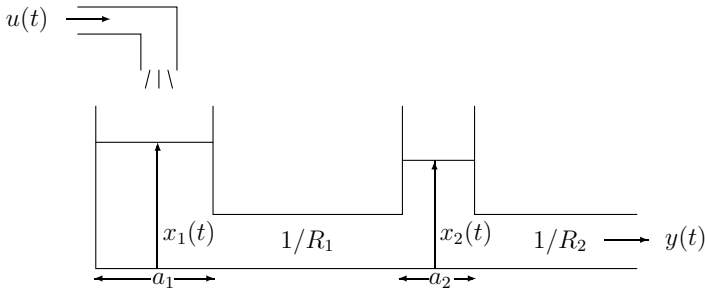


Figure 2.1.9: Fluid system

flow through the first orifice is proportional (with constant $1/R_1$) to $x_1(t) - x_2(t)$, and the flow through the second orifice is proportional with constant $1/R_2$ to $x_2(t)$. Specify all items of the corresponding dynamical system $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$.

3. (*Mass-spring system*) Consider the mechanical system illustrated in Figure 2.1.10. Two masses m_1, m_2 are suspended on ideal springs with stiffness coefficients k_1, k_2 , hanging from a fixed support. The outputs are the displacements $y_1(t), y_2(t)$ of the two masses from their equilibrium positions. The input is a piecewise continuous force $u(t)$ applied

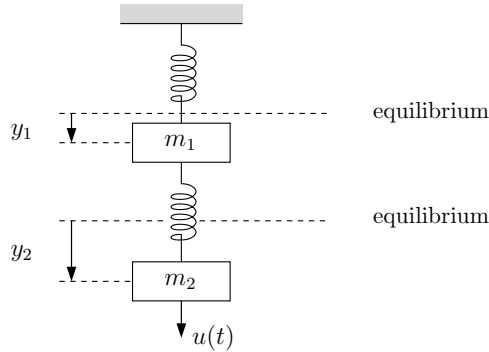


Figure 2.1.10: Mass-spring system

to the second mass. Assuming that the frictional resistances for the two masses in the surrounding medium are c_1 and c_2 , we obtain the following equations of motion

$$\begin{aligned} m_1 \ddot{y}_1 + c_1 \dot{y}_1 + k_1 y_1 &= k_2 (y_2 - y_1) \\ m_2 \ddot{y}_2 + c_2 \dot{y}_2 + k_2 (y_2 - y_1) &= u \end{aligned} \quad (39)$$

Introduce a suitable state vector and specify a differentiable dynamical system Σ describing the above mechanical system. If $u(t) = \bar{u} \in \mathbb{R}$ is constant determine the corresponding set of equilibrium states. (see *Driver* (1977), [138, pp.173-74])

4. (*RLC circuit*) Consider the electrical circuit illustrated in Figure 2.1.11. If y_1, y_2 are the currents across the resistors R_1, R_2 show that the equations of motion are of the same form as (39) but with $\dot{u}(t)$ instead of $u(t)$ on the right hand side of the second equation. Introduce the currents through the inductor and the charge of the capacitor

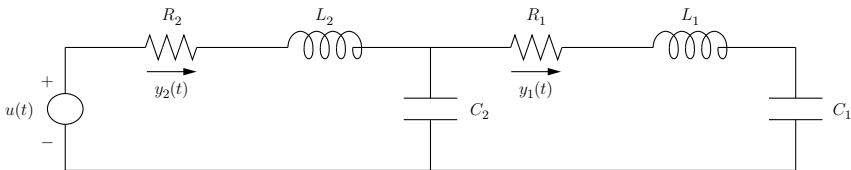


Figure 2.1.11: RLC-network

as state variables and specify the corresponding dynamical system Σ . Compare the state equation obtained with that of the previous exercise. Determine the equilibrium states corresponding to constant voltage $u(t) \equiv \bar{u}$. (see *Driver* (1977) [138, pp.177-178]).

5. (*Parity checker*) Construct an automaton $\mathcal{A} = (U, X, Y, \psi, \eta)$ with $U = Y = \mathbb{Z}_2$ and an initial state $x^0 \in X$ such that \mathcal{A} acts as a “parity check machine” when initialized at x^0 . This means it responds to an arbitrary finite sequence of zeros and ones with 0 if the number of ones is even and with 1 if the number of ones is odd (see *Birkhoff and Bartee* (1970) [60, pp.69-70]).

6. (*Coin operated dispenser*) Specify an automaton $\mathcal{A} = (U, X, Y, \psi, \eta)$ which models a candy machine that accepts two sorts of coins (nickels $N = 5c$ and dimes $D = 10c$). The

price of a candy is 15c. According to the amount of money inserted, it returns nothing or a piece of candy or a piece of candy plus change. Assume that the candy store within the dispenser is infinite (see *MacClamroch* (1980) [369, pp.192]).

7. (*Communication system*) Consider a communication system of the structure as shown in Figure 2.1.12. A binary message (sequence of zeros and ones) is firstly encoded, then transmitted by a communication channel and then decoded to obtain the received mes-

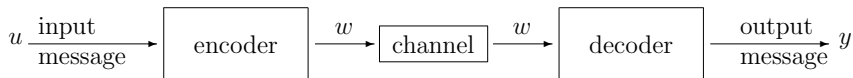


Figure 2.1.12: Communication system

sage. Assume that the channel transmits the signal w without noise and that the encoder algorithm is described by

$$w(t) = u(t) + u(t-1) + u(t-2), \quad t \in \mathbb{N} \quad (40)$$

where we set $u(t) = 0$ for $t < 0$.

- (i) Specify a dynamical system Σ_1 and an initial state $x(0) = x^0$ such that the corresponding input-output relation is identical with (40). Verify this for the input sequence $(0 \ 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1)$.
- (ii) Find a decoder algorithm which reconstructs the input message (such that $y(\cdot)$ is a shifted version of $u(\cdot)$).
- (iii) Describe this decoder by a dynamical system Σ_2 and an appropriate initial state $\bar{x}(0) = \bar{x}^0$.
- (iv) Specify a dynamical system Σ which describes the complete communication system, consisting of the encoder and the decoder coupled in series.

8. Sometimes the equations of motion of a differentiable system contains higher order derivatives of an internal variable z and/or derivatives of the control function u . The output equation may also contain derivatives of the internal variable z . In these cases one must seek a state vector x which enables these equations to be transformed into state and output equations of the prescribed form (see Definition 2.1.12). Find an appropriate state vector and determine the state and output equations corresponding to the following equations where $a_i, b_j \in \mathbb{R}, n \in \mathbb{N}, n \geq 1$ are given parameters

- (i) $\ddot{z}(t) + a_1 \dot{z}(t) + a_0 z(t) = u(t), \quad y(t) = \dot{z}(t),$
- (ii) $y^{(n)}(t) + a_{n-1} y^{(n-1)}(t) + \dots + a_1 y^{(1)}(t) + a_0 y(t) = u(t), \quad y^{(i)} = \frac{d^i y}{dt^i},$
- (iii) $\ddot{\bar{y}}(t) + a_2 \ddot{y}(t) + a_1 \dot{y}(t) = b_1 \dot{u}(t) + b_0 u(t).$

9. Find a suitable state vector and determine state and output equations for the discrete time systems described by the following higher order difference equations on $T = \mathbb{N}$.

- (i) $z(t+2) + a_1 z(t+1) + a_0 z(t) = u(t), \quad y(t) = z(t+1),$
- (ii) $y(t+n) + a_{n-1} y(t+n-1) + \dots + a_1 y(t+1) + a_0 y(t) = u(t),$

$$(iii) \quad y(t+n) + a_{n-1}y(t+n-1) + \dots + a_0y(t) = b_{n-1}u(t+n-1) + \dots + b_0u(t).$$

10. Examine which of the following equations describe a complete dynamical system Σ . Determine the dimension of Σ and determine whether Σ is linear and/or time-invariant.

- (i) $z(t) = u(t), \quad y(t) = \dot{z}(t), \quad t \in \mathbb{R},$
- (ii) $\ddot{z}(t) = u(t), \quad y(t) = \dot{z}(t), \quad t \in \mathbb{R},$
- (iii) $z(t) = \dot{u}(t), \quad y(t) = z(t), \quad t \in \mathbb{R},$
- (iv) $\dot{z}(t) = z(t)u(t), \quad y(t) = \dot{z}(t), \quad t \in \mathbb{R},$
- (v) $z(t) = (z(t-1) + z(t+1))/2, \quad y(t) = z(t), \quad t \in \mathbb{Z},$
- (vi) $z(t) = (u(t-1) + u(t+1))/2, \quad y(t) = z(t-1), \quad t \in \mathbb{Z},$
- (vii) $z(t+1) = z(t) + u(t-1), \quad y(t) = z(t-1), \quad t \in \mathbb{Z},$
- (viii) $z(t+1) = -z(t) + u(t+1), \quad y(t) = z(t)^2, \quad t \in \mathbb{Z},$
- (ix) $\dot{z}(t) = z(t)^{2/3}, \quad y(t) = z(t)^3, \quad t \in [0, \infty),$
- (x) $\ddot{z}(t) + t\dot{z}(t) + t^2z(t) = u(t), \quad y(t) = \ddot{z}(t), \quad t \in [0, \infty),$
- (xi) $z(t) = z(t-1) + z(t-2) + u(t), \quad y(t) = z(t), \quad t \in \mathbb{R},$
- (xii) $\dot{z}(t) = z(t-1) + \dot{z}(t-1) + u(t), \quad y(t) = z(t-1/2), \quad t \in \mathbb{R}.$

11. Consider the scalar delay system

$$\dot{x}(t) = x(t-1) + u(t), \quad y(t) = x(t), \quad t \geq 0. \quad (41)$$

- (i) Specify a dynamical system Σ described by these equations (see Example 2.1.25).
- (ii) Determine the solution $x(\cdot)$ corresponding to the control function $u(t) \equiv 0$ and the initial condition $x(t) = 1, \quad t \in [-1, 0]$ (give an explicit formula for $x(\cdot)$ on the intervals $[k, k+1], \quad k \in \mathbb{N}$). Show that (41) cannot be solved backwards in time.
- (iii) Solve (41) when $u(t) \equiv 1$ and $x(t) = 1, \quad t \in [-1, 0]$.

(cf. *Bellman and Cooke (1963)*)

12. Prove Gronwall's Lemma 2.1.18.

13. In Example 2.1.16 we saw that the equation $\dot{x}(t) = x(t)^2 + u(t)$ does not have a common interval of existence for all initial states $x^0 \in \mathbb{R}$ and constant controls $u(\cdot)$. Now introduce a small delay $\varepsilon > 0$

$$\dot{x}(t) = x(t-\varepsilon)^2 + u(t), \quad t \geq 0.$$

Specify $T, X, U, \mathcal{U}, \varphi$ for a dynamical system Σ (without outputs) described by this equation of motion. Determine whether Σ is complete and/or reversible.

14. (*Euler's equations*) The rotation of a rigid body around its centre of mass is described by the equations

$$\begin{aligned} I_1 \dot{\omega}_1(t) &= (I_2 - I_3) \omega_2(t) \omega_3(t) + u_1(t) \\ I_2 \dot{\omega}_2(t) &= (I_3 - I_1) \omega_1(t) \omega_3(t) + u_2(t) \\ I_3 \dot{\omega}_3(t) &= (I_1 - I_2) \omega_1(t) \omega_2(t) + u_3(t) \end{aligned}$$

where ω is the angular velocity in a coordinate system coinciding with the principal axes of the rigid body. I_1, I_2, I_3 are the principal moments of inertia and $u = [u_1, u_2, u_3]^\top$ is the applied torque, see *Goldstein* (1980) [194, pp.158]. Assume $I_1 = I_2$ (symmetry) and consider the free motion ($\tilde{u} \equiv 0$)

$$\begin{aligned}\tilde{\omega}_1(t) &= \cos[\omega(I_2 - I_3)t/I_2] \\ \tilde{\omega}_2(t) &= \sin[\omega(I_3 - I_2)t/I_2] \\ \tilde{\omega}_3(t) &= \omega_0\end{aligned}$$

where $\omega_0 > 0$ is given. Linearize the above system about $(\tilde{\omega}(\cdot), \tilde{u}(\cdot))$.

15. Case study: Fisheries model. Consider a simple Verhulst model (see Example 1.1.1.1) for the dynamics of a fish population $x(t)$ in a pond

$$\begin{aligned}\dot{x}(t) &= \alpha x(t)(K - x(t)) - u(t), \quad t \geq 0 \\ x(0) &= x^0 > 0\end{aligned}\tag{42}$$

where $u(t)$ is the harvesting rate, $K > 0$ the saturation level of the population and $\alpha > 0$ a constant. We require that $x(t) \geq 0$, $u(t) \geq 0$ for all t .

- (i) No harvesting. For $u(t) = 0$ determine an explicit formula for the solution of (42) on $[0, \infty)$ by separation of variables. Show that for all initial states $x^0 > 0$ the solution tends to the equilibrium state $\bar{x} = K$.
- (ii) Constant harvesting with over-exploitation. Show that a constant harvesting rate $u(t) \equiv \bar{u} > \alpha K^2/4$ leads to depletion in finite time for all initial conditions.
- (iii) Constant harvesting without over-exploitation. Assume that the harvesting rate $u(t) \equiv \bar{u}$ is constant but $\bar{u} < \alpha K^2/4$. Show that there exist two equilibrium states \bar{x}, \hat{x} , $0 < \bar{x} < \hat{x}$, and that for $x^0 > \bar{x}$ (42) admits a (unique) solution $x(t)$ on $[0, \infty)$ which tends to \hat{x} as $t \rightarrow \infty$. However, if $x^0 < \bar{x}$, show that depletion occurs in finite time.
- (iv) What happens if $u(t) \equiv \alpha K^2/4$, $t \geq 0$?
- (v) Constant effort harvesting. If a constant effort for harvesting is made the harvesting rate $u(t)$ will be proportional to $x(t)$ so $u(t) = e x(t)$. Show that depletion occurs in finite time if $e > \alpha K$. Thus assume $e < \alpha K$. Prove that there exist two equilibrium states 0 and \bar{x} and that for *any* initial state $x^0 > 0$ the corresponding solution of (42) tends to \bar{x} as $t \rightarrow \infty$. Determine the effort coefficient $e \in [0, \alpha K]$ which leads to an equilibrium state with maximal harvesting rate $e \bar{x}$. Discuss the result in comparison with the result obtained in (iv).
- (vi) Let \mathcal{U} be the set of all piecewise continuous control functions $u(\cdot) : [0, \infty) \rightarrow [0, u_1]$ where $0 < u_1 < \alpha K^2/4$. Determine $x_1 > 0$ such that for any initial state $x^0 \in X := [x_1, \infty)$ and any control function $u(\cdot)$ there exists a (unique) solution of (42) on $[0, \infty)$.

2.1.6 Notes and References

A general concept of a dynamical system was first formulated by *Kalman* (1963) [287] (see also the introduction of *Kalman et al.* (1969) [290]). A concept of equal generality from an input-output point of view was defined by *Zadeh and Desoer* (1963) [542]. *Sontag* (1998)

has given a state space oriented general definition allowing for local existence of solutions in [472]. A novel comprehensive framework for the theory of dynamical systems has been developed by *J. C. Willems* in his *behavioural approach* to dynamical systems, see [529] and *Polderman and Willems* (1997) [416]. Its most salient feature is that it does not make an a priori distinction between inputs and outputs.

The relationship between the theory of dynamical systems and the theory of automata has been emphasized by *Arbib* (1968) [16]. The emergence of automata theory which was stimulated by the development of information-processing technology dates back to the fifties. An interesting early reference is the volume on Automata Theory in the *Annals of Mathematics Studies* series edited by *Shannon and McCarthy* (1956) [461]. This volume contains contributions by some of the pioneers of the field, Shannon, v. Neumann, Kleene, and Moore. Other important early contributions which led to the concept of *finite state machine* (Definition 2.1.6 with finite input, state and output sets) were *Huffman* (1954) [270], *Mealy* (1955) [371] and the papers collected in [381]. Finite state machines were studied as mathematical models of switching and encoding networks in abstraction from hardware considerations. The theory was strongly influenced by earlier developments in logic and the theory of computability of recursive functions, in particular by the concept of a Turing machine (invented by *Turing* in 1935). A comprehensive treatise on automata is *Eilenberg* (1974) [147], a more recent reference is *Khoussainov and Nerode* (2001) [307]. Many elementary examples of differentiable control systems are described in *MacClamroch* (1980) [369] and in the excellent introduction of *Luenberger* (1979) [349]. The satellite model (Example 2.1.27) is discussed in *Brockett* (1970) [77].

For the existence and uniqueness results from the theory of differential equations, see *Dieudonné* (1970) [132], *Hale* (1980) [214] and *Amann* (1990) [11]. The generalized Gronwall inequality can be found in [11] under the additional assumption that $\beta(\cdot)$ is continuous on T . However it is easy to see that local integrability suffices. For the more standard version of Gronwall's Lemma see e.g. [541]. Differential equations where the RHS depends measurably on t are carefully dealt with in *Aulbach and Wanner* (1996) [27]. For the continuous dependence of solutions on parameters, see e.g. [11], [214]. Differentiability theorems on which Subsection 2.1.4 on Linearization is based can be found in [11].

Some results on nonlinear control systems can be found in *Lee and Markus* (1967) [336]. A differential geometric approach has been developed by *Isidori* (1989) [275] and *Nijmeyer and van der Schaft* (1990) [393]. A comprehensive recent textbook which presents different methods for the study and design of nonlinear control systems both in state space and in an input-output framework is *Sastry* (1999) [448]. Advanced treatments of the classical theory of dynamical systems are *Palis and de Melo* (1982) [404], *Arnold* (1983) [20], *Devaney* (1989) [131] and *Katok and Hasselblatt* (1995) [295].

An elementary introduction to delay equations is given in *Driver* (1977) [138]. Many examples and interesting results on delay equations can be found in *Bellman and Cooke* (1963) [45]. Another standard reference for functional differential equations is *Hale* (1977) [213].

2.2 Linear Systems

The assumption of linearity allows every state and output trajectory to be represented as a linear combination (“superposition”) of a fixed set of simpler trajectories. The use of this superposition principle for the analysis of linear systems is the central topic of this section.

We begin by considering *general* linear systems and show that every trajectory can be decomposed into a *free* motion which depends only on the initial state and a *forced* motion starting at zero which depends only on the control function. Then we specialize to the class of linear time-invariant finite dimensional systems described by differential and difference equations. Their free motions are analyzed in detail and we show that every free trajectory can be decomposed into generalized eigenmotions. The forced motions will be considered more thoroughly in the next section. We conclude with the study of a particular infinite dimensional system described by partial differential equations.

2.2.1 General Linear Systems

Let \mathbb{K} be an arbitrary field and $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ a \mathbb{K} -linear system, then for every t_0 , $t \in T$, $t \geq t_0$ and $\lambda_i \in \mathbb{K}$, $x_i \in X$, $u_i \in U$, $u_i(\cdot) \in \mathcal{U}$, $i = 1, \dots, k$ we have

$$\varphi(t; t_0, \sum_{i=1}^k \lambda_i x_i, \sum_{i=1}^k \lambda_i u_i(\cdot)) = \sum_{i=1}^k \lambda_i \varphi(t; t_0, x_i, u_i(\cdot)) \quad (1)$$

$$\eta(t, \sum_{i=1}^k \lambda_i x_i, \sum_{i=1}^k \lambda_i u_i) = \sum_{i=1}^k \lambda_i \eta(t, x_i, u_i). \quad (2)$$

These equations express the *superposition principle* for the state and the output. As a special case we obtain the so-called *decomposition principle*

$$\varphi(t; t_0, x^0, u(\cdot)) = \varphi(t; t_0, x^0, 0_U) + \varphi(t; t_0, 0_X, u(\cdot)). \quad (3)$$

This shows that every trajectory of a linear system can be decomposed into the sum of a *free motion* $t \mapsto \varphi(t; t_0, x^0, 0_U)$ which depends only on the initial state x^0 and a *forced motion* $t \mapsto \varphi(t; t_0, 0_X, u(\cdot))$ which depends only on the control $u(\cdot)$.

Again as a special case of (1), by setting $u_i(\cdot) = 0_U$, we get the *superposition law of free motions*

$$\varphi(t; t_0, \sum_{i=1}^k \lambda_i x_i, 0_U) = \sum_{i=1}^k \lambda_i \varphi(t; t_0, x_i, 0_U) \quad (4)$$

and, by setting $x_i = 0_X$, the *superposition law of forced motions*

$$\varphi(t; t_0, 0_X, \sum_{i=1}^k \lambda_i u_i(\cdot)) = \sum_{i=1}^k \lambda_i \varphi(t; t_0, 0_X, u_i(\cdot)). \quad (5)$$

It is easy to see that the general superposition principle for state trajectories (1) is equivalent to the decomposition principle (3) together with the superposition laws (4) and (5).

The decomposition law leads us to introduce the following two families of linear maps. For any pair of times $(t, t_0) \in T_{\geq}^2$ we define the *evolution operator* $\Phi(t, t_0) : X \rightarrow X$ by

$$\Phi(t, t_0)x = \varphi(t; t_0, x, 0_{\mathcal{U}}), \quad x \in X \quad (6)$$

and the *input-to-state map* $\Theta(t, t_0) : \mathcal{U} \rightarrow X$ by

$$\Theta(t, t_0)u(\cdot) = \varphi(t; t_0, 0_X, u(\cdot)), \quad u(\cdot) \in \mathcal{U}. \quad (7)$$

The two maps are linear because of (4), (5). $\Phi(t, t_0)$ associates with any state x the state $x(t)$ at time t resulting from the free motion of Σ starting at $x(t_0) = x$. $\Theta(t, t_0)$ maps any control function $u(\cdot)$ onto the state $x(t)$ to which Σ is steered at time t by $u(\cdot)$ from the initial state $x(t_0) = 0$. By (3) all trajectories $t \mapsto \varphi(t; t_0, x^0, u(\cdot))$ of Σ are completely determined by these two families of linear operators

$$\varphi(t; t_0, x^0, u(\cdot)) = \Phi(t, t_0)x^0 + \Theta(t, t_0)u(\cdot), \quad (t, t_0) \in T_{\geq}^2.$$

We shall see later that for all linear systems of practical importance the linear map $\Theta(t, t_0)$ can be expressed with the aid of the operators $\Phi(t, s)_{(t,s) \in T_{\geq}^2}$. Therefore it is particularly important to study the properties of the family $(\Phi(t, s))_{(t,s) \in T_{\geq}^2}$. The axioms (1.4) and (1.6) of a state transition map imply the following basic equations

$$\Phi(t, t) = I_X, \quad t \in T \quad (8)$$

$$\Phi(t_2, t_1) \circ \Phi(t_1, t_0) = \Phi(t_2, t_0), \quad t_0, t_1, t_2 \in T, \quad t_0 \leq t_1 \leq t_2. \quad (9)$$

A family $(\Phi(t, s))_{(t,s) \in T_{\geq}^2}$ of linear operators on X with these properties is called a *family of evolution operators* on X . If Σ is time-invariant we may fix $t_0 = 0$ and obtain a one-parameter family $(\Phi(t))_{t \in T_0}$ of linear operators $\Phi(t) : X \rightarrow X$ defined by

$$\Phi(t)x = \Phi(t, 0)x = \varphi(t; 0, x, 0_{\mathcal{U}}), \quad t \in T_0 = \{t \in T; t \geq 0\}. \quad (10)$$

Equations (8) and (9) then imply

$$\Phi(0) = I_X \quad (11)$$

$$\Phi(t) \circ \Phi(s) = \Phi(t + s), \quad s, t \in T, \quad t, s \geq 0. \quad (12)$$

A family $(\Phi(t))_{t \in T_0}$ of linear operators on X with these properties is called a *semigroup of linear operators* on X . The theory of operator semigroups on normed spaces provides a mathematical basis for the study of infinite dimensional time-invariant linear systems (see *Notes and References* and Section 1.6). The following finite-dimensional example relates the abstract notions introduced above to familiar concepts from the theory of linear differential equations.

Example 2.2.1. (Linear differentiable systems). Let $T \subset \mathbb{R}$ be an interval, $X = \mathbb{K}^n$, $U = \mathbb{K}^m$, $Y = \mathbb{K}^p$, \mathcal{U} any linear subspace of $L_{\text{loc}}^1(T, \mathbb{K}^m)$, e.g. $\mathcal{U} = PC(T, \mathbb{K}^m)$ and $A(\cdot) \in PC(T, \mathbb{K}^{n \times n})$, $B(\cdot) \in PC(T, \mathbb{K}^{n \times m})$, $C(\cdot) \in PC(T, \mathbb{K}^{p \times n})$, $D(\cdot) \in PC(T, \mathbb{K}^{p \times m})$. By Corollary 2.1.20, there exists a unique solution of the initial value problem

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \quad t \in T \\ x(t_0) &= x^0 \end{aligned} \quad (13)$$

for every $u(\cdot) \in \mathcal{U}$, $t_0 \in T$, $x^0 \in X$. Recall that the fundamental matrix $X(t, t_0)$ associated with (13) is, by definition, the solution of the matrix differential equation

$$\begin{aligned}\dot{X}(t) &= A(t)X(t), \quad t \in T \\ X(t_0) &= I_n.\end{aligned}\tag{14}$$

This means that the columns $x^j(t, t_0)$ of $X(t, t_0)$ solve the initial value problems

$$\begin{aligned}\dot{x}(t) &= A(t)x(t), \quad t \in T \\ x(t_0) &= e^j\end{aligned}$$

where e^j is the j -th column of I_n , $j \in \underline{n}$. The variation-of-constants formula gives the following explicit representation for the solution $x(t) = \varphi(t; t_0, x^0, u(\cdot))$ of (13)

$$\varphi(t; t_0, x^0, u(\cdot)) = X(t, t_0)x^0 + \int_{t_0}^t X(t, s)B(s)u(s) ds, \quad t \in T.\tag{15}$$

If we set

$$\eta(t, x, u) = C(t)x + D(t)u$$

then $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is a linear differentiable system (defined by the matrix-valued functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$). Since φ is defined on $T^2 \times X \times \mathcal{U}$, Σ is complete and reversible. Let us now determine the linear operators $\Phi(t, t_0)$, $\Theta(t, t_0)$ associated with Σ . As an immediate consequence of (6) and (15) we obtain for all $t_0, t \in T$

$$\Phi(t, t_0)x = X(t, t_0)x.$$

So the fundamental matrix of (13) is just the matrix representation of the evolution operator $\Phi(t, t_0)$ with respect to the standard basis of \mathbb{K}^n . In the sequel we shall use the same notation $\Phi(t, t_0)$ for both the linear operators and their matrix representations. Since $\Phi(t, t_0)\Phi(t_0, t) = I_n$, the operators $\Phi(t, t_0)$, $t, t_0 \in T$ are all invertible. From (7) and (15) we obtain for all $t_0, t \in T$

$$\Theta(t, t_0)u(\cdot) = \int_{t_0}^t \Phi(t, s)B(s)u(s)ds, \quad u(\cdot) \in \mathcal{U}.\tag{16}$$

This specifies the relation between the input-to-state operators $\Theta(t, t_0)$ and the evolution operators $\Phi(t, s)$ for the system Σ .

We conclude this example with a few words about the important special case where $T = \mathbb{R}$ and $A(t) \equiv A$, $B(t) \equiv B$, $C(t) \equiv C$, $D(t) \equiv D$ are independent of time. In this case the system equations are

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t), \quad t \in \mathbb{R} \\ y(t) &= Cx(t) + Du(t).\end{aligned}\tag{17}$$

Let e^{At} denote the matrix exponential defined by the absolutely converging series

$$e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k, \quad t \in \mathbb{R}.\tag{18}$$

Then the fundamental matrix has the form $X(t, t_0) = e^{A(t-t_0)}$ and so the state transition map is given by

$$\varphi(t; t_0, x^0, u(\cdot)) = e^{A(t-t_0)}x^0 + \int_{t_0}^t e^{A(t-s)}Bu(s) ds, \quad t \in \mathbb{R}.\tag{19}$$

This formula shows that the system Σ is time-invariant, with associated semigroup of linear operators $\Phi(t) = e^{At}$. Since Σ is reversible this semigroup can actually be extended to a one-parameter *group* $\Phi = (e^{At})_{t \in \mathbb{R}}$ of linear operators on \mathbb{K}^n . \square

In the following example we briefly discuss the discrete time counterpart of the previous example.

Example 2.2.2. (Linear difference systems). Let \mathbb{K} be an arbitrary field, $U = \mathbb{K}^m$, $X = \mathbb{K}^n$, $Y = \mathbb{K}^p$, $T \subset \mathbb{Z}$ a time-domain satisfying $t \in T \Rightarrow t+1 \in T$, $\mathcal{U} = U^T$, and $A(\cdot) = (A(t))_{t \in T}$, $B(\cdot) = (B(t))_{t \in T}$, $C(\cdot) = (C(t))_{t \in T}$, $D(\cdot) = (D(t))_{t \in T}$ sequences of $n \times n$, $n \times m$, $p \times n$, $p \times m$ matrices over \mathbb{K} . Consider the discrete time counterpart of the system equations (13)

$$\begin{aligned} x(t+1) &= A(t)x(t) + B(t)u(t), \quad t \in T \\ y(t) &= C(t)x(t) + D(t)u(t). \end{aligned} \quad (20)$$

It is easily verified that for every $u(\cdot) \in \mathcal{U}$, $t_0 \in T$, $x^0 \in X$ the difference equation in (20) admits a unique solution $x(t) = \varphi(t; t_0, x^0, u(\cdot))$ with $x(t_0) = x^0$, namely

$$\varphi(t; t_0, x^0, u(\cdot)) = \Phi(t, t_0)x^0 + \sum_{s=t_0}^{t-1} \Phi(t, s+1)B(s)u(s), \quad t \in T_{t_0} \quad (21)$$

where $\Phi(t, s) = I_n$ for $s = t \in T$ and

$$\Phi(t, s) = A(t-1)A(t-2) \dots A(s), \quad s, t \in T, \quad s < t.$$

If we set $\eta(t, x, u) = C(t)x + D(t)u$, then $\Sigma = (T, U, \mathcal{U}, X, Y, \varphi, \eta)$ is a discrete time linear time-varying dynamical system. The associated input-to-state operator $\Theta(t, t_0)$ can again be expressed in terms of the evolution operator $\Phi(t, s)$,

$$\Theta(t, t_0)(u(\cdot)) = \sum_{s=t_0}^{t-1} \Phi(t, s+1)B(s)u(s), \quad u(\cdot) \in \mathcal{U}, \quad t_0, t \in T, \quad t_0 < t.$$

In the time-invariant case where $A(t)$, $B(t)$, $C(t)$, $D(t)$ are constant matrices the system equations are

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \quad t \in \mathbb{Z} \\ y(t) &= Cx(t) + Du(t), \end{aligned} \quad (22)$$

so that the state transition map is given by

$$\varphi(t; t_0, x^0, u(\cdot)) = A^{(t-t_0)}x^0 + \sum_{s=t_0}^{t-1} A^{(t-1-s)}Bu(s).$$

It follows from this formula that Σ is time-invariant with an associated discrete semigroup of linear operators $(\Phi(t))_{t \in \mathbb{N}}$ given by

$$\Phi(t) = \Phi(t, 0) = A^t, \quad t \in \mathbb{N}. \quad (23)$$

In contrast with the differentiable system of Example 2.2.1, Σ is not necessarily reversible. It is reversible if and only if A is nonsingular. \square

In the next two subsections we will study the free motions of the time-invariant linear systems (17) and (22) in more detail.

2.2.2 Free Motions of Time-Invariant Linear Differential Systems

Let $A \in \mathbb{K}^{n \times n}$ be a given matrix. In this subsection we will study the state trajectories of the free system without output (see Remark 2.1.2) given by

$$\dot{x}(t) = Ax(t), \quad t \in \mathbb{R}. \quad (24)$$

First note that the origin $\bar{x} = 0$ is always a singular point of the vector field $x \mapsto Ax$ on \mathbb{K}^n and is, therefore, an equilibrium point of (24). More generally, $\bar{x} \in \mathbb{K}^n$ is an equilibrium point if and only if $A\bar{x} = 0$, i.e. $\ker A$ is the set of equilibria of (24).

We have seen in Example 2.2.1 that the trajectories of (24) (i.e. the free motions of (17)) are described by the group of linear operators

$$\Phi(t) = e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k, \quad t \in \mathbb{R}. \quad (25)$$

This group has the following basic properties.

Lemma 2.2.3. *If $A \in \mathbb{K}^{n \times n}$, then for every $s, t \in \mathbb{R}$ we have*

- (i) $\frac{d}{dt} e^{At} = A e^{At} = e^{At} A$
- (ii) $e^{A(t+s)} = e^{At} e^{As}$
- (iii) $(e^{At})^{-1} = e^{-At}$
- (iv) $e^{S^{-1}AS} = S^{-1} e^{At} S, \quad S \in \mathbf{GL}_n(\mathbb{K}).$

Proof: Properties (ii) and (iii) express the fact that $(e^{At})_{t \in \mathbb{R}}$ is a group of linear operators, (i) follows because e^{At} is the fundamental matrix of (24) at $t_0 = 0$ and (iv) follows from the series representation (25) since $(S^{-1}AS)^k = S^{-1}A^kS$, $k \in \mathbb{N}$ and the similarity action $A \mapsto S^{-1}AS$ is continuous on $\mathbb{K}^{n \times n}$. \square

Our aim is to show that every trajectory of (24) can be represented as a superposition of a finite number of relatively simple trajectories, the (generalized) *eigenmotions*. These eigenmotions are easily determined once a basis of generalized eigenvectors of A has been found. Before we make this more precise we recall some spectral results from Linear Algebra. Suppose $\mathbb{K} = \mathbb{C}$ so that \mathbb{C}^n is the state space and $A \in \mathbb{C}^{n \times n}$. Let $\sigma(A)$ denote the *spectrum of A* , i.e. the set of eigenvalues

$$\sigma(A) = \{\lambda \in \mathbb{C} ; \det(\lambda I_n - A) = 0\}.$$

$\sigma(A)$ is the set of roots of the characteristic polynomial of A

$$\chi_A(s) = \det(sI_n - A) = s^n + a_{n-1}s^{n-1} + \cdots + a_1s + a_0. \quad (26)$$

Factorizing $\chi_A(s) \in \mathbb{C}[s]$ according to the Fundamental Theorem of Algebra we obtain

$$\chi_A(s) = \prod_{j=1}^{\ell} (s - \lambda_j)^{m(\lambda_j)}, \quad \lambda_i \neq \lambda_j \text{ for } i \neq j. \quad (27)$$

$m(\lambda_j)$ is said to be the *algebraic multiplicity* of the eigenvalue λ_j while $\dim \ker(\lambda_j I_n - A) \leq m(\lambda_j)$ is its *geometric multiplicity*. The following well-known decomposition result is basic for our analysis.

Lemma 2.2.4 (Spectral Decomposition Lemma). *If $\lambda_1, \dots, \lambda_\ell$ are the distinct eigenvalues of $A \in \mathbb{C}^{n \times n}$ with algebraic multiplicities $m(\lambda_1), \dots, m(\lambda_\ell)$ then*

$$\mathbb{C}^n = \ker(\lambda_1 I_n - A)^{m(\lambda_1)} \oplus \dots \oplus \ker(\lambda_\ell I_n - A)^{m(\lambda_\ell)} \quad (28)$$

i.e. \mathbb{C}^n is the direct sum (see Definition A.1.19) of the generalized eigenspaces $\ker(\lambda_j I_n - A)^{m(\lambda_j)}$, $j \in \underline{\ell}$. Moreover $\dim \ker(\lambda_j I_n - A)^{m(\lambda_j)} = m(\lambda_j)$ for each $j \in \underline{\ell}$.

$z \in \mathbb{C}^n$ is said to be a *generalized eigenvector of order $m \geq 1$ of A* if

$$(\lambda I_n - A)^m z = 0 \quad \text{and} \quad (\lambda I_n - A)^{m-1} z \neq 0. \quad (29)$$

Hence the non-zero elements of $\ker(\lambda_j I_n - A)^{m(\lambda_j)}$ are the generalized eigenvectors of order $\leq m(\lambda_j)$. The projections corresponding to the decomposition (28)

$$\begin{aligned} P_j &: \mathbb{C}^n \longrightarrow \ker(\lambda_j I_n - A)^{m(\lambda_j)}, \quad j \in \underline{\ell} \\ x &= x^1 \oplus \dots \oplus x^\ell \mapsto x^j \end{aligned}$$

are called *eigenprojections* of A . The following properties of the P_j , $j \in \underline{\ell}$ are obvious from the definition

$$P_j^2 = P_j, \quad P_j P_k = 0 \quad \text{if } j \neq k, \quad \sum_{j=1}^{\ell} P_j = I_n. \quad (30)$$

Moreover

$$AP_j = P_j A = \lambda_j P_j + N_j, \quad j \in \underline{\ell} \quad (31)$$

where $N_j = (A - \lambda_j I_n)P_j$ is nilpotent. N_j is called the *eigennilpotent* corresponding to the eigenvalue λ_j of A . Adding up these equalities and making use of (30) we obtain the *spectral representation* of A

$$A = A \sum_{j=1}^{\ell} P_j = \sum_{j=1}^{\ell} (\lambda_j P_j + N_j). \quad (32)$$

If $N_j = 0$, λ_j is said to be *semi-simple*. A is diagonalizable if and only if every eigenvalue is semi-simple and in this case

$$A = A \sum_{j=1}^{\ell} P_j = \sum_{j=1}^{\ell} \lambda_j P_j. \quad (33)$$

We now return to the free motions of (24). An initial state $z \in \ker(\lambda I_n - A)^m$ gives rise to the following *generalized eigenmotion* of (24)

$$e^{At} z = e^{\lambda t} e^{(A - \lambda I)t} z = e^{\lambda t} \sum_{j=0}^{m-1} \frac{t^j}{j!} (A - \lambda I)^j z, \quad t \in \mathbb{R}. \quad (34)$$

The trajectory remains in the linear subspace spanned by $z, Az, \dots, A^{m-1}z$ for all $t \geq 0$. In particular, if z is an eigenvector, $Az = \lambda z$, then

$$e^{At} z = e^{\lambda t} z, \quad t \in \mathbb{R} \quad (35)$$

remains always in the one-dimensional complex subspace through z . These trajectories are called (complex) *eigenmotions* of the system (24).

As functions of time, any generalized eigenmotion of order m (i.e. starting at a generalized eigenvector of order m) is the product of an *exponential* $e^{\lambda t}$ and a vector polynomial $\sum_{j=0}^{m-1} \frac{t^j}{j!} (A - \lambda I_n)^j z \in \mathbb{C}^n[t]$ of degree $m-1$. If $\operatorname{Re} \lambda \neq 0$ the exponential part determines the long term behaviour of the trajectory. $\|e^{At}z\|$ tends to zero or infinity depending on whether $\operatorname{Re} \lambda < 0$ or $\operatorname{Re} \lambda > 0$.

Remark 2.2.5. If $\lambda_0 = 0 \in \sigma(A)$ the associated eigenvectors are equilibrium points of (24). If z is an associated generalized eigenvector of order m then the corresponding generalized eigenmotion depends polynomially on time, $z(t) = e^{At}z = \sum_{j=0}^{m-1} (1/j!) A^j z t^j$. \square

Since by Lemma 2.2.4 every initial state can be represented as a sum of generalized eigenvectors we obtain the following corollary.

Corollary 2.2.6. *Every trajectory of the free system (24) is a superposition of the generalized eigenmotions. More precisely, if P_1, \dots, P_ℓ are the eigenprojections of $A \in \mathbb{C}^{n \times n}$ corresponding to the distinct eigenvalues $\lambda_1, \dots, \lambda_\ell$ with algebraic multiplicities $m(\lambda_1), \dots, m(\lambda_\ell)$ then*

$$e^{At}x^0 = \sum_{j=1}^{\ell} e^{\lambda_j t} \sum_{k=0}^{m(\lambda_j)-1} \frac{t^k}{k!} (A - \lambda_j I_n)^k P_j x^0, \quad t \geq 0, x^0 \in \mathbb{C}^n. \quad (36)$$

In particular, if A is diagonalizable then

$$e^{At}x^0 = \sum_{j=1}^{\ell} e^{\lambda_j t} P_j x^0, \quad t \geq 0, x^0 \in \mathbb{C}^n. \quad (37)$$

The latter formula gives us a method for computing e^{At} in the diagonalizable case. If (z^1, \dots, z^n) is a basis of eigenvectors of A , $Az^i = \lambda_i z^i$ then $S = [z^1, \dots, z^n] \in \mathbf{GL}_n(\mathbb{C})$ satisfies

$$e^{At} = S \operatorname{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}) S^{-1}. \quad (38)$$

For the general case, recall that if $J(\lambda, m)$ is a Jordan block of order $m \in \mathbb{N}^*$, i.e.

$$J(\lambda, m) = \begin{bmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} \in \mathbb{C}^{m \times m}, \quad \lambda \in \mathbb{C}, \quad m \in \mathbb{N}, \quad (39)$$

then

$$e^{J(\lambda, m)t} = e^{\lambda t} \begin{bmatrix} 1 & t/1! & t^2/2! & \cdots & t^{m-1}/(m-1)! \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & t^2/2! \\ & 0 & & \ddots & t/1! \\ & & & & 1 \end{bmatrix}, \quad t \in \mathbb{R}, \quad \lambda \in \mathbb{C}, \quad m \in \mathbb{N}. \quad (40)$$

Now suppose that $S^{-1}AS$ is in Jordan canonical form

$$S^{-1}AS = \bigoplus_{j=1}^{\ell} \bigoplus_{k=1}^{k_j} J(\lambda_j, m_{jk}) \quad (41)$$

where $m(\lambda_j) = \sum_{k=1}^{k_j} m_{jk}$ is the algebraic multiplicity of the eigenvalue $\lambda_j \in \sigma(A)$ and \bigoplus denotes the direct sum of matrices, see Definition A.1.20. Then

$$e^{At} = S \left[\bigoplus_{j=1}^{\ell} \bigoplus_{k=1}^{k_j} e^{J(\lambda_j, m_{jk})t} \right] S^{-1}, \quad t \in \mathbb{R}. \quad (42)$$

Whilst these formulas are useful for analytical purposes (see Chapter 3) they should not be used for the numerical computation of e^{At} , see *Notes and References*.

In most applications where A is real, one is only interested in real state trajectories. What, then, is the significance of the above analysis?

Remark 2.2.7. From an operator theoretic point of view one has to distinguish between a linear map $L : \mathbb{R}^m \rightarrow \mathbb{R}^n$ and its *complexification* $L^{\mathbb{C}} : \mathbb{C}^m \rightarrow \mathbb{C}^n$ defined by

$$L^{\mathbb{C}}(x + iy) = Lx + iLy, \quad x, y \in \mathbb{R}^m. \quad (43)$$

However, if there is no risk of confusion, we use the same symbol for a matrix $L \in \mathbb{R}^{n \times m}$, the corresponding linear map $v \mapsto Lv$ from \mathbb{R}^m to \mathbb{R}^n and its complexification as a linear map from \mathbb{C}^m to \mathbb{C}^n . Where necessary, we distinguish between the kernels (resp. ranges) of L and $L^{\mathbb{C}}$ by using the notations $\ker_{\mathbb{K}} L$ (resp. $\text{im}_{\mathbb{K}} L$). \square

For the rest of this subsection we suppose that $A \in \mathbb{R}^{n \times n}$. The *real eigenmotions* or *modes* of the system (24) are obtained by taking the real and imaginary parts of the complex eigenmotions. If $\lambda \in \sigma(A)$ is real and z a real eigenvector for λ , then the associated eigenmotion (35) is real. Whereas if $\lambda = \gamma + i\omega \in \sigma(A)$ is non-real ($\omega \neq 0$) and $z = (z_i) \in \mathbb{C}^n$ is an associated eigenvector then $\bar{\lambda} = \gamma - i\omega \in \sigma(A)$ and the conjugate complex vector $\bar{z} = (\bar{z}_i) \in \mathbb{C}^n$ is an eigenvector of A for $\bar{\lambda}$. Choosing $\text{Re } z = (1/2)(z + \bar{z}) \in \mathbb{R}^n$ and $\text{Im } z = 1/(2i)(z - \bar{z}) \in \mathbb{R}^n$ as initial states we obtain the following *real eigenmotions* of (24)

$$e^{At}(\text{Re } z) = \text{Re}(e^{At}z) = \text{Re}(e^{\lambda t}z) = e^{\gamma t}[(\cos \omega t) \text{Re } z - (\sin \omega t) \text{Im } z] \quad (44)$$

$$e^{At}(\text{Im } z) = \text{Im}(e^{At}z) = \text{Im}(e^{\lambda t}z) = e^{\gamma t}[(\sin \omega t) \text{Re } z + (\cos \omega t) \text{Im } z]. \quad (45)$$

There is a qualitative difference between the modes corresponding to real and to non-real eigenvalues. In the real case we have a ‘one-dimensional’ trajectory along the real line $\mathbb{R}z$ which is contractive if $\lambda < 0$, constant if $\lambda = 0$ and expansive if $\lambda > 0$. In the complex case we have a ‘two-dimensional’ oscillatory motion in the plane spanned by $\text{Re } z, \text{Im } z \in \mathbb{R}^n$. This motion is contractive if $\text{Re } \lambda < 0$, and expansive if $\text{Re } \lambda > 0$. Some typical eigenmotions are shown in Figure 2.2.3.

