Lino Guzzella

# Modeling and Analysis of Dynamic Systems

"Systemmodellierung"

September 3, 2020

ETH Zürich

# Preface

This text supports the corresponding lecture offered in the fifth semester of the mechanical engineering undergraduate program at ETH.

Students are assumed to have a thorough understanding of continuous-time linear and time-invariant control systems, such as they are presented in the lecture "Regelungstechnik I." To better understand the examples discussed in class students must have followed a first course on dynamic mechanical systems, thermodynamics, fluid dynamics, and electrical engineering.

This text focuses on the three questions:

- How can a mathematical model of a dynamic system be derived based on physical first principles?
- How can the parameters of this model be determined based on measurements?
- How can the main properties of the system be inferred using the mathematical model derived above?

The emphasis is on modeling based on physical first principles, i.e., engineering science (mechanics, thermo and fluid dynamics, etc.) is used to formulate adequate mathematical models of the system which must be analyzed. In general, only a few experimental data points are necessary for parameter identification and model validation.

The advantages of such an approach, compared to purely experimental input/output methods (aka "system identification"), are the ability to include nonlinearities, the possibility for system optimization in the design phase before a device is built, and the general benefits of having system-internal structural information. The drawbacks are, obviously, the greater efforts needed to first formulate a suitable model.

*All models are wrong. Some are useful.* George Box, 1919-2013.

This text concentrates on "modeling for control," i.e., the models aimed for must be able to predict the dynamic behavior of a system without involving a high computational burden such that they can be used in real-time applications. This typically restricts models to systems of ordinary differential equations of small order. These models cannot exactly describe the behavior of the system. This fact must be included in all subsequent considerations and an appropriate upper bound for the expected model error is part of a complete system description.

With these models, the fundamental properties of the systems to be controlled can be discovered. Stability, controllability and observability are just a few examples of the questions that must be answered before the controller synthesis process can be started.

Control-oriented models can be used for the synthesis of feedback as well as feedforward control systems. Models for feedforward applications have to be more accurate than models for feedback controller design. Therefore, feedback-oriented models are typically derived by further simplifying feedforward-oriented models (say by linearization and order reduction).

The controller *synthesis* part (both feedback and feedforward) is *not* within the scope of this text. For these aspects, other lectures are offered in the graduate curriculum at ETH both in the mechanical and electrical engineering departments.

A last remark: Students who have followed the undergraduate courses "Regelungstechnik I" and "Regelungstechnik II" will see that some of the problems discussed there are taken up again in this lecture. However, in this course the problems are discussed in more detail.

Zurich, Summer 2020 Lino Guzzella

# Contents

# 1

# Introduction

## 1.1 Overview

### 1.1.1 Objectives

This text bridges the gap that is sometimes perceived to exist between the individual engineering science disciplines (such as mechanics, thermodynamics, etc.) and the control literature, which often assumes that a mathematical model of the system that is to be controlled is available. This model can be either formulated as a linear time-invariant system in the time domain

$$\begin{aligned} \dot{x}(t) &= A \cdot x(t) + B \cdot u(t), \qquad x(t) \in I\!R^n, u(t) \in I\!R^m \\ y(t) &= C \cdot x(t) + D \cdot u(t), \qquad y(t) \in I\!R^p \end{aligned} \tag{1.1}$$

or in the frequency domain

$$Y(s) = \left[D + C(sI - A)^{-1}B\right] U(s), \qquad Y(s) \in \mathbb{C}^p,\ U(s) \in \mathbb{C}^m \tag{1.2}$$

or as a nonlinear time-varying system in the time domain[1]

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t), t), \qquad x(t) \in I\!R^n, u(t) \in I\!R^m \\ y(t) &= g(x(t), u(t), t), \qquad y(t) \in I\!R^p \end{aligned} \tag{1.3}$$

The step from an unstructured technical (or sometimes non-technical) system to a clearly defined mathematical model is far from trivial. In many technical areas, a well-established set of such models already exists. Thus, one of the intentions of this text is to assemble some of these existing models in a unified way. Another, maybe more important, objective of this text is to propose a methodology that is useful to formulate mathematical models for which