Generalized real eigenmotions are obtained by taking the real and imaginary parts of generalized complex eigenmotions (34). If $\lambda \in \sigma(A)$ is real and z is a generalized real eigenvector of order m for λ then (34) is a generalized real eigenmotion, remaining for all $t \geq 0$ in the m -dimensional linear subspace

$$\text{span}_{\mathbb{R}}\{z, Az, \dots, A^{m-1}z\} \subset \mathbb{R}^n.$$

If $\lambda = \gamma + i\omega \in \sigma(A)$, $\omega \neq 0$ and $z \in \mathbb{C}^n$ is an associated generalized eigenvector of order m then we obtain two *generalized real eigenmotions* associated with the pair $\lambda, \bar{\lambda} \in \sigma(A)$ and the generalized eigenvector z

$$e^{At} \operatorname{Re} z = \operatorname{Re}(e^{At} z) = \operatorname{Re} e^{\lambda t} \sum_{j=0}^{m-1} \frac{t^j}{j!} (A - \lambda I_n)^j z \quad (46)$$

$$= e^{\gamma t} \sum_{j=0}^{m-1} \frac{t^j}{j!} [(\cos \omega t) \operatorname{Re}(A - \lambda I_n)^j z - (\sin \omega t) \operatorname{Im}(A - \lambda I_n)^j z], \quad t \geq 0,$$

$$e^{At} \operatorname{Im} z = \operatorname{Im}(e^{At} z) = \operatorname{Im} e^{\lambda t} \sum_{j=0}^{m-1} \frac{t^j}{j!} (A - \lambda I_n)^j z \quad (47)$$

$$= e^{\gamma t} \sum_{j=0}^{m-1} \frac{t^j}{j!} [(\cos \omega t) \operatorname{Im}(A - \lambda I_n)^j z + (\sin \omega t) \operatorname{Re}(A - \lambda I_n)^j z], \quad t \geq 0.$$

Both trajectories remain for all $t \geq 0$ in the $2m$ -dimensional linear subspace

$$\operatorname{span}_{\mathbb{R}}\{\operatorname{Re}(A - \lambda I_n)^j z, \operatorname{Im}(A - \lambda I_n)^j z; \quad j = 0, \dots, m-1\}.$$

Since A is real, its spectrum can be written in the form

$$\sigma(A) = \{\rho_1, \dots, \rho_r, \lambda_1, \dots, \lambda_c, \bar{\lambda}_1, \dots, \bar{\lambda}_c\} \quad (48)$$

where $\rho_i \in \mathbb{R}$, $i \in \mathcal{r}$ and $\lambda_i \in \mathbb{C} \setminus \mathbb{R}$, $i \in \mathcal{c}$. The algebraic multiplicities of λ_i and $\bar{\lambda}_i$ are the same. Moreover, if $z^1, \dots, z^{m(\lambda_i)}$ is a basis of $\ker(\lambda_i I_n - A)^{m(\lambda_i)}$ then $\bar{z}^1, \dots, \bar{z}^{m(\lambda_i)}$ is a basis of $\ker(\bar{\lambda}_i I_n - A)^{m(\bar{\lambda}_i)}$ and $\operatorname{Re} z^1, \dots, \operatorname{Re} z^{m(\lambda_i)}, \operatorname{Im} z^1, \dots, \operatorname{Im} z^{m(\lambda_i)} \in \mathbb{R}^n$ is a basis (over \mathbb{R}) of the $2m(\lambda_i)$ -dimensional real linear subspace

$$\mathbb{R}^n \cap [\ker(\lambda_i I_n - A)^{m(\lambda_i)} \oplus \ker(\bar{\lambda}_i I_n - A)^{m(\bar{\lambda}_i)}] = \ker_{\mathbb{R}}(|\lambda_i|^2 I_n - 2(\operatorname{Re} \lambda_i)A + A^2)^{m(\lambda_i)}.$$

The real version of the spectral decomposition (28) is

$$\mathbb{R}^n = \oplus_{i=1}^r \ker_{\mathbb{R}}(\rho_i I_n - A)^{m(\rho_i)} \oplus \oplus_{i=1}^c \ker_{\mathbb{R}}(|\lambda_i|^2 I_n - 2(\operatorname{Re} \lambda_i)A + A^2)^{m(\lambda_i)}.$$

As a consequence we obtain the following real version of Corollary 2.2.6.

Corollary 2.2.8. *Let $A \in \mathbb{R}^{n \times n}$, then every real trajectory $e^{At}x^0$, $x^0 \in \mathbb{R}^n$ of (24) is a superposition of generalized real eigenmotions (modes).*

In the following example we determine the real modes of a linear oscillator.

Example 2.2.9. (Oscillator). Consider the motion of a unit mass connected to a support by a spring immersed in a homogeneous medium (see Figure 2.2.1). The spring constant is taken to be ν^2 (where $\nu \geq 0$) and the friction forces are proportional to the velocity with the constant 2α . If ξ measures the displacement of the mass from equilibrium then (see Example 1.3.2)

$$\ddot{\xi}(t) + 2\alpha\dot{\xi}(t) + \nu^2\xi(t) = 0. \quad (49)$$

Introducing the state vector $x = [x_1, x_2]^\top = [\xi, \dot{\xi}]^\top$ we obtain the state space model

$$\dot{x}(t) = Ax(t), \quad A = \begin{bmatrix} 0 & 1 \\ -\nu^2 & -2\alpha \end{bmatrix}.$$

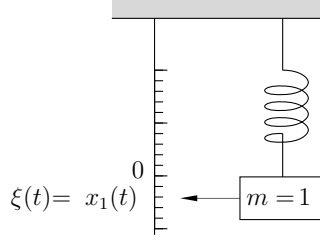


Figure 2.2.1: Mass-spring system

The eigenvalues of A are given by $\lambda_{1,2} = -\alpha \pm \sqrt{\alpha^2 - \nu^2}$ with corresponding eigenvectors

$$z^1 = \begin{bmatrix} 1 \\ \lambda_1 \end{bmatrix}, \quad z^2 = \begin{bmatrix} 1 \\ \lambda_2 \end{bmatrix}. \quad (50)$$

Clearly λ_1, λ_2 are real if and only if $|\alpha| \geq \nu$. In this case the eigenvalues have negative real parts (i.e. produce contracting eigenmotions) if and only if $\alpha > 0$. For $|\alpha| > \nu$ the real eigenmotions $x^1(t), x^2(t)$ starting at z^1 resp. z^2 are

$$x^1(t) = e^{At} z^1 = e^{(-\alpha + \sqrt{\alpha^2 - \nu^2})t} \begin{bmatrix} 1 \\ \lambda_1 \end{bmatrix}, \quad x^2(t) = e^{At} z^2 = e^{(-\alpha - \sqrt{\alpha^2 - \nu^2})t} \begin{bmatrix} 1 \\ \lambda_2 \end{bmatrix}$$

or in terms of ξ (the first coordinate)

$$\xi^{1,2}(t) = e^{(-\alpha \pm \sqrt{\alpha^2 - \nu^2})t}.$$

If $|\alpha| = \nu$, then $\lambda_1 = \lambda_2 = -\alpha$ and the corresponding generalized eigenspace is spanned by

$$z^1 = \begin{bmatrix} 1 \\ -\alpha \end{bmatrix}, \quad z^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

where z^1 is a real eigenvector and z^2 is a generalized eigenvector of second order. By (34) the corresponding (generalized) real eigenmotions are

$$x^1(t) = e^{-\alpha t} \begin{bmatrix} 1 \\ -\alpha \end{bmatrix}, \quad x^2(t) = e^{-\alpha t} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + t e^{-\alpha t} \begin{bmatrix} 1 \\ -\alpha \end{bmatrix}.$$

If $|\alpha| < \nu$ (hence $z^2 = \overline{z^1}$) then $\lambda_{1,2} = -\alpha \pm i\sqrt{\nu^2 - \alpha^2}$ and the corresponding real modes are by (44), (45) and (50)

$$\begin{aligned} x^1(t) &= e^{-\alpha t} \left(\cos(\sqrt{\nu^2 - \alpha^2} t) \begin{bmatrix} 1 \\ -\alpha \end{bmatrix} - \sin(\sqrt{\nu^2 - \alpha^2} t) \begin{bmatrix} 0 \\ \sqrt{\nu^2 - \alpha^2} \end{bmatrix} \right) \\ x^2(t) &= e^{-\alpha t} \left(\sin(\sqrt{\nu^2 - \alpha^2} t) \begin{bmatrix} 1 \\ -\alpha \end{bmatrix} + \cos(\sqrt{\nu^2 - \alpha^2} t) \begin{bmatrix} 0 \\ \sqrt{\nu^2 - \alpha^2} \end{bmatrix} \right). \end{aligned}$$

In terms of the first coordinates $\xi_i(t)$ of $x^i(t)$, $i = 1, 2$ this yields the following oscillatory eigenmotions corresponding to initial conditions $\xi_1(0) = 1$, $\dot{\xi}_1(0) = -\alpha$ and $\xi_2(0) = 0$, $\dot{\xi}_2(0) = \sqrt{\nu^2 - \alpha^2}$

$$\xi_1(t) = e^{-\alpha t} \cos(\sqrt{\nu^2 - \alpha^2} t), \quad \text{and} \quad \xi_2(t) = e^{-\alpha t} \sin(\sqrt{\nu^2 - \alpha^2} t).$$

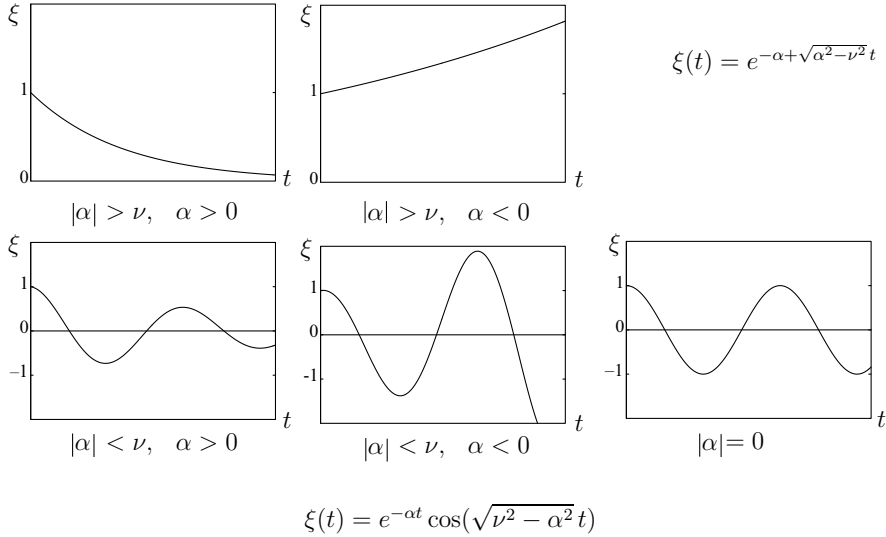


Figure 2.2.2: Eigenmotions of an oscillator

So if there is no damping ($\alpha = 0$) then the $\xi_i(\cdot)$ are periodic with period $2\pi/\nu$. If $\alpha > 0$ the oscillations die out whereas they are intensified if $\alpha < 0$. Some typical eigenmotions $\xi(\cdot)$ are shown in Figure 2.2.2. If we fix the coefficient $\nu > 0$ of the restoring force we observe a qualitative change as the damping $2\alpha > 0$ decreases from large values to zero. For $\alpha \geq \nu$ the eigenmotions converge monotonically to zero along the lines $\mathbb{R}z^i$ as $t \rightarrow \infty$ whereas, for $\alpha \in [0, \nu)$, the eigenmotions are oscillatory. \square

Example 2.2.10. (Inverted pendulum). Consider the cart pendulum system of Example 1.3.4 with no damping. The matrix A of the linearization about the upright position is of the form

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{32} & 0 & 0 \\ 0 & a_{42} & 0 & 0 \end{bmatrix} \quad (51)$$

where $a_{32} = M_0^{-1}m^2l^2g$, $a_{42} = M_0^{-1}(M + m)mgl$, see (1.3.32). Now

$$\det(\lambda I_n - A) = \lambda^2(\lambda - \sqrt{a_{42}})(\lambda + \sqrt{a_{42}}). \quad (52)$$

So A has the following eigenvalues and eigenvectors

$$\begin{aligned} \lambda_1 &= 0 & , & \quad z^1 = [1, 0, 0, 0]^\top \\ \lambda_2 &= \sqrt{a_{42}} & , & \quad z^2 = [1, a_{42}/a_{32}, \sqrt{a_{42}}, \sqrt{a_{42}} a_{42}/a_{32}]^\top \\ \lambda_3 &= -\sqrt{a_{42}} & , & \quad z^3 = [1, a_{42}/a_{32}, -\sqrt{a_{42}}, -\sqrt{a_{42}} a_{42}/a_{32}]^\top. \end{aligned}$$

By (52) we have $m(\lambda_1) = 2$, but the eigenspace of λ_1 is only one dimensional, so there is a generalized eigenvector $z^{1,2}$ of second order, for example $z^{1,2} = [0, 0, 1, 0]^\top$. All the eigenvalues are real and the corresponding eigenmotions are

$$e^{At} z^1 \equiv z^1, \quad e^{At} z^2 = e^{\sqrt{a_{42}}t} z^2, \quad e^{At} z^3 = e^{-\sqrt{a_{42}}t} z^3.$$

The first is an equilibrium state (upright position of the pendulum), the second is an expanding motion whilst the last is a contracting motion. The generalized eigenmode corresponding to the initial state $z^{1,2}$ is (see (34))

$$e^{At}z^{1,2} = z^{1,2} + tAz^{1,2} = z^{1,2} + tz^1.$$

As $t \rightarrow \infty$ we see that one eigenmode tends to zero, one is stationary and the other two both go to infinity (one exponentially and the other linearly). These results for the linearized model indicate (not unexpectedly) that the inverted pendulum has a relatively complicated dynamics which will probably be difficult to control. \square

By superposition of the eigenmotions of a system (24), different patterns of free motions around the origin can be generated depending on the numbers of contracting, expanding, stationary or periodic eigenmotions of the system. For a two dimensional linear system it is possible to give a complete classification of the flow patterns around the equilibrium state $\bar{x} = 0$. Let (z^1, z^2) be a basis of \mathbb{R}^2 such that the matrix representation of $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with respect to this basis is in real Jordan canonical form. There are three types of 2×2 matrices in real Jordan form

$$(i) \quad \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (ii) \quad \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \quad (iii) \quad \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}$$

where $\lambda_1, \lambda_2, \lambda, \alpha, \beta \in \mathbb{R}$. The corresponding free motions through any initial point $a \in \mathbb{R}^2$ are

$$(i) \quad \begin{bmatrix} a_1 e^{\lambda_1 t} \\ a_2 e^{\lambda_2 t} \end{bmatrix}, \quad (ii) \quad \begin{bmatrix} a_1 e^{\lambda t} + a_2 t e^{\lambda t} \\ a_2 e^{\lambda t} \end{bmatrix}, \quad (iii) \quad \begin{bmatrix} a_1 e^{\alpha t} \cos \beta t - a_2 e^{\alpha t} \sin \beta t \\ a_1 e^{\alpha t} \sin \beta t + a_2 e^{\alpha t} \cos \beta t \end{bmatrix}.$$

As a result we obtain for any $A \in \mathbb{R}^2$ a phase portrait of $\dot{x} = Ax$ around the origin which coincides qualitatively with exactly one of the patterns shown in Figure 2.2.3. The first six pictures correspond to the case (i), the next three pictures to case (ii) and the last six pictures to case (iii). Clearly the particular phase portrait will depend on the vectors z^1, z^2 and the magnitude of λ_1 and $\lambda_2, \lambda, \alpha$ and β .

In order to obtain a picture of the flow of a *nonlinear* time-invariant differentiable system with state space \mathbb{R}^2 , an important first step is to determine the phase portraits of its linearizations around each of its equilibrium states. In a second step global features have to be specified, such as limit cycles, connections between saddle points (separatrices), connections to infinity (unbounded orbits), periodic orbits etc. There is a rich qualitative theory of differential systems in the plane (see *Notes and References*). We will not develop this. Instead we conclude this subsection with an illustrative example showing the phase portrait of a simple nonlinear system with two distinct equilibria (a saddle point and a stable focus, see Figure 2.2.4).

Example 2.2.11. (Nonlinear oscillator). Consider an oscillator with nonlinear restoring force $\ddot{\xi} + 2\dot{\xi} + 5\xi + \xi^2 = 0$ or

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -5 & -2 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ -x_1^2(t) \end{bmatrix}. \quad (53)$$

Contrary to the linear oscillator (see Example 2.2.9), we have *two* equilibrium states $[0, 0]^\top$

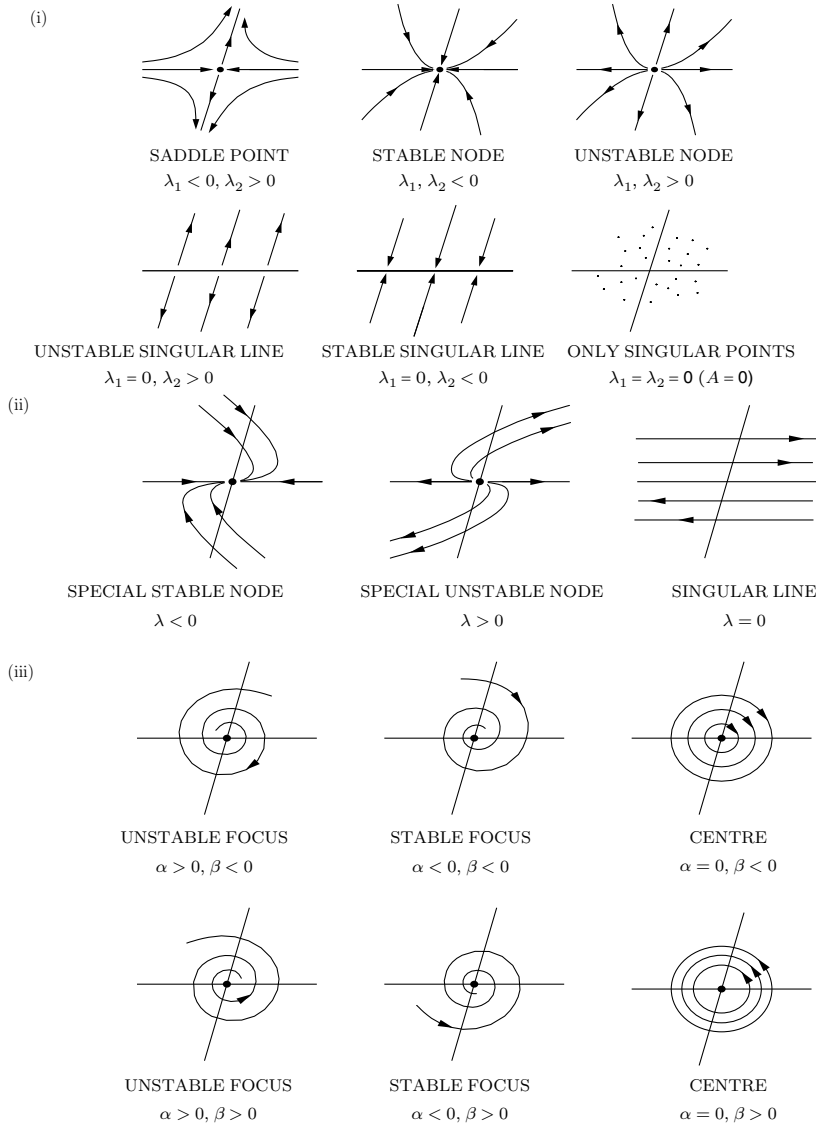


Figure 2.2.3: Phase portraits of two dimensional linear systems $\dot{x} = Ax$

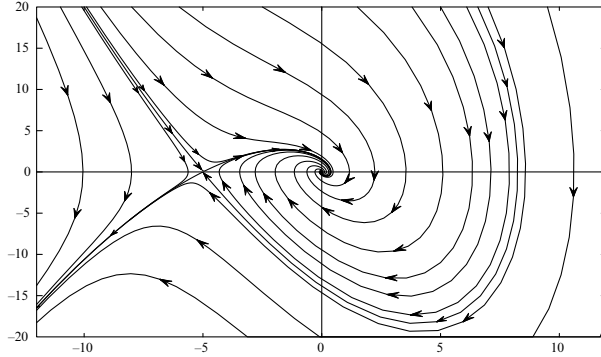


Figure 2.2.4: Phase portrait of the nonlinear oscillator (53) with two equilibria

and $[-5, 0]^\top$. The matrix of the linearization about $[0, 0]^\top$ is $A = \begin{bmatrix} 0 & 1 \\ -5 & -2 \end{bmatrix}$ which has eigenvalues $-1 \pm 2i$. Thus the local phase portrait of (53) around the origin has the form of a (nonlinear) *stable focus*. The matrix of the linearized system about $[-5, 0]^\top$ is $A = \begin{bmatrix} 0 & 1 \\ 5 & -2 \end{bmatrix}$ and this matrix has real eigenvalues $-1 \pm \sqrt{6}$ with corresponding eigenvectors $z^1 = [1, -1 + \sqrt{6}]^\top$, $z^2 = [1, -1 - \sqrt{6}]^\top$. Thus $[-5, 0]^\top$ is a *saddle point* of the nonlinear system (53). Figure 2.2.4 shows the actual trajectories of the nonlinear oscillator. \square

2.2.3 Free Motions of Time-Invariant Linear Difference Systems

As in the differentiable case, every free motion of a discrete time system (22) can be represented as a superposition of generalized eigenmotions. Therefore the analysis of the free system

$$x(t+1) = Ax(t) \quad (54)$$

reduces more or less to the spectral analysis of the operator A . If $z \in \ker(A - \lambda I)^m$ is a generalized eigenvector of $A \in \mathbb{C}^{n \times n}$ corresponding to the eigenvalue $\lambda \in \sigma(A)$, the associated generalized eigenmotion of (54) in \mathbb{C}^n is given by

$$\begin{aligned} A^t z &= [(A - \lambda I) + \lambda I]^t z = \sum_{\nu=0}^t \binom{t}{\nu} \lambda^{t-\nu} (A - \lambda I)^\nu z \\ &= \lambda^{t+1-m} \sum_{\nu=0}^{m-1} \binom{t}{\nu} \lambda^{m-1-\nu} (A - \lambda I)^\nu z, \quad t \geq m-1. \end{aligned} \quad (55)$$

If $\lambda = 0$ and z is an associated generalized eigenvector of order m then the free motions $A^t z$ ends at zero after m steps. This convergence to zero in finite time cannot occur in the continuous time case. However, if $\lambda \neq 0$ the generalized complex eigenmotion (55) is, as in the differentiable case, the product of an exponential term $(\lambda^t / \lambda^{m-1})$ in t and a vector polynomial $\sum_{\nu=0}^{m-1} \binom{t}{\nu} \lambda^{m-1-\nu} (A - \lambda I)^\nu z$ of degree $m-1$

in t .

In the case of an eigenvector ($m = 1$) we get the complex eigenmotion

$$A^t z = \lambda^t z, \quad t \in \mathbb{N}. \quad (56)$$

Now suppose $A \in \mathbb{R}^{n \times n}$. If $\lambda = r(\cos \theta + \imath \sin \theta) \in \sigma(A) \setminus \mathbb{R}$ and z is an associated eigenvector of A , the real eigenmotions corresponding to the pair of eigenvalues $\lambda, \bar{\lambda}$ and associated eigenvectors z, \bar{z} are

$$\begin{aligned} A^t(\operatorname{Re} z) &= \operatorname{Re} A^t z = r^t[(\cos \theta t) \operatorname{Re} z - (\sin \theta t) \operatorname{Im} z], \quad t \in \mathbb{N} \\ A^t(\operatorname{Im} z) &= \operatorname{Im} A^t z = r^t[(\cos \theta t) \operatorname{Im} z + (\sin \theta t) \operatorname{Re} z], \quad t \in \mathbb{N}. \end{aligned} \quad (57)$$

If z is a generalized eigenvector of order m of A for λ the associated generalized real eigenmotions for $t \geq m - 1$ are

$$\begin{aligned} A^t(\operatorname{Re} z) &= \sum_{\nu=0}^{m-1} \binom{t}{\nu} r^{t-\nu} [\cos(t-\nu)\theta \operatorname{Re}(A-\lambda I)^\nu z - \sin(t-\nu)\theta \operatorname{Im}(A-\lambda I)^\nu z] \\ A^t(\operatorname{Im} z) &= \sum_{\nu=0}^{m-1} \binom{t}{\nu} r^{t-\nu} [\cos(t-\nu)\theta \operatorname{Im}(A-\lambda I)^\nu z + \sin(t-\nu)\theta \operatorname{Re}(A-\lambda I)^\nu z]. \end{aligned} \quad (58)$$

As a discrete time counterpart to Corollaries 2.2.6 and 2.2.8 we obtain

Proposition 2.2.12. *Suppose $A \in \mathbb{K}^{n \times n}$, then very free motion of (54) in \mathbb{K}^n can be represented as a sum of generalized eigenmotions of the form (55) if $\mathbb{K} = \mathbb{C}$ and (58) if $\mathbb{K} = \mathbb{R}$. If A is diagonalizable over \mathbb{C} then all free motions of (54) in \mathbb{K}^n are superpositions of eigenmotions of the form (56) if $\mathbb{K} = \mathbb{C}$ and (57) if $\mathbb{K} = \mathbb{R}$.*

Example 2.2.13. (Fibonacci's model). In 1202 the mathematician L. Fibonacci (1180 - 1240) introduced a model of a fictitious rabbit population which is a simple example of a population model with age structure. Assume that a single pair of rabbits starts the population. They reproduce twice, once at time 1 and once at time 2, then die. At each reproduction they produce a new pair of rabbits, one male and one female. These will go on to reproduce twice etc.. The resulting dynamics of the population is summarized in Table 2.2.5.

Time	0	1	2	3	4	5	6
Age group-1	1	1	2	3	5	8	13
Age group-2	0	1	1	2	3	5	8

Table 2.2.5: Number of rabbits pairs

If we denote by $x_1(t)$ and $x_2(t)$ the numbers of rabbit pairs in the first and second age group at time t and set $x(t) = [x_1(t), x_2(t)]^\top$ we obtain the following population model

$$x(t+1) = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} x(t), \quad x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (59)$$

Alternatively, the evolution of $\xi(\cdot) = x_1(\cdot)$ can be described by the second order difference equation (Fibonacci Renewal Equation)

$$\xi(t+2) = \xi(t+1) + \xi(t), \quad t \in \mathbb{N}$$

with initial conditions $\xi(0) = \xi(1) = 1$. The resulting sequence

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \dots$$

is called the *Fibonacci sequence* and plays a role in various fields of mathematics. The eigenvalues of the matrix in (59) are $\lambda_{1,2} = (1 \pm \sqrt{5})/2$ and the corresponding real modes are

$$z^1(t) = \left(\frac{1 + \sqrt{5}}{2} \right)^t \begin{bmatrix} (1 + \sqrt{5})/2 \\ 1 \end{bmatrix}, \quad z^2(t) = \left(\frac{1 - \sqrt{5}}{2} \right)^t \begin{bmatrix} (1 - \sqrt{5})/2 \\ 1 \end{bmatrix}.$$

The solution of the initial value problem (59) is of the form $x(t) = \alpha_1 z^1(t) + \alpha_2 z^2(t)$ for some $(\alpha_1, \alpha_2) \in \mathbb{R}^2$. Since $x(0) = [1, 0]^\top$ we obtain $\alpha_1 = 1/\sqrt{5}$, $\alpha_2 = -1/\sqrt{5}$ and hence the following analytic expression for the Fibonacci sequence

$$\xi(t) = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^{t+1} - \left(\frac{1 - \sqrt{5}}{2} \right)^{t+1} \right].$$

□

2.2.4 Infinite Dimensional Systems

In Example 2.1.25 we showed that the state space for a delay equation is an infinite dimensional vector space. This is also the case for the partial differential equation described in Section 1.6. It is natural to ask therefore, whether there are results for infinite dimensional linear systems similar to those for time-invariant linear systems defined on finite dimensional vector spaces. The answer is often yes, but to develop these results in a general way is beyond the scope of this book. Instead we analyze the one-dimensional heat equation (taken from Section 1.6) in some detail and illustrate some of the main ideas and difficulties. As explained in Section 1.6 the evolution of the temperature in a heated metal bar of length ℓ with a fixed constant temperature at its ends can be determined via the equations

$$\frac{\partial \theta}{\partial t}(\xi, t) = k \frac{\partial^2 \theta}{\partial \xi^2}(\xi, t) + b(\xi)u(t), \quad t \in (0, \infty), \quad \xi \in (0, \ell) \quad (60a)$$

$$\theta(0, t) = \theta(\ell, t) = 0, \quad t \in (0, \infty), \quad (60b)$$

$$\theta(\xi, 0) = \theta_0(\xi), \quad 0 \leq \xi \leq \ell. \quad (60c)$$

The output equation is

$$y(t) = \int_0^\ell c(\xi)\theta(\xi, t)d\xi = \langle c(\cdot), \theta(\cdot, t) \rangle_{L^2} \quad (61)$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is the inner product on $L^2(0, \ell; \mathbb{R})$. We will show that in a suitable setting we can write this controlled partial differential equation and the output equation in the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad t \in (0, \infty), \quad x(0) = x^0, \\ y(t) &= Cx(t), \end{aligned} \quad (62)$$

where the state space is an infinite dimensional Hilbert space. Let $X = L^2(0, \ell; \mathbb{R})$ and denote by $\mathcal{D}(A)$ the linear space

$$\mathcal{D}(A) = \{x; x(\cdot) \in \mathcal{C}^2([0, \ell], \mathbb{R}), x(0) = x(\ell) = 0\}$$

where $\mathcal{C}^2([0, \ell]; \mathbb{R})$ is the vector space of twice continuously differentiable functions on $[0, \ell]$ (at 0, and ℓ one-sided derivatives are considered). $\mathcal{D}(A)$ can be viewed as a linear subspace of X and will be endowed with the corresponding L^2 norm. We will assume

$$b(\cdot), \theta_0(\cdot) \in \mathcal{D}(A), \quad c(\cdot) \in X \quad \text{and} \quad \mathcal{U} = \mathcal{C}(\mathbb{R}_+, \mathbb{R}).$$

The operators A, B, C in (62) are defined by

$$\begin{aligned} A &: \mathcal{D}(A) \rightarrow X, & (Az)(\xi) &= k \frac{d^2 z}{d\xi^2}(\xi), & \xi &\in (0, \ell), & z(\cdot) &\in \mathcal{D}(A) \\ B &: \mathbb{R} \rightarrow X, & (Bu)(\xi) &= b(\xi)u, & \xi &\in [0, \ell], & u &\in \mathbb{R} \\ C &: X \rightarrow \mathbb{R}, & Cz &= \langle c(\cdot), z(\cdot) \rangle_{L^2}, & & & z(\cdot) &\in X. \end{aligned} \quad (63)$$

Before we can discuss the relationship between the operator differential equation (62) and the partial differential equation (60) we need to define what is meant by a solution of these equations.

Definition 2.2.14. A continuous function $\theta(\cdot, \cdot) : [0, \ell] \times \mathbb{R}_+ \rightarrow \mathbb{R}$ is said to be a solution of (60) if all the derivatives

$$\frac{\partial \theta}{\partial t}(\cdot, \cdot), \quad \frac{\partial \theta}{\partial \xi}(\cdot, \cdot), \quad \frac{\partial^2 \theta}{\partial \xi^2}(\cdot, \cdot)$$

exist and are continuous on $(0, \ell) \times (0, \infty)$ and the three equations in (60) are satisfied.

Definition 2.2.15. A continuous function $x(\cdot) : \mathbb{R}_+ \rightarrow X$ is a solution of (62) if it is Fréchet differentiable on $(0, \infty)$ and satisfies (62) in X .

Let us now construct the solution of (60). At the beginning of this section we saw that for time-invariant finite dimensional systems both the free and forced motions can be determined via a semigroup $\Phi(t)$ which is constructed from the (generalized) eigenmotions. This suggests that we should examine the eigenvalues and associated eigenvectors of the operator A ,

$$A\psi = \lambda\psi, \quad \psi \in \mathcal{D}(A), \quad \psi \neq 0, \quad \lambda \in \mathbb{C} \quad (64)$$

or, equivalently, the nontrivial solutions of

$$k \frac{d^2 \psi}{d\xi^2}(\xi) = \lambda \psi(\xi), \quad \xi \in (0, \ell), \quad \psi(0) = \psi(\ell) = 0. \quad (65)$$

The differential equation in (65) has the general solution

$$\psi(\xi) = ae^{\sqrt{\lambda/k}\xi} + be^{-\sqrt{\lambda/k}\xi}, \quad a, b \in \mathbb{C}.$$

To satisfy the boundary condition $\psi(0) = 0$ we must have $a + b = 0$, hence $\psi(\xi) = 2a \sinh \sqrt{\lambda/k} \xi$. Now let $\sqrt{\lambda/k} \ell = \alpha + \imath\beta$ $\alpha, \beta \in \mathbb{R}$, then the second boundary condition $\psi(\ell) = 0$ implies

$$\sinh \sqrt{\lambda/k} \ell = \sinh(\alpha + \imath\beta) = \sinh \alpha \cos \beta + \imath \cosh \alpha \sin \beta = 0.$$

As a consequence we obtain $\alpha = 0$ and $\beta \in \mathbb{Z}\pi$, $\beta \neq 0$ (since $\beta = 0$ would yield the trivial solution). Hence $\sqrt{\lambda/k} \ell = \pm \imath n\pi$ for $n \in \mathbb{N}^*$. So there is an infinite sequence of eigenvalues and associated eigenvectors

$$\lambda_n = -k \frac{n^2 \pi^2}{\ell^2}, \quad \psi_n(\xi) = \sqrt{\frac{2}{\ell}} \sin \frac{n\pi\xi}{\ell}, \quad n \in \mathbb{N}^* \quad (66)$$

with corresponding eigenmotions

$$\theta_n(\xi, t) = \sqrt{\frac{2}{\ell}} \exp\left(-k \frac{n^2 \pi^2 t}{\ell^2}\right) \sin \frac{n\pi\xi}{\ell}, \quad t \geq 0, \quad n \in \mathbb{N}^*.$$

It is easily verified that the eigenvectors $\psi_n(\cdot) \in \mathcal{D}(A) \subset L^2(0, \ell; \mathbb{R})$ are orthogonal to each other in the Hilbert space $L^2(0, \ell; \mathbb{R})$

$$\langle \psi_n, \psi_m \rangle_{L^2} = \int_0^\ell \psi_n(\xi) \psi_m(\xi) d\xi = \delta_{mn}. \quad (67)$$

In fact it is known from the theory of Fourier series that the functions $\{\psi_n\}_{n \in \mathbb{N}^*}$ form an orthonormal basis of the Hilbert space $L^2(0, \ell; \mathbb{R})$ in the sense that (67) holds and any $z(\cdot) \in L^2(0, \ell; \mathbb{R})$ can be expressed in a unique way as an infinite linear combination of the ψ_n 's, viz $z(\xi) = \sum_{n=1}^\infty \alpha_n \psi_n(\xi)$. Here $\alpha_n = \langle z, \psi_n \rangle_{L^2}$, $n = 1, 2, \dots$, and the equality is to be interpreted in the sense of $L^2(0, \ell; \mathbb{R})$, i.e. $\lim_{N \rightarrow \infty} \|z(\cdot) - \sum_{n=1}^N \alpha_n \psi_n(\cdot)\|_{L^2} = 0$ (see Section A.3). If, in particular, $z(\cdot) \in \mathcal{D}(A)$, then

$$\alpha_n = \int_0^\ell z(\xi) \psi_n(\xi) d\xi = -\frac{\ell^2}{n^2 \pi^2} \int_0^\ell \frac{d^2 z}{d\xi^2}(\xi) \psi_n(\xi) d\xi, \quad n \in \mathbb{N}$$

on integrating by parts twice and using the fact that $z(0) = z(\ell) = 0$. Thus for $z(\cdot) \in \mathcal{D}(A)$, there exists a constant M such that

$$|\alpha_n| \leq \frac{M}{n^2}, \quad n \in \mathbb{N}^*. \quad (68)$$

Now, since the pre-Hilbert space $\mathcal{D}(A) \subset X$ has a basis consisting of eigenvectors of the operator A , if we mirror the development for time-invariant finite dimensional systems, we would expect the associated semigroup on $\mathcal{D}(A)$ to be given by the superposition of eigenmotions

$$(\Phi(t)z(\cdot))(\xi) = \sum_{n=1}^\infty e^{\lambda_n t} \langle z(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi). \quad (69)$$

Then the solution of the controlled equation (60) would be

$$\theta(\xi, t) = (\Phi(t)\theta_0(\cdot))(\xi) + \int_0^t (\Phi(t-s)b(\cdot))(\xi)u(s)ds$$

(see (19)) or more explicitly

$$\theta(\xi, t) = \sum_{n=1}^{\infty} e^{\lambda_n t} \langle \theta_0(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi) + \int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-s)} \langle b(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi) u(s) ds. \quad (70)$$

Using (68) with constants M_0 , resp. M_b for $\theta_0(\cdot)$ and $b(\cdot) \in \mathcal{D}(A)$, it is easy to see that the series in (70) is uniformly absolutely convergent in $(\xi, t) \in [0, \ell] \times [0, t_1]$ for arbitrary $t_1 > 0$ and

$$|\theta(\xi, t)| \leq \sqrt{\frac{2}{\ell}} M_0 \sum_{n=1}^{\infty} \frac{e^{\lambda_n t}}{n^2} + \frac{\sqrt{2\ell^3}}{k\pi^2} M_b \sup_{0 \leq s \leq t} |u(s)| \sum_{n=1}^{\infty} \frac{1 - e^{\lambda_n t}}{n^4}.$$

Therefore $\theta(\cdot, \cdot)$ is well defined and continuous on $[0, \ell] \times \mathbb{R}_+$.

Theorem 2.2.16. *Given $\theta_0 \in \mathcal{D}(A)$, $u(\cdot) \in \mathcal{U}$, then (60) has exactly one solution in the sense of Definition 2.2.14 and this solution is given by the function $\theta(\cdot, \cdot)$ defined by (70).*

Proof: It follows directly from the definition, from (70) and the above convergence result that $\theta(\cdot, \cdot)$ satisfies conditions (60b), (60c). In order to prove that $\theta(\cdot, \cdot)$ solves the partial differential equation (60a) one proceeds as follows. First it is shown that the partial derivatives $\frac{\partial \theta}{\partial t}(\xi, t)$ and $k \frac{\partial^2 \theta}{\partial \xi^2}(\xi, t)$ can be calculated from (70) term by term. This can be done by proving that the resulting series are uniformly absolutely convergent on $[0, \ell] \times [0, t_1]$ for arbitrary $t_1 > 0$ (making use of the estimate (68) as above). Then comparing the two series for $\frac{\partial \theta}{\partial t}(\xi, t)$ and $k \frac{\partial^2 \theta}{\partial \xi^2}(\xi, t)$ term by term it becomes clear that (60a) holds. We omit the details.

To prove uniqueness we assume that there is a second solution $\hat{\theta}(\cdot, \cdot)$ and set $e(\cdot, \cdot) = (\theta - \hat{\theta})(\cdot, \cdot)$. Then $e(\cdot, \cdot)$ must satisfy

$$\begin{aligned} \frac{\partial e}{\partial t}(\xi, t) &= k \frac{\partial^2 e}{\partial \xi^2}(\xi, t), & t \in (0, \infty), & \xi \in (0, \ell) \\ e(0, t) &= e(\ell, t) = 0, & t \in \mathbb{R}_+ \\ e(\xi, 0) &= 0, & \xi \in [0, \ell]. \end{aligned}$$

Consider the function $E(t) = \int_0^\ell e^2(\xi, t) d\xi$, then

$$\frac{dE}{dt}(t) = 2 \int_0^\ell e(\xi, t) \frac{\partial e}{\partial t}(\xi, t) d\xi = 2k \int_0^\ell e(\xi, t) \frac{\partial^2 e}{\partial \xi^2}(\xi, t) d\xi = -2k \int_0^\ell \left(\frac{\partial e}{\partial \xi}(\xi, t) \right)^2 d\xi$$

on integration by parts. So $E(t)$ is non-increasing, but we have $E(0) = 0$ and $E(t) \geq 0$. Hence $e(\cdot, \cdot) \equiv 0$ and the solution of (60) is unique. \square

Up until now we have analyzed (60) on the pre-Hilbert space $\mathcal{D}(A)$. For technical reasons it is advantageous to associate with (60) a dynamical system with the whole Hilbert space X as state space and an extended space of control functions. To achieve this suppose that $\theta_0 \in X$ and $u(\cdot) \in \hat{\mathcal{U}} = L^2_{\text{loc}}(\mathbb{R}_+, \mathbb{R})$. Then the series in (70) are still absolutely convergent for each $t \geq 0$ with respect to the norm of the Hilbert space $X = L^2(0, \ell; \mathbb{R})$. Hence the map

$$\varphi : \{(t, t_0) \in \mathbb{R}^2; t \geq t_0\} \times X \times \hat{\mathcal{U}} \rightarrow X, \quad (t; t_0, \theta_0, u(\cdot)) \mapsto \varphi(t; t_0, \theta_0, u(\cdot))(\cdot)$$

is well defined by $\varphi(t; t_0, \theta_0, u(\cdot))(\cdot) = z(\cdot)$ where for $\xi \in [0, \ell]$

$$z(\xi) = \sum_{n=1}^{\infty} e^{\lambda_n(t-t_0)} \langle \theta_0(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi) + \int_{t_0}^t \sum_{n=1}^{\infty} e^{\lambda_n(t-s)} \langle b(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi) u(s) ds. \quad (71)$$

This formula extends the solution formula (70) so that it is applicable for all $(\theta_0, u(\cdot)) \in X \times \hat{\mathcal{U}}$ instead of $(\theta_0, u(\cdot)) \in \mathcal{D}(A) \times \mathcal{U}$. The coordinates $\alpha_n(t)$, $n \in \mathbb{N}^*$ of $\varphi(t; t_0, \theta_0, u(\cdot))$ with respect to the basis $(\psi_n)_{n \in \mathbb{N}^*}$ of X are given by

$$\alpha_n(t) = e^{\lambda_n(t-t_0)} \langle \theta_0(\cdot), \psi_n(\cdot) \rangle_{L^2} + \int_{t_0}^t e^{\lambda_n(t-s)} \langle b(\cdot), \psi_n(\cdot) \rangle_{L^2} u(s) ds$$

and hence satisfy the differential equations

$$\dot{\alpha}_n(t) = \lambda_n \alpha_n(t) + \langle b(\cdot), \psi_n(\cdot) \rangle_{L^2} u(t).$$

Using this fact it is easy to see that φ satisfies the axioms of a state transition map (Definition 2.1.1). Therefore, defining the output map by $\eta(x, u) = \langle c, x \rangle_{L^2}$, we obtain a dynamical system $\Sigma = (\mathbb{R}, \mathbb{R}, \hat{\mathcal{U}}, X, \mathbb{R}, \varphi, \eta)$ with state space $X = L^2(0, \ell; \mathbb{R})$. This system is obviously linear and time-invariant. The associated operator semi-group (69) describing the free motions of the system is given by (69) where $z(\cdot)$ is now allowed to vary in X . The associated input-state map $\Theta(t, 0) : \hat{\mathcal{U}} \rightarrow X$ (see (7)) is given by

$$(\Theta(t, 0)u(\cdot))(\xi) = \int_0^t \sum_{n=1}^{\infty} e^{\lambda_n(t-s)} \langle b(\cdot), \psi_n(\cdot) \rangle_{L^2} \psi_n(\xi) u(s) ds, \quad \xi \in [0, \ell].$$

For initial states $x(0) = \theta_0 \in \mathcal{D}(A)$ and controls $u(\cdot) \in \mathcal{U}$, we have by Theorem 2.2.16

$$\varphi(t; 0, \theta_0, u(\cdot))(\cdot) = \theta(\cdot, t), \quad t \geq 0,$$

where $\theta(\cdot, \cdot)$ solves the partial differential equation (60). Hence $t \mapsto \varphi(t; 0, \theta_0, u(\cdot))(\cdot)$ describes the evolution of the temperature profile along the metal bar under the influence of the control $u(\cdot)$. A typical controlled temperature profile is shown in Figure 2.2.6. Here the initial temperature is zero along the bar and the object of the control is to steer the temperature to the profile $\theta(\xi, T) = \sin \pi \xi$, $\xi \in [0, \ell]$ in time $[0, 5]$. Since the process of heat propagation is not reversible we expect that the system Σ is not reversible. In fact the series in (71) do not converge in $X = L^2(0, \ell; \mathbb{R})$ for $t < t_0$. Moreover, it follows from an analysis of (71) that $\varphi(t; t_0, \theta_0, 0) \in \mathcal{D}(A)$ for arbitrary $t > 0$ and $\theta_0 \in X$, hence any free trajectory enters the dense subspace $\mathcal{D}(A) \subset X$ of smooth temperature profiles immediately after leaving the possibly discontinuous initial temperature profile θ_0 .

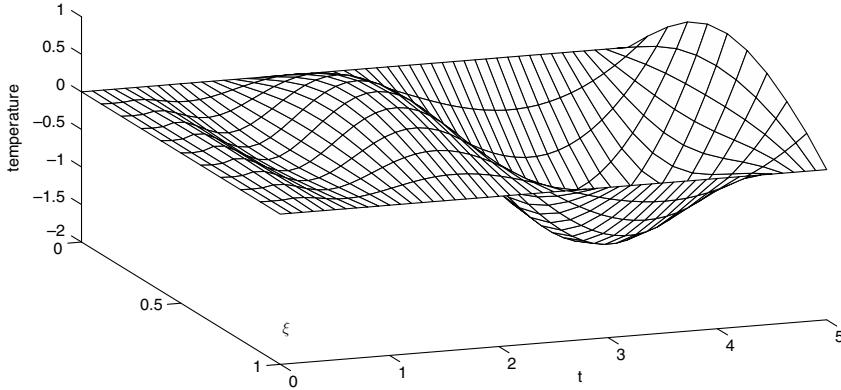


Figure 2.2.6: Evolution of the controlled temperature profile

Remark 2.2.17. We have used the abstract differential equation (62) only as a heuristic tool in order to obtain the expression (70) for a solution of the partial differential equation (60). In fact it is possible to show that, for every $x^0 \in X$, the semigroup $\Phi(t)$ on X yields a solution $\Phi(t)x^0$ of (62) for $u(t) \equiv 0$ in the sense of Definition 2.2.15, see *Notes and References*. Then using similar estimates as in the proof of Theorem 2.2.16 one can show that, for $\theta_0 \in \mathcal{D}(A)$, $u(\cdot) \in \mathcal{U}$, the solution $\theta(\cdot, \cdot)$ of (60) gives rise to a unique solution $t \mapsto x(t) = \theta(\cdot, t)$ of (62). For the more general initial condition $x(0) = \theta_0 \in X$ and arbitrary controls $u(\cdot) \in \mathcal{U} = L^2_{\text{loc}}(\mathbb{R}_+; \mathbb{R})$, it can be shown that the solution $t \mapsto x(t) = \varphi(t; t_0, \theta_0, u(\cdot))$ is a *mild solution* of (62), i.e. satisfies the variation-of-constants formula

$$x(t) = \Phi(t)x^0 + \int_0^t \Phi(t-s)bu(s)ds$$

where the integral is a Bochner integral [538].

However, one should not be misled by the formal analogy with the finite dimensional situation. There are essential differences between the theories of finite and of infinite dimensional linear systems, not all of which are illustrated by the above example.

- (i) As in the example the operator A is in general an *unbounded* linear operator and is not defined on the whole state space but only on a dense subspace $\mathcal{D}(A)$ of X .
- (ii) The initial value problem (62) does not necessarily have a differentiable solution and the mild solution given by the variation-of-constants formula will in general only be a solution of (62) in the sense of Definition 2.2.15 if $x^0 \in \mathcal{D}(A)$ and $u(\cdot)$ is sufficiently smooth.
- (iii) In contrast with the example the spectrum of the operator A will *not* in general consist of eigenvalues only (see Section A.4) and even if this is the case the generalized eigenvectors of A will *not* in general span the whole state space X .
- (iv) The operators B and C need *not* be *bounded* (e.g. if the control is acting through the boundary conditions or if point measurements are taken).

As a consequence of (iii) the semigroup Φ cannot always be constructed via the eigenmotions by using a series representation as in (69). The relationship between semigroups of operators $(\Phi(t))_{t \in \mathbb{R}_+}$ and their “infinitesimal generators” A must be put on another footing and criteria must be found under which a given unbounded linear operator $A : \mathcal{D}(A) \rightarrow X$

“generates” a semigroup $(\Phi(t))_{t \in \mathbb{R}_+}$. This is done in the theory of operator semigroups, see Remark 5.5.44 and *Notes and References*. \square

2.2.5 Exercises

1. Prove that a linear system with continuous coefficient matrices $A(t)$, $B(t)$, $C(t)$ and $D(t)$ is time-invariant if and only if the matrix functions are constant.

2. Determine the evolution operator $\Phi(t, t_0)$ of the time-varying scalar differential equation

$$\dot{x}(t) = a(t)x(t)$$

where $a(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is continuous. Show that the result can be generalized to vector differential equations

$$\dot{x}(t) = A(t)x(t)$$

where $A(\cdot)$ is a continuous $n \times n$ -matrix function such that $A(t)A(s) = A(s)A(t)$, $s, t \in \mathbb{R}$. Apply this to $A(t) = a(t)A$ where $a(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is continuous and $A \in \mathbb{R}^{n \times n}$.

3. If $A(\cdot) \in \mathcal{C}([t_0, t_1]; \mathbb{R}^{n \times n})$ is a continuous $n \times n$ -matrix function show that the solution $\Phi(t, t_0)$ of the matrix differential equation

$$\dot{X}(t) = A(t)X(t), \quad t_0 \leq t \leq t_1, \quad X(t_0) = I_n$$

can be obtained as the uniform limit of the recursive sequence

$$\begin{aligned} \Phi_0(t, t_0) &\equiv I_n, \quad t \in [t_0, t_1] \\ \Phi_k(t, t_0) &= I_n + \int_{t_0}^t A(s)\Phi_{k-1}(s, t_0)ds, \quad t \in [t_0, t_1] \end{aligned}$$

(cf. Hale (1980) [214, III.3]).

4. Find an appropriate state vector and determine matrices A, B, C, D for the discrete time linear system represented in the block diagram shown in Figure 2.2.7, where Δ is the unit time delay.

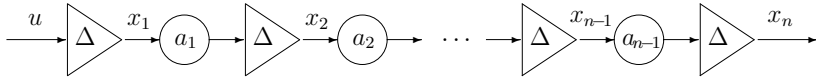


Figure 2.2.7: Block diagram of discrete time system

5. Consider the discrete time system (A, B, C, D) represented in Figure 2.2.7 (Ex. 4).

- Determine the spectrum of A .
- Compute A^t , $t \in \mathbb{N}$.
- Specify a basis of \mathbb{R}^n consisting of generalized eigenvectors of A .
- What can be said of the free motions of this system?

6. Compute A^t , $t \in \mathbb{N}$ and e^{At} , $t \in \mathbb{R}$ for the following matrices A where $\alpha, \beta > 0$

$$(i) \begin{bmatrix} 0 & \beta \\ \alpha & 0 \end{bmatrix}, \quad (ii) \begin{bmatrix} 0 & \beta \\ -\alpha & 0 \end{bmatrix}, \quad (iii) \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad (iv) \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

7. Compute the characteristic polynomial, the eigenvalues, the real modes and the free motions for the following systems and initial states

$$(i) \quad x(t+1) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} x(t), \quad x(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$(ii) \quad x(t+1) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x(t), \quad x(0) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

$$(iii) \quad \dot{x}(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ -2 & 0 & -1 \end{bmatrix} x(t), \quad x(0) = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

$$(iv) \quad \dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} x(t), \quad x(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

8. Let $\mathbb{Z}_3 = \{0, 1, 2\}$ be the field of integers modulo 3. Consider the finite linear machine over $\mathbb{K} = \mathbb{Z}_3$ represented in Figure 2.2.8 where Δ is the unit delay operator. The summer

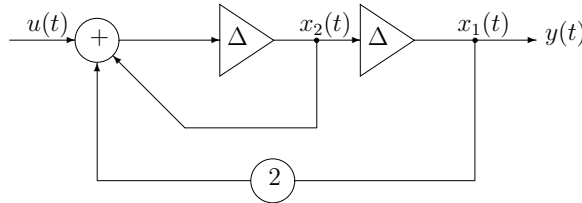


Figure 2.2.8: Finite linear machine

adds mod 3 and the gain multiplies mod 3.

(i) Determine the system equations

$$x(t+1) = Ax(t) + Bu(t), \quad t \in \mathbb{N}; \quad y(t) = Cx(t) + Du(t).$$

(ii) Show that $A^2 - A + I = 0$.

(iii) Determine the fundamental matrix A^t , $t \in \mathbb{N}$.

(Note that all calculations have to be carried out in \mathbb{Z}_3)

9. Consider the electrical circuit represented in Figure 1.4.4. If the voltage source $e(t)$ is taken as input $u(t)$ and the charge on the capacitor is $q(t)$ it is shown in Example 1.4.4 that

$$L\ddot{q}(t) + R\dot{q}(t) + (1/C)q(t) = u(t).$$

(i) If the current $y(t)$ through the inductor together with the charge on the capacitor are state variables and the output is the current through the inductor specify the matrices A , B , C describing the system.

(ii) Determine the eigenvalues, eigenvectors and real eigenmodes of the system in terms of the parameters L , $R \geq 0$, $C > 0$.

(iii) Under which conditions do oscillating eigenmotions occur? Express the frequency in terms of R , C , L . What happens if $R = 0$? What happens if C approaches 0? What happens if $L = 0$?

(Driver (1977), pp.119-122 [138])

2.2.6 Notes and References

The analysis of free motions of linear differentiable systems is contained in most books on ordinary differential equations. For basic results we refer to the introductory textbooks *Driver* (1977) [138], *Polking et al.* (2001) [417] and the books recommended in the *Notes and References* of Section 2.1. For analogous results on difference equations see *Agarwal* (1992) [5] and *Kelley and Peterson* (2001) [298]. The results from Linear Algebra required for the spectral analysis of linear time-invariant systems can be found in most textbooks on Matrix Theory. Standard references which contain many results which are useful in different areas of Linear Systems Theory are *Gantmacher* (1959) [182] and *Horn and Johnson* (1985) [264].

For the numerical problems involved in computing the exponential of a matrix, see *Mohler and Van Loan* (1978) [378] and Section 2.5.

An excellent textbook on control theoretic aspects of linear differentiable systems is *Brockett* (1970) [77] which also contains many results on *time-varying* linear systems as does the book of *Rugh* (1993) [442]. For discrete time linear control systems see *Ogata* (1987) [396], *Franklin et al.* (1998) [169], and a comprehensive book which treats continuous time and discrete time systems in parallel is *Oppenheim et al.* (1997) [399].

Some results and a number of references concerning the qualitative theory of differential systems in the plane can be found in Section 3.1.

Example 2.2.13 (Fibonacci's model) is discussed in *Hoppenstead* (1982) [262]. Many interesting examples of a non-technological kind are described in the introductory text by *Luenberger* (1979) [349]. Linear models for real life mechanical, electrical and electromechanical control systems are discussed in *Franklin et al.* (1986) [168].

A classical reference for the theory of operator semigroups which also contains the necessary functional analytic foundations is *Hille and Phillips* (1957) [231]. Other references are *Pazy* (1983) [406] and the graduate textbook of *Engel and Nagel* (2000) [152], the latter one contains many interesting examples and applications. A comprehensive introduction to the theory of infinite dimensional linear systems is the excellent text book by *Curtain and Zwart* (1995) [116] and a concise introduction to infinite dimensional as well as finite dimensional control systems is given in *Zabczyk* (1992) [541].

2.3 Linear Systems: Input–Output Behaviour

In this section we study the forced motions of systems of the form (2.17) (resp. (2.22)) and begin by explaining how they can be represented as superpositions of trajectories generated by impulse controls. Then we analyze the input-output behaviour of time-invariant linear systems in time domain. Norms on the input and output function spaces are introduced and conditions are given under which the input-output map of a system is a *bounded linear operator* between these normed spaces. Systems with this property are called *input-output stable*. In the second subsection Laplace and Fourier transforms are used to obtain a frequency domain representation of the input-output behaviour in terms of transfer matrices. Under a diagonalizability assumption a dyadic decomposition of these matrices is constructed and it is shown how the transfer matrix is related to the response of the system to sinusoidal input signals. Finally the relationship between the time domain and frequency domain representations is discussed and leads in the case of input-output stable systems to a computable formula for the norm of the input-output operator.

2.3.1 Input-Output Behaviour in Time Domain

In many applications only the input-output behaviour of a system is of interest. We have seen that in state space theory a *pointwise* approach is possible because the instantaneous state of the system at time t contains all the necessary information needed to determine the effect of the past inputs upon the present output. In general the present output of a dynamical system is not determined by its present input alone, but by the whole control function on the preceding time interval. So if the concept of state is dropped, a *functional* viewpoint must be adopted which looks at the signal as a whole rather than its values at certain points in time. The analysis of the input-output behaviour of a system involves the investigation of the dependence of the output *function* (“output signal”) on the input *function* (“input signal”). This is why functional analytical methods become important in this context (see Section A.3 and Section A.4). In this subsection we first initialize the linear state space systems (2.17) and (2.22) at $x(0) = 0$ and study the input-output behaviour on the time domain $T = \mathbb{R}_+$ (resp. \mathbb{N}). Then, under a stability assumption, we consider the input-output behaviour on the time domain $T = \mathbb{R}$ (resp. \mathbb{Z}). Throughout the section we assume that $U = \mathbb{K}^m$, $X = \mathbb{K}^n$ and $Y = \mathbb{K}^p$ for some given $m, n, p \in \mathbb{N}^*$.

Impulse Responses

For the systems (2.17) and (2.22) initialized at $x(0) = 0$ the dependence of the state and output trajectories on the control function is described by the *input-state map*

$$u(\cdot) \mapsto x(\cdot; u(\cdot)) = \varphi(\cdot; 0, 0, u(\cdot)), \quad u(\cdot) \in \mathcal{U} \quad (1)$$

and the *input-output map*

$$u(\cdot) \mapsto y(\cdot; u(\cdot)) = C\varphi(\cdot; 0, 0, u(\cdot)) + Du(\cdot), \quad u(\cdot) \in \mathcal{U}. \quad (2)$$

on the time domain $T = \mathbb{R}_+$ (resp. \mathbb{N}). We assume that in the discrete time case $\mathcal{U} = U^{\mathbb{N}}$ and in the continuous time case $\mathcal{U} = L_{\text{loc}}^1(\mathbb{R}_+; U)$. Obviously (1) is a special case of (2) with $Y = X$, $C = I_X$, $D = 0$ and thus it is sufficient to study (2). Let us first consider the discrete time case (2.22) with $T = \mathbb{N}$. Since $x(0) = 0$ we have

$$y(t; u(\cdot)) = Du(t) + \sum_{s=0}^{t-1} CA^{t-s-1}Bu(s), \quad u(\cdot) \in \mathcal{U}, \quad t \in \mathbb{N}. \quad (3)$$

(3) can be written in a more concise form using the *convolution of sequences*, see Section A.3.

Definition 2.3.1 (Convolution of sequences). If $g = (g(t))_{t \in \mathbb{N}}$, $v = (v(t))_{t \in \mathbb{N}} \in \mathbb{K}^{\mathbb{N}}$ are sequences over a field \mathbb{K} the *convolution* $y = g * v$ is defined to be the sequence $y = (y(t))_{t \in \mathbb{N}} \in \mathbb{K}^{\mathbb{N}}$ given by

$$y(t) = \sum_{s=0}^t g(t-s)v(s), \quad t \in \mathbb{N}. \quad (4)$$

It is an easy exercise to prove that the set $\mathbb{K}^{\mathbb{N}}$ of scalar sequences is a commutative ring with respect to the operations $+$ and $*$. In fact, $\mathbb{K}^{\mathbb{N}}$ is even an integral domain (i.e. has no zero divisors) and has the sequence $(1, 0, 0, \dots)$ as a unit element. Considering matrices and vectors with entries in this ring we define the convolution $\mathcal{G} * u$ of a sequence $\mathcal{G} = (\mathcal{G}(t))_{t \in \mathbb{N}}$ of $p \times m$ -matrices and a sequence $u = (u(t))_{t \in \mathbb{N}}$ of m -vectors to be the sequence of p -vectors given by

$$(\mathcal{G} * u)(t) = \sum_{s=0}^t \mathcal{G}(t-s)u(s) = \sum_{s=0}^t \mathcal{G}(s)u(t-s), \quad t \in \mathbb{N}. \quad (5)$$

Using this notation (3) can be written in the form

$$y(t) = (\mathcal{G} * u)(t), \quad t \in \mathbb{N} \quad (6)$$

where

$$\mathcal{G} = (D, CB, CAB, \dots, CA^{t-1}B, \dots) \in (\mathbb{K}^{p \times m})^{\mathbb{N}}. \quad (7)$$

The matrices $CA^{t-1}B$, $t \geq 1$ occurring in this sequence are called the *Markov parameters* of the system (2.22). Together with the matrix D of direct input-output coupling they completely determine the input-output behaviour of (2.22) under the condition that the system is initialized at $x(0) = 0$. The sequence \mathcal{G} can be viewed as a *weighting pattern* which determines the present output $y(t)$ as the weighted sum over the past and present inputs $u(s)$, $s \leq t$, see (5).

The sequence (7) can be obtained (at least theoretically) by a series of experiments as follows: Let (e^1, \dots, e^m) be the standard basis of \mathbb{K}^m and suppose the following input signals are applied to the system

$$w^j(t) = \begin{cases} e^j & \text{if } t = 0, \quad j \in \underline{m} \\ 0 & \text{if } t > 0 \end{cases}. \quad (8)$$

This particular test signal is called the j^{th} *unit impulse*. By (6) the corresponding output sequences are

$$y(t; u^j(\cdot)) = (\mathcal{G} * u^j)(t) = \mathcal{G}(t)e^j = \mathcal{G}^j(t)$$

where $\mathcal{G}^j(t)$ is the j^{th} column vector of $\mathcal{G}(t)$. Because of this property \mathcal{G} is also called the *impulse response* of the system (2.22). We see, therefore, that if we were able to take exact measurements of the output signals (state trajectories for the special case $C = I_X$, $D = 0$) corresponding to the m test signals $u^1(\cdot), \dots, u^m(\cdot)$, the input-output map (2) (resp. input-to-state map (1)) of the system would be completely determined.

Example 2.3.2. Consider the discrete time scalar system

$$\begin{aligned} x(t+1) &= ax(t) + bu(t), & x(0) &= 0 \\ y(t) &= x(t) \end{aligned} \tag{9}$$

where $a, b \in \mathbb{R}$, $b \neq 0$ are given. Solving this equation for $u(\cdot) = (1, 0, 0, \dots)$ we obtained the impulse response

$$g = (0, b, ab, a^2b, \dots).$$

Let us compute the *step response* of (9) which is the output $\bar{y}(\cdot)$ corresponding to the constant input $\bar{u}(t) \equiv 1$. We get from (5)

$$\bar{y}(t; u(\cdot)) = \sum_{s=1}^t a^{s-1}b = \begin{cases} tb & \text{if } a = 1 \\ \frac{1-a^t}{1-a}b & \text{if } a \neq 1 \end{cases}, \quad t \in \mathbb{N}.$$

Now $\bar{y}(t; u(\cdot)) - \bar{y}(t-1; u(\cdot)) = g(t)$ for $t \geq 1$. Furthermore $\lim_{t \rightarrow \infty} |\bar{y}(t; u(\cdot))| = \infty$ if $|a| > 1$ or $a = 1$, and $\lim_{t \rightarrow \infty} \bar{y}(t; u(\cdot)) = b(1-a)^{-1}$ if $|a| < 1$. Note that $\bar{x} = b(1-a)^{-1}$ is just the equilibrium state of the system (9) for the control $\bar{u}(t) \equiv 1$. \square

A similar analysis can be carried out for continuous time systems with $T = \mathbb{R}_+$. For simplicity we assume that there is no direct input-output coupling so that $D = 0$, then the output function of (2.17) corresponding to an input $u(\cdot) \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K}^m)$ and zero initial condition is given by

$$y(t; u(\cdot)) = C\varphi(t; 0, 0, u(\cdot)) = \int_0^t Ce^{A(t-s)}Bu(s)ds, \quad t \in \mathbb{R}_+. \tag{10}$$

Recall the following definition, see Section A.3.

Definition 2.3.3 (Convolution of functions). If $g, v \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K})$ the convolution of g and v is defined almost everywhere by

$$(g * v)(t) = \int_0^t g(t-s)v(s)ds, \quad \text{a.e. } t \in \mathbb{R}_+. \tag{11}$$

The integral in (11) exists almost everywhere and defines a locally integrable function. $L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K})$ is in fact a commutative ring over the field \mathbb{K} with respect to the operations $+$ and $*$. Considering matrices and vectors with entries in this ring we define the convolution of a matrix function $\mathcal{G}(\cdot) \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K}^{p \times m})$ and a vector function $u(\cdot) \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K}^m)$ by

$$(\mathcal{G} * u)(t) = \int_0^t \mathcal{G}(t-s)u(s)ds = \int_0^t \mathcal{G}(s)u(t-s)ds, \quad \text{a.e. } t \in \mathbb{R}_+. \tag{12}$$

Using this notation (10) can be written

$$y(t) = (\mathcal{G} * u)(t), \quad t \geq 0$$

where \mathcal{G} is the continuous $p \times m$ -matrix function (called the convolution kernel of (2.17) with $D = 0$) given by

$$\mathcal{G}(t) = Ce^{At}B, \quad t \geq 0. \quad (13)$$

Again $\mathcal{G}(t)$ can be viewed as a *weighting pattern*. However we run into some theoretical difficulties in attempting to mirror the interpretation of $\mathcal{G}(t)$ as an impulse response. In order to see why this is the case, let us consider a *test signal* of the form $u(\cdot) = v(\cdot)e^j$ where $v(\cdot) \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K})$. At time t the i^{th} component of the corresponding output $y(t, v(\cdot)e^j)$ is

$$y_i(t, v(\cdot)e^j) = \int_0^t \mathcal{G}_{ij}(t-s)v(s)ds, \quad t \geq 0$$

and we would like this to equal $\mathcal{G}_{ij}(t)$. This means that $v(\cdot)$ must have the property

$$\mathcal{G}_{ij}(t) = \int_0^t \mathcal{G}_{ij}(t-s)v(s)ds, \quad t \geq 0. \quad (14)$$

However, there is no function $v(\cdot)$ satisfying (14) for a non-zero $\mathcal{G}_{ij}(t)$. In fact the commutative ring $L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K})$ does not have a unit element. This is an important difference between the discrete and continuous time cases. To get around this difficulty we show that it is possible to use piecewise continuous input signals which result in outputs *approximating* the components $\mathcal{G}_{ij}(t)$ of $\mathcal{G}(t)$.

Lemma 2.3.4. *Let $(u_k(\cdot))_{k \in \mathbb{N}}$ be a sequence of non-negative integrable functions on \mathbb{R}_+ such that*

$$\int_0^\infty u_k(s)ds = 1 \quad \text{and} \quad u_k|[\alpha_k, \infty) \equiv 0, \quad k \in \mathbb{N} \quad (15)$$

where $\alpha_k \searrow 0$ as $k \rightarrow \infty$. Then for every $f \in \mathcal{C}(\mathbb{R}_+; \mathbb{R})$ we have

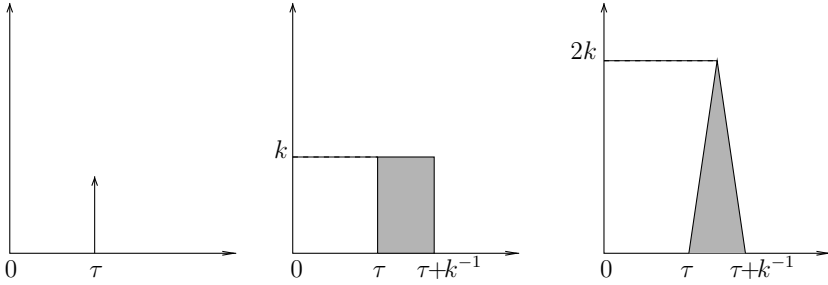
$$f(t) = \lim_{k \rightarrow \infty} (f * u_k)(t), \quad t \in \mathbb{R}_+ \quad (16)$$

uniformly on compact intervals in \mathbb{R}_+ .

Proof: The statement follows from the uniform continuity of f on compact intervals $I \subset \mathbb{R}_+$ since

$$|(f * u_k)(t) - f(t)| \leq \int_0^t |f(t-s) - f(t)|u_k(s)ds \leq \sup_{0 \leq s \leq \alpha_k} |f(t-s) - f(t)|, \quad t \in I.$$

□

Figure 2.3.1: The Dirac impulse δ_τ and approximations

Now suppose that a sequence $(u_k(\cdot))_{k \in \mathbb{N}}$ of input signals is chosen to satisfy (15) where $\alpha_k \searrow 0$ as $k \rightarrow \infty$. Because of (16) such sequences are called *approximate identities*. If the input is $u_k(\cdot)e^j$, the i^{th} component of the corresponding output will approximate $\mathcal{G}_{ij}(\cdot)$ uniformly on compact intervals.

$$\mathcal{G}_{ij}(t) = \lim_{k \rightarrow \infty} y_i(t; u_k(\cdot)e^j) = \lim_{k \rightarrow \infty} (\mathcal{G}_{ij} * u_k)(t), \text{ uniformly on compact intervals.} \quad (17)$$

Typical candidates for the test functions $u_k(\cdot)$, $k \in \mathbb{N}$ are shown on the right in Figure 2.3.1. They can be viewed as approximations of the so-called *Dirac impulse* δ_0 which is not a function but the unit point measure at 0. Because of (17) the $p \times m$ -matrix function (13) is called the *impulse response* of the differentiable system (2.17) (with $D = 0$).

Remark 2.3.5. For every $\tau \in \mathbb{R}_+$, the *Dirac impulse* δ_τ (on \mathbb{R}_+) at time τ is the unit point measure at τ , defined by

$$\int f(s) \delta_\tau(ds) = f(\tau), \quad f \in C(\mathbb{R}_+; \mathbb{K}). \quad (18)$$

Although δ_τ has no density with respect to the Lebesgue measure it is usual in the control literature to write

$$\int f(s) \delta_\tau(s) ds \text{ or } \int f(s) \delta(\tau - s) ds \text{ or } \int f(\tau - s) \delta(s) ds \text{ instead of } \int f(s) \delta_\tau(ds).$$

Note, however, that this is only a suggestive notation and does not mean that δ_τ is a function. The general definition of convolution between a measure and a function (see Remark A.3.16) implies that for all $f \in C(\mathbb{R}_+; \mathbb{K})$, $\tau \in \mathbb{R}_+$

$$(\delta_\tau * f)(t) = (f * \delta_\tau)(t) = \begin{cases} f(t - \tau), & t \in [\tau, \infty) \\ 0, & t \in [0, \tau) \end{cases}.$$

Hence δ_τ acts as a forward shift on f via the convolution. In particular, $\delta_0 * f = f * \delta_0 = f$ for $f \in C(\mathbb{R}_+, \mathbb{K})$. This, together with Lemma 2.3.4 explains why we may regard δ_0 as the limit of the above sequences $(u_k(\cdot))$. Graphically the Dirac impulse δ_τ is represented by a vertical arrow of length 1 at τ (see Figure 2.3.1). \square

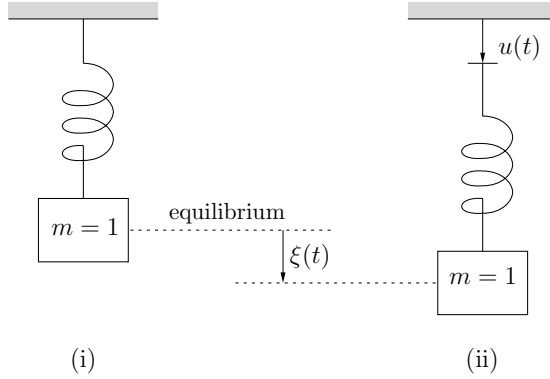


Figure 2.3.2: (i) Free linear oscillator at rest (ii) Forced linear oscillator in motion

Example 2.3.6. (Forced linear oscillator). Consider the mass-spring system described in Example 2.2.9 but suppose now that the support can be moved vertically. Let $u(t)$ denote the displacement of the support from some nominal reference point (see Figure 2.3.2). The displacement $\xi(t)$ of the mass from its equilibrium position (under the control $u(\cdot) = 0$) is taken as the output. Now since the restoring force is $\nu^2(\xi(t) - u(t))$ instead of $\nu^2\xi(t)$ we obtain the following equation of motion

$$\ddot{\xi}(t) + 2\alpha\dot{\xi}(t) + \nu^2\xi(t) = \nu^2u(t), \quad y(t) = \xi(t), \quad t \in \mathbb{R}_+.$$

The system matrices of the corresponding state space model are (see Example 2.2.9)

$$A = \begin{bmatrix} 0 & 1 \\ -\nu^2 & -2\alpha \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \nu^2 \end{bmatrix}, \quad C = [1, 0].$$

The (scalar) impulse response of this system is $Ce^{At}B$, $t \geq 0$, and hence equal to the first coordinate of the free motion starting at $x^0 = [0, \nu^2]^\top$. To compute it we have to represent x^0 as a linear combination of the eigenvectors $z^1 = [1, \lambda_1]^\top$, $z^2 = [1, \lambda_2]^\top$ where $\lambda_{1,2} = -\alpha \pm \sqrt{\alpha^2 - \nu^2}$. A short calculation yields

$$x^0 = \frac{\nu^2}{2\sqrt{\alpha^2 - \nu^2}}(z^1 - z^2).$$

Hence the impulse response is

$$\mathcal{G}(t) = \frac{\nu^2}{2\sqrt{\alpha^2 - \nu^2}}(e^{\lambda_1 t} - e^{\lambda_2 t}), \quad t \geq 0.$$

For $|\alpha| < \nu$ this response is oscillating with frequency $\omega = \sqrt{\nu^2 - \alpha^2}$

$$\mathcal{G}(t) = \nu^2 e^{-\alpha t} \left[\frac{e^{i\omega t} - e^{-i\omega t}}{2i\omega} \right] = (\nu^2/\omega) e^{-\alpha t} \sin \omega t.$$

□

The fact that the impulse response completely determines the input-output behaviour (resp. forced motions) of the system is, as in the discrete time case, an

immediate consequence of the explicit formula for the output (10). However, it is instructive to explain this fact directly by the basic properties of time-invariance and linearity (superposition principle). For the sake of simplicity, we suppose that $m = p = 1$, i.e. Σ is a *single input single output* (siso) system.

For discrete time systems Σ of the form (2.22) the situation is simple. The output value $y(t)$ only depends on the input values $u(s)$, $s \in [0, t] \cap \mathbb{N}$. On the finite time set $[0, t] \cap \mathbb{N}$, $u(\cdot)$ can be represented as linear combination of shifted unit impulses (8). Hence, by linearity and time-invariance of Σ , $y(t)$ is completely determined if the system responses to the unit impulses are known.

For differentiable systems Σ of the form (2.17) (with $D = 0$) an additional property is used. For any $t_1 > 0$ the restriction of the output $y(\cdot; u(\cdot))|_{[0, t_1]} \in \mathcal{C}([0, t_1]; \mathbb{K})$ depends continuously on the restriction of the input $u(\cdot)|_{[0, t_1]} \in L^1(0, t_1; \mathbb{K})$. Indeed, it follows immediately from (10) that, for arbitrary $u(\cdot)$, $v(\cdot) \in \mathcal{U}$ and any fixed $t_1 > 0$

$$\sup_{t \in [0, t_1]} |y(t; u(\cdot)) - y(t; v(\cdot))| \leq K \int_0^{t_1} |u(s) - v(s)| ds, \quad (19)$$

where $K = \sup_{0 \leq s \leq t_1} |Ce^{As}B|$.

Now consider for any *continuous* $u(\cdot) \in \mathcal{U}$ the step function approximation

$$v_k(t) = \sum_{j=0}^{k-1} u(jt_1/k)(1/k)S_{jt_1/k}w_k(t), \quad t \in [0, t_1] \quad (20)$$

where $S_{jt_1/k}$ is the forward shift by $\tau = jt_1/k$ and $w_k(\cdot)$ is the approximate Dirac impulse (see Figure 2.3.3)

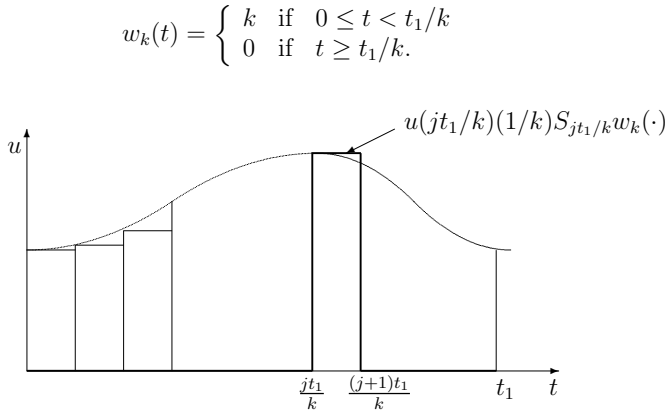


Figure 2.3.3: Step function approximation to an input

From the definition of the Riemann integral it is known that

$$\int_0^{t_1} |u(t) - v_k(t)| dt \rightarrow 0 \quad \text{as } k \rightarrow \infty$$

and hence by (19) $y(t; v_k(\cdot)) \rightarrow y(t; u(\cdot))$ as $k \rightarrow \infty$.

By linearity and time invariance of Σ the system's response to $v_k(\cdot)$ is the following superposition of shifted system responses to the approximate Dirac impulses $w_k(\cdot)$

$$y(t; v_k(\cdot)) = \sum_{j=0}^{k-1} u(jt_1/k)(1/k)S_{jt_1/k}y(t; w_k(\cdot)), \quad t \in [0, t_1].$$

Since by Lemma 2.3.4 $y(t; w_k(\cdot)) \rightarrow \mathcal{G}(t)$ uniformly on compact intervals as $k \rightarrow \infty$, we finally obtain, for any $t_1 > 0$,

$$\sup_{t \in [0, t_1]} |y(t; u(\cdot)) - \sum_{j=0}^{k-1} u(jt_1/k)(1/k)S_{jt_1/k}\mathcal{G}(t)| \rightarrow 0 \quad \text{as } k \rightarrow \infty. \quad (21)$$

Thus making use of only the three basic properties of linearity, time invariance and continuity, we see that the output signal $y(t; u(\cdot))$ corresponding to a continuous control function $u(\cdot) \in \mathcal{U}$ can be approximated by linear combinations of shifted impulse responses with coefficients determined by $u(\cdot)$.

Input-Output Operators in Time Domain

The input-output behaviour of the system (2.17)¹ with time domain $T = \mathbb{R}_+$, initialized at $x(0) = 0$, is described by the *input-output operator*

$$\begin{aligned} \mathbb{L}_+ : L_{\text{loc}}^1(\mathbb{R}_+; \mathbb{K}^m) &\rightarrow L_{\text{loc}}^1(\mathbb{R}_+; \mathbb{K}^p) \\ (\mathbb{L}_+ u)(t) &= Du(t) + \int_0^t Ce^{A(t-\tau)}Bu(\tau) d\tau = Du(t) + (\mathcal{G} * u)(t), \quad t \in \mathbb{R}_+. \end{aligned} \quad (22)$$

Its counterpart for the discrete time system (2.22) on the time domain $T = \mathbb{N}$ is

$$\begin{aligned} \mathbb{L}_+ : \ell_{\text{loc}}^1(\mathbb{N}; \mathbb{K}^m) &= (\mathbb{K}^m)^{\mathbb{N}} \rightarrow \ell_{\text{loc}}^1(\mathbb{N}; \mathbb{K}^p) = (\mathbb{K}^p)^{\mathbb{N}} \\ (\mathbb{L}_+ u)(t) &= Du(t) + \sum_{k=0}^{t-1} CA^{t-k-1}Bu(k) = (\mathcal{G} * u)(t), \quad t \in \mathbb{N}. \end{aligned} \quad (23)$$

Here the continuous and discrete time convolution kernels are given by

$$\mathcal{G}(t) = Ce^{At}B, \quad t \in \mathbb{R}_+ \quad \text{and} \quad \mathcal{G}(t) = CA^{t-1}B, \quad t \in \mathbb{N}^*, \quad \mathcal{G}(0) = D. \quad (24)$$

Remark 2.3.7. Note that in the discrete time case the feedthrough matrix is determined by the convolution kernel \mathcal{G} whereas we have seen that in the continuous time case it is not possible to express the direct input-output coupling via convolution with a (locally) integrable convolution kernel. However, if we allow for a Dirac impulse in the convolution kernel and set $\mathcal{G}(t) = \delta_0(t)D + Ce^{At}B$, $t \in \mathbb{R}_+$, then we may write the first equation in (40) as $y(t) = (\mathcal{G} * u)(t)$. More general convolution kernels involving *measures* or, more generally, *distributions* on \mathbb{R}_+ are considered in the literature, see *Notes and References*. □

¹Now we allow for an arbitrary direct input-output coupling $D \in \mathbb{K}^{p \times m}$.

If S_τ , $\tau \in \mathbb{R}_+$ are the forward shift operators for the time domain $T = \mathbb{R}_+$ (see (1.28)), then

$$(\mathbb{L}_+ S_\tau u)(t) = Du(t-\tau) + \int_\tau^t C e^{A(t-s)} Bu(s-\tau) ds = Du(t-\tau) + \int_0^{t-\tau} C e^{A(t-\tau-s)} Bu(s) ds$$

and $(\mathbb{L}_+ S_\tau u)(t) = 0$ for $t < \tau$. Hence the continuous time input-output operator \mathbb{L}_+ is *time-invariant* in the sense that it commutes with the forward shift operators S_τ , $\tau \in \mathbb{R}_+$

$$(\mathbb{L}_+ S_\tau u)(\cdot) = (S_\tau \mathbb{L}_+ u)(\cdot), \quad u(\cdot) \in L^1_{\text{loc}}(\mathbb{R}_+; \mathbb{K}^m). \quad (25)$$

Similarly it can be shown that the discrete time input-output operator \mathbb{L}_+ commutes with the discrete time forward shift operators S_τ , $\tau \in \mathbb{N}$.

Up until now we have discussed the input-output behaviour of the linear systems (2.17) (resp. (2.22)) without comparing the sizes of the input and output signals. For such a comparison, which is of great importance in applications, we need to introduce *norms* on the signal spaces. Suppose that \mathbb{K}^m and \mathbb{K}^p are provided with arbitrary norms $\|\cdot\|_{\mathbb{K}^m}$ and $\|\cdot\|_{\mathbb{K}^p}$ and $\mathbb{K}^{p \times m}$ with the corresponding operator norm, see Definition A.1.4². Then, for any fixed $q \in [1, \infty]$, the size of an input or output signals can be measured by their L^q -norm (resp. ℓ^q -norm), see Definitions A.3.10 and A.3.1). However, the system response $t \mapsto y(t; u(\cdot))$ is not necessarily in $L^q(\mathbb{R}_+; \mathbb{K}^p)$ (resp. $\ell^q(\mathbb{N}; \mathbb{K}^p)$) if $u(\cdot) \in L^q(\mathbb{R}_+; \mathbb{K}^m)$ (resp. $u(\cdot) \in \ell^q(\mathbb{N}; \mathbb{K}^m)$). Thus, in general, the input-output operator \mathbb{L}_+ may transform an input signal of finite L^q -norm (resp. ℓ^q -norm) into an output signal of infinite L^q -norm (resp. ℓ^q -norm). The system is called *input-output stable* or, more precisely, *L^q -stable* (resp. *ℓ^q -stable*) if this cannot happen. The following proposition gives a sufficient condition for L^q -stability (resp. ℓ^q -stability).³

Proposition 2.3.8. *Suppose*

$$\sigma(A) \subset \mathbb{C}_- \quad (\text{resp. } \sigma(A) \subset \mathbb{D}). \quad (26)$$

Then the continuous (resp. discrete) time input-output operator \mathbb{L}_+ given by (22) (resp. (23)) defines a bounded linear operator from $L^q(\mathbb{R}_+; \mathbb{K}^m)$ to $L^q(\mathbb{R}_+; \mathbb{K}^p)$ (resp. $\ell^q(\mathbb{N}; \mathbb{K}^m)$ to $\ell^q(\mathbb{N}; \mathbb{K}^p)$).

Proof: The linearity of \mathbb{L}_+ follows directly from the definition (see (22), (23)). By assumption there exists $\omega > 0$ such that $\text{Re } \lambda < -\omega$ (resp. $|\lambda| < e^{-\omega}$) for all $\lambda \in \sigma(A)$. We will see later (Lemma 3.3.19) that this implies

$$\|e^{At}\| \leq M_\omega e^{-\omega t}, \quad t \in \mathbb{R}_+, \quad \|A^t\| \leq M_\omega e^{-\omega t}, \quad t \in \mathbb{N} \quad (27)$$

for a suitable constant $M_\omega \geq 1$. As a consequence, we have $\mathcal{G} \in L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})$ (resp. $\mathcal{G} \in \ell^1(\mathbb{N}; \mathbb{K}^{p \times m})$). It then follows from (22) and the convolution inequality (A.3.24) that $y = \mathbb{L}_+ u \in L^q(\mathbb{R}_+; \mathbb{K}^p)$ for all $u \in L^q(\mathbb{R}_+; \mathbb{K}^m)$ and

$$\|\mathbb{L}_+ u\|_{L^q(\mathbb{R}_+; \mathbb{K}^p)} \leq (\|D\|_{\mathcal{L}(\mathbb{K}^m, \mathbb{K}^p)} + \|\mathcal{G}\|_{L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})}) \|u\|_{L^q(\mathbb{R}_+; \mathbb{K}^m)}, \quad u \in L^q(\mathbb{R}_+; \mathbb{K}^m).$$

²In the following we only make the norm explicit in the statements of theorems or propositions or where the particular norm used may be unclear

³We will see in the next chapter that condition (26) is equivalent to the asymptotic stability of the system (22) (resp. (23)), see Section 3.3.

Hence \mathbb{L}_+ is a bounded linear operator from $L^q(\mathbb{R}_+; \mathbb{K}^m)$ into $L^q(\mathbb{R}_+; \mathbb{K}^p)$ (see Section A.4). An analogous inequality holds in the discrete time case and this concludes the proof. \square

In applications the square of the L^2 -norm (resp. ℓ^2 -norm) of a signal can often be interpreted as a measure of its energy, see Section 1.4. The previous proposition (with $q = 2$) implies that, if (26) holds, the system (2.17) (resp. (2.22)) transforms finite energy input signals into finite energy output signals.