[1] There is no complete frequency-domain theory for nonlinear systems. In Section 5.2.6 a theory is presented that allows to combine frequency-domain approaches with *static* nonlinearities.

no standard approach is at hand. Unfortunately, such a methodology cannot be "algorithmic" in the sense that it would offer a recipe that, when exactly followed, is guaranteed to produce the best possible model. The process of abstraction from a complex and partially unknown system to a deterministic mathematical model is simply not amenable to such a high degree of formalization. Only experience, gained by solving many problems, can help in that situation.

The focus below will be on "control-oriented models," i.e., models which capture a system's main static and dynamic phenomena without creating an excessive computational burden.[2] The main reason for that requirement is that the models are assumed to be used in real-time loops (for instance, in feedback control systems) or repeatedly in numerical computations (for instance, when optimizing the system parameters or feedforward control signals).

The model synthesis is based on physical first principles (for instance, the first and second laws of thermodynamics, the Lagrange equations in mechanics or the Biot-Savart law in electromagnetics). Compared to experimental methods (for instance correlation methods), this approach has two major benefits:

- The models obtained are able to *extrapolate* the system behavior, i.e., they can be useful beyond the operating conditions used in model validation.
- The models can be formulated even if the real system is not available (system still in planning phase or experiments not possible).

Once such a mathematical model exists for a first system, the adaptation of that model to minor system modifications is relatively easy. Subsequent controller designs, which are based on the system model, can then be carried out (almost) automatically. This time-saving approach is regarded by many as the most important advantage of a model-based procedure.

In most cases, some experiments will be necessary for parameter identification and model validation. Therefore, sometimes the term "gray-box models" is used in order to distinguish this class of models from those derived completely experimentally ("black-box models") and those which require no experiments at all ("white-box models").

Several important issues are not discussed in this text. For instance, all models and signals are assumed to be deterministic such that stochastic effects cannot be included, and little is said about input/output models that are derived by processing the available data of the system's response to known excitations ("System Identification"). Readers interested in these aspects are referred to [11].

[2] Of course, this is a qualitative argument and "computational burden" is very much a function of computer speed. Nevertheless, cost and physical constraints will always force engineers to husband computational power.

### 1.1.2 Why Use Models?

Mathematical models are important for many problem settings,[3] three of which will be discussed in some detail in this subsection.

#### System Analysis and Synthesis

When a system is to be designed for the first time, several fundamental questions arise that are closely related to control actions, such as:

- What are the optimal system parameters (regarding performance, safety, economy, etc.)?
- Is the system inherently stable or can it be stabilized by an appropriate feedback action? If yes, what are the "best" (cost, performance, etc.) actuator and sensor configurations?
- What happens if a sensor or an actuator fails and how can the system's robustness be increased?

If the system is not available for experimentation, a mathematical model must be used to answer these questions.

*Example 1.1 (Geostationary Satellite).*

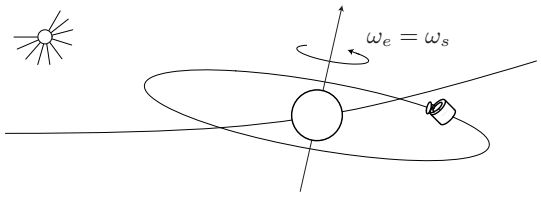



**Fig. 1.1.** Geostationary satellite.

Communication satellites are preferably "parked" on geostationary orbits, where the satellite's circular orbit in the equatorial plane has the same angular velocity $\omega_s$ as the angular velocity $\omega_e$ of the earth. However, the moon's and sun's gravitational pull and other forces will disturb the ideal circular motion of the satellite. A controller stabilizing this circular reference trajectory is therefore necessary. The question is then, which geometric thruster configuration is best? What is the minimum thruster size and how much propellant must be carried onboard for a desired minimum lifetime of the satellite? What sensors are necessary for stabilization? What happens if a sensor or an actuator fails?

[3] One can even argue that a given technical system is not fully understood until a mathematical model of the system is formulated.

### Feedforward Control Systems

The synthesis of feedforward controllers is a major step in the design of many technical systems. Some of the questions that arise are:

- What are the control signals that yield optimal system behavior (shortest cycle time, lowest fuel consumption, etc.)?
- How can the system's response be improved (speed, precision, etc.)?
- How much is lost when trading optimality for safety, reliability, etc.?

A quantitative answer to these question can be found only if a mathematical model of the system is available for a computer-assisted analysis.

*Example 1.2 (Assembly Robot).*

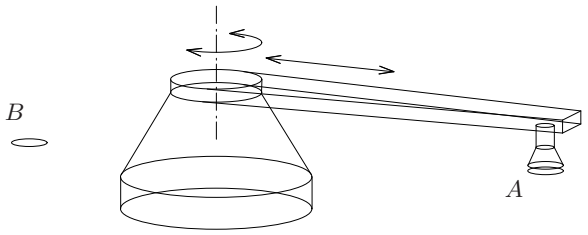



**Fig. 1.2.** "Pick and place" assembly robot.

In many industrial processes robots are used to automate assembly tasks. To maximize the performance of such systems one must find the commands to the joint's electric motors that produce minimum transfer times between the place $A$, where the robot picks up an object, and the location $B$, where this item is placed in another machine for assembly.

### Feedback Control Systems

Feedback controllers are necessary in order to compensate for the errors in the feedforward commands and to reject unmeasurable disturbances acting on the system. If the system is inherently unstable, stabilization is another important issue. Typical questions that must be answered are:

- How can system stability be maintained for a given set of expected modeling errors?
- How can a specified disturbance rejection be guaranteed for disturbances acting in specific frequency bands?
- What are the minimum and maximum bandwidths that a controller must attain for a specific system in order for stability and performance requirements to be guaranteed?

*Example 1.3 (Magnetic Bearing).*

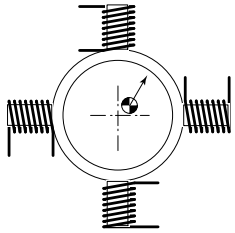


**Fig. 1.3.** Cross-section of a magnetic bearing.

Magnetic bearings have several advantages over standard journal or ball bearings, such that they are used, e.g., in high-performance vacuum pumps. Some problems must be solved to make their operation possible: How can the current in the coils of the stator be controlled such that the rotor keeps a desired minimum distance from the stator? How can the closed-loop system's response be made insensitive to disturbances (periodic forces introduced by the rotor not being perfectly balanced)? What is the influence of varying system parameters (coil temperature, shaft tolerances, etc.)?

Models are always limited in their scope to a specific application area that must be defined at the outset of the model developing process. A model that works in many applications is (in general) not the best model for any of these problem settings.

Moreover, all models inherently reflect the system's behavior only approximately. It is of paramount importance to be aware of this when applying models for specific tasks. The estimation and quantification of the expected model errors is, therefore, an important part of all model developing processes.

# 2

# System Modeling for Control

## 2.1 Some Definitions

This section introduces the most important definitions used in this text. Generic definitions of systems and signals and their properties require a rather advanced mathematical background (see, for instance, [17]) which is not assumed to be available at this point. Therefore, in the following, a pragmatic approach is chosen, where the salient properties of a signal or system class are defined by examples.

Signals are abstract information carriers that can be used without affecting the source that generated these signals. In this course signals are assumed to be real-valued continuous functions of the time $t$. These functions can be scalar or multidimensional. Input signals drive systems, output signals contain information about the system state variables (see below). In practice, physical entities (voltage, pressure, etc.) are used to approximate signals.

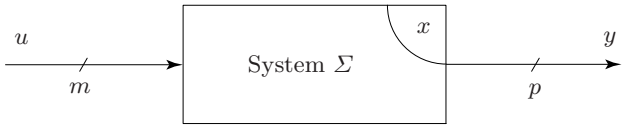



**Fig. 2.1.** General definition of a system, input $u(t) \in \mathbb{R}^m$, output $y(t) \in \mathbb{R}^p$, internal state variable $x(t) \in \mathbb{R}^n$.