So far we have described the input-output behaviour on \mathbb{R}_+ (resp. \mathbb{N}) by setting the initial state to be zero at time $t_0 = 0$. We will now show that under assumption (26) we can do without fixing an initial condition and study the input-output behaviour of the systems (2.17) (resp. (2.22)) on the extended time domain $T = \mathbb{R}$ (resp. \mathbb{Z}). First we note that for any $t_0 \in \mathbb{R}$ the input-output behaviour of the continuous time system (2.17) with fixed initial state $x(t_0) = x^0$ at time $t_0 \in \mathbb{R}$ and control functions $u(\cdot) \in L^q(t_0, \infty; \mathbb{K}^m)$ is given by

$$\begin{aligned} y(t) &= y(t; t_0, x^0, u(\cdot)) = Du(t) + Ce^{A(t-t_0)}x^0 + \int_{t_0}^t Ce^{A(t-s)}Bu(s)ds \\ &= y(t - t_0; 0, x^0, S_{t_0}u(\cdot)) = Ce^{A(t-t_0)}x^0 + \mathbb{L}_+(S_{t_0}u(\cdot)|\mathbb{R}_+)(t - t_0), \quad t \geq t_0. \end{aligned} \quad (28)$$

It follows from (28) and the convolution inequality (A.3.24) that $y(\cdot; t_0, x^0, u(\cdot)) \in L^2(t_0, \infty; \mathbb{K}^p)$ for every $u(\cdot) \in L^2(t_0, \infty; \mathbb{K}^m)$. Moreover if $D = 0$ the output function $y(t)$ tends to zero as $t \rightarrow \infty$. To prove this last result we need the following lemma.

Lemma 2.3.9. *Suppose $t_0 \in \mathbb{R}$ and $y(\cdot) : [t_0, \infty) \rightarrow \mathbb{C}^p$ is absolutely continuous such that $y(\cdot), \dot{y}(\cdot) \in L^2(t_0, \infty; \mathbb{C}^p)$, then $y(t) \rightarrow 0$ as $t \rightarrow \infty$.*

Proof: For the usual inner product on \mathbb{C}^p and $t \geq t_1 \geq t_0$, we have

$$\int_{t_1}^t [\langle \dot{y}(s), y(s) \rangle + \langle y(s), \dot{y}(s) \rangle] ds = \|y(t)\|_2^2 - \|y(t_1)\|_2^2.$$

Now since all norms on \mathbb{C}^p are equivalent $y(\cdot), \dot{y}(\cdot) \in L^2(t_0, \infty; \mathbb{C}^p)$ where \mathbb{C}^p is normed with the 2-norm. Hence $\langle \dot{y}(\cdot), y(\cdot) \rangle \in L^1(t_0, \infty; \mathbb{C})$ and so, for any given $\varepsilon > 0$, there exists $t_\varepsilon \geq t_0$, such that for all $t_1 \geq t_\varepsilon$

$$\left| \|y(t)\|_2^2 - \|y(t_1)\|_2^2 \right| = \left| \int_{t_1}^t [\langle \dot{y}(s), y(s) \rangle + \langle y(s), \dot{y}(s) \rangle] ds \right| < \varepsilon, \quad t \geq t_1. \quad (29)$$

On the other hand since $y(\cdot) \in L^2(t_0, \infty; \mathbb{C}^p)$, there exists $t_1^\varepsilon \geq t_\varepsilon$ such that $\|y(t_1^\varepsilon)\|_2 < \varepsilon$, hence choosing $t_1 = t_1^\varepsilon$ in (29) we get $\|y(t)\|_2^2 < 2\varepsilon$ for $t \geq t_1^\varepsilon$. \square

Proposition 2.3.10. *Suppose that $D = 0$ and $\sigma(A) \subset \mathbb{C}_-$. Then there exists a constant K such that for all $t_0 \in \mathbb{R}$, $x^0 \in \mathbb{K}^n$, $u(\cdot) \in L^2(t_0, \infty; \mathbb{K}^m)$*

$$\|y(t; t_0, x^0, u(\cdot))\|_{\mathbb{K}^p} \leq K [\|x^0\|_{\mathbb{K}^n} + \|u(\cdot)\|_{L^2(t_0, \infty; \mathbb{K}^m)}], \quad t \geq t_0, \quad (30)$$

where $y(\cdot) = y(\cdot; t_0, x^0, u(\cdot)) : [t_0, \infty) \rightarrow \mathbb{K}^p$ is the associated output function of the system (2.17) defined by (28). Moreover $y(\cdot), \dot{y}(\cdot) \in L^2(t_0, \infty; \mathbb{K}^p)$ and $y(t) \rightarrow 0$ as $t \rightarrow \infty$.

Proof: Since $\|e^{At}\| \leq Me^{-\omega t}$ for some $M \geq 1$, $\omega > 0$ we obtain from (28)

$$\begin{aligned} \|y(t; t_0, x^0, u(\cdot))\| &\leq \|C\|Me^{-\omega(t-t_0)}\|x^0\| + \|C\|\|B\|M \int_{t_0}^t e^{-\omega(t-s)}\|u(s)\| ds \\ &\leq \|C\|M\|x^0\| + (M\|C\|\|B\|/\sqrt{2\omega})\|u(\cdot)\|_{L^2(t_0, \infty; \mathbb{K}^m)}, \quad t \geq t_0 \end{aligned}$$

by the Cauchy-Schwarz inequality (A.3.20). This implies the inequality (30) for a suitably large constant $K > 0$. Now it follows from (28), (27) and the convolution inequality (A.3.24) that $y(\cdot) \in L^2(t_0, \infty; \mathbb{K}^p)$. Applying the same argument with $C = I_n$ we obtain $x(\cdot) \in L^2(t_0, \infty; \mathbb{K}^n)$. Moreover $y(\cdot) = Cx(\cdot)$ is absolutely continuous and since $\dot{y}(t) = C\dot{x}(t) = CAx(t) + CBu(t)$ for $t \geq 0$, we get $\dot{y}(\cdot) \in L^2(t_0, \infty; \mathbb{K}^p)$. Thus $y(t) \rightarrow 0$ as $t \rightarrow \infty$ by Lemma 2.3.9. \square

Remark 2.3.11. Applying the above proposition with $C = I_n$ we see that if $\sigma(A) \subset \mathbb{C}_-$, then $\varphi(t; t_0, x^0, u(\cdot)) \rightarrow 0$ as $t \rightarrow \infty$, for all initial states $x^0 \in \mathbb{K}^n$ and input functions $u(\cdot) \in L^2(t_0, \infty; \mathbb{K}^m)$. \square

Remark 2.3.12. The condition $D = 0$ is not needed in the discrete time case. The reason is that $\|u(t)\| \leq \|u(\cdot)\|_{\ell^2(t_0, \infty; \mathbb{K}^m)}$ for all $t \geq t_0$. We leave it to the reader (see Ex. 4) to prove the following discrete time counterpart of Proposition 2.3.10.

Suppose $\sigma(A) \subset \mathbb{D}$. Then there exists a constant K such that for all $t_0 \in \mathbb{Z}$, $x^0 \in \mathbb{K}^n$

$$\|y(t; t_0, x^0, u(\cdot))\| \leq K [\|x^0\| + \|u(\cdot)\|_{\ell^2(t_0, \infty; \mathbb{K}^m)}], \quad u(\cdot) \in \ell^2(t_0, \infty; \mathbb{K}^m), \quad t \in \mathbb{Z}, \quad t \geq t_0 \quad (31)$$

where $y(\cdot) = y(\cdot; t_0, x^0, u(\cdot))$ is the corresponding output function of the discrete time system (2.22) given by

$$y(t) = y(t; t_0, x^0, u(\cdot)) = CA^{t-t_0}x^0 + Du(t) + \sum_{k=t_0}^{t-1} CA^{t-k-1}Bu(k) \quad (32)$$

Moreover $y(\cdot) \in \ell^2(t_0, \infty; \mathbb{K}^p)$ and $y(t) \rightarrow 0$ as $t \rightarrow \infty$. \square

Now let $t_0 \rightarrow -\infty$ in (28). Because of (27), we see that as t_0 goes back, the influence of the initial state on the output $y(t)$ gets less and less. This leads us to define the *input-output operator* of (2.17) with time domain $T = \mathbb{R}$ by

$$\begin{aligned} \mathbb{L} &: L^q(\mathbb{R}; \mathbb{K}^m) \rightarrow L^q(\mathbb{R}; \mathbb{K}^p) \\ (\mathbb{L}u)(t) &= Du(t) + \int_{-\infty}^t Ce^{A(t-s)}Bu(s) ds = Du(t) + (\mathcal{G} * u)(t) \quad t \in \mathbb{R}, \quad (33) \end{aligned}$$

where $\mathcal{G} * u$ is to be understood as a convolution of two functions defined on \mathbb{R} , see (A.3.23). In the discrete time case we define the *input-output operator* of (2.22) with the time domain $T = \mathbb{Z}$ by

$$\begin{aligned} \mathbb{L} &: \ell^q(\mathbb{Z}; \mathbb{K}^m) \rightarrow \ell^q(\mathbb{Z}; \mathbb{K}^p) \\ (\mathbb{L}u)(t) &= Du(t) + \sum_{s=-\infty}^{t-1} CA^{t-s-1}Bu(s) = (\mathcal{G} * u)(t), \quad t \in \mathbb{Z} \quad (34) \end{aligned}$$

where the convolution is as in (A.3.6). In both cases the convolution kernel \mathcal{G} defined by (24) is trivially extended to \mathbb{R} (resp. \mathbb{Z}) by setting $\mathcal{G}(t) = 0$ for $t < 0$. The assumption (26) then implies

$$\mathcal{G} \in L^1(\mathbb{R}; \mathbb{K}^{p \times m}), \quad (\text{resp. } \mathcal{G} \in \ell^1(\mathbb{Z}; \mathbb{K}^{p \times m})). \quad (35)$$

As a consequence of Propositions A.3.14 and A.3.3 the input-output operator \mathbb{L} is well defined, linear and bounded in both the continuous and the discrete time case, see Corollary 2.3.16.

Remark 2.3.13. Suppose $t_0 \in \mathbb{R}$ and $v(\cdot) \in L^q(-\infty, t_0; \mathbb{K}^m)$, then

$$x^0 = \int_{-\infty}^{t_0} e^{A(t_0-s)} B v(s) ds \quad (36)$$

is well defined because of (27). In fact since $\|e^{At}\| \leq M e^{-\omega t}$ for some $M \geq 1$, $\omega > 0$ the function $t \mapsto \|e^{At}\|$ is L^{q^*} -integrable on \mathbb{R}_+ , where $q^* \in [1, \infty]$ is the conjugate exponent of q , and therefore $\|e^{A(t_0-s)}\| \|Bu(s)\|$ is integrable on $(-\infty, t_0]$ by the Hölder inequality (A.3.21) with $r = 1$. Now let $u(\cdot) \in L^q(t_0, \infty; \mathbb{K}^m)$ be arbitrary and denote by $u_v(\cdot) \in L^q(\mathbb{R}; \mathbb{K}^m)$ the extension of $u(\cdot)$ to \mathbb{R} by $u_v(t) = v(t)$ for $t < t_0$. Then for $t \geq t_0$,

$$\int_{-\infty}^t C e^{A(t-s)} B u_v(s) ds = C e^{A(t-t_0)} x_0 + \int_{t_0}^t C e^{A(t-s)} B u(s) ds, \quad u(\cdot) \in L^q(t_0, \infty; \mathbb{K}^m), \quad (37)$$

where x^0 is given by (36). Hence we recover from the input-output operator \mathbb{L} on the time domain \mathbb{R} the expression (28) for the input-output behaviour of the system (2.17) at the initial state $x(t_0) = x^0 \in \mathbb{K}^n$. An analogous result holds in the discrete time case.

In the first section we described how the internal state $x(t)$ at any time t incorporates the total effect of all past controls. This is again illustrated by (37) in combination with (36). These formulas show that, in order to predict the future output, once the state at time $x(t_0) = x^0$ (36) is known, one may forget about the previous control values $u(t)$, $t < t_0$. \square

We have seen above that under the assumption (26) the input-output behaviour of a state space system with the time domain \mathbb{R} (resp. \mathbb{Z}) can be described by a suitable convolution kernel. If one is only interested in the input-output behaviour of the system, it suffices to know the convolution kernel and one may forget about the state. Discarding the internal dynamics the state space system is reduced to a *black box model* or *input-output system*. An input-output system is basically just a map which associates with any input signal the corresponding output signal. In the remainder of this subsection we will make this concept more precise and consider especially those input-output systems whose behaviour is described by convolution kernels.

Definition 2.3.14. Let $T \subset \mathbb{R}$, $U, \mathcal{U} \subset U^T$, $Y, \mathcal{Y} \subset Y^T$ be non-empty sets and $\mathbb{G} : \mathcal{U} \rightarrow \mathcal{Y}$ a *causal* map, i.e. \mathbb{G} satisfies

$$\forall t \in T \cap (-\infty, t_1] : u(t) = v(t) \quad \Rightarrow \quad (\mathbb{G}u)(t_1) = (\mathbb{G}v)(t_1) \quad (38)$$

for all $t_1 \in T$, $u(\cdot), v(\cdot) \in \mathcal{U}$. Then the sextuple $(T, U, \mathcal{U}, \mathbb{G}, Y, \mathcal{Y})$ is said to be an *input-output system* with time domain T , set of input values U , set of output values Y , set of input signals \mathcal{U} , set of output signals \mathcal{Y} and input-output operator \mathbb{G} .

Input-output systems are represented by blockdiagrams as in Figure 2.3.4. If there is no risk of confusion, an input-output system is denoted by $(\mathcal{U}, \mathbb{G}, \mathcal{Y})$.

Every complete state space system Σ (see Definition 2.1.3) together with an initial

Figure 2.3.4: Input-output system with input-output operator \mathbb{G}

condition $x(t_0) = x^0$ where $(t_0, x^0) \in T \times X$ is fixed, defines an input-output system with time domain T_{t_0} and input-output operator (1.8).

An input-output system is said to be *linear* if $U, \mathcal{U}, Y, \mathcal{Y}$ are vector spaces over some field \mathbb{K} and $\mathbb{G} : \mathcal{U} \rightarrow \mathcal{Y}$ is \mathbb{K} -linear. If Σ is a linear state space system with zero initial state $x(t_0) = 0$ then the corresponding input-output system with time domain T_{t_0} is linear.

An input-output system is said to be *time-invariant* if T contains 0 and is closed under addition, \mathcal{U}, \mathcal{Y} are forward shift-invariant and \mathbb{G} commutes with the forward shift operators S_τ , $\tau \in T$, $\tau \geq 0$, compare Definition 2.1.24.

A wide class of time-invariant linear input-output systems can be described by convolution kernels. In particular, most input-output systems which are given by *linear* ordinary, partial or delay differential equations with time-invariant parameters are *convolution systems*. An input-output system $(T, U, \mathcal{U}_+, \mathbb{G}_+, Y, \mathcal{Y}_+)$ with time domain $T = \mathbb{R}_+$ (resp. $T = \mathbb{N}$), input space $U = \mathbb{K}^m$, output space $Y = \mathbb{K}^p$ is called a convolution system if $\mathcal{U}_+, \mathcal{Y}_+$ are linear subspaces of $L_{\text{loc}}^q(\mathbb{R}_+, \mathbb{K}^m)$, $L_{\text{loc}}^q(\mathbb{R}_+, \mathbb{K}^p)$ for some $1 \leq q \leq \infty$ (resp. linear subspaces of $(\mathbb{K}^m)^\mathbb{N}$, $(\mathbb{K}^p)^\mathbb{N}$) and if the input-output operator $\mathbb{G}_+ : \mathcal{U}_+ \rightarrow \mathcal{Y}_+$ is of the form

$$\begin{aligned} (\mathbb{G}_+ u)(t) &= Du(t) + \int_0^t \mathcal{G}(t-s)u(s) ds = Du(t) + (\mathcal{G} * u)(t), \quad t \in \mathbb{R}_+, u(\cdot) \in \mathcal{U}_+ \\ (\mathbb{G}_+ u)(t) &= \mathcal{G}(0)u(t) + \sum_{s=0}^{t-1} \mathcal{G}(t-s)u(s) = (\mathcal{G} * u)(t), \quad t \in \mathbb{N}, u(\cdot) \in \mathcal{U}_+. \end{aligned} \quad (39)$$

Here $D \in \mathbb{K}^{p \times m}$ is a given *feedthrough matrix* and $\mathcal{G} \in L_{\text{loc}}^1(\mathbb{R}_+; \mathbb{K}^{p \times m})$ (resp. $\mathcal{G} \in (\mathbb{K}^{p \times m})^\mathbb{N}$) is a given *convolution kernel*. Convolution systems with time domain $T = \mathbb{R}$ (resp. $T = \mathbb{Z}$) are defined in a similar way by equations of the form

$$\begin{aligned} (\mathbb{G}u)(t) &= Du(t) + \int_{-\infty}^t \mathcal{G}(t-s)u(s) ds = Du(t) + (\mathcal{G} * u)(t), \quad t \in \mathbb{R}, u(\cdot) \in \mathcal{U} \\ (\mathbb{G}u)(t) &= \mathcal{G}(0)u(t) + \sum_{s=-\infty}^{t-1} \mathcal{G}(t-s)u(s) = (\mathcal{G} * u)(t), \quad t \in \mathbb{Z}, u(\cdot) \in \mathcal{U} \end{aligned} \quad (40)$$

where \mathcal{U} is a linear subspace of $L^q(\mathbb{R}; \mathbb{K}^m)$ (resp. $\ell^q(\mathbb{Z}; \mathbb{K}^m)$). But for these equations to make sense one needs to assume that the convolution kernel is integrable (resp. summable), i.e. \mathcal{G} satisfies assumption (35) where $\mathcal{G}(t) = 0$ for all $t < 0$ (resp. all $t \in \mathbb{Z}$, $t < 0$). In the next proposition S_τ denotes the shift operator (1.28) for the corresponding time domains $T = \mathbb{R}_+, \mathbb{N}, \mathbb{R}, \mathbb{Z}$. Note that, for all $\tau \in \mathbb{R}$ (resp. $\tau \in \mathbb{Z}$), the shift S_τ is a norm preserving automorphism of $L^q(\mathbb{R}; \mathbb{K}^k)$ (resp. $\ell^q(\mathbb{Z}; \mathbb{K}^k)$) and for all $\tau \in \mathbb{R}_+$ (resp. $\tau \in \mathbb{N}$) the forward shift S_τ is a norm preserving automorphism of $L^q(\mathbb{R}_+; \mathbb{K}^k)$ (resp. $\ell^q(\mathbb{N}; \mathbb{K}^k)$).

Proposition 2.3.15. *Suppose that $\mathcal{G} \in L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})$ and $D \in \mathbb{K}^{p \times m}$ (resp. $\mathcal{G} \in \ell^1(\mathbb{N}; \mathbb{K}^{p \times m})$). Then, for arbitrary $1 \leq q \leq \infty$,*

(i) \mathbb{G}_+ defined by (39) yields a bounded linear operator from $\mathcal{U}_+ = L^q(\mathbb{R}_+; \mathbb{K}^m)$ to $\mathcal{Y}_+ = L^q(\mathbb{R}_+; \mathbb{K}^p)$ (resp. $\mathcal{U}_+ = \ell^q(\mathbb{N}; \mathbb{K}^m)$ to $\mathcal{Y}_+ = \ell^q(\mathbb{N}; \mathbb{K}^p)$) which is time-invariant, i.e. commutes with the shift operators S_τ , $\tau \in \mathbb{R}_+$ (resp. $\tau \in \mathbb{N}$).

(ii) \mathbb{G} defined by (40) yields a bounded linear operator from $\mathcal{U} = L^q(\mathbb{R}; \mathbb{K}^m)$ to $\mathcal{Y} = L^q(\mathbb{R}; \mathbb{K}^p)$ (resp. $\mathcal{U} = \ell^q(\mathbb{Z}; \mathbb{K}^m)$ to $\mathcal{Y} = \ell^q(\mathbb{Z}; \mathbb{K}^p)$) which is time-invariant, i.e. commutes with the shift operators S_τ , $\tau \in \mathbb{R}$ (resp. $\tau \in \mathbb{Z}$).

$(\mathcal{U}_+, \mathbb{G}_+, \mathcal{Y}_+)$ and $(\mathcal{U}, \mathbb{G}, \mathcal{Y})$ are time-invariant linear L^q -stable (resp. ℓ^q -stable) input-output systems. Moreover,

$$\begin{aligned} \|\mathbb{G}_+\|_{\mathcal{L}(L^q(\mathbb{R}_+; \mathbb{K}^m), L^q(\mathbb{R}_+; \mathbb{K}^p))} &= \|\mathbb{G}\|_{\mathcal{L}(L^q(\mathbb{R}; \mathbb{K}^m), L^q(\mathbb{R}; \mathbb{K}^p))} \leq \|D\|_{\mathcal{L}(\mathbb{K}^m, \mathbb{K}^p)} + \|\mathcal{G}\|_{L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})} \\ (\text{resp. } \|\mathbb{G}_+\|_{\mathcal{L}(\ell^q(\mathbb{N}; \mathbb{K}^m), \ell^q(\mathbb{N}; \mathbb{K}^p))} &= \|\mathbb{G}\|_{\mathcal{L}(\ell^q(\mathbb{Z}; \mathbb{K}^m), \ell^q(\mathbb{Z}; \mathbb{K}^p))} \leq \|\mathcal{G}\|_{\ell^1(\mathbb{N}; \mathbb{K}^{p \times m})}). \end{aligned}$$

Proof: We only prove the statements for the continuous time case, the proof for the discrete time case is similar.

It follows from the integrability of the kernel and Proposition A.3.14 that the output signals $y(\cdot) = (\mathcal{G} * u)(\cdot)$ are q -integrable for all $u(\cdot) \in \mathcal{U}_+$ and all $u(\cdot) \in \mathcal{U}$. By the same proposition it follows that $\mathbb{G}_+ : \mathcal{U}_+ \rightarrow \mathcal{Y}_+$ and $\mathbb{G} : \mathcal{U} \rightarrow \mathcal{Y}$ are bounded linear operators satisfying the above inequality. The time-invariance of \mathbb{G}_+ , \mathbb{G} is shown in exactly the same way as the time-invariance of \mathbb{L}_+ , see (25). This proves (i) and (ii) and the statement thereafter (which is equivalent to (i) and (ii)).

It only remains to prove that \mathbb{G} and \mathbb{G}_+ have the same operator norm. Since the Banach spaces $L^q(\mathbb{R}_+; \mathbb{K}^k)$ can be embedded isometrically into the Banach spaces $L^q(\mathbb{R}; \mathbb{K}^k)$, $k = m, p$ (by trivial extension), we have $\|\mathbb{G}_+\| \leq \|\mathbb{G}\|$. To prove the converse inequality, suppose $u(\cdot) \in L^q(\mathbb{R}_+; \mathbb{K}^m)$, let $\tau \leq 0$ and define the shifted signal $u_\tau(\cdot) \in L^q(\mathbb{R}; \mathbb{K}^m)$ by $u_\tau(t) = u(t - \tau)$, $t \geq \tau$, $u_\tau(t) = 0$, $t < \tau$. In particular, $u_0(\cdot)$ is the trivial extension of $u(\cdot)$ to \mathbb{R} and $u_\tau(\cdot) = (S_\tau u_0)(\cdot)$. The subspace $\mathcal{U}_\infty \subset L^q(\mathbb{R}; \mathbb{K}^m)$ of all the shifted u_τ where $u(\cdot) \in L^q(\mathbb{R}_+; \mathbb{K}^m)$ and $\tau < 0$, is dense in $L^q(\mathbb{R}; \mathbb{K}^m)$. Since \mathbb{G} commutes with the shift operator S_τ and \mathbb{G}_+ has the same operator norm as the restriction of \mathbb{G} to $L^q(\mathbb{R}_+; \mathbb{K}^m) \subset L^q(\mathbb{R}; \mathbb{K}^m)$, the restriction of \mathbb{G} to the normed subspace \mathcal{U}_∞ has the same operator norm as \mathbb{G}_+ . Hence it follows from the continuity of \mathbb{G} and the density of \mathcal{U}_∞ in $L^q(\mathbb{R}; \mathbb{K}^m)$ that \mathbb{G} and \mathbb{G}_+ have the same norm. \square

We have seen in the proof of Proposition 2.3.8 that the weighting pattern (13) and (7) of the state space systems (2.17) and (2.22) initialized at $x(0) = 0$ satisfies condition (35) if $\sigma(A) \subset \mathbb{C}_-$ and $\sigma(A) \subset \mathbb{D}$. Applying the previous proposition to these weighting patterns we obtain

Corollary 2.3.16. *Suppose $\sigma(A) \subset \mathbb{C}_-$ (resp. $\sigma(A) \subset \mathbb{D}$) then the input-output operator \mathbb{L} of the state space system (2.17) (resp. (2.22)) defined by (33) (resp. (34)) is a time-invariant bounded linear operator from $L^q(\mathbb{R}; \mathbb{K}^m)$ to $L^q(\mathbb{R}; \mathbb{K}^p)$ (resp. from $\ell^q(\mathbb{Z}; \mathbb{K}^m)$ to $\ell^q(\mathbb{Z}; \mathbb{K}^p)$) and satisfies*

$$\begin{aligned} \|\mathbb{L}_+\|_{\mathcal{L}(L^q(\mathbb{R}_+; \mathbb{K}^m), L^q(\mathbb{R}_+; \mathbb{K}^p))} &= \|\mathbb{L}\|_{\mathcal{L}(L^q(\mathbb{R}; \mathbb{K}^m), L^q(\mathbb{R}; \mathbb{K}^p))} \leq \|D\|_{\mathcal{L}(\mathbb{K}^m, \mathbb{K}^p)} + \|\mathcal{G}\|_{L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})} \\ (\text{resp. } \|\mathbb{L}_+\|_{\mathcal{L}(\ell^q(\mathbb{N}; \mathbb{K}^m), \ell^q(\mathbb{N}; \mathbb{K}^p))} &= \|\mathbb{L}\|_{\mathcal{L}(\ell^q(\mathbb{Z}; \mathbb{K}^m), \ell^q(\mathbb{Z}; \mathbb{K}^p))} \leq \|\mathcal{G}\|_{\ell^1(\mathbb{N}; \mathbb{K}^{p \times m})}). \end{aligned}$$

2.3.2 Transfer Functions

In the previous subsection we studied the input-output behaviour of linear state space systems in the *time domain*, i.e. state trajectories and input and output signals were considered as functions of time, and the system's input-output behaviour was modelled as a mapping between spaces of these time functions. We have seen that these mappings can be described by convolution kernels and this suggests that transform techniques may be a useful tool for their analysis (see Section A.3). In fact, Fourier and Laplace transforms have been used to describe the input-output behaviour of electrical circuits since the early decades of the past century. Via these transforms, convolution is converted into multiplication, and so the convolution operator of a system is transformed into a multiplication operator, determined by the Laplace transform of the convolution kernel, the so-called *transfer function* (or *transfer matrix*). A variety of graphical design methods has been developed in terms of these transfer functions, see *Notes and References*.

Transform techniques are based on the idea of representing continuous time signals as superpositions of harmonic oscillations. The variables of Fourier and Laplace transforms are interpreted as frequencies and therefore the analysis of the input-output behaviour of linear systems using these methods is called *frequency domain analysis*. In this subsection we give a brief introduction to some basic notions of this field. For a summary on transforms, see Section A.3.

Throughout the subsection it is assumed that all finite dimensional vector spaces \mathbb{K}^m , \mathbb{K}^p are equipped with Euclidean norms and $\mathbb{K}^{p \times m}$ with the corresponding operator norm (spectral norm).

Signal Transforms

We first define the Fourier transform for signals defined on $T = \mathbb{R}$ and the Laplace transform for signals defined on $T = \mathbb{R}_+$. Then we define the discrete Fourier transform of discrete time signals on $T = \mathbb{Z}$ and the \mathbf{z} -transform of signals defined on $T = \mathbb{N}$.

The *Fourier transform* of a function $u(\cdot) \in L^1(\mathbb{R}; \mathbb{K}^m)$ is defined by

$$\tilde{u}(\omega) = (\mathcal{F}u)(\omega) := \int_{-\infty}^{\infty} u(t)e^{-i\omega t} dt, \quad \omega \in \mathbb{R}.$$

For every $u(\cdot) \in L^1(\mathbb{R}; \mathbb{K}^m)$ the Fourier transform $\tilde{u}(\omega)$ is continuous in $\omega \in \mathbb{R}$ and tends to 0 as $|\omega| \rightarrow \pm\infty$ by Riemann's Lemma, see Proposition A.3.28. Note that if u takes its values in \mathbb{R}^m , then $\overline{\tilde{u}(\omega)} = \tilde{u}(-\omega)$.

We also need to consider the Fourier transform of signals $u(\cdot) \in L^2(\mathbb{R}; \mathbb{K}^m)$. There is an initial difficulty to be overcome since $L^2(\mathbb{R}; \mathbb{K}^m) \not\subset L^1(\mathbb{R}; \mathbb{K}^m)$. However (see Plancherel's Theorem A.3.33) the Fourier transforms $\tilde{u}_N(\cdot)$ of the truncated functions $u_N(\cdot) = u(\cdot)1_{[-N, N]} \in L^1(\mathbb{R}; \mathbb{K}^m)$ converge in $L^2(\mathbb{R}; \mathbb{C}^m)$ to a limit $\tilde{u}(\cdot)$ called the *Fourier-Plancherel transform* of $u(\cdot)$,

$$\tilde{u}(\cdot) = \lim_{N \rightarrow \infty} \tilde{u}_N(\cdot) \text{ in } L^2(\mathbb{R}; \mathbb{C}^m), \quad \tilde{u}_N(\omega) = \int_{-N}^N u(t)e^{-i\omega t} dt, \quad \omega \in \mathbb{R}. \quad (41)$$

Note that for $u(\cdot) \in L^1(\mathbb{R}; \mathbb{K}^m)$, the Fourier transform $\tilde{u}(\omega)$ is defined pointwise for every $\omega \in \mathbb{R}$, whereas for $u(\cdot) \in L^2(\mathbb{R}; \mathbb{K}^m)$ the Fourier-Plancherel transform $\tilde{u}(\cdot)$ is

only determined as an element of $L^2(\mathbb{R}; \mathbb{C}^m)$, i.e. almost everywhere.

For $u(\cdot) \in L^1_{loc}(\mathbb{R}_+; \mathbb{K}^m)$, $\alpha \in \mathbb{R}$ we set $u_\alpha(\cdot) : t \rightarrow e^{-\alpha t}u(t)$, $t \in \mathbb{R}_+$ and define

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

Functions belonging to $\mathcal{E}_\alpha(\mathbb{K}^m)$ for some $\alpha \in \mathbb{R}$ are called *Laplace transformable* and all signals occurring in control theory belong to this class. The *Laplace transform* (see Definition A.3.17) of $u(\cdot) \in \mathcal{E}_\alpha(\mathbb{K}^m)$ is defined by

$$\hat{u}(s) = (\mathcal{L}u)(s) := \int_0^\infty u(t)e^{-st}dt, \quad \operatorname{Re} s \geq \alpha. \quad (42)$$

$\hat{u}(\cdot)$ is continuous and bounded on the closed set $\{s \in \mathbb{C}; \operatorname{Re} s \geq \alpha\}$. It is analytic on $\{s \in \mathbb{C}; \operatorname{Re} s > \alpha\}$ and will be identified with its analytic extensions to complex domains containing this set. Note that if u takes its values in \mathbb{R}^m , then $\overline{\hat{u}(s)} = \hat{u}(\bar{s})$. Now suppose $u(\cdot) \in \mathcal{E}_\alpha(\mathbb{K}^m)$. Then for every $\beta \geq \alpha$ the Laplace transform $\hat{u}(\beta + i\omega)$ on the vertical line $\{\beta + i\omega; \omega \in \mathbb{R}\}$ can be expressed by the Fourier transform of the integrable function $u_\beta(\cdot)$ where $u_\beta(t) = u(t)e^{-\beta t}$, $t \geq 0$ and $u_\beta(t) = 0$, $t < 0$

$$(\mathcal{L}u)(\beta + i\omega) = \int_0^\infty (u(t)e^{-\beta t})e^{-i\omega t}dt = (\mathcal{F}u_\beta)(\omega), \quad \omega \in \mathbb{R}. \quad (43)$$

Hence $\hat{u}(\beta + i\omega) \rightarrow 0$ as $|\omega| \rightarrow \infty$ by Riemann's Lemma.

The counterpart of the Fourier transform for discrete time signals on \mathbb{Z} (two-sided sequences) is the *discrete Fourier transform*. For an arbitrary summable two-sided sequence $u(\cdot) \in \ell^1(\mathbb{Z}; \mathbb{K}^m)$ it is given by

$$\tilde{u}(\theta) = (\mathcal{F}_D u)(\theta) := \sum_{t=-\infty}^\infty u(t)e^{-it\theta}, \quad \theta \in [-\pi, \pi]. \quad (44)$$

The series converges uniformly for $\theta \in [-\pi, \pi]$ and its limit $\tilde{u}(\cdot) : [-\pi, \pi] \rightarrow \mathbb{C}^m$ is continuous with $\tilde{u}(-\pi) = \tilde{u}(\pi)$. Note that if $u(\cdot)$ takes its values in \mathbb{R}^m , then $\overline{\tilde{u}(\theta)} = \tilde{u}(-\theta)$.

For signals $u(\cdot) \in \ell^2(\mathbb{Z}; \mathbb{K}^m)$, we have a similar difficulty to that of the continuous time case since $\ell^2(\mathbb{Z}; \mathbb{K}^m) \not\subset \ell^1(\mathbb{Z}; \mathbb{K}^m)$. However since the functions $\psi_t(\theta) := (2\pi)^{-1/2}e^{-it\theta}$, $t \in \mathbb{Z}$ form an orthonormal basis of the Hilbert space $L^2(-\pi, \pi; \mathbb{C}^m)$ (see Example A.4.6) the series in (44) converges in $L^2(-\pi, \pi; \mathbb{C}^m)$ to some function $\tilde{u}(\cdot) \in L^2(-\pi, \pi; \mathbb{C}^m)$, for every sequence $u(\cdot) \in \ell^2(\mathbb{Z}; \mathbb{K}^m)$. Note again that for $u(\cdot) \in \ell^1(\mathbb{Z}; \mathbb{K}^m)$, the discrete Fourier transform $\tilde{u}(\theta)$ is defined pointwise for every $\theta \in [-\pi, \pi]$, whereas for $u(\cdot) \in \ell^2(\mathbb{Z}; \mathbb{K}^m)$ the transform $\tilde{u}(\cdot)$ is only determined as an element of $L^2(-\pi, \pi; \mathbb{C}^m)$, i.e. almost everywhere.

The counterpart of the Laplace transform for discrete time signals on \mathbb{N} is the *z-transform* which associates with any one-sided sequence $u(\cdot) : \mathbb{N} \rightarrow \mathbb{K}^m$, the formal power series in z^{-1}

$$\hat{u}(z) = (\mathcal{Z}u)(z) = \sum_{t=0}^\infty u(t)z^{-t}, \quad (45)$$

see Definition A.3.5. If the sequence $(u(t))$ is exponentially bounded this formal power series defines a complex analytic function $\hat{u}(\cdot)$ on some neighbourhood of ∞ . For $u(\cdot) \in (\mathbb{K}^m)^\mathbb{N}$, $\gamma > 0$ we set $u_\gamma(t) = u(t)\gamma^{-t}$, $t \in \mathbb{N}$ and define

$$\mathcal{S}_\gamma(\mathbb{K}^m) = \{u(\cdot) \in (\mathbb{K}^m)^\mathbb{N}; u_\gamma(\cdot) \in \ell^1(\mathbb{N}; \mathbb{K}^m)\}, \quad \gamma > 0.$$

Then if $u(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^m)$ the series on the RHS of (45) is absolutely convergent for all $z \in \mathbb{C}$, $|z| \geq \gamma$ and defines a continuous function $\hat{u}(\cdot)$ on $\{z \in \mathbb{C}; |z| \geq \gamma\}$. This function is analytic on $\{z \in \mathbb{C}; |z| > \gamma\}$ and it will be identified with its analytic extensions to complex domains containing this set. We will use the same symbol $\hat{u}(z)$ to denote the *formal power series* (45) and the *associated complex analytic function*. Note that if $u(\cdot)$ takes its values in \mathbb{R}^m , then $\hat{u}(z) = \hat{u}(\bar{z})$.

Now suppose $u(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^m)$ for some $\gamma > 0$. Then the holomorphic function given by the \mathbf{z} -transform (45) may be viewed as the frequency domain representation of the discrete time signal $u(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^m)$. In fact, on any circle $\{re^{i\theta}; \theta \in [-\pi, \pi]\}$ with $r \geq \gamma$ the function $\hat{u}(z)$ defined by (45) can be expressed as the discrete Fourier transform of the summable sequence $u_r(\cdot)$ where $u_r(t) = u(t)r^{-t}$, $t \in \mathbb{N}$ and $u_r(t) = 0$, $t \in \mathbb{Z} \setminus \mathbb{N}$,

$$(\mathcal{Z}u)(re^{i\theta}) = \sum_{t=0}^{\infty} (u(t)r^{-t})e^{-it\theta} = (\mathcal{F}_D u_r)(\theta), \quad \theta \in [-\pi, \pi]. \quad (46)$$

Transfer Matrices

We now turn from the representation of signals to the representation of input-output behaviours in frequency domain, and begin our discussion for convolution systems with time domain $T = \mathbb{R}_+$ (resp. \mathbb{N}) and input-output operator described by (39). Suppose that for some $\alpha \in \mathbb{R}$ (resp. $\gamma > 0$), $\mathcal{G}(\cdot) \in \mathcal{E}_\alpha(\mathbb{K}^{p \times m})$ and $u(\cdot) \in \mathcal{E}_\alpha(\mathbb{K}^m)$ (resp. $\mathcal{G}(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^{p \times m})$ and $u(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^m)$). Then it follows from (A.3.14) and (A.3.32) that the corresponding output signal $y(\cdot) = (\mathcal{G} * u)(\cdot)$ is in $\mathcal{E}_\alpha(\mathbb{K}^p)$ (resp. $\mathcal{S}_\gamma(\mathbb{K}^p)$). Taking the Laplace transform (resp. \mathbf{z} -transform) of both sides of equation (39) we obtain (see Theorem A.3.21 (resp. Theorem A.3.7))⁴

$$\hat{y}(s) = G(s)\hat{u}(s), \quad \operatorname{Re} s \geq \alpha, \quad \hat{y}(z) = G(z)\hat{u}(z), \quad |z| \geq \gamma \quad (47)$$

where $G(s)$ (resp. $G(z)$) is defined as follows.

Definition 2.3.17. Suppose that $D \in \mathbb{K}^{p \times m}$ and for some $\alpha \in \mathbb{R}$ (resp. $\gamma > 0$), $\mathcal{G}(\cdot) \in \mathcal{E}_\alpha(\mathbb{K}^{p \times m})$ (resp. $\mathcal{G}(\cdot) \in \mathcal{S}_\gamma(\mathbb{K}^{p \times m})$). Then the Laplace transform of $D\delta_0(t) + \mathcal{G}(\cdot)$ (resp. \mathbf{z} -transform of $\mathcal{G}(\cdot)$)

$$G(s) = D + \int_0^\infty \mathcal{G}(t)e^{-st}dt, \quad \operatorname{Re} s \geq \alpha, \quad G(z) = \mathcal{G}(0) + \sum_{t=1}^{\infty} \mathcal{G}(t)z^{-t}, \quad |z| \geq \gamma, \quad (48)$$

is called the *transfer matrix* of the input-output system described by (39) or (40).

The convolution kernel \mathcal{G} is uniquely determined in the continuous time case (almost everywhere) by its Laplace transform and in the discrete time case by its \mathbf{z} -transform (see Theorems A.3.19 and A.3.8). So the transfer matrix completely determines the

⁴Note that in the discrete time the conditions $\mathcal{G} \in \mathcal{S}_\gamma(\mathbb{K}^{p \times m})$, $u \in \mathcal{S}_\gamma(\mathbb{K}^m)$ are not needed if an interpretation of the \mathbf{z} -transform as a *function* on $\{z \in \mathbb{C}; |z| \geq \gamma\}$ is not required. The algebraic \mathbf{z} -transform can be applied to arbitrary signals and convolution kernels on \mathbb{N} and yields *formal power series* with matrix and vector coefficients, respectively, see Subsection A.3.1. The algebraic \mathbf{z} -transform converts the convolution of time functions into the multiplication of formal power-series.

input-output operator of the convolution system described by (39) or (40). Going back to the input-output behaviour of a state space system of the form (24) we can obtain an explicit expression for the associated transfer matrix by using the Laplace transform of $(e^{At})_{t \in \mathbb{R}_+}$ (resp. \mathbf{z} -transform of $(A^t)_{t \in \mathbb{N}}$). However, it is also possible to obtain this expression directly from the system equations as we show in the following example.

Example 2.3.18. Consider the state space system of the form (2.17)

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t), \quad t \in \mathbb{R}_+ \\ y(t) &= Cx(t) + Du(t).\end{aligned}\tag{49}$$

with initial state $x(0) = 0$, and let $u(\cdot)$ be a Laplace transformable input function. Since $\|e^{At}\| \leq e^{\|A\|t}$, $t \geq 0$ the convolution kernel $t \mapsto e^{At}$ is exponentially bounded. Hence it follows from (A.3.32) that the state and output trajectories

$$t \mapsto x(t) = (e^{A \cdot} * Bu(\cdot))(t), \quad t \mapsto y(t) = Du(t) + Cx(t)$$

are Laplace transformable. Applying the Laplace transform to (49) we obtain

$$\begin{aligned}s\hat{x}(s) &= A\hat{x}(s) + B\hat{u}(s) \\ \hat{y}(s) &= C\hat{x}(s) + D\hat{u}(s), \quad \operatorname{Re} s \geq \alpha\end{aligned}$$

for some suitably large α . Therefore $\hat{y}(s) = (D + C(sI_n - A)^{-1}B)\hat{u}(s)$ and so the transfer matrix of the above system is given by

$$G(s) = D + C(sI_n - A)^{-1}B.\tag{50}$$

Note that this matrix-valued function is defined on $\rho(A) = \mathbb{C} \setminus \sigma(A)$. Since $(sI_n - A)^{-1} = \det(sI_n - A)^{-1} \operatorname{adj}(sI_n - A)$ where the adjugate $\operatorname{adj}(sI_n - A)$ is a polynomial matrix whose entries are of degree $\leq n - 1$, we see that the transfer function of the time-invariant linear system (2.17) is a proper rational matrix, i.e. a matrix with entries $g_{ij}(s)$ satisfying

$$g_{ij} = \frac{p_{ij}}{q_{ij}}, \quad p_{ij}, q_{ij} \in \mathbb{K}[s], \quad \deg p_{ij} \leq \deg q_{ij}, \quad i = 1, \dots, p, \quad j = 1, \dots, m.$$

If $D = 0$, $G(s)$ is strictly proper rational, i.e. its entries satisfy $\deg p_{ij} < \deg q_{ij}$. □

Remark 2.3.19. It is shown in realization theory that, conversely, for every proper rational matrix $G(s) \in \mathbb{K}^{p \times m}(s)$ there exists a time-invariant linear state space system of the form (49) whose transfer-matrix is $G(s)$. □

In the next example we present a system with an irrational transfer function.