Mathematical models of dynamic systems can be subdivided into two broad classes:

1. parametric models (PM); and
2. non-parametric models (NPM).

Parametric models are compact descriptions of the system behavior using concise mathematical formulations (ODE, PDE, transfer functions, etc.) for which only a few numeric parameters are required. Predicting the system output for arbitrar inputs is possible, but is numerically demanding. Nonparametric models are mathematically much simpler (tables, lists, artificial neural networks or just logical expressions), but require a large number of data points that capture the system's response to fixed inputs. Starting with these known input/output relations, the system's response to other excitations can be computed (for instance, the steady-state output of a system which is driven by periodic inputs when the system's frequency response is known). Nonparametric models can be also *rule-based*, in which case the system behavior is described only qualitatively (for instance, in fuzzy logic formulations). Examples of PM and NPM models are shown below.

## 2.2 Overview of Modeling Methods

In engineering sciences, modeling is a central topic and several levels of modeling accuracy can be found. Coarse models are used in the first phases of a planning or optimization project. Later, when the structure of a system becomes better known, more precise models are developed. At the end, precise models such as finite-element models (FEM) of mechanical structures or electromagnetic devices are used that describe the system as a continuum. More recently, even discrete-particle models have been developed for microscopic systems.

### 2.2.1 Parametric and Nonparametric Models

As mentioned, a first important distinction is whether a model is "parametric" or "nonparametric." An example is used to explain the difference between these two approaches.

*Example 2.1 (Mechanical Oscillator).*

A damped spring and mass oscillator, as illustrated in Figure 2.2, can be described by the second-order differential equation

$$m\ddot{y}(t) + d\dot{y}(t) + ky(t) = F(t) \tag{2.1}$$

where $y$ is the position of the mass, $\dot{y}$ its velocity, and $\ddot{y}$ its acceleration, respectively. The constants $m$, $d$ and $k$ represent the mass, the damping and the spring constant in this system. Equation (2.1) is a typical example of a

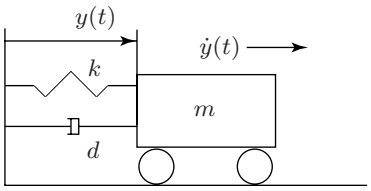
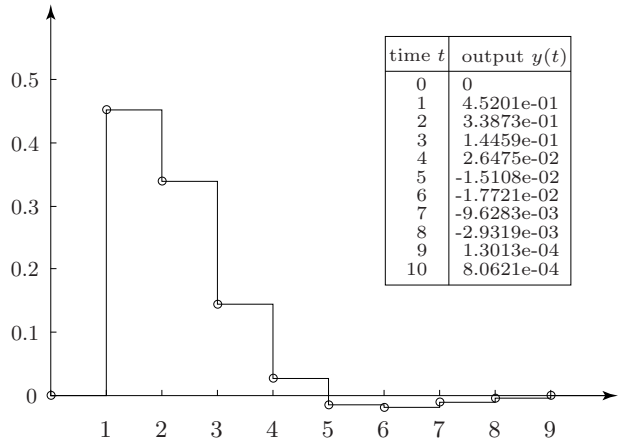



**Fig. 2.2.** Damped mechanical oscillator.

parametric model. Only a few (in this case three) physical parameters are needed to describe the system's behavior.

The same system can be described by its impulse response $\sigma$, stored in the form of a plot or a table (see Figure 2.3). These many parameters are obtained by stimulating the system with an input $u(t) \approx \delta(t)$ that approximates[1] an impulse and by measuring the position of the mass $y(t)$ at times $t = k \cdot T$, with $T$ a constant sampling period. Using a convolution operation (see Equation (4.43) in Section 4.3), the output $y(t)$ can then be computed for *arbitrary* inputs $u(t)$ using the following approximation

$$y(t) = \sigma(t) * u(t) = \int_0^t \sigma(t-\rho)\, u(\rho)\, d\rho \approx \sum_{i=1}^{t/T} \sigma(t - i\,T)\, u(i\,T)\, T$$




| time $t$ | output $y(t)$ |
|---|---|
| 0 | 0 |
| 1 | 4.5201e-01 |
| 2 | 3.3873e-01 |
| 3 | 1.4459e-01 |
| 4 | 2.6475e-02 |
| 5 | -1.5108e-02 |
| 6 | -1.7721e-02 |
| 7 | -9.6283e-03 |
| 8 | -2.9319e-03 |
| 9 | 1.3013e-04 |
| 10 | 8.0621e-04 |