Example 2.3.20. Consider the delay differential system

$$\dot{x}(t) = A_0x(t) + A_1x(t - h) + Bu(t), \quad y(t) = Cx(t), \quad t > 0,$$

where $(A_0, A_1, B, C) \in \mathbb{K}^{n \times n} \times \mathbb{K}^{n \times n} \times \mathbb{K}^{n \times m} \times \mathbb{K}^{p \times n}$, $h > 0$ are given and the initial state is zero: $x(\tau) = 0$, $\tau \in [-h, 0]$. A state space description of such a system has been presented in Example 2.1.25. Constructing the state trajectory by successive application of the variation-of-constant formula (see (1.30)) one can show that if the input u is exponentially

bounded then so will $x(\cdot)$ and $y(\cdot)$ be exponentially bounded. Hence we may take the Laplace transform to obtain (see Proposition A.3.20)

$$s\hat{x}(s) = A_0\hat{x}(s) + A_1e^{-hs}\hat{x}(s) + B\hat{u}(s), \quad \hat{y}(s) = C\hat{x}(s), \quad \operatorname{Re} s \geq \alpha$$

for some suitably large α . So the transfer matrix of the above system is given by

$$G(s) = C(sI_n - A_0 - e^{-hs}A_1)^{-1}B, \quad \operatorname{Re} s \geq \alpha.$$

Note that this matrix function is no longer rational. Applying the inverse Laplace transform (see Theorem A.3.19) we conclude that the input-output behaviour of the above delay system can be described in time domain by a convolution kernel

$$y(t) = (\mathcal{G} * u)(t), \quad \text{where } \mathcal{G} = \mathcal{L}^{-1}(G).$$

The kernel \mathcal{G} can be determined as follows. Let $t \mapsto \Phi(t) \in \mathbb{K}^{n \times n}$ be the fundamental solution of the delay equation, i.e. the matrix solution of the initial value problem

$$\dot{\Phi}(t) = A_0\Phi(t) + A_1\Phi(t-h), \quad t \geq 0; \quad \Phi(s) = 0, \quad s \in [-h, 0), \quad \Phi(0) = I_n$$

(Existence and uniqueness of the solution follow as in Example 2.1.25.) Then, see [213], the state trajectories with initial function zero are given by

$$x(t) = \int_0^t \Phi(t-\tau)Bu(\tau)d\tau, \quad t \geq 0.$$

Hence the input-output behaviour (starting at zero) will be described by a convolution operator with kernel $\mathcal{G}(t) = C\Phi(t)B$, $t \geq 0$. \square

In the next example we derive a formula for the transfer matrix of the discrete time system (2.22).

Example 2.3.21. Consider the state space system of the form (2.22)

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \quad t \in \mathbb{N} \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{51}$$

with initial state $x(0) = 0$. Applying the algebraic \mathbf{z} -transform we obtain for arbitrary input sequences $u(\cdot) : \mathbb{N} \rightarrow \mathbb{K}^m$

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

and hence

$$\hat{y}(z) = G(z)\hat{u}(z) \quad \text{where} \quad G(z) = D + C(zI_n - A)^{-1}B. \tag{52}$$

We know from Example 2.3.18 that $G(z)$ is proper rational and defined on $\rho(A)$. The corresponding formal power series in $\mathbb{K}^{p \times m}[[z^{-1}]]$ can be obtained by expressing the proper rational functions $g_{ij}(z) = q_{ij}(z)/p_{ij}(z)$ (via long division) in the form $g_{ij}(z) = \sum_{k=0}^{\infty} \gamma_k^{ij} z^{-k}$ (the Laurent expansion of $g_{ij}(z)$ at ∞ , see Section A.2). \square

Examples 2.3.18 and 2.3.21 show that the transfer functions of the continuous time system (2.17) and the discrete time system (2.22) coincide. This makes it possible to transfer results concerning the input-output behaviour from one class of systems to the other.

Dyadic Decomposition of Transfer Matrices

We now examine how the input and output signals of a state space system of the form (2.17) and (2.22) are coupled via the internal dynamics of the system. In order to simplify the analysis we will only consider the (generic) case where the system matrix A is diagonalizable. Then there exists a basis v^1, \dots, v^n of \mathbb{C}^n consisting of eigenvectors of A . Let $V_j = \mathbb{C}v^j$, $j \in \underline{n}$ and let $\lambda_1, \dots, \lambda_n$ be the corresponding (not necessarily distinct) eigenvalues of A . If $\tilde{P}_i : \mathbb{C}^n \rightarrow \mathbb{C}^n$ is the canonical projection from $\mathbb{C}^n = \bigoplus_{j=1}^n V_j$ onto V_i , $\sum_{j=1}^n \alpha_j v^j \mapsto \alpha_i v^i$, then these projections \tilde{P}_i , $i = 1, \dots, n$ have the properties given in (2.30), (2.33) with $\ell = n$ and \tilde{P}_i instead of P_i . In particular

$$(sI_n - A)^{-1} = \sum_{i=1}^n (s - \lambda_i)^{-1} \tilde{P}_i, \quad s \in \rho(A).$$

If (w^1, \dots, w^n) is the biorthogonal basis of (v^1, \dots, v^n) , i.e. $w^{j*}v^i = \delta_{ij}$ (Kronecker symbol), then $\tilde{P}_i = v^i w^{i*}$. Hence the transfer matrix for (2.17) and (2.22) is

$$\begin{aligned} G(s) &= D + C(sI_n - A)^{-1}B = D + \sum_{i=1}^n (s - \lambda_i)^{-1} C \tilde{P}_i B = D + \sum_{i=1}^n (s - \lambda_i)^{-1} (Cv^i)(w^{i*}B) \\ &= D + \sum_{i=1}^n c^i (s - \lambda_i)^{-1} b^{i*} \end{aligned}$$

where $c^i = Cv^i \in \mathbb{C}^p$ and $b^i = B^*w^i \in \mathbb{C}^m$. This representation is called the *dyadic*

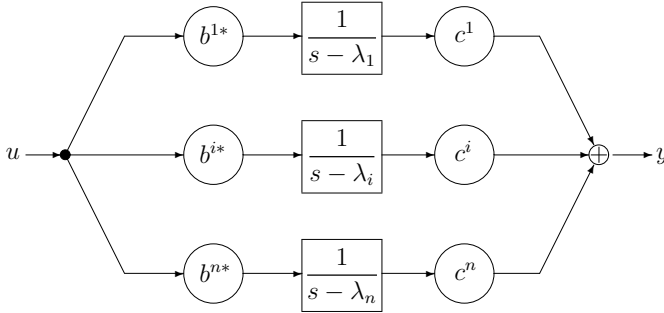


Figure 2.3.5: Dyadic decomposition of the transfer function with $D = 0$

decomposition of the transfer function for the system (A, B, C, D) and is illustrated in Figure 2.3.5. The components b_1^i, \dots, b_m^i of b^i specify the intensity by which the m inputs excite the i -th eigenmode of the system. Whereas the components c_1^i, \dots, c_p^i of c^i determine the intensity by which the i -th eigenmode influences the p outputs. If $b^i = 0$, no input will affect the i -th eigenmode and if $c^i = 0$, the i -th eigenmode will not affect the output. So, in both cases the i -th eigenmode is not important for the input-output behaviour of the system (A, B, C, D) initialized at $x(0) = 0$.

Interpretation of the Transfer Function: Response to Sinusoidal Inputs

In the previous subsection we showed that the weighting pattern $\mathcal{G}(t)$ of a continuous time convolution system can be approximately obtained by testing the input-output

system with approximations to Dirac impulses. We will now show that the transfer function of such a system can be determined approximately by applying harmonic test signals of various frequencies to the system and interpolating the results. In contrast to the impulse approximations, this procedure is not only of theoretical but also of some practical importance, see *Notes and References*.

For simplicity we only consider the real scalar case with time domain $T = \mathbb{R}_+$ (resp. $T = \mathbb{N}$) and integrable (resp. summable) convolution kernel. Then the associated transfer function $g(s)$ is defined and continuous on the closed right half-plane $\overline{\mathbb{C}_+}$ (resp. on $\overline{\mathbb{D}_+}$) and hence also on the imaginary axis $i\mathbb{R}$ (resp. the unit circle $\partial\mathbb{D}$).

Proposition 2.3.22. *Let $g(\cdot)$ be the transfer function of a real scalar convolution system (40) with integrable kernel $\mathcal{G}(\cdot) \in L^1(\mathbb{R}_+; \mathbb{R})$ (resp. summable kernel $\mathcal{G}(\cdot) \in \ell^1(\mathbb{N}; \mathbb{R})$) and a real feedthrough coefficient $D \in \mathbb{R}$. Then, for every $\omega \in \mathbb{R}$ (resp. $\theta \in [-\pi, \pi]$), the system response to the input signal $u(t) = \sin \omega t$, $t \in \mathbb{R}_+$ (resp. $u(t) = \sin \theta t$, $t \in \mathbb{N}$) approximates for large t the “steady state response”*

$$y^{ss}(t) = |g(i\omega)| \sin(\omega t + \varphi(\omega)), \quad t \geq 0 \quad (\text{resp. } y^{ss}(t) = |g(e^{i\theta})| \sin(\theta t + \varphi(\theta)), \quad t \in \mathbb{N})$$

where $\varphi(\omega)$ (resp. $\varphi(\theta)$) is an argument function of $g(i\omega)$ (resp. $g(e^{i\theta})$).

Proof: We only prove the proposition for the continuous time case. The discrete time case is left to the reader, see Ex. 8. First note that

$$y^{ss}(t) = |g(i\omega)| \operatorname{Im} e^{i(\omega t + \varphi(\omega))} = |g(i\omega)| \operatorname{Im} (e^{i\varphi(\omega)} e^{i\omega t}) = \operatorname{Im} (g(i\omega) e^{i\omega t}), \quad t \geq 0.$$

The system response to $u(t) = \sin \omega t = \operatorname{Im} e^{i\omega t}$ is $y(t) = D \sin \omega t + (\mathcal{G} * u)(t)$ for $t \in \mathbb{R}_+$. On the other hand $\operatorname{Im} (g(i\omega) e^{i\omega t}) = \operatorname{Im} (D e^{i\omega t} + \hat{\mathcal{G}}(i\omega) e^{i\omega t}) = D \sin \omega t + \operatorname{Im} (\hat{\mathcal{G}}(i\omega) e^{i\omega t})$. Therefore

$$\begin{aligned} |y(t) - \operatorname{Im} (g(i\omega) e^{i\omega t})| &= \left| \int_0^t \mathcal{G}(\tau) \sin \omega(t - \tau) d\tau - \operatorname{Im} \int_0^\infty \mathcal{G}(\tau) e^{-i\omega\tau} d\tau e^{i\omega t} \right| \\ &= \left| \operatorname{Im} \int_t^\infty \mathcal{G}(\tau) e^{i\omega(t-\tau)} d\tau \right|, \quad \text{as } \mathcal{G}(\cdot) \text{ is real} \\ &\leq \left| \int_t^\infty \mathcal{G}(\tau) e^{i\omega(t-\tau)} d\tau \right| \leq \int_t^\infty |\mathcal{G}(\tau)| d\tau \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned} \tag{53}$$

Since $y^{ss}(t) = \operatorname{Im} (g(i\omega) e^{i\omega t})$ this concludes the proof. \square

In the continuous time case the steady state response $y^{ss}(t)$ is a harmonic oscillation of the same frequency as the harmonic input signal but amplified by $|g(i\omega)|$ and phase shifted by $\arg g(i\omega)$. Consequently, the values of the transfer function g on the imaginary axis can be determined experimentally (sometimes with great accuracy) by measuring for each interesting frequency ω the magnitude and phase shift (with respect to the phase of the input signal) of the steady state response. This procedure can, in principle, also be applied to multivariable convolution systems by feeding harmonic inputs successively into each of the m input channels (keeping the other input channels at zero) and measuring the amplifications and phase shifts of the system's responses on each of p output channels. Analogous results hold for the discrete time case.

Frequency Responses

We have just seen that the transfer function $g(s)$ of a scalar convolution system (40) with integrable kernel on $T = \mathbb{R}_+$ (resp. summable kernel on $T = \mathbb{N}$) is defined and continuous on the imaginary axis $\imath\mathbb{R}$ (resp. the unit circle $\partial\mathbb{D}$). $g(\cdot)$ is completely determined by its values on the imaginary axis (resp. the unit circle), see Proposition A.3.41 (resp. Proposition A.3.45). Thus the restriction of $g(s)$ to $\imath\mathbb{R}$ (resp. $\partial\mathbb{D}$) determines the input-output behaviour of the convolution system. This motivates the following definition.

Definition 2.3.23. Let $g(\cdot)$ be the transfer function of a real scalar convolution system with integrable kernel \mathcal{G} on \mathbb{R}_+ (resp. summable kernel \mathcal{G} on \mathbb{N}). Then the complex valued function $\omega \rightarrow g(\imath\omega)$ (resp. $\theta \rightarrow g(e^{\imath\theta})$) on \mathbb{R} (resp. $[-\pi, \pi]$) is called the *complex frequency response*. The real valued function $\omega \rightarrow |g(\imath\omega)|$ (resp. $\theta \rightarrow |g(e^{\imath\theta})|$) is called the *amplitude (gain) response* and any (continuous) argument function $\omega \rightarrow \arg g(\imath\omega)$ (resp. $\theta \rightarrow \arg g(e^{\imath\theta})$) the *phase response* of the continuous (resp. discrete) time scalar convolution system (40).

The importance of these concepts for classical control theory, results from the fact that three of the four most prominent classical analysis and design techniques for linear siso systems (Nyquist, Bode, Nichols chart and root locus methods) are based on graphical representations of the frequency response. In particular, Nyquist's method proceeds from the *polar plot* $\{g(\imath\omega); \omega \in \mathbb{R}\}$ of the complex frequency response, and Bode's method proceeds from the graphs of the amplitude and the phase responses, see *Notes and References*.

Remark 2.3.24. Classical techniques have been developed for siso systems. Clearly they can be applied individually to represent graphically the influence of the j -th input channel on the i -th output of a multivariable system (2.17) or (2.22) described by the entry $g_{ij}(s)$ of the associated transfer matrix $G(s) = (g_{ij}(s)) \in \mathbb{K}^{p \times m}(s)$. However, these graphical methods are, in general, not suitable for analyzing the input-output behaviour of a multivariable system as a whole. \square

Before illustrating the concept of *complex frequency response* by the *polar plots* of some simple siso systems let us make some general remarks concerning these plots.

- (i) The transfer functions of stable real siso systems satisfy $\overline{g(\imath\omega)} = g(-\imath\omega)$ for all $\omega \in \mathbb{R}$, so that their polar plots are symmetric with respect to the real axis, and hence need only be computed for $\omega \geq 0$.
- (ii) It is usual to indicate the orientation of the polar plot by an arrow showing the direction in which $g(\imath\omega)$ evolves as ω is increasing. If $g(s)$ and $h(s)$ are two transfer functions satisfying $h(s) = g(-s)$ then g and h have the same polar plots but they have reverse orientations.
- (iii) For a continuous time siso convolution system with integrable kernel \mathcal{G} and $D = 0$, the amplitude response $|g(\imath\omega)|$ tends to zero as $|\omega| \rightarrow \infty$ and so the harmonic inputs with large frequencies will be attenuated by the system, see (53). On the contrary, siso systems whose amplitude response is constant (i.e. $|g(\imath\omega)| = c > 0$ for all $\omega \in \mathbb{R}$) amplify/dampen harmonic inputs by the

same factor c for all frequencies. They are called *all-pass* functions and in the rational case are characterized by the property of pole-zero symmetry with respect to the imaginary axis: if s_0 is a pole of g , then $-\bar{s}_0$ is a zero.

Example 2.3.25. The transfer function of any first order real siso system is of the form $d + b/(s + a)$. If $a \neq 0$ the corresponding polar plot is a circle since $|g(i\omega) - d - b/(2a)| = |b|/|2a|$ for all $\omega \in \mathbb{R}$. For $a = 0$ the polar plot is a vertical line since $g(i\omega) = d - ib/\omega$. In Figure 2.3.25 we illustrate, by their polar plots for $\omega \geq 0$, the frequency responses of five simple real siso systems (with and without delay). (The full polar plots are then obtained by adding their reflections about the real axis). Consider the transfer functions

$$g_1(s) = 1/(s^2 + 2s + 5), \quad g_2(s) = 1/s(s^2 + 2s + 5), \quad g_3(s) = (s + 1)/(s^2 + 2s + 5),$$

$$g_4(s) = e^{-s}/(1 + s), \quad g_5(s) = 1/(s + 1 + e^{-s}).$$

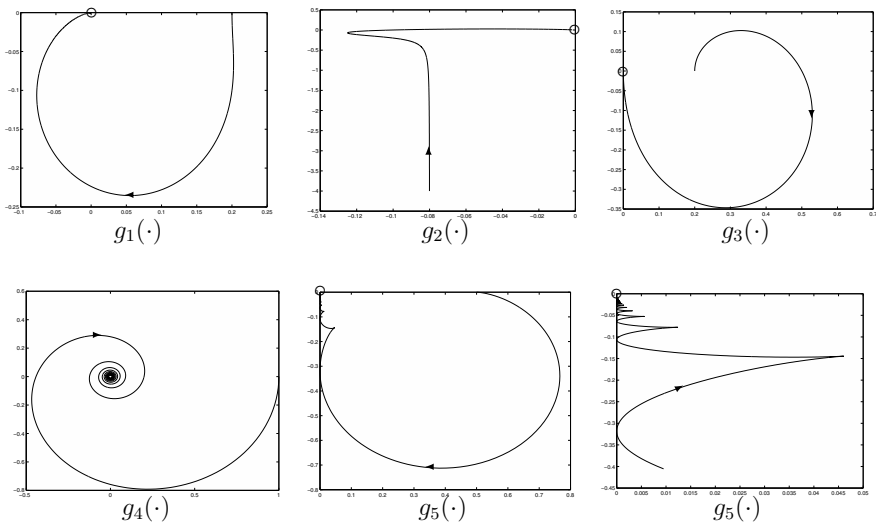


Figure 2.3.6: Polar plots of the transfer functions $g_1(\cdot), \dots, g_5(\cdot)$

It is easy to construct state space systems of the form (49) whose transfer functions are $g_1(s)$, $g_2(s)$ and $g_3(s)$, see Ex. 11. $g_4(s)$ and $g_5(s)$ are the transfer functions of the delay systems

$$\dot{x}(t) = -x(t) + u(t-1), \quad y(t) = x(t) \quad \text{and} \quad \dot{x}(t) = -x(t) - x(t-1) + u(t), \quad y(t) = x(t).$$

In the six pictures of Figure 2.3.6 the origin in the complex plane is indicated by \circ . The pictures were produced via the `plot` command in MATLAB with a frequency band, in the main, from 0 to 100. The exceptions are the polar plot of g_2 (with frequency range $[0.5, 6]$) and the right hand plot of g_5 which is a zoom in (for higher frequencies) of the left hand one. g_1 has two complex poles at $-1 \pm 2i$ and its plot is a typical one for a second order system. g_2 is obtained from g_1 by adding a pole at the origin. Obviously the polar plot will be unbounded if there is a pole of the transfer function on the imaginary axis. Here $\operatorname{Re} g_2(i\omega) \rightarrow -0.08$ and $\operatorname{Im} g_2(i\omega) \rightarrow -\infty$ as $\omega \rightarrow 0$ (see Ex. 10). g_3 is obtained from g_1

by adding a zero at $s = -1$. The polar plot is quite different to that of g_1 . For small positive values of ω we have $\text{Im } g_3(i\omega) \geq 0$ and for all ω we have $\text{Re } g_3(i\omega) \geq 0$, whereas this is not the case for g_1 . In fact g_3 is analytic on $\overline{\mathbb{C}_+}$ and maps this closed right half-plane into itself. Transfer functions with this property are called *positive real*. g_4 is the transfer function of a first order system where there is a delay of one unit in the input. The effect of the delay is to change the non-delayed plot of a circle to that of a spiral. Such behaviour is exhibited even if the delay is very small. This shows that in designing controls (or analyzing stability) by frequency domain methods one should be careful about neglecting delays. g_5 is another transfer function of a system with a delay, this time not in the control but in the state. For $\omega = 0$ we have $g_5(0) = 1/2$ which lies on the polar plot of $g(s) = 1/(s+2)$ which is a circle $\partial D((1/4, 0), 1/4)$ of radius $1/4$ around the centre $(1/4, 0)$. Then at $\omega = \pi$ it hits the imaginary axis $i\mathbb{R}$ for the first time. At $\omega = 2\pi$ it returns to the circle $\partial D((1/4, 0), 1/4)$ and at $\omega = 3\pi$ it hits $i\mathbb{R}$ again. The process is continued at multiples of π as seen in the right hand figure for g_5 . Again this type of behaviour would also occur if the delay was small.

The transfer functions g_1 , g_3 , g_4 and g_5 are all in $H^\infty(\mathbb{C}_+; \mathbb{C})$, i.e. they are continuous and bounded on $\overline{\mathbb{C}_+}$ and analytic in \mathbb{C}_+ , see Ex. 11. We will see in the next subsection that this implies that the corresponding convolution systems are L^2 -stable. g_2 is not L^2 -stable. \square

2.3.3 Relationship Between Input–Output Operators and Transfer Matrices

In this subsection we consider convolution systems in both time domain and frequency domain and clarify the relationship between their input-output operators and transfer matrices. The technical development relies on Section A.3. Throughout the subsection it is assumed that all finite dimensional vector spaces \mathbb{K}^p , \mathbb{K}^m are equipped with their standard Euclidean norms $\|\cdot\|_2$ and $\mathbb{K}^{p \times m}$ with the corresponding operator norm $\|\cdot\|_{2,2}$.

We suppose that an integrable convolution kernel $\mathcal{G}(\cdot) \in L^1(\mathbb{R}_+; \mathbb{C}^{p \times m})$ (resp. $\mathcal{G}(\cdot) \in \ell^1(\mathbb{N}; \mathbb{C}^{p \times m})$) and a feedthrough matrix $D \in \mathbb{C}^{p \times m}$ are given and first consider the associated convolution system with the time domain $T = \mathbb{R}_+$ (resp. \mathbb{N}). This convolution system is described by $(\mathcal{U}_+, \mathbb{G}_+, \mathcal{Y}_+)$ where $\mathcal{U}_+ = L^q(\mathbb{R}_+; \mathbb{C}^m)$, $\mathcal{Y}_+ = L^q(\mathbb{R}_+; \mathbb{C}^p)$ (resp. $\mathcal{U}_+ = \ell^q(\mathbb{N}; \mathbb{C}^m)$, $\mathcal{Y}_+ = \ell^q(\mathbb{N}; \mathbb{C}^p)$) and the input-output operator $\mathbb{G}_+ : \mathcal{U}_+ \rightarrow \mathcal{Y}_+$ is of the form (39), see Proposition 2.3.15. In order to describe the relationship between \mathbb{G}_+ and the transfer matrix $G(\cdot)$ we introduce the Hardy spaces $H^q(\mathbb{C}_+; \mathbb{C}^n)$ (resp. $H^q(\mathbb{D}_+; \mathbb{C}^n)$).

Definition 2.3.26. For $1 \leq q \leq \infty$ denote by $H^q(\mathbb{C}_+; \mathbb{C}^n)$ the space of all analytic functions $v(\cdot)$ on \mathbb{C}_+ with values in \mathbb{C}^n satisfying $\|v(\cdot)\|_{H^q(\mathbb{C}_+; \mathbb{C}^n)} < \infty$ where

$$\|v(\cdot)\|_{H^q(\mathbb{C}_+; \mathbb{C}^n)} = \begin{cases} \sup_{\alpha > 0} \left(\int_{-\infty}^{\infty} \|v(\alpha + i\omega)\|_2^q d\omega \right)^{1/q}, & \text{if } 1 \leq q < \infty \\ \sup_{s \in \mathbb{C}_+} \|v(s)\|_2, & \text{if } q = \infty. \end{cases} \quad (54)$$

For some properties of these vector spaces see Subsection A.3.4. It is known that $H^q(\mathbb{C}_+; \mathbb{C}^n)$ provided with the norm (54) is a Banach space.

Definition 2.3.27. For $1 \leq q \leq \infty$ denote by $H^q(\mathbb{D}_+; \mathbb{C}^n)$ the space of all analytic functions $v(\cdot)$ on \mathbb{D}_+ with values in \mathbb{C}^n satisfying $\|v(\cdot)\|_{H^q(\mathbb{D}_+; \mathbb{C}^n)} < \infty$ where

$$\|v(\cdot)\|_{H^q(\mathbb{D}_+; \mathbb{C}^n)} = \begin{cases} \sup_{r>1} \left(\int_{-\pi}^{\pi} \|v(re^{i\theta})\|_2^q d\theta \right)^{1/q}, & \text{if } 1 \leq q < \infty \\ \sup_{z \in \mathbb{D}_+} \|v(z)\|_2, & \text{if } q = \infty. \end{cases} \quad (55)$$

Again it is known that $H^q(\mathbb{D}_+; \mathbb{C}^n)$ provided with this norm (55) is a Banach space. We now specialize to the case where $q = 2$ so that $\mathcal{U}_+ = L^2(\mathbb{R}_+; \mathbb{C}^m)$, (resp. $\mathcal{U}_+ = \ell^2(\mathbb{N}; \mathbb{C}^m)$), and $\mathcal{Y}_+ = L^2(\mathbb{R}_+; \mathbb{C}^p)$, (resp. $\mathcal{Y}_+ = \ell^2(\mathbb{N}; \mathbb{C}^p)$). Let $G(s) = (D + \mathcal{L}\mathcal{G})(s)$ (resp. $G(z) = \mathcal{Z}\mathcal{G}(z)$) be the transfer matrix of the convolution system $(\mathcal{U}_+, \mathbb{G}_+, \mathcal{Y}_+)$, see Definition 2.3.17. Since $\mathcal{G}(\cdot) \in L^1(\mathbb{R}_+; \mathbb{C}^{p \times m})$, its Laplace transform is analytic on \mathbb{C}_+ , bounded and continuous on $\overline{\mathbb{C}_+}$ so that $G(\cdot) \in H^\infty(\mathbb{C}_+; \mathbb{C}^{p \times m})$. Similarly, in the discrete time case, $G(z)$ is analytic on \mathbb{D}_+ , bounded and continuous on $\overline{\mathbb{D}_+}$ so that $G(\cdot) \in H^\infty(\mathbb{D}_+; \mathbb{C}^{p \times m})$. Moreover by Proposition A.3.41 and Proposition A.3.45 we have

$$\|G(\cdot)\|_{H^\infty(\mathbb{C}_+; \mathbb{C}^{p \times m})} = \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2}, \quad \|G(\cdot)\|_{H^\infty(\mathbb{D}_+; \mathbb{C}^{p \times m})} = \max_{\theta \in [-\pi, \pi]} \|G(e^{i\theta})\|_{2,2}. \quad (56)$$

If $u(\cdot) \in L^2(\mathbb{R}_+; \mathbb{C}^m)$ (resp. $u(\cdot) \in \ell^2(\mathbb{N}; \mathbb{C}^m)$), then $u(\cdot) \in \mathcal{E}_\alpha(\mathbb{C}^m)$ (resp. $\mathcal{S}_\gamma(\mathbb{C}^m)$) for every $\alpha > 0$ (resp. $\gamma > 1$). Therefore by (47)

$$\mathcal{L}((\mathbb{G}_+ u)(\cdot))(s) = G(s)\hat{u}(s), \quad \text{Re } s > 0, \quad \mathcal{Z}((\mathbb{G}_+ u)(\cdot))(z) = G(z)\hat{u}(z), \quad |z| > 1 \quad (57)$$

for every $u(\cdot) \in L^2(\mathbb{R}_+; \mathbb{C}^m)$ (resp. $u(\cdot) \in \ell^2(\mathbb{N}; \mathbb{C}^m)$). From (54) and (55) we get

$$\begin{aligned} \|G(\cdot)w(\cdot)\|_{H^2(\mathbb{C}_+; \mathbb{C}^p)} &\leq \|G(\cdot)\|_{H^\infty(\mathbb{C}_+; \mathbb{C}^{p \times m})} \|w(\cdot)\|_{H^2(\mathbb{C}_+; \mathbb{C}^m)}, \quad w(\cdot) \in H^2(\mathbb{C}_+; \mathbb{C}^m) \\ \|G(\cdot)w(\cdot)\|_{H^2(\mathbb{D}_+; \mathbb{C}^p)} &\leq \|G(\cdot)\|_{H^\infty(\mathbb{D}_+; \mathbb{C}^{p \times m})} \|w(\cdot)\|_{H^2(\mathbb{D}_+; \mathbb{C}^m)}, \quad w(\cdot) \in H^2(\mathbb{D}_+; \mathbb{C}^m). \end{aligned} \quad (58)$$

This shows that pointwise multiplication of an H^2 -function by $G(s)$ yields an H^2 -function. Let $M_G^+ : H^2(\mathbb{C}_+; \mathbb{C}^m) \rightarrow H^2(\mathbb{C}_+; \mathbb{C}^p)$ be the associated multiplication operator for the continuous time case defined by

$$(M_G^+ w)(s) = G(s)w(s), \quad w(\cdot) \in H^2(\mathbb{C}_+; \mathbb{C}^m), \quad s \in \mathbb{C}_+, \quad (59)$$

and $M_G^+ : H^2(\mathbb{D}_+; \mathbb{C}^m) \rightarrow H^2(\mathbb{D}_+; \mathbb{C}^p)$ be its counterpart for the discrete time case defined by

$$(M_G^+ w)(z) = G(z)w(z), \quad w(\cdot) \in H^2(\mathbb{D}_+; \mathbb{C}^m), \quad z \in \mathbb{D}_+. \quad (60)$$

From (57) we see that the following diagrams commute

$$\begin{array}{ccc} L^2(\mathbb{R}_+; \mathbb{C}^m) & \xrightarrow{\mathbb{G}_+} & L^2(\mathbb{R}_+; \mathbb{C}^p) \\ \mathcal{L} \downarrow & & \downarrow \mathcal{L} \\ H^2(\mathbb{C}_+; \mathbb{C}^m) & \xrightarrow{M_G^+} & H^2(\mathbb{C}_+; \mathbb{C}^p) \end{array}, \quad \begin{array}{ccc} \ell^2(\mathbb{N}; \mathbb{C}^m) & \xrightarrow{\mathbb{G}_+} & \ell^2(\mathbb{N}; \mathbb{C}^p) \\ z \downarrow & & \downarrow z \\ H^2(\mathbb{D}_+; \mathbb{C}^m) & \xrightarrow{M_G^+} & H^2(\mathbb{D}_+; \mathbb{C}^p) \end{array}.$$

By Theorem A.3.43 the normalized \mathbf{z} -transform $(2\pi)^{-1/2}\mathcal{Z}$ is an isometry between the two ℓ^2 and H^2 spaces in the second diagram. Similarly by Theorem A.3.47 the normalized Laplace transform $(2\pi)^{-1/2}\mathcal{L}$ is an isometry between the two L^2 and H^2 spaces in the first diagram. Hence

$$\begin{aligned}\|\mathbb{G}_+\|_{\mathcal{L}(L^2(\mathbb{R}_+;\mathbb{C}^m), L^2(\mathbb{R}_+;\mathbb{C}^p))} &= \|M_G^+\|_{\mathcal{L}(H^2(\mathbb{C}_+;\mathbb{C}^m), H^2(\mathbb{C}_+;\mathbb{C}^p))}, \\ \|\mathbb{G}_+\|_{\mathcal{L}(\ell^2(\mathbb{N};\mathbb{C}^m), \ell^2(\mathbb{N};\mathbb{C}^p))} &= \|M_G^+\|_{\mathcal{L}(H^2(\mathbb{D}_+;\mathbb{C}^m), H^2(\mathbb{D}_+;\mathbb{C}^p))}.\end{aligned}\quad (61)$$

We will see later in Theorem 2.3.28 that the operator norms (61) of the multiplication operators M_G^+ can be computed by maximizing $\|G(s)\|_{2,2}$ on the imaginary axis and the unit circle, respectively.

We now turn to the case where $T = \mathbb{R}$ (resp. \mathbb{Z}) and consider the convolution system $(\mathcal{U}, \mathbb{G}, \mathcal{Y})$ where $\mathcal{U} = L^2(\mathbb{R}; \mathbb{C}^m)$, $\mathcal{Y} = L^2(\mathbb{R}; \mathbb{C}^p)$ (resp. $\mathcal{U} = \ell^2(\mathbb{Z}; \mathbb{C}^m)$, $\mathcal{Y} = \ell^2(\mathbb{Z}; \mathbb{C}^p)$) and the input-output operator \mathbb{G} is given by (40). Again it follows from the convolution inequalities (A.3.24) and (A.3.7) that $\mathbb{G} : \mathcal{U} \rightarrow \mathcal{Y}$ is a bounded linear operator. Since the control functions may admit non-zero values on $(-\infty, 0)$ the Laplace transform \mathcal{L} is no longer applicable to these signals. Instead we make use of the Fourier-Plancherel transform defined by (41) and the discrete Fourier transform defined by (44) where the limit of the series is to be understood in $L^2(-\pi, \pi; \mathbb{C}^m)$. We extend $\mathcal{G}(\cdot)$ trivially to \mathbb{R} (resp. \mathbb{Z}) by setting $\mathcal{G}(t) = 0$ for $t < 0$. By (43) (with $\beta = 0$) and (46) (with $r = 1$) the Fourier transform (resp. discrete Fourier transform) of this extension is

$$\mathcal{F}(\mathcal{G})(\omega) = (\mathcal{L}\mathcal{G})(i\omega) = G(i\omega) - D, \quad \mathcal{F}_D(\mathcal{G})(\theta) = (\mathcal{Z}\mathcal{G})(e^{i\theta}) = G(e^{i\theta}).$$

Hence by Proposition A.3.35 (iii) (resp. Proposition A.3.39) the Fourier-Plancherel transform (resp. discrete Fourier transform) of the output signal $y(\cdot) = (\mathbb{G}u)(\cdot)$, $u(\cdot) \in \mathcal{U}$ is given by

$$\begin{aligned}\tilde{y}(\omega) &= (\mathcal{F}(Du + \mathcal{G} * u))(\omega) = G(i\omega)\tilde{u}(\omega), \quad \text{a.e. } \omega \in \mathbb{R}, \\ \tilde{y}(\theta) &= (\mathcal{F}_D(\mathcal{G} * u))(\theta) = G(e^{i\theta})\tilde{u}(\theta), \quad \text{a.e. } \theta \in [-\pi, \pi]\end{aligned}\quad (62)$$

where $\tilde{u}(\cdot) \in L^2(\mathbb{R}; \mathbb{C}^m)$ (resp. $\tilde{u}(\cdot) \in L^2(-\pi, \pi; \mathbb{C}^m)$) denotes the Fourier-Plancherel transform (resp. discrete Fourier transform) of $u(\cdot)$. Now for $w(\cdot) \in L^2(\mathbb{R}; \mathbb{C}^m)$ (resp. $w(\cdot) \in L^2(-\pi, \pi; \mathbb{C}^m)$) we have

$$\begin{aligned}\|G(i\cdot)w(\cdot)\|_{L^2(\mathbb{R}; \mathbb{C}^p)} &\leq \max_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} \|w(\cdot)\|_{L^2(\mathbb{R}; \mathbb{C}^m)}, \\ \|G(e^{i\cdot})w(\cdot)\|_{L^2(-\pi, \pi; \mathbb{C}^p)} &\leq \max_{\theta \in [-\pi, \pi]} \|G(e^{i\theta})\|_{2,2} \|w(\cdot)\|_{L^2(-\pi, \pi; \mathbb{C}^m)}.\end{aligned}\quad (63)$$

Hence the multiplication operators

$$\begin{aligned}M_G : L^2(\mathbb{R}; \mathbb{C}^m) &\rightarrow L^2(\mathbb{R}; \mathbb{C}^p), \quad (M_G w)(\omega) = G(i\omega)w(\omega), \quad \text{a.e. } \omega \in \mathbb{R}, \\ M_G : L^2(-\pi, \pi; \mathbb{C}^m) &\rightarrow L^2(-\pi, \pi; \mathbb{C}^p), \quad (M_G w)(\theta) = G(e^{i\theta})w(\theta), \quad \text{a.e. } \theta \in [-\pi, \pi]\end{aligned}\quad (64)$$

are well defined, linear and bounded. The equations in (62) imply that the following

diagrams commute

$$\begin{array}{ccc}
 L^2(\mathbb{R}; \mathbb{C}^m) & \xrightarrow{\mathbb{G}} & L^2(\mathbb{R}; \mathbb{C}^p) \\
 \mathcal{F} \downarrow & & \downarrow \mathcal{F} \\
 L^2(\mathbb{R}; \mathbb{C}^m) & \xrightarrow{M_G} & L^2(\mathbb{R}; \mathbb{C}^p)
 \end{array}
 \qquad
 \begin{array}{ccc}
 \ell^2(\mathbb{Z}; \mathbb{C}^m) & \xrightarrow{\mathbb{G}} & \ell^2(\mathbb{Z}; \mathbb{C}^p) \\
 \mathcal{F}_D \downarrow & & \downarrow \mathcal{F}_D \\
 L^2(-\pi, \pi; \mathbb{C}^m) & \xrightarrow{M_G} & L^2(-\pi, \pi; \mathbb{C}^p)
 \end{array}$$

Since by Theorem A.3.33 the normalized Fourier transform $(2\pi)^{-1/2}\mathcal{F}$ is an isometry between the two L^2 spaces on the left diagram and by Remark A.3.38 the normalized discrete Fourier transform $(2\pi)^{-1/2}\mathcal{F}_D$ is an isometry between the two ℓ^2 and L^2 spaces of the right diagram, it follows that \mathbb{G} and M_G have the same norm,

$$\begin{aligned}
 \|\mathbb{G}\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{C}^m), L^2(\mathbb{R}; \mathbb{C}^p))} &= \|M_G\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{C}^m), L^2(\mathbb{R}; \mathbb{C}^p))}, \\
 \|\mathbb{G}\|_{\mathcal{L}(\ell^2(\mathbb{Z}; \mathbb{C}^m), \ell^2(\mathbb{Z}; \mathbb{C}^p))} &= \|M_G\|_{\mathcal{L}(L^2(-\pi, \pi; \mathbb{C}^m), L^2(-\pi, \pi; \mathbb{C}^p))}.
 \end{aligned} \tag{65}$$

The following theorem links these results with the corresponding ones for \mathbb{G}_+ given in (61). It will be used in Section 5.3 to characterize the complex stability radius.

Theorem 2.3.28. *Suppose $\mathcal{G}(\cdot) \in L^1(\mathbb{R}_+; \mathbb{K}^{p \times m})$ (resp. $\mathcal{G}(\cdot) \in \ell^1(\mathbb{N}; \mathbb{K}^{p \times m})$) and $D \in \mathbb{K}^{p \times m}$, $\mathbb{G} \in \mathcal{L}(L^2(\mathbb{R}; \mathbb{K}^m), L^2(\mathbb{R}; \mathbb{K}^p))$ (resp. $\mathbb{G} \in \mathcal{L}(\ell^2(\mathbb{Z}; \mathbb{K}^m), \ell^2(\mathbb{Z}; \mathbb{K}^p))$), $\mathbb{G}_+ \in \mathcal{L}(L^2(\mathbb{R}_+; \mathbb{K}^m), L^2(\mathbb{R}_+; \mathbb{K}^p))$ (resp. $\mathbb{G} \in \mathcal{L}(\ell^2(\mathbb{N}; \mathbb{K}^m), \ell^2(\mathbb{N}; \mathbb{K}^p))$) are the input-output operators defined by (39) (resp. (40)) and $G(\cdot)$ is the associated transfer matrix (48), then*

$$\begin{aligned}
 \|G\|_{H^\infty} &= \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} = \|\mathbb{G}_+\|_{\mathcal{L}(L^2(\mathbb{R}_+; \mathbb{K}^m), L^2(\mathbb{R}_+; \mathbb{K}^p))} = \|\mathbb{G}\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{K}^m), L^2(\mathbb{R}; \mathbb{K}^p))}, \\
 \|G\|_{H^\infty} &= \max_{\theta \in [-\pi, \pi]} \|G(e^{i\theta})\|_{2,2} = \|\mathbb{G}_+\|_{\mathcal{L}(\ell^2(\mathbb{N}; \mathbb{K}^m), \ell^2(\mathbb{N}; \mathbb{K}^p))} = \|\mathbb{G}\|_{\mathcal{L}(\ell^2(\mathbb{Z}; \mathbb{K}^m), \ell^2(\mathbb{Z}; \mathbb{K}^p))}.
 \end{aligned} \tag{66}$$

Proof: The proof is for the continuous time case. If \mathbb{G} is a real convolution operator we have $\|\mathbb{G}\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{R}^m), L^2(\mathbb{R}; \mathbb{R}^p))} = \|\mathbb{G}\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{C}^m), L^2(\mathbb{R}; \mathbb{C}^p))}$, so we need only consider the case $\mathbb{K} = \mathbb{C}$. By (56)

$$\|G\|_{H^\infty(\mathbb{C}_+; \mathbb{C}^{p \times m})} = \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2}.$$

So because of (65) and Proposition 2.3.15 it remains to prove that

$$\|M_G\|_{\mathcal{L}(L^2(\mathbb{R}; \mathbb{C}^m), L^2(\mathbb{R}; \mathbb{C}^p))} = \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2}.$$

By (63) the inequality \leq holds. To prove the converse inequality let $\varepsilon > 0$ be arbitrary. Then there exist $\omega_0 \in \mathbb{R}$, $\delta > 0$ and $u \in \mathbb{C}^m$, $\|u\|_2 = 1$ such that

$$\|G(i\omega)u\|_2 \geq \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} - \varepsilon, \quad \omega \in [\omega_0 - \delta, \omega_0 + \delta]. \tag{67}$$

This follows from the fact that $G(\cdot)$ is continuous and bounded on $\overline{\mathbb{C}_+}$: One first chooses ω_0 such that

$$\|G(i\omega_0)\|_{2,2} \geq \sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} - \varepsilon/2,$$

then $u \in \mathbb{C}^m$, $\|u\|_2 = 1$ such that $\|G(i\omega_0)u\|_2 = \|G(i\omega_0)\|_{2,2}$ (see Definition A.1.4) and finally $\delta > 0$ such that

$$\|G(i\omega)u - G(i\omega_0)u\|_2 \leq \varepsilon/2, \quad \omega \in [\omega_0 - \delta, \omega_0 + \delta].$$

Define $\tilde{u}(\cdot) \in L^2(\mathbb{R}; \mathbb{C}^m)$ by $\tilde{u}(\omega) = u/\sqrt{2\delta}$ for $\omega \in [\omega_0 - \delta, \omega_0 + \delta]$ and $\tilde{u}(\omega) = 0$ otherwise. Then $\|\tilde{u}\|_{L^2(\mathbb{R}; \mathbb{C}^m)} = 1$ and by (67)

$$\int_{-\infty}^{\infty} \|(M_G \tilde{u})(\omega)\|_2^2 d\omega = \frac{1}{2\delta} \int_{\omega_0 - \delta}^{\omega_0 + \delta} \|G(i\omega)u\|_2^2 d\omega \geq \left(\sup_{\omega \in \mathbb{R}} \|G(i\omega)\|_{2,2} - \varepsilon \right)^2.$$

This completes the proof of the continuous time case. The proof for the discrete time case is similar, see Ex. 13. \square

The following example illustrates how the norm of the input-output operator of a state space system (49) (resp. (51)) can be determined by applying (66).

Example 2.3.29. Consider the oscillator described in Example 2.3.6,

$$\ddot{\xi}(t) + 2\alpha\dot{\xi}(t) + \nu^2\xi(t) = \nu^2u(t), \quad y(t) = \xi(t), \quad t \in \mathbb{R}_+.$$

We assume $\alpha > 0$, $\nu \neq 0$ so that the corresponding state space system satisfies $\sigma(A) \subset \mathbb{C}_-$. The transfer function is $g(s) = \nu^2/(s^2 + 2\alpha s + \nu^2)$ and a simple calculation gives

$$\sup_{\omega \in \mathbb{R}} |g(i\omega)| = \begin{cases} 1 & \text{if } \nu^2 \leq 2\alpha^2 \\ \frac{\nu^2}{2\alpha\sqrt{\nu^2 - \alpha^2}} & \text{if } \nu^2 > 2\alpha^2. \end{cases}$$

The discrete time counterpart is

$$\xi(t+2) + 2\alpha\xi(t+1) + \nu^2\xi(t) = \nu^2u(t), \quad y(t) = \xi(t), \quad t \in \mathbb{N}.$$

The corresponding state space system satisfies $\sigma(A) \subset \mathbb{D}$ if $1 > \nu^2 > |2\alpha| - 1$. The transfer function is as above and an easy calculation gives

$$\max_{\theta \in [-\pi, \pi]} |g(e^{i\theta})| = \begin{cases} \frac{|\nu|^3}{(1 - \nu^2)[\nu^2 - \alpha^2]^{1/2}} & \text{if } |\alpha|(1 + \nu^2) < 2\nu^2 \\ \frac{\nu^2}{1 - 2|\alpha| + \nu^2} & \text{if } |\alpha|(1 + \nu^2) \geq 2\nu^2. \end{cases} \quad \square$$

2.3.4 Exercises

1. Compute the impulse response $\mathcal{G} = (\mathcal{G}(t))_{t \in \mathbb{N}}$ and the *step response* (corresponding to the constant input $u(t) \equiv 1$, $t \in \mathbb{N}$) for the discrete time system (A, B, C) of Ex. 2.2.4 and the discrete time system (A, B, C, D) of Ex. 2.2.8.

2. Calculate the impulse response for the continuous time system (49) with

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

Find also the response $y^k(\cdot)$ to the input $u^k(\cdot)$ where

$$u^k(t) = \begin{cases} k & \text{if } t \in [0, 1/k) \\ 0 & \text{if } t \geq 1/k \end{cases}$$

and show that $y^k(\cdot)$ converges uniformly on compact intervals to the impulse response.

3. Consider the input-output relation of a siso convolution system of the form

$$y(t) = \int_0^t \mathcal{G}(t - \tau)u(\tau)d\tau, \quad t \geq 0,$$

where $\mathcal{G} : \mathbb{R}_+ \rightarrow \mathbb{R}$ is continuous. Let $\bar{y}(t)$, $t \geq 0$ be the *step response* (corresponding to $u(t) \equiv 1$, $t \geq 0$). Show that $\bar{y}(\cdot)$ is differentiable on $[0, \infty)$ and its derivative is the impulse response $\mathcal{G}(\cdot)$. Formulate and prove an analogous result for discrete time systems.

4. Prove the discrete time counterpart of Proposition 2.3.10 as stated in Remark 2.3.12.

5. Consider the scalar system

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \dots + a_0y(t) = b_{n-1}u^{(n-1)}(t) + \dots + b_0u(t), \quad y^{(i)} = \frac{d^i y}{dt^i}. \quad (68)$$

Find an appropriate state space model (A, B, C) for this system, see Ex. 2.1.8 and Ex. 2.1.9. Show that the impulse response $\mathcal{G}(t) = Ce^{At}B$ is a quasi-polynomial of the form $\mathcal{G}(t) = \sum_{i=1}^{\ell} p_i(t)e^{\lambda_i t}$ where the p_i are polynomials and $\lambda_1, \dots, \lambda_{\ell}$ are the distinct roots of the characteristic polynomial $p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_0$. Find an explicit expression for the transfer function and verify your answer by applying the Laplace transform to (68) with zero initial conditions.

6. Construct a dyadic decomposition of the system (A, B, C) where

$$A = \begin{bmatrix} 1 & 4 \\ 1 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad C = I_2.$$

7. Calculate the impulse response $\mathcal{G}(t)$ and the transfer function $g(s)$ for the continuous time system (A, B, C, D) where

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [1, 0], \quad D = 1.$$

Use MATLAB to plot the gain and frequency responses. Determine $\|g\|_{H^\infty(\mathbb{C}_+; \mathbb{C})}$.

8. Prove Proposition 2.3.22 for the discrete time case.

9. Find the convolution kernels \mathcal{G}_i corresponding to the transfer functions g_1, \dots, g_5 given in Example 2.3.25. In the case of g_5 you need only compute $\mathcal{G}_5(t)$ for $t \in [0, 3]$.

10. If, as in Example 2.3.25, $g_1(s) = 1/(s^2 + 2s + 5)$ and $x(\omega) = \operatorname{Re} g_1(i\omega)$, $y(\omega) = \operatorname{Im} g_1(i\omega)$ find an algebraic equation in x, y for the polar plot of $g_1(\cdot)$. Carry out the same programme for $g_2(s) = 1/s(s^2 + 2s + 5)$ and show that $\operatorname{Re} g_2(i\omega) \rightarrow -0.08$ and $\operatorname{Im} g_2(i\omega) \rightarrow -\infty$ as $\omega \rightarrow 0$.

11. Prove that the transfer functions g_1, g_3, g_4 and g_5 given in Example 2.3.25 are in $H^\infty(\mathbb{C}_+; \mathbb{C})$ and calculate their H^∞ -norms. Find corresponding state space models for the scalar transfer functions g_1, g_2 and g_3 .

12. Consider the transfer function

$$G(s) = \frac{1}{s^2 + 2s + 2} \begin{bmatrix} s+1 & +1 \\ -1 & s+1 \end{bmatrix}.$$

Show that $\|G\|_{H^\infty(\mathbb{C}_+; \mathbb{C}^{2 \times 2})} = 1$. For any $\varepsilon \in (0, 1)$, construct a function $\tilde{u}(\cdot) \in L^2(\mathbb{R}; \mathbb{C}^2)$ with $\|\tilde{u}\|_{L^2(\mathbb{R}; \mathbb{C}^2)} = 1$ such that $\|M_G \tilde{u}\|_{L^2(\mathbb{R}; \mathbb{C}^2)}^2 \geq (1 - \varepsilon)^2$, see the proof of Theorem 2.3.28.

13. Prove Theorem 2.3.28 for the discrete time case.

14. Consider the electrical circuit of Ex. 2.2.9 with input the driving voltage u and output the charge q on the capacitor. Determine the impulse response of the system. Let $u(t) = \sin \omega_0 t$ where $\omega_0 > 0$ is given. Specify conditions in terms of R, L, C under which the system admits a periodic trajectory with period $\omega_0 > 0$ (substitute $q(t) = a \cos \omega_0 t + b \sin \omega_0 t$ in the differential equation). Show that if these conditions are satisfied and $R > 0$ every trajectory of the system with initial state $x^0 \neq 0$ (under the control $u(t) = \sin \omega_0 t$) will approach this periodic solution as $t \rightarrow \infty$.

2.3.5 Notes and References

The seminal monograph of *Desoer and Vidyasagar* (1975) [130] on input-output systems is still a standard reference. More recent textbooks which contain chapters on input-output systems are *Delchamps* (1988) [125], *Sontag* (1998) [472] and *Sastry* (1999) [448]. In [130] one can find details of a convolution algebra which allows for Dirac impulses in the convolution kernel, see also [116].

For background material on convolutions, \mathbf{z} and Laplace transforms and Fourier transforms see the books recommended in Section A.3. An excellent introductory textbook which covers much of the material of this section, written from an engineering point of view and aimed at undergraduates is *Kwakernaak and Sivan* (1991) [322].

The Hardy space H^q play a role in systems theory in the context of robust control and H^∞ theory, see *Zhou et al.* (1996) [546].

Frequency response methods spread rapidly in the 1930's after the appearance of *Nyquist's* classical paper on feedback amplifier stability (1932) [395] which arose from problems of long distance telephony. By the early 1950's frequency domain methods dominated the analysis and design of automatic control systems. Nyquist's method proceeds from a modification of the *polar plot* of the complex frequency response. Bode's method proceeds from the graphs of amplitude and phase response and the Nichols chart combines the two Bode plots into a plot of the gain in decibels against phase shift in degrees, see e.g. *Macfarlane* (1979) [356]. Other standard references are [322], [168]. A recent book on system identification via frequency domain methods is *Pintelon and Schoukens* (2001) [413].

2.4 Transformations and Interconnections of Linear State Space Systems

In this section we only consider continuous time systems of the form (2.17) which we denote by the shorthand notation $\Sigma = (A, B, C, D)$. All the definitions and results can also be applied to discrete time systems of the form (2.22). The first subsection is concerned with showing that the systems (A, B, C, D) form a category and we specify some standard constructions for this category (subsystems, quotient systems, direct sum, ...). In the second subsection we introduce the basic coupling schemes for two systems $\Sigma_i = (A_i, B_i, C_i, D_i)$ $i = 1, 2$ and discuss the general form of a composite linear time-invariant system. As usual we do not distinguish notationally between a linear map and the matrix representing it with respect to a given basis. In Subsection 2.4.1 a coordinate free interpretation of A, B, C, D as linear maps between vector spaces will prevail. However in applications these maps will be described by matrices with respect to given bases of the input, state and output spaces.

2.4.1 Morphisms and Standard Constructions

Changes of bases in the state space X , the input space U and/or the output space Y of a system (A, B, C, D) lead to transformations of the matrices A, B, C, D . Hence a given physical system may be modelled by different quadruples of the form (A, B, C, D) . This raises the question – which conditions render two systems to be isomorphic or similar? Two vector spaces (groups) are called isomorphic if there exists a linear isomorphism (resp. group isomorphism) between them. Using the terminology of category theory, isomorphisms are *invertible morphisms* and morphisms are *structure preserving* “maps” between structured objects of a given class. We do not intend to explore these generalities, but just to mention that the concepts of “isomorphism” and “(homo)morphism” are of fundamental importance in the construction of a mathematical theory.

Morphisms between dynamical systems can be defined in various ways. One possible definition would be:

If $\Sigma_i = (T, U_i, \mathcal{U}_i, X_i, Y_i, \varphi_i, \eta_i)$, $i = 1, 2$ are two dynamical systems of a given class \mathcal{S} then a morphism from Σ_1 to Σ_2 is a triple (ρ, τ, σ) consisting of maps $\rho : U_1 \rightarrow U_2$, $\tau : X_1 \rightarrow X_2$, $\sigma : Y_1 \rightarrow Y_2$ which have certain properties (such as smoothness, linearity etc.) depending on the specific class \mathcal{S} . Moreover the following three conditions must be satisfied for all t , $t_0 \in T$, $u \in U_1$, $u(\cdot) \in \mathcal{U}_1$, $x \in X_1$.

$$u(\cdot) \in \mathcal{U}_1 \Rightarrow \rho \circ u(\cdot) \in \mathcal{U}_2 \quad \text{and} \quad T_{t_0, x, u(\cdot)}^{\Sigma_1} \subset T_{t_0, \tau(x), \rho \circ u(\cdot)}^{\Sigma_2} \quad (1)$$

$$\tau(\varphi_1(t; t_0, x, u(\cdot))) = \varphi_2(t; t_0, \tau(x), \rho \circ u(\cdot)), \quad t \in T_{t_0, x, u(\cdot)}^{\Sigma_1}, \quad u(\cdot) \in \mathcal{U}_1 \quad (2)$$

$$\sigma(\eta_1(t, x, u)) = \eta_2(t, \tau(x), \rho(u)). \quad (3)$$

Other definitions may allow for transformations of the time and for more general mappings between \mathcal{U}_1 and \mathcal{U}_2 . We do not go into further details at this general level, but now give a precise definition for time-invariant linear finite dimensional systems.

Definition 2.4.1. (Linear system morphism). Consider two finite dimensional linear systems $\Sigma_i = (A_i, B_i, C_i, D_i)$ with input space U_i , state space X_i , output space Y_i ($i = 1, 2$). A triple (R, T, S) of linear maps $R : U_1 \rightarrow U_2$, $T : X_1 \rightarrow X_2$, $S : Y_1 \rightarrow Y_2$ is called a *linear system morphism* from Σ_1 to Σ_2 (with the notation $(R, S, T) \in \text{Mor}(\Sigma_1, \Sigma_2)$ or $(R, S, T) : \Sigma_1 \mapsto \Sigma_2$) if

$$A_2T = TA_1, \quad B_2R = TB_1, \quad C_2T = SC_1, \quad D_2R = SD_1, \quad (4)$$

i.e. the following diagrams commute

$$\begin{array}{ccccccc} U_1 & \xrightarrow{B_1} & X_1 & \xrightarrow{A_1} & X_1 & \xrightarrow{C_1} & Y_1 \\ R \downarrow & & \downarrow T & & \downarrow T & & \downarrow S \\ U_2 & \xrightarrow{B_2} & X_2 & \xrightarrow{A_2} & X_2 & \xrightarrow{C_2} & Y_2 \end{array} \quad \begin{array}{ccc} U_1 & \xrightarrow{D_1} & Y_1 \\ R \downarrow & & \downarrow S \\ U_2 & \xrightarrow{D_2} & Y_2 \end{array}$$

If we represent the system (A, B, C, D) by the linear map

$$\Sigma : X \times U \rightarrow X \times Y, \quad \begin{bmatrix} x \\ u \end{bmatrix} \mapsto \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}, \quad (5)$$

then conditions (4) can be expressed equivalently by one compound equation

$$\begin{bmatrix} T & 0 \\ 0 & S \end{bmatrix} \Sigma_1 = \Sigma_2 \begin{bmatrix} T & 0 \\ 0 & R \end{bmatrix}. \quad (6)$$

In other words the matrices $T \oplus S = \text{diag}(T, S)$ and $T \oplus R = \text{diag}(T, R)$ intertwine the linear maps Σ_1 and Σ_2 . Intertwining operators play an important role in system theory.

Remark 2.4.2. It is a simple matter to verify that the class of all time-invariant finite dimensional linear systems together with the morphisms defined above form a category if the composition of two morphisms is defined in the obvious way

$$(R_2, T_2, S_2) \circ (R_1, T_1, S_1) = (R_2R_1, T_2T_1, S_2S_1).$$

□

A morphism $(R, T, S) \in \text{Mor}(\Sigma_1, \Sigma_2)$ is called a *(linear system) isomorphism* if it admits a left and right inverse in the sense of the above composition, and this is the case if and only if $R : U_1 \rightarrow U_2$, $T : X_1 \rightarrow X_2$, $S : Y_1 \rightarrow Y_2$ are vector space isomorphisms.

In terms of matrix representations a linear system isomorphism describes changes of bases in the input, state and output spaces. In fact, consider system equations of the form (2.17)

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad t \in \mathbb{R} \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

where $(A, B, C, D) \in \mathbf{L}_{n,m,p}(\mathbb{K}) := \mathbb{K}^{n \times n} \times \mathbb{K}^{n \times m} \times \mathbb{K}^{p \times n} \times \mathbb{K}^{p \times m}$, $n, m, p \geq 1$ and suppose we introduce new bases (v^1, \dots, v^m) in $U = \mathbb{K}^m$, (z^1, \dots, z^n) in $X = \mathbb{K}^n$, and (w^1, \dots, w^p) in $Y = \mathbb{K}^p$. The coordinate vectors of $u \in \mathbb{K}^m$, $x \in \mathbb{K}^n$, and $y \in \mathbb{K}^p$ with respect to the new bases are given by

$$\hat{u} = R^{-1}u, \quad \hat{x} = T^{-1}x, \quad \hat{y} = S^{-1}y \quad (7)$$

where

$$R = [v^1, \dots, v^m] \in \mathbf{GL}_m(\mathbb{K}), \quad T = [z^1, \dots, z^n] \in \mathbf{GL}_n(\mathbb{K}), \quad S = [w^1, \dots, w^p] \in \mathbf{GL}_p(\mathbb{K}).$$

In terms of the new coordinate vectors the system equations read

$$\begin{aligned} \dot{\hat{x}}(t) &= \hat{A}\hat{x}(t) + \hat{B}\hat{u}(t), \quad t \in \mathbb{R} \\ \hat{y}(t) &= \hat{C}\hat{x}(t) + \hat{D}\hat{u}(t) \end{aligned} \quad (8)$$

where

$$\hat{A} = TAT^{-1}, \quad \hat{B} = TBR^{-1}, \quad \hat{C} = SCT^{-1}, \quad \hat{D} = SDR^{-1}. \quad (9)$$

In applications the external variables often represent physical quantities such as current, velocity, temperature. Linear transformations of the input and output vectors would destroy this physical interpretation and so usually one does not consider coordinate transformations in the input and output spaces. If only linear coordinate transformations $\hat{x} = T^{-1}x$ in the state space are allowed, the system equations are transformed into

$$\begin{aligned} \dot{\hat{x}}(t) &= TAT^{-1}\hat{x}(t) + TBu(t) \\ y(t) &= CT^{-1}\hat{x}(t) + Du(t). \end{aligned} \quad (10)$$

This leads us to the more restrictive class of *similarity transformations*

$$T \cdot (A, B, C, D) = (TAT^{-1}, TB, CT^{-1}, D), \quad T \in \mathbf{GL}_n(\mathbb{K}). \quad (11)$$

Definition 2.4.3. (Isomorphism, Similarity). Two finite dimensional linear systems $\Sigma_i = (A_i, B_i, C_i, D_i)$, $i = 1, 2$ are said to be

- (i) *isomorphic* if there exists a linear system isomorphism $(R, S, T) : \Sigma_1 \mapsto \Sigma_2$,
- (ii) *similar* if $U_1 = U_2$, $Y_1 = Y_2$, and there exists a linear isomorphism $T : X_1 \rightarrow X_2$ satisfying

$$A_2 = TA_1T^{-1}, \quad B_2 = TB_1, \quad C_2 = C_1T^{-1}, \quad D_1 = D_2. \quad (12)$$

The input-output operator of a linear system will in general change under arbitrary linear system isomorphisms, but not under similarity transformations.

Proposition 2.4.4. *The input-output operator and the transfer matrix of a linear system (2.17) are invariant under similarity transformations.*

Proof: It suffices to show the invariance of the transfer matrix. But this follows immediately from (12) since

$$C_2(sI - A_2)^{-1}B_2 + D_2 = C_1T^{-1}(sI - TA_1T^{-1})^{-1}TB_1 + D_1 = C_1(sI - A_1)^{-1}B_1 + D_1.$$

□

We now introduce the concepts of *subsystem* and *quotient system*.

Definition 2.4.5. (Subsystem). $\Sigma_1 = (A_1, B_1, C_1, D_1)$ is called a *subsystem* of $\Sigma_2 = (A_2, B_2, C_2, D_2)$ if there exist linear injections $R : U_1 \rightarrow U_2$, $T : X_1 \rightarrow X_2$, $S : Y_1 \rightarrow Y_2$ such that (4) holds. In this case (R, T, S) is called a *system embedding*.

Suppose a system (A, B, C, D) is given. If $U_1 \subset U$, $X_1 \subset X$, $Y_1 \subset Y$ are linear subspaces such that

$$BU_1 \subset X_1, \quad AX_1 \subset X_1, \quad CX_1 \subset Y_1, \quad DU_1 \subset Y_1 \quad (13)$$

then the system (A_1, B_1, C_1, D_1) obtained by restricting A, B, C, D to X_1, U_1, X_1, U_1 respectively is a subsystem of (A, B, C, D) . The embedding is given by (R, T, S) where $R: U_1 \rightarrow U$, $T: X_1 \rightarrow X$, $S: Y_1 \rightarrow Y$ are the canonical injections.

Now let $\hat{R}: U \rightarrow U/U_1$, $\hat{T}: X \rightarrow X/X_1$, $\hat{S}: Y \rightarrow Y/Y_1$ be the natural projections. Then there exist linear maps $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ which make the following diagrams commute, and the maps are uniquely determined by this property.

$$\begin{array}{ccccccc} U & \xrightarrow{B} & X & \xrightarrow{A} & X & \xrightarrow{C} & Y \\ \hat{R} \downarrow & & \downarrow \hat{T} & & \downarrow \hat{T} & & \downarrow \hat{S} \\ U/U_1 & \xrightarrow{\hat{B}} & X/X_1 & \xrightarrow{\hat{A}} & X/X_1 & \xrightarrow{\hat{C}} & Y/Y_1 \end{array} \quad (14)$$

Definition 2.4.6. (Quotient system). Suppose (A, B, C, D) is a linear system with input space U , state space X , output space Y and $U_1 \subset U$, $X_1 \subset X$, $Y_1 \subset Y$ are linear subspaces such that (13) holds. If Σ_1 denotes the corresponding subsystem of Σ then the linear system $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ with input space $\hat{U} = U/U_1$, state space $\hat{X} = X/X_1$ and output space $\hat{Y} = Y/Y_1$ defined by (14) is called the *quotient system of Σ by Σ_1* and is denoted by Σ/Σ_1 .

If

$$\hat{R}: U \rightarrow \hat{U} = U/U_1, \quad \hat{T}: X \rightarrow \hat{X} = X/X_1, \quad \hat{S}: Y \rightarrow \hat{Y} = Y/Y_1$$

are the canonical projections then the quotient system $\hat{\Sigma} = \Sigma/\Sigma_1$ is uniquely determined by the property that $(\hat{R}, \hat{T}, \hat{S})$ is a linear system morphism from Σ to $\hat{\Sigma}$. This morphism is called the *canonical system projection* from Σ to $\hat{\Sigma}$.

Now suppose that U_2, X_2, Y_2 are algebraic complements of U_1, X_1, Y_1 in U, X, Y respectively, then A, B, C, D have the following representations with respect to the decompositions $U = U_1 \oplus U_2$, $X = X_1 \oplus X_2$, $Y = Y_1 \oplus Y_2$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} \\ 0 & D_{22} \end{bmatrix}. \quad (15)$$

It is straightforward to verify that the isomorphisms

$$U_2 \cong U/U_1, \quad X_2 \cong X/X_1, \quad Y_2 \cong Y/Y_1$$

induced by the restriction of R, T, S to U_2, X_2, Y_2 respectively define a system isomorphism between $\Sigma_2 = (A_{22}, B_{22}, C_{22}, D_{22})$ and $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$.

Example 2.4.7. Consider the mass-spring-damper system Σ shown in Figure 2.4.1. The masses m_1, m_2 slide on a horizontal surface without friction. The stiffness coefficients of the springs are k_1, k_2 and the damping coefficient is c . The outputs are the displacements $y_1(t), y_2(t)$ of the two masses from some given equilibrium positions and the inputs $u_1(t)$,

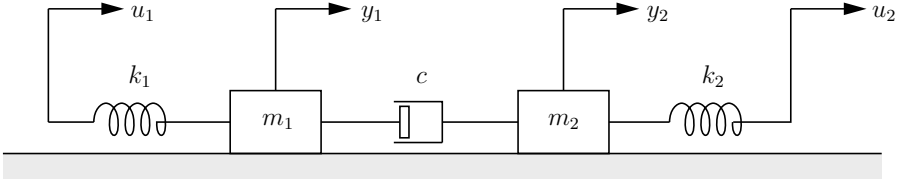


Figure 2.4.1: Mass-spring-damper system

$u_2(t)$ are the displacements of the outer ends of the springs from their corresponding rest positions. The equations of motion of this mechanical system are given by

$$\begin{aligned} m_1 \ddot{y}_1 &= k_1(u_1 - y_1) + c(\dot{y}_2 - \dot{y}_1) \\ m_2 \ddot{y}_2 &= k_2(u_2 - y_2) - c(\dot{y}_2 - \dot{y}_1). \end{aligned} \quad (16)$$

If we define the state variables to be $x_1 = y_1$, $x_2 = \dot{y}_1$, $x_3 = y_2$, $x_4 = \dot{y}_2$ we obtain the time-invariant linear system (A, B, C, D) where $C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, $D = 0$ and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -k_1/m_1 & -c/m_1 & 0 & c/m_1 \\ 0 & 0 & 0 & 1 \\ 0 & c/m_2 & -k_2/m_2 & -c/m_2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ k_1/m_1 & 0 \\ 0 & 0 \\ 0 & k_2/m_2 \end{bmatrix}.$$

Now suppose $k_1/m_1 = k_2/m_2$ and define

$$X_1 = \{x \in \mathbb{R}^4; x_1 = x_3, x_2 = x_4\}, \quad U_1 = \{u \in \mathbb{R}^2; u_1 = u_2\}, \quad Y_1 = \{y \in \mathbb{R}^2; y_1 = y_2\}$$

then it is easy to verify the conditions (13) are satisfied. This means that if the initial state lies in X_1 and the same control is applied to both spring ends then the distance between the two masses remains constant ($y_1(t) = y_2(t)$), so there is no actual interaction between the masses via the damper. The subsystem $\Sigma_1 = (A_1, B_1, C_1, 0)$ obtained by the restriction of A , C , and B to X_1 and U_1 , respectively, describes simultaneously the motions of two “decoupled” mass spring systems. To obtain a matrix representation of A_1 , B_1 , C_1 we choose the basis vectors $z^1 = [1, 0, 1, 0]^\top$, $z^2 = [0, 1, 0, 1]^\top$, $z^3 = [1, 0, 0, 0]^\top$, $z^4 = [0, 1, 0, 0]^\top$ in X , $v^1 = [1, 1]^\top$, $v^2 = [1, 0]^\top$ in U and $w^1 = [1, 1]^\top$, $w^2 = [1, 0]^\top$ in Y . With respect to these bases A , B , C have the following matrix representations (see (9))

$$\begin{aligned} A &\sim \begin{bmatrix} 0 & 1 & 0 & 0 \\ -k_1/m_1 & 0 & 0 & c/m_2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -k_1/m_1 & -(c/m_1 + c/m_2) \end{bmatrix}, \quad B \sim \begin{bmatrix} 0 & 0 \\ k_1/m_1 & 0 \\ 0 & 0 \\ 0 & k_1/m_1 \end{bmatrix}, \\ C &\sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \end{aligned} \quad (17)$$

X_1 is spanned by z^1, z^2 , U_1 by v^1 and Y_1 by w^1 . Hence

$$A_1 = \begin{bmatrix} 0 & 1 \\ -k_1/m_1 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ k_1/m_1 \end{bmatrix}, \quad C_1 = [1, 0].$$

Equivalently the subsystem can be described by the following second order differential equation

$$\ddot{y}^1(t) + (k_1/m_1)y^1(t) = (k_1/m_1)u^1(t).$$

Since $k_1/m_1 = k_2/m_2$, the above equation is equivalent to

$$(m_1 + m_2)\ddot{y}^1(t) + (k_1 + k_2)(y^1(t) - u^1(t)) = 0$$

and hence describes the motion of the “aggregated” mass-spring system shown in Figure 2.4.2.

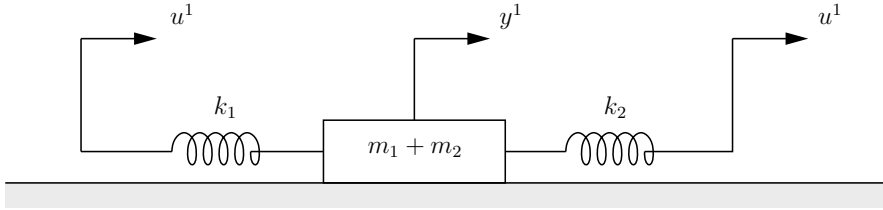


Figure 2.4.2: Aggregated mass-spring system

Now let us consider the quotient system $\hat{\Sigma} := \Sigma/\Sigma_1$. We choose as bases for the quotient spaces \hat{X} , \hat{U} , \hat{Y} the vectors (equivalence classes) $z^3 + X_1$, $z^4 + X_1$ and $v^2 + U_1$, $w^2 + Y_1$ respectively. Then by (17) the matrix representation of the quotient system $\hat{\Sigma} = \Sigma/\Sigma_1$ is given by

$$\hat{A} = \begin{bmatrix} 0 & 1 \\ -k_1/m_1 & -(c/m_1 + c/m_2) \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 0 \\ k_1/m_1 \end{bmatrix}, \quad \hat{C} = [1, 0].$$

This yields a second order differential equation

$$m_1 \frac{d^2 \hat{y}}{dt^2} + c \frac{m_1 + m_2}{m_2} \frac{d\hat{y}}{dt} + k_1 \hat{y} = k_1 \hat{u}. \quad (18)$$

It can be interpreted as follows: Suppose that for given initial conditions $y_1(0) = \hat{y}(0)$, $\dot{y}_1(0) = \dot{\hat{y}}(0)$ and a control $u_1(\cdot) = \hat{u}(\cdot)$, the initial values $y_2(0)$, $\dot{y}_2(0)$ for the second mass and control $u_2(\cdot)$ are chosen in such a way that the centre of mass remains at equilibrium. Then $m_1 y_1(t) + m_2 y_2(t) \equiv 0$ and substituting for $y_2(t)$ in the first equation of (16) we see that $y_1(\cdot)$ satisfies (18). Hence (18) describes the equation of motion of the first mass under the assumption that the centre of mass of Σ remains at rest. $\hat{\Sigma}$ admits an analogous interpretation with respect to the second mass. \square

Note that if a subsystem Σ_1 of Σ and the corresponding quotient system $\hat{\Sigma} = \Sigma/\Sigma_1$ are known, it is not, in general possible to reconstruct the complete system Σ from them. Whilst A_{11} and A_{22} can be reconstructed from Σ_1 and $\hat{\Sigma}$ respectively, this is not the case for A_{12} . Hence Σ is not, in general, the direct sum of Σ_1 and $\hat{\Sigma}$ in the sense of the following definition.

Definition 2.4.8. (Direct sum). Let $\Sigma_i = (A_i, B_i, C_i, D_i)$ be systems with state space X_i , input space U_i and output space Y_i , $i \in \underline{N}$. The direct sum is a system (A, B, C, D) with state space X , input space U and output space Y given by

$$\begin{aligned} X &= \prod_{i=1}^N X_i, & U &= \prod_{i=1}^N U_i, & Y &= \prod_{i=1}^N Y_i, \\ A &= \bigoplus_{i=1}^N A_i, & B &= \bigoplus_{i=1}^N B_i, & C &= \bigoplus_{i=1}^N C_i, & D &= \bigoplus_{i=1}^N D_i. \end{aligned} \quad (19)$$

2.4.2 Composite Systems

The direct sum is a trivial way of building a composite system from a collection of systems. In fact it is just a collection of uncoupled systems. Hence in a direct sum each subsystem can be studied independently of the other subsystems. This is not the case if the subsystems $\Sigma_i, i \in \underline{N}$ are *interconnected* within the composite system $\bar{\Sigma}$. In many areas of application one encounters large scale systems which are made up of complex arrays of many interconnected subsystems. The purpose of this subsection is to provide a general framework for describing such systems where $\Sigma_i = (A_i, B_i, C_i, D_i), i \in \underline{N}$. In order to do this we introduce various interconnection schemes and also the general form of a composite time-invariant linear system.

The following four examples illustrate the most important ways that a system can be interconnected with other systems or with itself (feedback). We denote by $\bar{U}, \bar{X}, \bar{Y}$ the input, state and output spaces of the composite system. When \bar{X} is not defined explicitly it is understood that $\bar{X} = X_1 \times X_2$.

Example 2.4.9. (Series connection). A *series connection* of Σ_1 and Σ_2 is obtained when the input of Σ_2 is connected to the output of Σ_1 . Thus it requires $U_2 = Y_1$. The input and output of the composite system is u^1 and y^2 respectively and $\bar{U} = U_1, \bar{Y} = Y_2$. The composite system $\bar{\Sigma} = (\bar{A}, \bar{B}, \bar{C}, \bar{D})$ is given by

$$\bar{A} = \begin{bmatrix} A_1 & 0 \\ B_2 C_1 & A_2 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad \bar{C} = [0, C_2], \quad \bar{D} = D_2 D_1. \quad (20)$$

The transfer matrix of the series connection is simply the product of the individual transfer matrices

$$\bar{G}(s) = \bar{C}(sI - \bar{A})^{-1} \bar{B} + \bar{D} = G_2(s)G_1(s) \quad (21)$$

where $G_i(s), i = 1, 2$ are the transfer function matrices of the subsystems connected in series. Thus multiplication of transfer matrices corresponds to series connection of the respective systems.

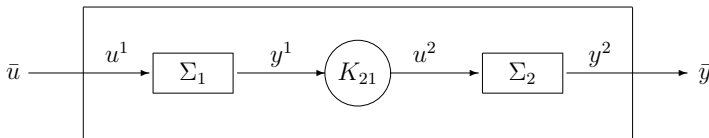


Figure 2.4.3: Series connection

If a direct coupling of the two subsystems in series is not possible because $Y_1 \neq U_2$ an *adapter* or *coupling matrix* $K_{21} : Y_1 \rightarrow U_2$ can be used to connect y^1 with u^2 , viz. $u^2 = K_{21}y^1$, see Figure 2.4.3. The corresponding state space equation and transfer matrix are obtained by replacing C_1 by $K_{21}C_1$ in (20) and $G_1(s)$ by $K_{21}G_1(s)$ in (21). \square

Example 2.4.10. (Parallel connection). A *parallel connection* of Σ_1, Σ_2 is obtained if both systems have the same input, and the output of the composite system is the sum of the individual outputs (see the left hand figure in Figure 2.4.4).

In this case $\bar{U} = U_1 = U_2, \bar{Y} = Y_1 = Y_2, u^1 = u^2 = \bar{u}$ and $\bar{\Sigma}$ is described by

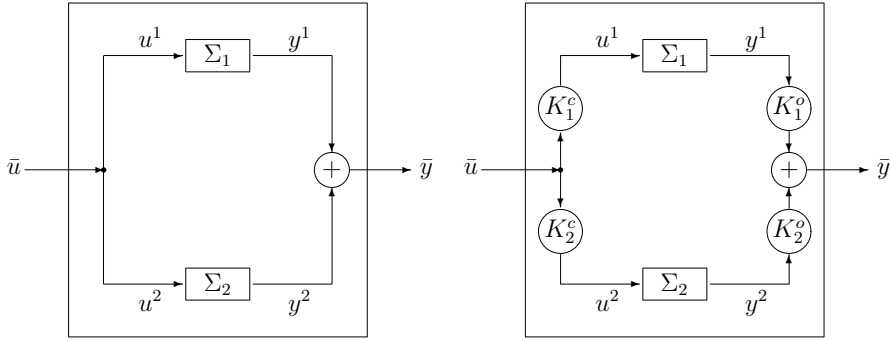


Figure 2.4.4: Parallel connection and extended parallel connection

$$\bar{A} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \bar{C} = [C_1, C_2], \quad \bar{D} = D_1 + D_2. \quad (22)$$

The transfer matrix of Σ is given by $\bar{G}(s) = G_1(s) + G_2(s)$. Thus the addition of transfer matrices corresponds to the parallel connection of the respective systems. If Σ_1, Σ_2 do not have the same input and output spaces an extended parallel connection can be obtained by setting

$$u^1 = K_1^c \bar{u}, \quad u^2 = K_2^c \bar{u}, \quad \bar{y} = K_1^o y^1 + K_2^o y^2$$

where $K_i^c : \bar{U} \rightarrow U_i$ and $K_i^o : Y_i \rightarrow \bar{Y}$, $i = 1, 2$, are called *input and output coupling matrices* (see the right hand figure in Figure 2.4.4). The corresponding system equation and transfer matrix are easily determined. \square

The third basic way of coupling two systems is via feedback. This configuration is of fundamental importance in control where a central problem is that of producing a desired input-output behaviour of a given system by feedback.

Example 2.4.11. (Dynamic output feedback). The feedback interconnection of

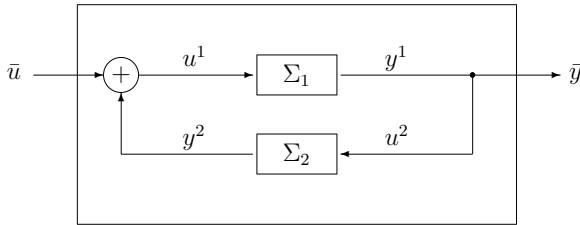


Figure 2.4.5: Dynamic output feedback

two systems Σ_1, Σ_2 connects the output of Σ_1 to the input of Σ_2 and the output of Σ_2 to the input of Σ_1 according to the formulas $u^2 = y^1$ and $u^1 = y^2 + \bar{u}$ (see Figure 2.4.5). \bar{u} is considered as the input of the feedback system $\bar{\Sigma}$ and $\bar{y} = y^1$ as the output. Clearly, this interconnection presupposes that $Y_1 = U_2$ and $Y_2 = U_1$ (otherwise coupling matrices K_{ij} , $i, j = 1, 2$, $i \neq j$ must be used). But in contrast with the previous interconnections

transformation $u = Fx + \bar{u}$ (see Figure 2.4.6). In this case $\bar{U} = U$, $\bar{Y} = Y$, $\bar{X} = X$ and the feedback system is described by

$$\bar{A} = A + BF, \quad \bar{B} = B, \quad \bar{C} = C + DF, \quad \bar{D} = D. \quad (25a)$$

In the special case when $F = KC$, $D = 0$ we obtain *static output feedback* where the *output* of the system Σ is connected with the input of the same system via $u = Ky + \bar{u}$ (see Figure 2.4.6), so

$$\bar{A} = A + BKC, \quad \bar{B} = B, \quad \bar{C} = C. \quad (25b)$$

The respective transfer matrices of these feedback configurations are

$$\begin{aligned} \bar{G}(s) &= [C + DF](sI - A - BF)^{-1}B + D, \\ \bar{G}(s) &= C(sI - A - BKC)^{-1}B = G(s)(I - KG(s))^{-1}, \end{aligned} \quad (26)$$

where $G(s)$ is the transfer matrix of Σ . □

Let us now proceed to describe the general form of a composite system obtained by connecting finitely many subsystems $\Sigma_i = (A_i, B_i, C_i, D_i)$, $i \in \underline{N}$ via constant coupling matrices. All of the above examples (for $N = 2$) are special cases of the connection scheme shown in Figure 2.4.7.

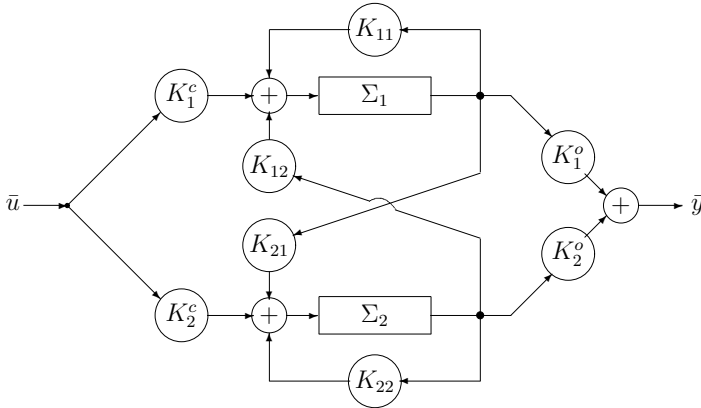


Figure 2.4.7: General composite system of two subsystems

For arbitrary $N \geq 1$ a general linear, time-invariant connection scheme for the systems $\Sigma_1, \Sigma_2, \dots, \Sigma_N$ is given by the $(N + 1)^2$ matrices

$$K_i^c : \bar{U} \rightarrow U_i, \quad K_{ij} : Y_j \rightarrow U_i, \quad K_i^o : Y_i \rightarrow \bar{Y}, \quad D_o : \bar{U} \rightarrow \bar{Y}$$

($i, j \in \underline{N}$) where $\bar{U} = \prod_{i=1}^N U_i$ is the input space, $\bar{Y} = \prod_{i=1}^N Y_i$ the output space and $\bar{X} = \prod_{i=1}^N X_i$ is the state space of the composite system.

The compound matrix $K = (K_{ij})_{i,j \in \underline{N}}$ is called the *matrix of interconnections* between the subsystems. If $K_{ij} \neq 0$ the output of Σ_j exercises an influence on the input of Σ_i . The matrices K_{ii} describe static feedback loops for the subsystems Σ_i . Since, in general, not all subsystems are connected with every other subsystem,

many of the matrices (K_{ij}) $i, j \in \underline{N}$ will be zero matrices.

The matrices

$$K^c = \begin{bmatrix} K_1^c \\ \vdots \\ K_N^c \end{bmatrix}, \quad K^o = [K_1^o \dots K_N^o]$$

are called matrices of *input couplings* and *output couplings* respectively. K^o contains the matrix coefficients which specify those linear combinations of the subsystem's outputs which yield the output of the composite system. D^o describes the direct input-output coupling of $\bar{\Sigma}$,

$$\bar{y} = \sum_{i=1}^N K_i^o y^i + D^o \bar{u} = \sum_{i=1}^N K_i^o C_i x^i + \sum_{i=1}^N K_i^o D_i u^i + D_o \bar{u}. \quad (27)$$

The inputs u^i of Σ_i are obtained by adding the terms $K_{ij} y^j$ from each of the subsystems plus the terms $K_i^c \bar{u}$ from the external control \bar{u} ,

$$u^i = \sum_{j=1}^N K_{ij} y^j + K_i^c \bar{u} = \sum_{j=1}^N K_{ij} C_j x^j + \sum_{j=1}^N K_{ij} D_j u^j + K_i^c \bar{u}. \quad (28)$$

In the general case when input-output couplings are present (28) is an implicit formula for the u^i 's $i \in \underline{N}$ in terms of x^1, \dots, x^N and \bar{u} . So not every connection scheme is feasible and there will only exist unique solutions if the following *well-posedness condition* is satisfied by K

$$\det(I - K \operatorname{diag}(D_1, \dots, D_N)) \neq 0. \quad (29)$$

Let (A, B, C, D) be the direct sum of the systems $\Sigma_1, \dots, \Sigma_N$ as described by equation (19), then if (29) holds the unique solution of (28) can be expressed in the form

$$u = (I - KD)^{-1}(KCx + K^c \bar{u}) \quad (30)$$

where $u \in \bar{U}$, $x \in \bar{X}$ are the vectors with components u^i and x^i . Substituting (30) in the system equations of Σ_i and in (27) we see that the composite system is described by

$$\begin{aligned} \bar{A} &= A + B(I - KD)^{-1}KC, & \bar{B} &= BK^c \\ \bar{C} &= K^o C + K^o D(I - KD)^{-1}KC, & \bar{D} &= K^o D(I - KD)^{-1}K^c + D^o. \end{aligned} \quad (31)$$

If there are no direct input-output couplings in the subsystems, condition (29) is trivially satisfied and (31) simplifies to

$$\begin{aligned} \bar{A} &= \begin{bmatrix} A_1 + B_1 K_{11} C_1 & B_1 K_{12} C_2 & \dots & B_1 K_{1N} C_N \\ B_2 K_{21} C_1 & A_2 + B_2 K_{22} C_2 & \dots & B_2 K_{2N} C_N \\ \vdots & & & \vdots \\ B_N K_{N1} C_1 & \dots & & A_N + B_N K_{NN} C_N \end{bmatrix} \\ \bar{B} &= BK^c, \quad \bar{C} = K^o C, \quad \bar{D} = K^o DK^c + D^o. \end{aligned} \quad (32)$$

The interconnection structure of a composite system $\bar{\Sigma}$ can be represented by a directed graph with $(N+2)$ nodes denoted by \bar{U} , \bar{Y} and Σ_i ($i \in \underline{N}$), directed edges $\Sigma_j \rightarrow \Sigma_i$, $\bar{U} \rightarrow \Sigma_i$, $\Sigma_i \rightarrow \bar{Y}$ for those $i, j \in \underline{N}$ with $K_{ij} \neq 0$, $K_i^c \neq 0$, $K_i^o \neq 0$.

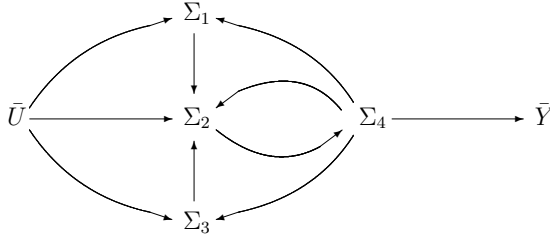


Figure 2.4.8: Directed graph

Example 2.4.13. If $N = 4$ and the interconnection structure is given by the matrices

$$K = \begin{bmatrix} 0 & 0 & 0 & K_{14} \\ K_{21} & 0 & K_{23} & K_{24} \\ 0 & 0 & 0 & K_{34} \\ 0 & K_{42} & 0 & 0 \end{bmatrix}, \quad K^c = \begin{bmatrix} K_1^c \\ K_2^c \\ K_3^c \\ 0 \end{bmatrix}, \quad K^o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ K_4^o \end{bmatrix}$$

the representation as a directed graph is shown in Figure 2.4.8. \square

A more detailed picture of the input and output couplings is obtained if the vertices $\bar{u}_1^i, \dots, \bar{u}_m^i, \bar{y}_1^i, \dots, \bar{y}_p^i$ are introduced. An edge $\bar{u}_j^i \rightarrow \Sigma_i$ or $\Sigma_i \rightarrow \bar{y}_k^i$ is drawn whenever the j^{th} column of K_i^c or the k^{th} row K_i^o is non-zero.

In particular any system $\Sigma = (A, B, C, D)$ with input space $\bar{U} = \mathbb{K}^m$, state space $\bar{X} = \mathbb{K}^n$ and output space $\bar{Y} = \mathbb{K}^p$ can always be represented as an interconnection of *integrators* (resp. *unit delays* in the discrete time case)

$$\Sigma_i : \dot{x}_i = u_i, \quad y_i = x_i \quad (i = 1, \dots, n). \quad (33)$$

Indeed if we define

$$K^c = B, \quad K = A, \quad K^o = C, \quad D^o = D \quad (34)$$

it is easy to verify that the resultant composite system is identical with Σ . The graph (with separate representation of each input and output channel) associated with the interconnection scheme (34) is called the *system graph* of Σ .

Example 2.4.14. The matrices describing the overhead crane of Example 1.3.4 have the following structure

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{32} & 0 & 0 \\ 0 & a_{42} & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where a_{32} , a_{42} , b_3 , b_4 are determined by physical parameters. The corresponding system graph is shown in Figure 2.4.9. \square

Note that the system graph will in general be altered by a similarity transformation. Hence system graphs are only meaningful for the analysis of a system if the quantities and subsystems represented by the vertices correspond to real physical parts of the system, and if their interconnection is of importance in the overall analysis of the system. When this is the case, only system isomorphisms which do not destroy the structure of the graph can be allowed.

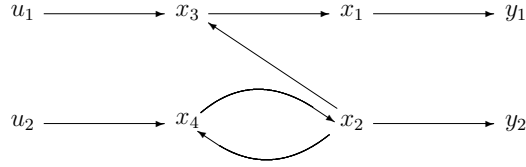


Figure 2.4.9: System graph for the overhead crane

2.4.3 Exercises

1. Consider the *RLC*-circuit described in Ex. 2.9
 - (i) Derive state space models of this circuit with state variables
 - (a) x_1 = current i through inductor, x_2 = charge q of capacitor
 - (b) $x_1 = i$, x_2 = voltage v across inductor
 - (c) x_1 = voltage across capacitor, $x_2 = i$.
 - (ii) Show that the linear systems (A, B, C) obtained in (i) are all similar.
 - (iii) Let $R = 3$, $L = 1$, $C = 0.5$. Define a state vector $x = [x_1, x_2]^\top$ for which the corresponding system matrix A is diagonal.
 - (iv) For which triples (R, L, C) are the resulting dynamical systems of (i) similar to the one defined in (iii).

2. Let

$$A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [0, 1].$$

Examine whether or not the triple (A, B, C) is similar or isomorphic to $(\hat{A}, \hat{B}, \hat{C})$ where

- (i) $\hat{A} = \begin{bmatrix} 1 & -2 \\ 0 & -1 \end{bmatrix}$, $\hat{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\hat{C} = [1, 1]$.
- (ii) $\hat{A} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$, $\hat{B} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\hat{C} = [-1, 0]$.
- (iii) $\hat{A} = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}$, $\hat{B} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\hat{C} = [-1, 0]$.
- (iv) $\hat{A} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}$, $\hat{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\hat{C} = [2, 4]$.

3. Let (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ be similar.

- (i) Show by means of an example that $A = \hat{A}$, $B = \hat{B}$ does not imply $C = \hat{C}$.
- (ii) Specify conditions for A , B under which $A = \hat{A}$, $B = \hat{B}$ does imply $C = \hat{C}$.

4. Show that the system $\Sigma_1 = (A_1, B_1, C_1)$ where

$$A_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & -2 \\ 0 & -3 \end{bmatrix}, \quad C_1 = [0, 1]$$

is a subsystem of $\Sigma = (A, B, C)$ where

$$A = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & 4 \\ 1 & -1 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & -2 \\ -1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix}.$$

Determine the quotient system Σ/Σ_1 .

5. Let $\Sigma_i = (A_i, B_i, C_i, D_i)$ be two time-invariant linear systems. Prove

- (i) if Σ_1 is a subsystem of Σ_2 , then $\sigma(A_1) \subset \sigma(A_2)$,
- (ii) if Σ_2 is a quotient system of Σ_1 , then $\sigma(A_2) \subset \sigma(A_1)$.

6. Let (R, T, S) be a morphism from Σ to $\hat{\Sigma}$. Specify conditions under which, given Σ , the linear maps R, T, S uniquely determine the system $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$.

7. Extend the set $\text{Mor}(\Sigma, \hat{\Sigma})$ by allowing in addition state feedback transformations $F : X \rightarrow \hat{U}$. Find a counterpart of (6) for these feedback morphisms $(R, T, S, F) : \Sigma \rightarrow \hat{\Sigma}$. Define a composition rule $(R, T, S, F) \circ (\bar{R}, \bar{T}, \bar{S}, \bar{F})$ and determine necessary and sufficient conditions for $(R, T, S, F) \in \text{Mor}_{\text{feedback}}(\Sigma, \hat{\Sigma})$ to be a *feedback isomorphism*.

8. Draw the graphs of the linear systems in

- (i) Example 2.1.27,
- (ii) Exercises 1.1, 1.2, 1.3, 1.4, 1.9(i) and 2.4

2.4.4 Notes and References

More details concerning categories of time-invariant linear systems can be found in *Prätzel-Wolters* (1983) [419]. System morphisms in the context of abstract realization theory are studied in *Sontag* (1990) [472].

The field of large scale systems and decentralized control has generated considerable interest amongst control theorists, see the special issue of *IEEE Transactions Automatic Control* (1978) [24], *Siljak* (1991)[465], the collection of papers edited by Leondes [339], [338] and the informative Control Handbook edited by *Levine* (1996) [342].

2.5 Sampling and Approximation: Relations Between Continuous and Discrete Time Systems

The practical implementation of a particular control scheme on a physical plant often involves both continuous and discrete time signals. Such systems are called *hybrid-time* or *sampled-data* systems and can arise, for example, when a digital computer is used to control a continuous time process. In these systems it is necessary to have interfaces which convert (continuous time) *analog* signals into (discrete time) *digital* signals (*A/D-converter, sampler*) and *digital* signals into *analog* ones (*D/A-converter, hold*).

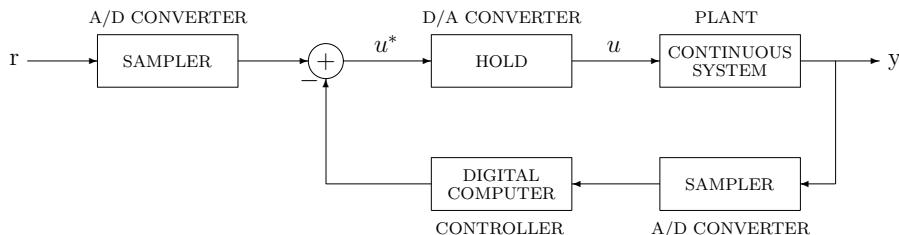


Figure 2.5.1: Digital control of a continuous time plant

The use of a digital computer as a controller implies two types of discretizations a) discretization of *time* and b) discretization of the system *parameters* and *variables* (quantization). Quantization effects arise because real numbers have to be stored and processed using a finite number of digits. The problem of how the various round off errors interact and propagate in a given feedback algorithm needs to be investigated by methods which have been developed in the field of Numerical Analysis and are outside the scope of this book (see *Notes and References*).

In this section we first examine the relationship between discrete and continuous time *signals* and prove a sampling theorem which specifies conditions under which a continuous time signal can be completely restored from its sampled values. Then we go on to examine the relationship between discrete and continuous time *systems*, neglecting quantization errors. We begin by describing the *sampling* of a differentiable system, i.e. the conversion of a continuous time system into a discrete time one by connecting it in series with a hold element and a sampler. We then discuss in some detail the use of numerical integration methods for the *approximation* of continuous time systems by discrete time systems. This is obviously of great importance for the digital simulation of continuous time processes and is relevant in many areas of control and communication where analog devices are replaced by “equivalent” digital devices. We use Euler’s method as the simplest numerical integration scheme to explain some basic concepts. Some higher order methods (single and multistep) are also briefly described and their convergence properties are illustrated by an example with strong oscillations in the control. Whereas Numerical Analysis usually considers the approximation of single trajectories, for system theoretic purposes it is

more important to consider the approximation of *differentiable systems* by *discrete time systems* with controls and initial states which are not fixed. At the end of the section we point out some specific difficulties related to this problem.

2.5.1 A/D- and D/A-Conversion of Signals

A *sampler* associates with each continuous time signal $f(\cdot)$ on $[0, \infty)$ a sequence $f^* = (f(t_k))_{k \in \mathbb{N}}$ of values of f at given sampling instants $t_k \in [0, \infty)$, $k \in \mathbb{N}$. We will find it useful to represent this discrete time signal by a series of impulses

$$f^*(t) = \sum_{k \in \mathbb{N}} f(t_k) \delta(t - t_k) = \sum_{k \in \mathbb{N}} f(t_k) \delta_{t_k}(t) \quad (1)$$

where $\delta(t - t_k) = \delta_{t_k}(t)$ is the Dirac impulse at t_k . Graphically (1) is represented by a sequence of vertical arrows of lengths $f(t_k)$ symbolizing the impulse $f(t_k) \delta(t - t_k)$ (see Figure 2.5.3). Usually equidistant sequences $t_k = k\tau$, $k \in \mathbb{N}$ are chosen. In this case

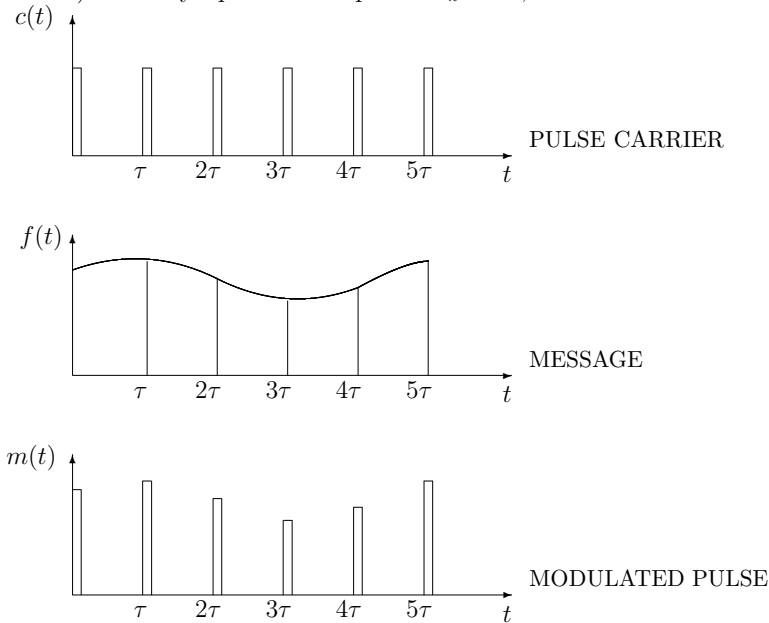
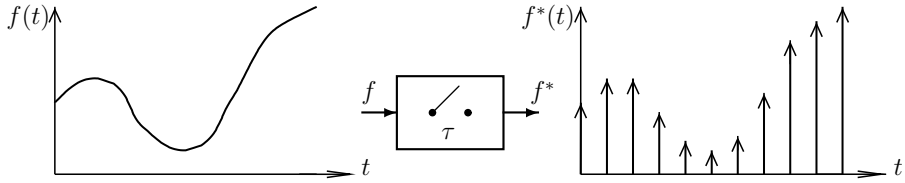


Figure 2.5.2: Pulse amplitude modulation

$\tau > 0$ is called the *sampling period*, $2\pi/\tau$ the *sample rate*. In communication theory frequent use is made of pulse amplitude modulation (see *Notes and References*). A continuous time signal (“message”) $f(\cdot)$ modulates the amplitude of a unit pulse train $c(\cdot)$ of period τ (“the carrier”) to give a modulated pulse $m(t) = f(t)c(t)$ (see Figure 2.5.2). The representation (1) corresponds to an *impulse modulation model* for the sampler where the pulses are idealized to have “infinitely small” width and the carrier signal is a train of impulses $\sum_{k \in \mathbb{N}} \delta_{k\tau}(\cdot)$ which are modulated by f to yield the sampled signal $f^*(\cdot) = \sum_{k \in \mathbb{N}} f(k\tau) \delta_{k\tau}(\cdot)$.

Figure 2.5.3: Ideal sampler with sampling period τ

A *hold* is a device which transforms a series of impulses into a continuous time signal $f(\cdot)$ by a given extrapolation formula. If a k^{th} -order polynomial is used for the extrapolation it is called a k^{th} -order *hold*. The simplest and most widely used is the zero-order hold H_τ^0 defined by $f(t) = f(k\tau)$ if $k\tau \leq t < (k+1)\tau$ (see Figure 2.5.4). A first order hold H_τ^1 (see Figure 2.5.5) associates with f^* the piecewise linear signal

$$f(t) = f(k\tau) + [f(k\tau) - f((k-1)\tau)](t - k\tau)/\tau, \quad k\tau \leq t < (k+1)\tau.$$

The continuous time signals produced by the zero and first order holds are both

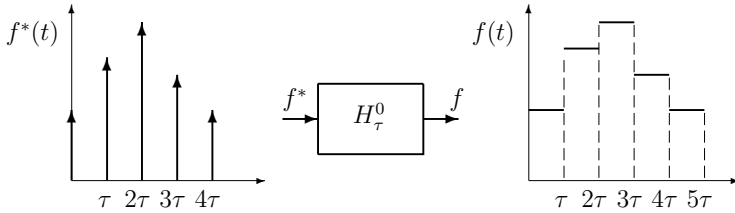


Figure 2.5.4: Zero order hold

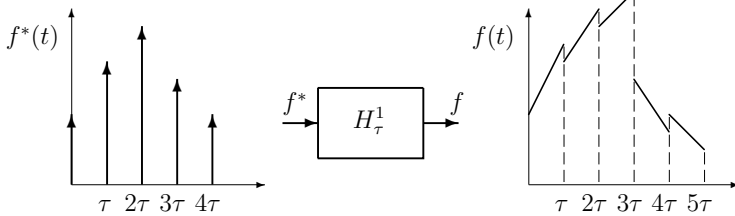


Figure 2.5.5: First order hold

discontinuous. A continuous signal may be produced by the following formula which yields a piecewise linear *interpolation* of the discrete time signal, i.e. a function $f(t)$ whose values at the sampling times $k\tau$ coincide with those of the discrete time signal,

$$f(t) = f(k\tau) + [f((k+1)\tau) - f(k\tau)](t - k\tau)/\tau, \quad k\tau \leq t < (k+1)\tau. \quad (2)$$

Unfortunately, this formula is not implementable since it requires the knowledge of $f((k+1)\tau)$ at time $t < (k+1)\tau$: the operator $f^* \mapsto f$ defined by (2) is not causal. To obtain causality it is necessary to introduce a delay τ , i.e.

$$f(t) = f((k-1)\tau) + [f(k\tau) - f((k-1)\tau)](t - k\tau)/\tau, \quad k\tau \leq t < (k+1)\tau. \quad (3)$$

This is a *delayed first order interpolator* since its value at the sampling time $k\tau$ coincides with the value of the discrete time signal at the previous sampling time $(k-1)\tau$.

Remark 2.5.1. Representing discrete time signals $(f(k\tau))_{k \in \mathbb{N}}$ as series of impulses $f^* = \sum_{k \in \mathbb{N}} f(k\tau)\delta_{k\tau}$ allows us to describe hold elements via convolutions. Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be piecewise continuous and continuous on $\mathbb{R} \setminus \mathbb{N}\tau$, and zero on $(-\infty, 0)$. Then the convolution $f^* * h$ is almost everywhere defined and given by

$$(f^* * h)(t) = \sum_{k \in \mathbb{N}} f(k\tau)(\delta_{k\tau} * h)(t) = \sum_{k=0}^n f(k\tau)h(t - k\tau), \quad n\tau < t < (n+1)\tau.$$

It is easily verified (Ex. 1) that the zero order hold is described by the convolution kernel $h = 1_{[0, \tau)}$ and the first order hold is described by the piecewise linear convolution kernel

$$h : t \mapsto (1 + t/\tau)1_{[0, \tau)}(t) + (1 - t/\tau)1_{[\tau, 2\tau)}(t).$$

□

An A/D -converter is a system which accepts continuous time signals (for example, voltages) and produces a sequence of binary numbers which represent the signal values at the sampling times. In practice it contains a sampler-and-hold element and the converter proper. The function of the sampler-and-hold is to sample the signal and hold its value long enough to allow the converter to code it by binary numbers and store it in a register. If we neglect the difference between the sampled signal value and its binary code the A/D -converter may be modelled as a sampler. Similarly the D/A -converter may be modelled as a hold.

2.5.2 The Sampling Theorem

When a continuous time signal is sampled there will in general be a loss of information and one would expect that the amount lost depends in some way on the rate of sampling. An analysis of this problem is obviously important in communication systems where continuous time signals are encoded, processed, transmitted and stored in a digital fashion. It is also important for the analysis and design of automatic control systems where a continuous time plant is regulated by digital controls. It is not a priori clear that a theoretically well behaved continuous time control law will actually perform well when implemented (digitally) on a computer. The same applies to the converse design methodology where the continuous time system is first discretized and then discrete time controls are designed for the discrete model. In both cases essential information may be lost either by sampling the output (sampled observations) or by discretizing the system.

In the following we determine conditions under which a continuous time signal $v(\cdot) : \mathbb{R} \rightarrow \mathbb{C}$ can be completely reconstructed from its sampled values $v(k\tau)$, $k \in \mathbb{Z}$. Mathematically this is an interpolation problem.

A sampler with sampling period $\tau > 0$ cannot distinguish between a signal $v : t \mapsto v(t)$ and a signal $w : t \mapsto v(t) + \sin(2\pi t/\tau)$. This indicates that it might be useful to represent the signal as a superposition of harmonic oscillations. In order to explain how this can be done we will need to use some results on Fourier series and Fourier

transforms (see Sections A.3 and A.4).

It is well known that, for every $l > 0$, the functions $\psi_k : \theta \mapsto e^{ik\pi\theta/l}$, $k \in \mathbb{Z}$ form an orthonormal basis of the Hilbert space $L^2(-l, l; \mathbb{C})$ provided with the inner product

$$\langle u(\cdot), w(\cdot) \rangle = \frac{1}{2l} \int_{-l}^l u(\theta) \overline{w(\theta)} d\theta, \quad u(\cdot), w(\cdot) \in L^2(-l, l; \mathbb{C}). \quad (4)$$

In other chapters of this book the Hilbert space $L^2(-l, l; \mathbb{C})$ is provided with an inner product without the scalar $1/2l$. We have chosen not to do this here since the use of the inner product in (4) simplifies some of the formulas.

Every function $u(\cdot) \in L^2(-l, l; \mathbb{C})$, $l > 0$ is the sum of its Fourier series in $L^2(-l, l; \mathbb{C})$ (see Example A.4.6 and Theorem A.4.7)

$$u(\cdot) = \sum_{k \in \mathbb{Z}} c_k \psi_k(\cdot); \quad \psi_k(\theta) = e^{ik\pi\theta/l}, \quad \theta \in [-l, l]; \quad c_k = \frac{1}{2l} \int_{-l}^l u(\theta) e^{-ik\pi\theta/l} d\theta, \quad k \in \mathbb{Z} \quad (5)$$

where the two-sided sequence $(c_k)_{k \in \mathbb{Z}}$ of Fourier coefficients $c_k = \langle u(\cdot), e^{ik\pi(\cdot)/l} \rangle$ belongs to $\ell^2(\mathbb{Z}; \mathbb{C})$. Note that the sequence of harmonic oscillations $c_k \psi_k(\cdot)$, though summable in $L^2(-l, l; \mathbb{C})$, is not necessarily pointwise summable for all $\theta \in [-l, l]$.

It follows from (5) that the restriction of every signal $v(\cdot) \in L^2(\mathbb{R}; \mathbb{C})$ to any finite interval $[-l, l]$, $l > 0$ is almost everywhere equal, *on this interval*, to the sum of its Fourier series in $L^2(-l, l; \mathbb{C})$. But it is not possible, in general, to represent the signal $v(\cdot)$ on the *whole real axis* as a superposition of a *countable* set of harmonic oscillations $t \mapsto e^{ik\pi t/l}$. However, under an additional condition $v(\cdot) \in L^1(\mathbb{R}; \mathbb{C})$ can be represented as an integral over *all* harmonic oscillations $e^{i\omega t}$, $\omega \in \mathbb{R}$, with the Fourier transform $\tilde{v}(\cdot)$ as a density function. More precisely the Fourier transform $\tilde{v}(\cdot) : \mathbb{R} \rightarrow \mathbb{C}$ is defined by

$$\tilde{v}(\omega) = (\mathcal{F}v)(\omega) = \int_{-\infty}^{\infty} v(t) e^{-i\omega t} dt, \quad \omega \in \mathbb{R}. \quad (6)$$

Although $\tilde{v}(\cdot)$ is continuous it may not be integrable on \mathbb{R} . But when this is the case then

$$v(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{v}(\omega) e^{i\omega t} d\omega, \quad \text{a. e. } t \in \mathbb{R}, \quad (7)$$

see Theorem A.3.29. If $v(\cdot) \in L^1(\mathbb{R}; \mathbb{C})$ is also continuous then equality holds in (7) for *all* $t \in \mathbb{R}$.

In the sampling theorem we will be concerned with signals $v(\cdot)$ of finite energy, i.e. $v(\cdot) \in L^2(\mathbb{R}; \mathbb{C})$. Since the Lebesgue measure of \mathbb{R} is infinite, $L^2(\mathbb{R}; \mathbb{C})$ is not contained in $L^1(\mathbb{R}; \mathbb{C})$ and so the definition (6) of the Fourier transform is not directly applicable. However by Plancherel's Theorem A.3.33 for any given $v(\cdot) \in L^2(\mathbb{R}; \mathbb{C})$ the sequence of functions

$$\tilde{v}_N(\omega) = \int_{-N}^N v(t) e^{-i\omega t} dt, \quad \omega \in \mathbb{R}, \quad N \in \mathbb{N}$$

converges in $L^2(\mathbb{R}; \mathbb{C})$. Its limit is again denoted by $\mathcal{F}v$ or \tilde{v} and is called the Fourier-Plancherel transform of v . Let

$$v_N(t) = \frac{1}{2\pi} \int_{-N}^N \tilde{v}(\omega) e^{i\omega t} d\omega, \quad t \in \mathbb{R}, \quad N \in \mathbb{N}$$

then by the inversion result in Plancherel's Theorem A.3.33 $v_N(\cdot)$ converges to $v(\cdot)$ in $L^2(\mathbb{R}; \mathbb{C})$.

$v(\cdot) \in L^2(\mathbb{R}; \mathbb{C})$ is said to be of *limited bandwidth* if there exists $\omega_0 < \infty$ such that

$$\tilde{v}(\omega) = 0 \quad \text{for all } \omega \in \mathbb{R}, \quad |\omega| > \omega_0. \quad (8)$$

The smallest $\omega_0 \geq 0$ with this property is called the *bandwidth* of $v(\cdot)$ and is denoted by ω_v . The following theorem shows that it is possible to reconstruct the signal $v(\cdot)$ from its sampled values $v(k\tau)$, $k \in \mathbb{Z}$ if $v(\cdot)$ is of limited bandwidth and the sampling frequency $2\pi/\tau$ is at least twice the bandwidth ω_v of the signal. This reconstruction will be carried out via a series of sinc functions where the function $\text{sinc} : \mathbb{C} \rightarrow \mathbb{C}$ is defined by

$$\text{sinc} : z \mapsto \begin{cases} z^{-1} \sin z & , \quad z \neq 0 \\ 1 & , \quad z = 0 \end{cases}. \quad (9)$$

$\text{sinc}(\cdot)$ is an entire analytic function on \mathbb{C} with the *globally* convergent power series

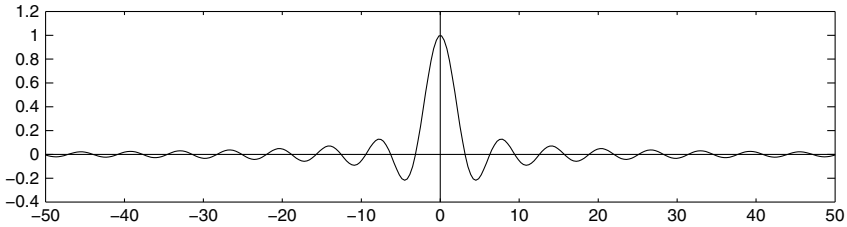


Figure 2.5.6: The sinc function

expansion

$$\text{sinc}(z) = \sum_{k=0}^{\infty} (-1)^k z^{2k} / (2k+1)!. \quad (10)$$

Moreover we note without proof that (see [479] and Ex. 3)

$$\sum_{k \in \mathbb{Z}} |\text{sinc}[t - k\pi]|^2 = 1, \quad t \in \mathbb{R}. \quad (11)$$

Theorem 2.5.2. (Sampling Theorem). Suppose $v(\cdot) \in L^2(\mathbb{R}; \mathbb{C})$ is a continuous function of limited bandwidth $\omega_v < \infty$ and τ is chosen such that

$$0 < \tau < \pi/\omega_v. \quad (12)$$

Then the sequence of functions $(v(k\tau) \text{sinc}[(\pi/\tau)(\cdot) - k\pi])_{k \in \mathbb{Z}}$ is absolutely summable in $L^2(\mathbb{R}; \mathbb{C})$ with sum $v(\cdot)$. Moreover, this sequence is pointwise absolutely summable, uniformly in $t \in \mathbb{R}$, and

$$v(t) = \sum_{k \in \mathbb{Z}} v(k\tau) \text{sinc}[(\pi/\tau)t - k\pi], \quad t \in \mathbb{R}. \quad (13)$$

Proof: Since v is of limited bandwidth we have $\tilde{v}(\cdot) \in L^1(\mathbb{R}; \mathbb{C})$. Using the continuity of $v(\cdot)$ we obtain from (7) and (12)

$$v(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{v}(\omega) e^{i\omega t} d\omega = \frac{1}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} \tilde{v}(\omega) e^{i\omega t} d\omega, \quad t \in \mathbb{R}. \quad (14)$$

In particular

$$v(k\tau) = \frac{1}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} \tilde{v}(\omega) e^{i\omega k\tau} d\omega, \quad k \in \mathbb{Z}. \quad (15)$$

The restriction $u(\cdot) = \tilde{v}(\cdot)|[-\pi/\tau, \pi/\tau]$ of $\tilde{v}(\cdot)$ to the interval $[-\pi/\tau, \pi/\tau]$ is square integrable. Hence, on this interval, it is the sum of its Fourier series in $L^2(-\pi/\tau, \pi/\tau; \mathbb{C})$ (see (5) with $l = \pi/\tau$)

$$u(\cdot) = \sum_{k \in \mathbb{Z}} c_k \psi_k(\cdot), \quad \psi_k(\theta) = e^{ik\tau\theta}, \quad \theta \in [-\pi/\tau, \pi/\tau], \quad c_k = \frac{\tau}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} \tilde{v}(\theta) e^{-ik\tau\theta} d\theta \quad (16)$$

where $(c_k)_{k \in \mathbb{Z}} \in \ell^2(\mathbb{Z}; \mathbb{C})$. It follows from (15) that $c_k = \tau v(-k\tau)$ for $k \in \mathbb{Z}$ and so $\sum_{k \in \mathbb{Z}} |v(k\tau)|^2 < \infty$. Let ψ_k^e be the trivial extension of ψ_k to \mathbb{R} , i.e. $\psi_k^e(\omega) = e^{ik\tau\omega} 1_{[-\pi/\tau, \pi/\tau]}(\omega)$. The two-sided sequence $(\psi_k^e(\cdot))_{k \in \mathbb{Z}}$ forms an orthogonal family in the Hilbert space $L^2(\mathbb{R}; \mathbb{C})$ with $\|\psi_k^e(\cdot)\|_{L^2(\mathbb{R}; \mathbb{C})}^2 = 2\pi/\tau$. Hence (see Proposition A.4.3) $(c_k \psi_k^e(\cdot))_{k \in \mathbb{Z}} = (\tau v(-k\tau) \psi_k^e(\cdot))_{k \in \mathbb{Z}}$ is absolutely summable in $L^2(\mathbb{R}; \mathbb{C})$ and since $\tilde{v}(\cdot)$ vanishes outside the interval $[-\pi/\tau, \pi/\tau]$, we obtain from (16) that

$$\tilde{v}(\cdot) = \sum_{k \in \mathbb{Z}} \tau v(-k\tau) \psi_k^e(\cdot) = \sum_{k \in \mathbb{Z}} \tau v(k\tau) e^{-ik\tau(\cdot)} 1_{[-\pi/\tau, \pi/\tau]}(\cdot) \quad (17)$$

in $L^2(\mathbb{R}; \mathbb{C})$. Applying the inverse Fourier-Plancherel transform to the function $1_{[-\pi/\tau, \pi/\tau]}(\cdot)$ we get, see (45)

$$\mathcal{F}^{-1}(1_{[-\pi/\tau, \pi/\tau]}(\cdot))(t) = (1/\tau) \operatorname{sinc}((\pi/\tau)t), \quad t \in \mathbb{R}.$$

and by the time shifting property of the Fourier-Plancherel transform we have

$$\mathcal{F}^{-1}(\tau e^{-ik\tau(\cdot)} 1_{[-\pi/\tau, \pi/\tau]}(\cdot))(t) = \operatorname{sinc}[(\pi/\tau)t - k\pi], \quad t \in \mathbb{R}, \quad k \in \mathbb{Z}. \quad (18)$$

(see Proposition A.3.35). Since the inverse Fourier-Plancherel transform is a bounded linear operator on $L^2(\mathbb{R}; \mathbb{C})$ it follows from (14), (17) and (18) that

$$v(\cdot) = \sum_{k \in \mathbb{Z}} v(k\tau) \operatorname{sinc}[(\pi/\tau)(\cdot) - k\pi] \quad (19)$$

where the sequence $(v(k\tau) \operatorname{sinc}[(\pi/\tau)(\cdot) - k\pi])_{k \in \mathbb{Z}} = (v(k\tau) \mathcal{F}^{-1}(\tau \psi_k^e(-(\cdot))))_{k \in \mathbb{Z}}$ is absolutely summable in $L^2(\mathbb{R}; \mathbb{C})$ since we have shown above that the sequence $(\tau v(k\tau) \psi_k^e(-(\cdot)))_{k \in \mathbb{Z}}$ is absolutely summable in $L^2(\mathbb{R}; \mathbb{C})$. This proves the first statement of the theorem.

Now $\sum_{k \in \mathbb{Z}} |v(k\tau)|^2 < \infty$ and it follows easily from (9) that the function $t \mapsto \sum_{k \in \mathbb{Z}} |\operatorname{sinc}[(\pi/\tau)t - k\pi]|^2$ is bounded on \mathbb{R} , see Ex. 3. Hence by the Cauchy-Schwarz inequality in $\ell^2(\mathbb{Z}; \mathbb{C})$ (see (A.3.3)) the sequences $(v(k\tau) \operatorname{sinc}[(\pi/\tau)t - k\pi])_{k \in \mathbb{Z}}$, $t \in \mathbb{R}$ are absolutely summable, uniformly in $t \in \mathbb{R}$. As a consequence the sum of the series in (13) is continuous in $t \in \mathbb{R}$ and equals $v(t)$ almost everywhere (see Proposition A.3.11). Since $v(\cdot)$ is continuous by assumption, we finally obtain the equality in (13) for all $t \in \mathbb{R}$. \square

The equation (13) is an explicit interpolation formula for the values of the time signal in between the sampling instants. Note, however that the sampled values $v(k\tau)$ of the past $k\tau < t$ as well as the future $k\tau > t$ are needed in order to compute the value of $v(\cdot)$ at time t . Hence the interpolation (13) *cannot* be implemented as a causal system with the sampled signal as input and the reconstructed signal as output.

Remark 2.5.3. (i) If the signal $v(t)$ is real then each term in the series on the RHS of (13) is also real.

(ii) If $t \in \mathbb{R}$ is replaced by $z \in \mathbb{C}$ in (14) then $v(\cdot)$ can be extended to a continuous function $v_{\mathbb{C}}(\cdot) : \mathbb{C} \rightarrow \mathbb{C}$ by

$$v_{\mathbb{C}}(z) = \frac{1}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} \tilde{v}(\omega) e^{i\omega z} d\omega, \quad z \in \mathbb{C}.$$

$v_{\mathbb{C}}(\cdot)$ is analytic on \mathbb{C} as can be shown by applying e.g. the theorems of Morera and Fubini. So we see that there would be no restriction in assuming that $v(\cdot) = v_{\mathbb{C}}(\cdot)|_{\mathbb{R}}$ is real analytic in the statement of the sampling theorem. Now the extension $v_{\mathbb{C}}(\cdot)$ satisfies the following exponential estimate

$$|v_{\mathbb{C}}(z)| \leq e^{\pi|z|/\tau} \frac{1}{2\pi} \int_{-\pi/\tau}^{\pi/\tau} |\tilde{v}(\omega)| d\omega = C e^{\pi|z|/\tau}, \quad z \in \mathbb{C}. \quad (20)$$

By a theorem of Paley-Wiener the fact that $v_{\mathbb{C}}(\cdot)$ is analytic on \mathbb{C} and satisfies the inequality $|v_{\mathbb{C}}(z)| \leq C e^{\pi|z|/\tau}$, $z \in \mathbb{C}$ for some constant C is actually equivalent to $v(\cdot) \in L^2(\mathbb{R}, \mathbb{C})$ being of limited bandwidth $[-\pi/\tau, \pi/\tau]$, see *Notes and References*.

(iii) Let $l > 0$ and define the Paley-Wiener space

$$PW(l) = \{v(\cdot) \in L^2(\mathbb{R}; \mathbb{C}); \tilde{v}(\omega) = 0 \text{ for all } \omega \in \mathbb{R}, |\omega| > l\}.$$

Then $PW(l)$ is a closed subspace of $L^2(\mathbb{R}; \mathbb{C})$. In the above proof (see (18)) we have seen that for $l = \pi/\tau$ the 2-sided sequence of sinc functions $(\text{sinc}[l(\cdot) - k\pi])_{k \in \mathbb{Z}}$ form an orthogonal family in $L^2(\mathbb{R}, \mathbb{C})$. In fact it is a real orthogonal *basis* for the space $PW(l)$ and the formula (19) is just the expansion of v with respect to this basis. \square

2.5.3 Sampling Continuous Time Systems

Consider a series connection of a zero-hold, a continuous time system Σ of the form (2.17) and a sampler. We write $(u^\tau(k))_{k \in \mathbb{N}}$ for the input sequence in order to indicate that the input value $u^\tau(k)$ is fed into the hold at time $k\tau$. The corresponding sampled states are given by

$$x^\tau(k) = e^{Ak\tau} x_0 + \int_0^{k\tau} e^{A(k\tau-s)} B u(s) ds.$$

Now $u(t) = u^\tau(k)$, $t \in [k\tau, (k+1)\tau)$, so the evolution of the sampled states is described by the difference equation

$$x^\tau(k+1) = e^{A\tau} x^\tau(k) + \left(\int_0^\tau e^{A\tau-s} B ds \right) u^\tau(k).$$

The discrete time system $\Sigma^{(\tau)} = (e^{A\tau}, \int_0^\tau e^{A\tau-s} B ds, C, D)$ is called the *sampled system* obtained from Σ by sampling at times $k\tau$, $k \in \mathbb{N}$. Note that the system matrix $e^{A\tau}$ of the sampled system is always nonsingular. This is a distinctive feature of discrete time systems obtained from sampling continuous time systems of the form (2.17).

Remark 2.5.4. If (i) only sampled times $k\tau$, $k \in \mathbb{N}$ are considered, (ii) only step inputs are used as controls, (iii) quantization errors are neglected, then the sampled system *exactly* reproduces the state and output values of the corresponding continuous time system at the times $t = k\tau$, $k \in \mathbb{N}$. \square

Example 2.5.5. The linearized equations of motion of the inverted pendulum as described in Example 1.3.4 are

$$\dot{x}(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & a_{32} & 0 & 0 \\ 0 & a_{42} & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix} u(t). \quad (21)$$

If the crane is controlled by a digital computer so that the force on the crane only changes at discrete instants $k\tau$, $k \in \mathbb{N}$, $\tau > 0$, then the sampled state trajectories $(x^\tau(k))_{k \in \mathbb{N}}$ are described by

$$x^\tau(k+1) = A_\tau x^\tau(k) + B_\tau u^\tau(k)$$

where

$$A_\tau = e^{A\tau} = \begin{bmatrix} 1 & a_{32} a_{42}^{-1} (\cosh(\sqrt{a_{42}}\tau) - 1) & \tau & a_{32} a_{42}^{-3/2} (\sinh(\sqrt{a_{42}}\tau) - \sqrt{a_{42}}\tau) \\ 0 & \cosh(\sqrt{a_{42}}\tau) & 0 & a_{42}^{-1/2} \sinh(\sqrt{a_{42}}\tau) \\ 0 & a_{32} a_{42}^{-1/2} \sinh(\sqrt{a_{42}}\tau) & 1 & a_{32} a_{42}^{-1} (\cosh(\sqrt{a_{42}}\tau) - 1) \\ 0 & \sqrt{a_{42}} \sinh(\sqrt{a_{42}}\tau) & 0 & \cosh(\sqrt{a_{42}}\tau) \end{bmatrix}$$

$$B_\tau = \int_0^\tau e^{As} B ds = \begin{bmatrix} b_4 a_{32} a_{42}^{-2} (\cosh(\sqrt{a_{42}}\tau) - 1) + \tau^2 (b_3 - b_4 a_{32} a_{42}^{-1})/2 \\ b_4 a_{42}^{-1} (\cosh(\sqrt{a_{42}}\tau) - 1) \\ b_4 a_{32} a_{42}^{-3/2} \sinh(\sqrt{a_{42}}\tau) + \tau (b_3 - b_4 a_{32} a_{42}^{-1}) \\ b_4 a_{42}^{-1/2} \sinh(\sqrt{a_{42}}\tau) \end{bmatrix}$$

\square

In general the sampled system will not yield information about the state values of the continuous time system between the sampling times. The dynamics of the continuous time system, particularly the location of the eigenvalues of A , and the frequency spectrum of the control and disturbance signals will determine which sampling rates are necessary to obtain sufficient information about the trajectories of the continuous plant from the sampled system. For example, if the feedback system shown in Figure 2.5.1 is required to track signals $r(t)$ having spectral content up to a frequency ω_0 , then the sampling theorem enables us to specify an absolute lower bound on the sampling frequency $2\pi\tau^{-1} \geq 2\omega_0$. However in practice the sampling times have to be considerably higher (5-20 times this theoretical lower bound) depending upon dynamic characteristics of the closed loop system, such as its bandwidth, and additional performance requirements, e.g. reducing the delays between reference input and system response, see *Notes and References*.

2.5.4 Approximation of Continuous Systems by Discrete Systems

Control and communication systems are increasingly making use of digital rather than analog devices. Very often a feedback controller for a continuous plant is designed in continuous time and then a discrete time “equivalent” is implemented on the computer. In communication engineering the development of integrated circuit technology has generated a trend to replace analog by digital filters. A *filter* is a device which passes “desirable” frequency components of an input function (the useful signal) and rejects all others (noise). There are well established techniques for the design of analog filters, usually time invariant linear circuits, which meet prescribed performance specifications. Thus a common method in digital filter design is to first design a good analog filter and then approximate it by a digital filter. There are two aspects to the approximation problem:

- (i) one must find a good discrete-time approximation $\Sigma^{(\tau)}$ of the continuous time system *at the sampling instants* and implement the discrete time system by a digital device,
- (ii) the discrete time system $\Sigma^{(\tau)}$ must be converted into a continuous time system $\bar{\Sigma}_\tau$ by extrapolating the output values of $\Sigma^{(\tau)}$ between the sampling instants.

This latter conversion is usually carried out with a sampler and hold as illustrated in Figure 2.5.7. Note that if $\Sigma^{(\tau)}$ is a time-invariant linear system the resulting system

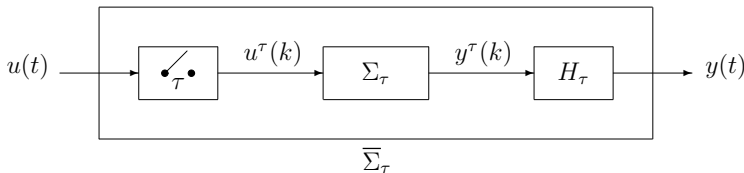


Figure 2.5.7: Conversion of a discrete time system into a continuous time system

$\bar{\Sigma}_\tau$ is linear but not time-invariant since $\bar{\Sigma}_\tau$ will only be invariant with respect to time shifts which are multiples of τ . If the sample period $\tau > 0$ cannot be made small, due to measurement costs or technical reasons, the choice of the extrapolating hold H_τ may be crucial for the performance of the continuous time system $\bar{\Sigma}_\tau$ as an approximation of the original system Σ .

In the following we will deal mainly with problem (i) and present various approximation schemes derived from numerical integration methods. However, we start with the ideal theoretical solution of the approximation problem for the system $\Sigma^{(\tau)}$.

Throughout this subsection $\|\cdot\|$ is an arbitrary norm on \mathbb{R}^n and matrices are normed by the corresponding operator norm. We do not distinguish between these norms unless their use is unclear.

Sample and hold method

If Σ is given by the matrices (A, B, C, D) , we have shown in the previous subsection, that under the assumption of Remark 2.5.4 approximation errors at the sampling

instants can be avoided if we choose the sampled system

$$\Sigma^{(\tau)} = (e^{A\tau}, \int_0^\tau e^{As} B ds, C, D)$$

as the discrete time approximation of Σ . The resulting continuous time system $\bar{\Sigma}_\tau$ (Figure 2.5.7) is called the *hold equivalent* approximation of Σ . The approximation error $x(t) - \bar{x}^\tau(t)$ is zero for $t \in \tau\mathbb{N}$ and is reduced to the extrapolation error between the sampling instants. The extrapolation error depends on the sampling period and the chosen hold element.

Example 2.5.6. Let Σ be the scalar system

$$\begin{aligned}\dot{x}(t) &= ax(t) + bu(t), \quad a > 0 \\ y(t) &= x(t).\end{aligned}\tag{22}$$

Then $\Sigma^{(\tau)}$ has the form

$$\begin{aligned}x^\tau(k+1) &= e^{a\tau}x^\tau(k) + b a^{-1}(e^{a\tau} - 1)u^\tau(k) \\ y^\tau(k) &= x^\tau(k).\end{aligned}$$

Since $\Sigma^{(\tau)}$ reproduces exactly the state trajectory of the continuous time system Σ at the instants $k\tau$, the approximation error of the hold equivalent system only depends on the hold element in Figure 2.5.7. For the zero hold we have

$$x^\tau(t) = x^\tau(k) \quad \text{for } t \in [k\tau, (k+1)\tau) = I_k(\tau)$$

and the corresponding approximation error is

$$\begin{aligned}|x(t) - x^\tau(t)| &= |(e^{at} - 1)x(k\tau) + \int_{k\tau}^t e^{a(t-s)} bu^\tau(k) ds|, \quad t \in I_k(\tau) \\ &\leq |e^{a\tau} - 1||x(k\tau)| + |b a^{-1}(e^{a\tau} - 1)||u^\tau(k)|.\end{aligned}$$

□

It should be noted that the sampled system $\Sigma^{(\tau)}$ requires the computation of $e^{A\tau}$ and $\int_0^\tau e^{As} B ds$ and so can only be implemented approximately on the computer. This leads us to consider more direct approximating procedures.

Euler's method

Euler's method is the simplest integration method. Although it is numerically inefficient (see Table 2.5.12) we feel it is worthwhile discussing in some detail since it illustrates the concepts and problems which are typical for all finite difference methods. *An important advantage of these methods is that they are applicable to nonlinear as well as linear systems.* Since the output map of the approximating discrete time system is usually chosen to be the same as that of the continuous time system we disregard the output map in the following analysis. Consider a differentiable system Σ (as in Definition 2.1.12) with equation of motion

$$\dot{x}(t) = f(t, x(t), u(t)), \quad t \in T.\tag{23}$$

We assume that $U = \mathbb{R}^m$, $X = \mathbb{R}^n$, $T = [a, b] \subset \mathbb{R}$ is compact, $f : T \times X \times U \rightarrow X$ is continuous and satisfies for a given control $u(\cdot) \in \mathcal{U}$ and all $t \in T$ a Lipschitz condition

$$\|f(t, x, u(t)) - f(t, \bar{x}, u(t))\| \leq L\|x - \bar{x}\|, \quad x, \bar{x} \in \mathbb{R}^n \quad (24)$$

where L may depend upon the control $u(\cdot)$.

For a given step-size $\tau > 0$ and initial time $t_0 \in T$ we let $t_k = t_0 + k\tau$ and write N_τ for the largest natural number $N \leq (b - t_0)/\tau$. The *approximate* value of $x(t_k)$ will be denoted by $x^\tau(k)$ and the control value at time $k\tau$ by $u^\tau(k)$.

Euler's method associates with the differential equation (23) (and initial time $t_0 \in T$) the difference equation

$$x^\tau(k+1) = x^\tau(k) + \tau f(t_k, x^\tau(k), u^\tau(k)), \quad (25)$$

with the domain $T_\tau = \{k \in \mathbb{N}; 0 \leq k \leq N_\tau\}$. Equation (25) is obtained if the derivative in (23) is replaced by the difference quotient

$$[x(t+\tau) - x(t)]/\tau \quad \text{at } t = t_k.$$

Alternatively Euler's method can be interpreted in terms of *numerical integration*. In fact, integration of (23) over $[t_k, t_{k+1}]$ yields

$$x(t_{k+1}) = x(t_k) + \int_{t_k}^{t_{k+1}} f(t, x(t), u(t)) dt. \quad (26)$$

Euler's method is obtained by approximating the above integral according to the "forward rectangular rule"

$$\int_{t_k}^{t_{k+1}} f(t, x(t), u(t)) dt \approx \tau f(t_k, x(t_k), u(t_k)).$$

Example 2.5.7. For linear time invariant system (2.17), the discretized version (25) has the special form

$$x^\tau(k+1) = (I + \tau A)x^\tau(k) + \tau B u^\tau(k), \quad k \in \mathbb{N}. \quad (27)$$

Comparing this with the sampled system $\Sigma^{(\tau)}$ we see that the matrix exponential $e^{A\tau}$ is approximated by $I + A\tau$ while $\int_0^\tau e^{At} B dt$ is approximated by τB .

Now let $u(t) \equiv 0$, $t_0 = 0$, $x(0) = x^\tau(0) = x^0$ and $\tau_N = t/N$ for a fixed $t > 0$, $N \in \mathbb{N}$. With respect to this step size the state $x(t) = e^{At}x^0$ of the continuous time system corresponds to the state $x^{\tau_N}(N)$ of the discrete time system (27). The approximation error is

$$\|x(t) - x^{\tau_N}(N)\| = \|e^{At}x^0 - (I + tA/N)^N x^0\|.$$

As in the scalar case it is easy to show

$$\|e^{At} - (I + tA/N)^N\| \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Hence

$$\lim_{N \rightarrow \infty} \|x(t) - x^{\tau_N}(N)\| = 0.$$

So for any given initial state the corresponding free motion of (27) approximates (pointwise) the trajectory of the continuous time system at times $k\tau$ as $\tau \rightarrow 0$. \square

The following theorem shows that the convergence result of the preceding example is true for arbitrary differential systems (23) if suitable smoothness assumptions are made.

Theorem 2.5.8. *Assume that $x(\cdot)$ is a twice continuously differentiable trajectory of (23) corresponding to some control function $u(\cdot)$ with $\|\ddot{x}(t)\| \leq \gamma$ for $t \in [t_0, b]$. Let $x^\tau(\cdot)$ be the trajectory of (25) corresponding to the control function $u^\tau(\cdot) : k \mapsto u(t_k)$ and initial state $x^\tau(0) = x(t_0) + e^0$ where e^0 is an initial error. If f satisfies (24) then*

$$\|x(t_k) - x^\tau(k)\| \leq e^{Lk\tau} \|e^0\| + \tau\gamma(e^{Lk\tau} - 1)/(2L), \quad 0 \leq k \leq N_\tau. \quad (28)$$

In particular if $\|x(t_0) - x^\tau(0)\| \leq c_1\tau$ for some constant c_1 , then

$$\max_{k \leq N_\tau} \|x(t_k) - x^\tau(k)\| \leq c\tau \quad (29)$$

for some constant $c \geq 0$.

Proof: Since $x(\cdot)$ is twice continuously differentiable, we have

$$x(t_{k+1}) = x(t_k) + \tau\dot{x}(t_k) + R(t_k). \quad (30)$$

where

$$\|R(t_k)\| \leq (\tau^2/2) \max_{t_k \leq t \leq t_{k+1}} \|\ddot{x}(t)\| \leq \tau^2\gamma/2. \quad (31)$$

Let $e_k = x(t_k) - x^\tau(k)$, $k \leq N_\tau$, then from (30)

$$\begin{aligned} x(t_{k+1}) &= x(t_k) + \tau f(t_k, x(t_k), u(t_k)) + R(t_k) \\ x^\tau(k+1) &= x^\tau(k) + \tau f(t_k, x^\tau(k), u^\tau(k)), \quad u^\tau(k) = u(t_k). \end{aligned}$$

So

$$e_{k+1} = e_k + \tau[f(t_k, x(t_k), u(t_k)) - f(t_k, x^\tau(k), u^\tau(k))] + R(t_k).$$

Using (24) and (31) we obtain $\|e_{k+1}\| \leq (1 + \tau L)\|e_k\| + \tau^2\gamma/2$. Therefore

$$\begin{aligned} \|e_k\| &\leq (1 + \tau L)^k \|e^0\| + \sum_{i=0}^{k-1} (1 + \tau L)^i \tau^2\gamma/2 \\ &= (1 + \tau L)^k \|e^0\| + \tau\gamma[(1 + \tau L)^k - 1]/(2L) \end{aligned}$$

for $k \leq N_\tau$. Since $(1 + \tau L)^k \leq e^{k\tau L}$, the theorem is proved. \square

The inequality (29) indicates that the approximation error should be halved when τ is halved (see Table 2.5.8). Although this characterizes the convergence rate of Euler's integration method, the explicit estimate (28) is much too conservative in most applications (see Table 2.5.8).

Note that the preceding theorem only describes the approximation of the trajectory $x(t)$ by the discrete time trajectory $x^\tau(k)$ at the instants $t_k = t_0 + k\tau$. In order to obtain an approximation on the whole interval $[t_0, t_0 + N_\tau]$ the integration method must be combined with an extrapolation procedure, as in Example 2.5.6. As a result the overall approximation error is a combination of integration errors and

extrapolation errors. For instance, if a zero hold is used for extrapolation, the overall error is

$$\|x(t) - \bar{x}^\tau(t)\| = \|x(t) - x^\tau(k)\| \leq \|x(t) - x(k\tau)\| + \|x(k\tau) - x^\tau(k)\| \leq c_2\tau + c\tau,$$

where $t_0 \leq t \leq t_0 + N_\tau$, c is any constant such that (29) is satisfied and

$$c_2 = \max\{\|\dot{x}(t)\|, t_0 \leq t \leq t_0 + N_\tau\}.$$

Example 2.5.9. Let us consider the same scalar equation as in Example 2.5.6 on a given time interval $[0, N_\tau]$. In order to apply Theorem 2.5.8 we require a bound on $|\ddot{x}(t)|$. If $u(\cdot)$ is differentiable, we have

$$\ddot{x}(t) = a\dot{x}(t) + b\dot{u}(t) = a^2x(t) + abu(t) + b\dot{u}(t).$$

Hence we can choose

t	Exact solution	Error: $e(t) = x(t) - x^\tau(t/\tau)$			Error bound (28)
	$x(t)$	$\tau = 0.2$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.05$
2.0	2.631	-0.015	-0.007	-0.004	0.007
4.0	3.967	-0.024	-0.012	-0.006	0.017
8.0	5.956	-0.033	-0.016	-0.008	0.049
$u(t) \equiv 1$					
t	$x(t)$	$\tau = 0.2$	$\tau = 0.1$	$\tau = 0.05$	$\tau = 0.05$
2.0	2.624	-0.050	-0.013	-0.005	1.403
4.0	3.954	-0.089	-0.022	-0.009	3.118
8.0	5.934	-0.141	-0.034	-0.013	7.774
$u(t) = 1 + \sin 8\pi t$					

Table 2.5.8: Euler's method applied to $\dot{x} = -0.1x + u$, $x(0) = 1$

$$\gamma = a^2 \max_{0 \leq t \leq N_\tau} |x(t)| + |ab| \max_{0 \leq t \leq N_\tau} |u(t)| + |b| \max_{0 \leq t \leq N_\tau} |\dot{u}(t)|$$

where

$$\max_{0 \leq t \leq N_\tau} |x(t)| \leq e^{aN_\tau} |x^0| + |b| a^{-1} (e^{aN_\tau} - 1) \max_{0 \leq t \leq N_\tau} |u(t)|. \quad (32)$$

We see that the upper bound (28) for the integration error depends through γ not only on the control function $u(\cdot)|[0, N_\tau]$ but also on its rate of change $\dot{u}(\cdot)|[0, N_\tau]$. This dependence is illustrated in Table 2.5.8. In particular the integration error may become large if the control function changes rapidly even if τ is small. If a zero order hold is used for interpolation, the overall approximation error is

$$\begin{aligned} |x(t) - \bar{x}^\tau(t)| &\leq e^{aN_\tau} |e^0| + \gamma (2a)^{-1} (e^{aN_\tau} - 1) \tau \\ &\quad + [a \max_{0 \leq t \leq N_\tau} |x(t)| + |b| \max_{0 \leq t \leq N_\tau} |u(t)|] \tau. \end{aligned}$$

Here the third term (which represents an upper bound for the extrapolation error) can be estimated in terms of $|x^0|$ and $\max_{0 \leq t \leq N_\tau} |u(t)|$ using (32).

If $u(\cdot)$ is only piecewise differentiable with jumps at some $k\tau$, $k \leq N$, the preceding analysis must be applied successively to each interval on which $u(\cdot)$ is differentiable. \square

Single and multi-step methods

Euler's method is a typical *single-step method*. These methods are characterized by the property that they only require knowledge of the present approximate state $x^\tau(k)$ in order to compute the next value $x^\tau(k+1)$. The general form of an explicit single-step method is

$$x^\tau(k+1) = x^\tau(k) + \tau F(t_k, x^\tau(k); \tau, f, u). \quad (33)$$

Here, for any given $t \in T$, $z \in \mathbb{R}^n$, $F(t, z; \tau, f, u)$ is a specific approximation of the difference quotient $\tau^{-1}[x(t+\tau) - z]$ where $x(\cdot)$ is the exact solution of (23) with $x(t) = z$ and control function $u(\cdot)$. In the special case of Euler's method we have $F(t, z; \tau, f, u) = f(t, z, u(t))$.

A $(\nu+1)$ -step method requires the values $x^\tau(k), \dots, x^\tau(k-\nu)$ in order to compute $x^\tau(k+1)$. The general form of such a multi-step method for $k \geq \nu$ is

$$x^\tau(k+1) = \sum_{j=0}^{\nu} a_j x^\tau(k-j) + \tau F(t_k, x^\tau(k+1), x^\tau(k), \dots, x^\tau(k-\nu); \tau, f, u), \quad (34)$$

where a_0, \dots, a_ν are given constants. The method is called *explicit* if F does not depend upon $x^\tau(k+1)$; otherwise it is called *implicit*. Hence in implicit methods it is necessary to solve (34) for $x^\tau(k+1)$ at each step. Nevertheless implicit methods are often more efficient than explicit ones.

Note that whereas single-step methods are self starting, multi-step methods need to be initialized by a single-step method.

In the following we will briefly describe some results for single-step methods and also *linear* multi-step methods of the following form

$$x^\tau(k+1) = \sum_{j=0}^{\nu} a_j x^\tau(k-j) + \tau \sum_{j=-1}^{\nu} b_j f(t_{k-j}, x^\tau(k-j), u(t_{k-j})), \quad k \geq \nu \quad (35)$$

where $a_0, \dots, a_\nu, b_{-1}, \dots, b_\nu$ are given constants. For the sake of simplicity we treat the scalar case. The expression (35) is a $(\nu+1)$ -step method if $a_\nu \neq 0$ or $b_\nu \neq 0$. It is *explicit* if $b_{-1} = 0$ and *implicit* if $b_{-1} \neq 0$.

Example 2.5.10. (Trapezoidal and Heun's method). The trapezoidal method is defined by

$$x^\tau(k+1) = x^\tau(k) + (\tau/2)[f(t_k, x^\tau(k), u(t_k)) + f(t_{k+1}, x^\tau(k+1), u(t_{k+1}))]. \quad (36)$$

It is an *implicit* single-step method and is called the *trapezoidal method* since if f does not depend on x it reduces to the trapezoidal rule for numerical integration (see Figure 2.5.9). This implicit method can be converted into an explicit method by using Euler's method to predict $x^\tau(k+1)$ and then substituting this in the RHS of (36). As a result we obtain the so-called *Heun method*

$$x^\tau(k+1) = x^\tau(k) + (\tau/2)[f(t_k, x^\tau(k), u(t_k)) + f(t_{k+1}, x^\tau(k) + \tau f(t_k, x^\tau(k), u(t_k)), u(t_{k+1}))].$$

This is an explicit single-step method of the form (33) with

$$F(t, z; \tau, f, u) = (1/2)[f(t, z, u(t)) + f(t+\tau, z+\tau f(t, z, u(t)), u(t+\tau))].$$

It is a simple example of a *predictor-corrector algorithm* with Euler's method as predictor and the trapezoidal method as corrector. At each step it requires two evaluations of f . \square

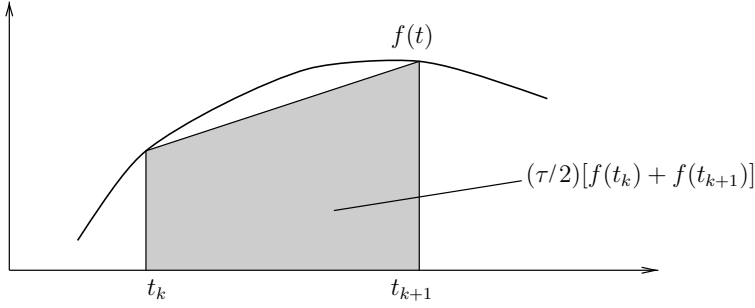


Figure 2.5.9: Trapezoidal method

Example 2.5.11. (Midpoint method). The *midpoint method* is defined by

$$x^\tau(k+1) = x^\tau(k-1) + 2\tau f(t_k, x^\tau(k), u(t_k)), \quad k \geq 1.$$

It is an explicit two-step method which requires one evaluation of f at each step and

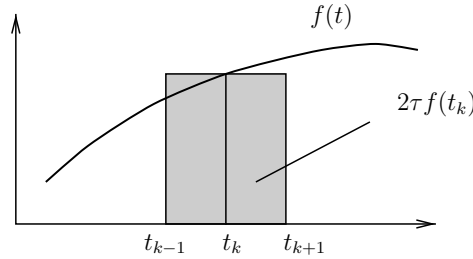


Figure 2.5.10: Midpoint rule

corresponds to the midpoint rule of numerical integration (see Figure 2.5.10). The value of $x^\tau(1)$ must be provided by a single step method. \square

Example 2.5.12. (Classical Runge-Kutta method). The *Runge-Kutta method* is defined by

$$x^\tau(k+1) = x^\tau(k) + (\tau/6)[f_1^\tau(k) + 2f_2^\tau(k) + 2f_3^\tau(k) + f_4^\tau(k)]$$

where

$$\begin{aligned} f_1^\tau(k) &= f(t_k, x^\tau(k), u(t_k)), \\ f_2^\tau(k) &= f(t_k + \tau/2, x^\tau(k) + (\tau/2)f_1^\tau(k), u(t_k + \tau/2)), \\ f_3^\tau(k) &= f(t_k + \tau/2, x^\tau(k) + (\tau/2)f_2^\tau(k), u(t_k + \tau/2)), \\ f_4^\tau(k) &= f(t_k + \tau, x^\tau(k) + \tau f_3^\tau(k), u(t_{k+1})). \end{aligned}$$

It is an explicit single-step method requiring 4 evaluations of f per step. \square

Euler's method, the trapezoidal and midpoint methods are special cases of numerical methods which are based on numerical integration procedures. The general idea of their construction is as follows:

Let $j \in \mathbb{N}$ be given and integrate (23) from t_{k-j} to t_{k+1} , $k \geq j$ to obtain

$$x(t_{k+1}) = x(t_{k-j}) + \int_{t_{k-j}}^{t_{k+1}} f(t, x(t), u(t))dt, \quad k \geq j. \quad (37)$$

Determine the polynomial $P(t)$ of a given degree $\nu \geq 0$ which coincides with the integrand $f(t, x(t), u(t))$ at the $(\nu + 1)$ points t_i , $i \leq k + 1$. The integral of $P(t)$ over the interval $[t_{k-j}, t_{k+1}]$ is then used as an approximation for the integral in (37). To illustrate the above scheme we derive the explicit method of Adams-Bashforth which is frequently used in practice.

Example 2.5.13. (Adams-Bashforth method). We choose $j = 0$ and replace the integrand in (37) by the interpolating polynomial $P_\nu(t)$ of degree ν in t satisfying

$$P_\nu(t_i) = f(t_i, x(t_i), u(t_i)), \quad i = k - \nu, \dots, k.$$

Using the Lagrange's interpolation formula, we have

$$P_\nu(t) = \sum_{i=0}^{\nu} f(t_{k-i}, x(t_{k-i}), u(t_{k-i})) L_i(t)$$

where

$$L_i(t) = \prod_{\ell=0, \ell \neq i}^{\nu} \frac{t - t_{k-\ell}}{t_{k-i} - t_{k-\ell}}.$$

This yields the following *Adams-Bashforth integration formula*

$$x^\tau(k+1) = x^\tau(k) + \tau[\beta_{\nu 0} f_k + \beta_{\nu 1} f_{k-1} + \dots + \beta_{\nu \nu} f_{k-\nu}]$$

where

$$\begin{aligned} \beta_{00} &= 1 \\ \beta_{\nu i} &= \frac{1}{\tau} \int_{t_k}^{t_{k+1}} L_i(t) dt = \int_0^1 \prod_{\ell=0, \ell \neq i}^{\nu} \frac{r + \ell}{-i + \ell} dr, \quad i = 0, \dots, \nu, \\ f_\ell &= f(t_\ell, x^\tau(\ell), u(t_\ell)). \end{aligned}$$

Some values of $\beta_{\nu i}$ and the corresponding formulas are given in Table 2.5.11. □

i	0	1	2	3	$x^\tau(k+1) =$
β_{0i}	1				$x^\tau(k) + \tau f_k$
β_{1i}	$\frac{3}{2}$	$-\frac{1}{2}$			$x^\tau(k) + (\tau/2) [3f_k - f_{k-1}]$
β_{2i}	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		$x^\tau(k) + (\tau/12) [23f_k - 16f_{k-1} + 5f_{k-2}]$
β_{3i}	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	$x^\tau(k) + (\tau/24) [55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}]$

Table 2.5.11: Adams-Bashforth formulas

The above method can also be used to obtain implicit multi-step methods which are combined with a suitable explicit method (predictor) to yield efficient predictor-corrector algorithms. A popular example of this type is *Milne's method*.

Example 2.5.14. (Milne's method). Choose $j = 1$ in (37) and use a quadratic polynomial $P(t)$ which interpolates $f(t, x(t), u(t))$ on the nodes t_{k-1}, t_k, t_{k+1} . The result for $k \geq 1$ is

$$x^\tau(k+1) = x^\tau(k-1) + (\tau/3)[f_{k-1} + 4f_k + f_{k+1}], \text{ where } f_k = f(t_k, x^\tau(k), u(t_k)). \quad (38)$$

This is the corrector of Milne's method and if f does not depend on x it reduces to *Simpson's rule for numerical integration*. The predictor of Milne's method is obtained by using a quadratic interpolation to the integrand at t_{k-2}, t_{k-1}, t_k in (37)

$$x^\tau(k+1) = x^\tau(k-3) + (4\tau/3)[2f_{k-2} - f_{k-1} + 2f_k], \quad k \geq 3. \quad (39)$$

Thus in the k^{th} step of Milne's method $x^\tau(k+1)$ is computed via (38) where $f_{k+1} = f(t_{k+1}, x^\tau(k+1), u(t_{k+1}))$ and $x^\tau(k+1)$ is evaluated by (39). \square

Note that not all the methods we have presented define discrete time dynamical systems, in the sense of Definition 2.1.1. Indeed in the trapezoidal, Heun, Runge-Kutta and Milne methods (corrector plus predictor) the state $x^\tau(k+1)$ is influenced by the control value $u^\tau(k+1)$. This direct input-state coupling contradicts the axiom of causality. On the other hand if Euler's method, Milne's predictor, the midpoint rule or Adams-Bashforth methods are used for the discretization of a differentiable system (23), then discrete time dynamical systems are always obtained. However the discrete time systems obtained by the application of *multi-step* methods do not have the same state space as the corresponding continuous time system. If the method has a memory of length ν , i.e. if $x^\tau(k+1)$ is determined as function of $f_{k-\nu}, \dots, f_k$, then since the past values of x^τ and u^τ have to be "stored" in the state vector the state space of the resulting discrete time system is $X^{\nu+1} \times U^\nu$. This may give rise to certain "instability phenomena" which are observed in multi-step methods such as the midpoint rule and Milne's method. If these are applied for example to linear systems of the form (2.17), rounding errors can incite *unstable parasitic* oscillations of the discretized model which do not correspond to any eigenmotion of the continuous time system (see Ex. 10, 11 and Subsection 3.3.3).

In Theorem 2.5.8 we showed that the approximation error for Euler's method can be bounded by $c\tau$ as $\tau \rightarrow 0$. This is expressed by saying that Euler's method is of order 1. More generally, a particular method is said to be *of order p* if it yields for all initial value problems

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)), \quad t \in T = [a, b] \\ x(t_0) &= x^0 \end{aligned} \quad (40)$$

(where f has continuous and bounded derivatives up to order p on $T \times \mathbb{R}$) approximate solutions $x^\tau(\cdot)$ such that the global approximation error is order p . This means

$$\max_{0 \leq k \leq N_\tau} |x(t_k) - x^\tau(k)| = O(\tau^p) \text{ as } \tau \rightarrow 0,$$

whenever the initial errors tend to zero with order p

$$\max_{0 \leq i \leq \nu} |x(t_i) - x^\tau(i)| = O(\tau^p) \text{ as } \tau \rightarrow 0.$$

It can be shown that the Heun and midpoint methods are of order 2, the classical Runge-Kutta and Milne methods are of order 4. A linear multi-step method (35) is of order $p \geq 1$ if and only if $a_j \geq 0 \quad j = 0, \dots, \nu$ and

$$\sum_{j=0}^{\nu} a_j = 1 \quad \text{and} \quad \sum_{j=0}^{\nu} j a_j + \sum_{j=1}^{\nu} b_j = 1 \tag{41a}$$

$$\sum_{j=0}^{\nu} (-1)^j a_j + i \sum_{j=-1}^{\nu} (-j)^{i-1} b_j = 1 \quad \text{for } i = 2, \dots, p \tag{41b}$$

(see *Atkinson* (1989) [26]). The following example illustrates how the order of a method is reflected in the reduction of the approximation error with diminishing stepsize.

Example 2.5.15. Table 2.5.12 displays the approximation errors $e_k = x(t_k) - x^\tau(k)$ for a variety of methods applied to the initial value problem

$$\dot{x}(t) = -0.1x(t) + 10(1 + \sin 10\pi t), \quad x(0) = 1.$$

All the multistep methods are initialized with “exact” values. Observe the behaviour of

			Euler (order 1)	Heun (order 2)	Milne (order 4)	Runge-K. (order 4)	Adams-B. (order 4)
τ	t_k	$x(t_k)$	e_k	e_k	e_k	e_k	e_k
0.2	5.0	39.829	1.218684	1.573321	1.541310	-0.689172	1.546893
	10.0	63.382	5.220485	5.649258	5.650283	-2.144811	5.631805
	20.0	86.342	17.319951	17.631950	17.645748	-6.221867	17.624505
0.1	5.0	39.829	-0.315121	-0.123519	-0.618127	0.005845	-0.798817
	10.0	63.382	-0.459733	-0.196185	-2.742330	0.009268	-0.983663
	20.0	86.342	-0.550889	-0.258111	-12.553442	0.012160	-1.422443
0.05	5.0	39.829	-0.121190	-0.026503	0.027124	0.000289	0.203423
	10.0	63.382	-0.172322	-0.042103	0.026776	0.000456	0.307603
	20.0	86.342	-0.200045	-0.055414	0.028282	0.000594	0.424276
0.001	5.0	39.829	-0.001891	-0.000010	0.000000	0.000007	0.000000
	10.0	63.382	-0.002604	-0.000016	0.000000	0.000008	0.000000
	20.0	86.342	-0.002896	-0.000021	0.000000	0.000006	0.000000
0.0005	5.0	39.829	-0.000943	-0.000002	0.000000	0.000007	0.000000
	10.0	63.382	-0.001298	-0.000004	0.000000	0.000008	0.000000
	20.0	86.342	-0.001442	-0.000005	0.000000	0.000006	0.000000

Table 2.5.12: Approximation errors and their dependence on the step size τ

the errors as τ is halved. In accordance with the sampling theorem for $\tau = 0.2 > \pi/\omega_u = \pi/10\pi$ no approximation is achieved. All the methods except those of Milne and 4-step Adams-Bashforth yield reasonable first approximation for $\tau = 0.1 = \pi/\omega_u$. The next halving ($\tau = 0.05$) yields an improvement of the approximation which is better than the orders of the various methods predict. For $\tau = 0.001 \rightarrow \tau = 0.0005$ the magnitude of the errors reduces more or less as the order predicts, with the exception of the Runge-Kutta method for which the error reduction rate deteriorates more and more. This is due to the increasing influence of the rounding errors (see Ex. 9). \square

Small errors in the initial state and rounding errors may eventually lead to large errors in the solution if they incite unbounded eigenmotions of the discretized system. In Subsection 3.3.3 we will briefly discuss numerical stability properties of the above methods and we will see that (theoretically) very accurate methods such as Milne's may produce "unstable" discretized systems although the differential system itself is "stable". The selection of an adequate numerical method for a concrete initial value problem relies very much on experience and is still something of an art.

Additional problems arise when we apply theorems of Numerical Analysis, not to the problem of approximating a single solution of a differential equation, but to the much more complex problem of approximating a differential dynamical *system* by a discrete time one. We conclude this section by pointing out some specific difficulties in this context.

Dependence on the control functions

We have seen in Example 2.5.9 and Table 2.5.8 that the error bounds depend not only on the magnitude of the control function but also on the magnitude of its derivative. So we cannot expect that there exists a step-size τ which will yield good approximations for arbitrary control functions with values in a prescribed set. For example the Sampling Theorem suggests that the frequency spectrum of the input signals should be small outside $[-\pi/\tau, \pi/\tau]$.

The following example shows that bang-bang jumps of the control may cause considerable deviations and lead to oscillations around the exact solution which remain, even after the control $u(\cdot)$ has been switched off ($u(t) \equiv 0, t \geq t_1$).

Example 2.5.16. Consider the controlled harmonic oscillator without damping

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u, \quad x(0) = \begin{bmatrix} 5/2 \\ 0 \end{bmatrix}. \quad (42)$$

Let $\tau = 0.05$ and choose

$$u(t) = \begin{cases} (-1)^k 20 & \text{if } k \leq t < k+1; \quad k = 0, 1, 2, 3 \\ 0 & \text{if } t \geq 4 \end{cases}.$$

For $t \geq 4$ the solution of (42) should coincide with the periodic free motion

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x(4) = \begin{bmatrix} -11.4 \\ -28.2 \end{bmatrix}. \quad (43)$$

Figure 2.5.13 shows the "exact" solution curve of (42) for $0 \leq t \leq 30$ (as obtained by a Runge-Kutta method with step size $\tau/10 = 0.005$). It also shows the approximate solution curves (computed with step size $\tau = 0.05$ over the same time interval) by *Euler's method*, the *4-step method of Adams-Bashforth*, the *midpoint rule*, the *predictor of Milne's method* and the complete *Milne method* (predictor and corrector). All the multistep methods were initialized with accurate initial values (computed by Runge-Kutta's method with step size $\tau/10$).

Apart from the various deviations from the true solution (which are particularly large for the predictor of Milne's method) two facts are remarkable. If the same methods are applied to solve the initial value problem (43), all of them, with the exception of Euler's method, track the true circular solution very precisely. They do not show the dramatic

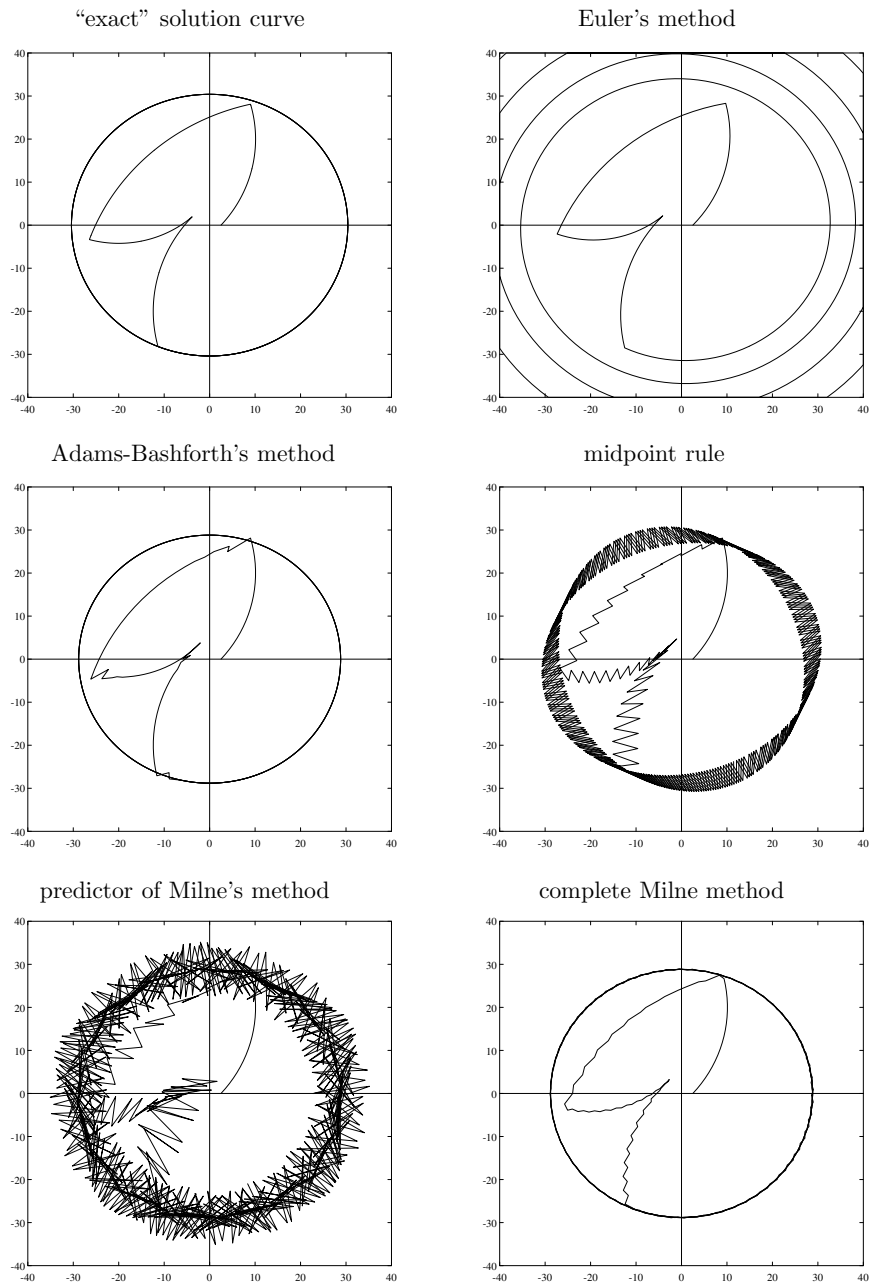


Figure 2.5.13: Approximation of a bang-bang controlled trajectory by various integration schemes

deviations which must, therefore, be caused by the large initial control oscillations. On the other hand Euler's method, which does not reproduce the periodic behaviour of the true solution, is very robust with respect to the effects of the control and shows for $t \geq 4$ the same qualitative behaviour when applied to (42) or (43). \square

Advanced controls

We have seen above that several higher order methods (such as the methods of Heun, Runge-Kutta, all implicit multi-step methods) produce difference equations with direct input state couplings when applied to a controlled differential equation (23). To obtain a discrete-time dynamical system we have to introduce the *shifted* function $\bar{u}^\tau : \mathbb{N} \rightarrow \mathbb{R}$, $\bar{u}^\tau(k) = u(t_{k+1})$ instead of $u^\tau(k) = u(t_k)$ as input signal. The following example shows that it is essential to feed the discrete time systems generated by these methods with the anticipated control value $\bar{u}^\tau(k) = u(t_{k+1})$ at time t_k . Otherwise the order of convergence will not be preserved.

Example 2.5.17. Consider

$$\dot{x}(t) = -2x(t) + u(t), \quad x(0) = 1, \quad u(t) = t^2, \quad t \geq 0. \quad (44)$$

Application of Heun's method yields the following discrete time system

$$x^\tau(k+1) = (1 - 2\tau + 2\tau^2)x^\tau(k) + (\tau/2)u^\tau(k-1) + (\tau/2)(1 - 2\tau)u^\tau(k) \quad (45)$$

where $u^\tau(k) := u((k+1)\tau)$. Table 2.5.14 shows the difference between the solution of (44)

Step size		Exact solution	Heun	Heun (non-advanced control)
	t_k	$x(t_k)$	e_k	e_k
$\tau = 0.2$	1.000	0.352	-0.018	0.079
	5.000	10.250	-0.013	0.867
$\tau = 0.1$	1.000	0.352	-0.004	0.049
	5.000	10.250	-0.003	0.442
$\tau = 0.05$	1.000	0.352	-0.001	0.026
	5.000	10.250	-0.001	0.223

Table 2.5.14: Errors for advanced and non-advanced controls

and the solution of (45) with control $u^\tau(k) = (k+1)^2\tau^2$ and initial value $x^\tau(0) = 1$. The table also gives the approximation error when the non-advanced control $u^\tau(k) = k^2\tau^2$ is used in (45). A comparison of the results in this table shows that the use of non-advanced control not only deteriorates the approximation but changes the *order* of convergence. The same is true for the Runge-Kutta method. If an analogous time shift is applied to the control function in the difference equation obtained from (44) via the Runge-Kutta method, the resulting discretization approximates the solutions of (44) with order 1 instead of 4. \square

2.5.5 Exercises

1. Prove the statements at the end of Remark 2.5.1.

2. Determine the convolution kernels which describe the first order interpolator (2) and the delayed first order interpolator (3).

3. (i) Prove that the function $\theta \mapsto \sum_{k \in \mathbb{Z}} |\text{sinc}[(\theta - k\pi)]|^2$ is bounded on \mathbb{R} .
 (ii) Prove (11).

4. Prove that under conditions of the Sampling Theorem $v(\cdot)$ is analytic on \mathbb{R} and can be extended analytically to \mathbb{C} . Show that this extension satisfies the exponential estimate (20) (see Remark 2.5.3).

5. Let $t_0 < t_1 < t_2$ and f_0, f_1, f_2 be real numbers. Determine the quadratic polynomial $P(t)$ with $P(t_i) = f_i, i = 0, 1, 2$. Apply this to obtain a second order hold H_τ^2 .

Suppose $u : \mathbb{R} \rightarrow \mathbb{R}$ is three times continuously differentiable, $u(t) = 0$ for $t \leq 0$ and $|u(t)| \leq M$ for $t \in \mathbb{R}$. Denote by $\bar{u}(\cdot)$ the function obtained by applying H_τ^2 to the sampled signal $\sum_{k \in \mathbb{Z}} u(k\tau)\delta(t - k\tau)$. Show that $|u(t) - \bar{u}(t)| \leq M\tau^3$.

Find the impulse response and the step response of H_τ^2 .

6. Determine the sampled system $\Sigma^{(\tau)}$ corresponding to the continuous time system

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \quad y(t) = [0, 1] x(t).$$

7. Describe the sampling of a continuous time system (A, B, C, D) with a first order hold instead of a zero order hold by

- (i) determining the equations of motion as higher order difference equations.
 (ii) Finding a state space model of the sampled system (hint: it is a $(n + m)$ -dimensional system).
 (iii) Determining the class of controls for which the sampled system exactly reproduces the state values of the continuous time system at the sample times $k\tau$.

8. Derive an error estimation analogous to that given in Example 2.5.6 if a first order hold is used instead of a zero order hold for the second hold element (see Figure 2.5.5).

9. Suppose the conditions in Theorem 2.5.8 hold and Euler's method is applied to

$$\dot{x}(t) = f(t, x(t), u(t)), \quad x(t_0) = x^0.$$

Because of rounding errors the operations performed at each step are actually

$$\tilde{x}^\tau(k+1) = \tilde{x}^\tau(k) + \tau f(t_k, \tilde{x}^\tau(k), u(t_k)) + \rho_k$$

where ρ_k is the so-called *local rounding error*. Assume $\tilde{x}^\tau(t_0) = x^0$ and $|\rho_k| \leq \rho$ for all $k \geq 0$.

- (i) Show the following estimate for the total error (approximation and quantization)

$$|x(t_k) - \tilde{x}^\tau(k)| \leq (\tau\gamma/2 + \rho/\tau)(e^{Lk\tau} - 1)/L, \quad 0 \leq k \leq N_\tau.$$

- (ii) Discuss the behaviour of the above error bound as a function of τ .

- (iii) Apply Euler's method to the initial value problem

$$\dot{x}(t) = -x^2(t), \quad x(0) = 1.$$

Find the exact solution, tabulate the errors $|x(3) - \tilde{x}^\tau(3/\tau)|$ for $\tau = 10^{-k}$, $k = 2, \dots, 8$ and interpret the table with reference to (ii).

10. Case study: Instability of the midpoint rule.

- (i) Write a computer program to solve the initial value problem

$$\dot{x}(t) = ax(t) + b, \quad x(0) = x^0$$

by the midpoint rule. Your program should be for arbitrary reals a , b , x^0 , stepsize $\tau > 0$ and interval $[0, \bar{t}]$, $\bar{t} > 0$.

- (ii) Choose $a = -2$, $b = 1$, $x^0 = 1$, $\tau = 0.02$ and $\bar{t} = 50$. Print the true solution values $x(k\tau)$ and the errors $x(k\tau) - x^\tau(k)$ for $k = m \cdot 100 + r$ where $m = 0, 1, 2, 4, 8, 16$ and $r = 1, \dots, 10$.
- (iii) Determine the discrete time system obtained by discretizing the equation from (i) via the midpoint rule. Find its eigenvalues and explain the results of (ii). (A systematic analysis will be given later in Subsection 3.3.3).

11. Case study: Impulse response.

- (i) Write a computer program to solve $\dot{x} = Ax + bu$, $x(0) = x^0$ with $A \in \mathbb{R}^{2 \times 2}$, $b \in \mathbb{R}^2$ by Euler's method, Heun's method and the midpoint rule. Your program should be for an arbitrary matrix A , control of the form $\alpha \cdot 1_{[0, \beta]}$, $\alpha \in \mathbb{R}$, $\beta > 0$, initial states $x^0 \in \mathbb{R}^2$, stepsize τ and interval $[0, \bar{t}]$, $\bar{t} > 0$.
- (ii) Use Euler's method to solve the above initial value problem with

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x^0 = 0, \quad u = \frac{1}{10\tau} 1_{[0, 10\tau]}, \quad \tau = 0.1, 0.01, 0.001, \quad \bar{t} = 10$$

(print solution values at $t = 0.1, 0.2, \dots, 1.0$ and $2.0, \dots, 10.0$).

- (iii) Repeat (ii) with $u = (1/10\tau^2)(1_{[0, 10\tau]} - 1_{[10\tau, 20\tau]})$.
- (iv) Compare the solutions of (ii), (iii) with the solutions of the corresponding homogeneous equation $\dot{x} = Ax$ with initial values $x(0) = b = [0, 1]^\top$ and $x(0) = Ab = [1, 0]^\top$ (if available use a graphical display). Interpret the result of (ii) by means of Lemma 2.3.4 on the approximations of the Dirac impulse. Analyze the result of (iii) in the same way.
- (v) Predict what will happen if the midpoint rule is used instead of Euler's method in (ii). Solve (ii) by means of Heun's method and by the midpoint rule (use a graphical display if available).
- (vi) Determine the discrete time systems $x^\tau(k+1) = A_\tau x^\tau(k) + b_\tau u^\tau(k)$ obtained by each of the three methods in (i). Compare the spectra of A_τ with the spectrum of A . What happens if $\tau \rightarrow 0$? Determine the behaviour of the true solution $x(k\tau)$ and the approximate solutions $x^\tau(k)$ as $k \rightarrow \infty$. Try to explain the phenomena observed in (v).

2.5.6 Notes and References

There was very little work carried out on the analysis of sampled-data systems until 1950 when research was motivated by the first use of digital computers in control systems. Frequency domain analysis was used in the first generation of textbooks dedicated exclusively to sampled-data systems, *Franklin and Ragazzini* (1958) [170], *Jury* (1958) [283]. Later textbooks put much more emphasis on the state space approach, *Kuo* (1980) [321], *Ackermann* (1985) [3] and *Oppenheim et al.* (1997) [399]. Recently there has been an upsurge in the analysis and design of hybrid systems, formed when continuous time and discrete time systems are interconnected, in the context of H^∞ theory, see *Dullerud* (1996) [140] and the references therein.

Basic problems concerning the relationship between continuous time and discrete time signals and systems, e.g. sample rate selection, effects of quantization errors, approximation errors, are neglected in most control theoretic textbooks. Our guide in this important area has been the book by *Franklin and Powell* (1998) [169], which gives a good introduction from an engineering point of view, see also the final chapter in *Franklin et al.* (1986) [168], Chapters 9 and 10 in *Kwakernaak and Sivan* (1991) [322] and *Franklin et al.* (1998) [169]. These references also contain further information about the A/D and D/A conversion of signals.

In the context of interpolation theory *Whittaker* (1915) [519] proved that

$$f(z) = \sum_{k=-\infty}^{\infty} f(k\tau) \operatorname{sinc}[\pi(z - k\tau)/\tau], \quad z \in \mathbb{C}$$

for every analytic function $f : \mathbb{C} \rightarrow \mathbb{C}$ with $|f(z)| \leq Ce^{\pi|z|/\tau}$, see the monograph by *Stenger* (1993) [479]. By Theorem X in *Paley and Wiener* (1934) [403] we have seen in Remark 2.5.3 that these conditions are equivalent to the ones in Theorem 2.5.2. *Shannon* (1948) [460] was the first to state the sampling theorem in form we have given it and he also recognized its basic importance for communication theory. *Nyquist* (1928) [394] also made early contributions to the field and the sampling rate $1/\tau$ per second is known as the Nyquist sampling rate. For generalizations of the sampling theorem to irregular spaced samples see *Beutler* (1961) [53]. *Higgins* (1996) [228] contains many interesting historical remarks on the sampling theorem and shows how it plays a role in different areas of mathematics and engineering. Nowadays there are a great variety of sampling results available in the literature and a comprehensive theory is gradually evolving, see e.g. *Benedetto* (1992) [48].

Applications to signal processing and communication are discussed in the well-known introductory textbook of *Kwakernaak and Sivan* (1991) [322]. For further reading on communication systems we refer to *Benedetto et al.* (1987) [49] and *Carlson* (1986) [90]. The difficulties involved in computing the exponential of a matrix are discussed in the paper *Moler and van Loan* (1978) [378], and the update *Moler and van Loan* (2003) [379]. Numerical methods for solving ordinary differential equation are presented in most textbooks on numerical mathematics, see e.g. *Stoer and Bulirsch* (1993). A detailed study can be found in the two volumes *Hairer et al.* (1993) and (1996) [210], [211]. In the control engineering literature numerical integration methods are described as recipes for digital simulation of continuous time systems. However, one must be cautious since the approximation of continuous time systems by discrete time systems is complicated by the fact that instead of determining a fixed solution, it is necessary to consider the system behaviour for a variety of controls and initial states.

Mathematical Systems Theory I
Modelling, State Space Analysis, Stability and
Robustness

Hinrichsen, D.; Pritchard, A.J.

2005, XVI, 804 p. 180 illus., Softcover

ISBN: 978-3-642-03940-9