**Fig. 2.3.** Impulse response of a damped mechanical oscillator, $T = 1$ s.

As mentioned above, in this text only parametric models will be discussed. Nonparametric models, although useful in certain circumstances, have several

[1] One choice is $u(t) = 1/T, \forall t \in [0, T)$ and all other $u(t) = 0$.

drawbacks that render them less attractive for the purposes of this course. Particularly, they require the system to be accessible for experiments, they cannot predict the behavior of modified systems and they are not useful for systematic design optimizations.

### 2.2.2 Forward and Backward Formulations

A second important distinction is whether a model formulation is *forward* or *backward*, a notion that is closely associated with causality. In engineering there is a general agreement on what the causes and what the effects are. This causality is defined by the generally accepted physical laws, the concepts of which engineers have become so familiar with that the causality chain usually appears to be obvious.

*Example 2.2 (Cruise Control).*

The relation between traction force $F(t)$ and velocity $v(t)$ of a car traveling on an even road is given by Newton's law

$$m\frac{d}{dt}v(t) = -\{k_0 + k_1 v(t)^2\} + F(t) \tag{2.2}$$

where all driving resistances are given by the term in curly brackets.

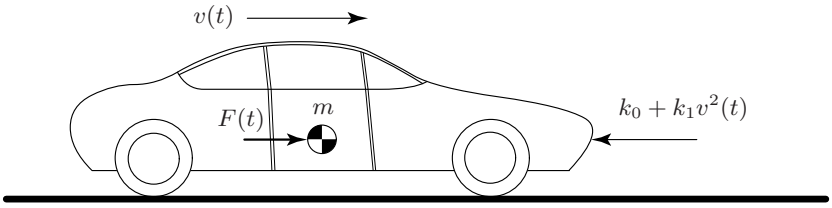



**Fig. 2.4.** Car moving in a plane.

Assuming the traction force $F(t)$ to be the system's input (as commanded by the driver) and assuming the actual fuel massflow consumed by the engine to be given by

$$\overset{*}{m}(t) = \{\mu + \epsilon F(t)\}\, v(t) \tag{2.3}$$

the car's fuel consumption can be computed by solving the differential equation (2.2), then inserting the resulting car speed and the input (force) into (2.3) and integrating the fuel massflow over the driving cycle. This is a typical forward formulation. Notice that a driver is essential in this loop.

If the driving pattern $v(t_i) = v_i$, $i = 1, \ldots, N$, $t_i - t_{i-1} = T$ is known at the outset (say a test cycle prescribed by a governmental authority), then a backward approach is possible. By inverting the causality chain, one can ask what force is necessary to follow that driving pattern. Using (2.2), the answer to that question can be found to be approximately

$$F(t_i) \approx m\frac{v(t_i) - v(t_{i-1})}{T} + k_0 + k_1 \left(\frac{v(t_i) + v(t_{i-1})}{2}\right)^2 \tag{2.4}$$

Inserting the resulting force $F(t_i)$ and the known speed $v(t_i)$ into (2.3) and summing up all resulting values

$$\sum_{i=1}^{N} \overset{*}{m}(t_i)\, T \tag{2.5}$$

a good approximation of the fuel consumed in the cycle is obtained (errors on the order of only a few percent).

Feedback-control-oriented models are (almost) always lumped parameter-models (ordinary differential equations or transfer functions) and are formulated always as forward equations, i.e., causality must be respected. Feedforward control-oriented models can be derived using both forward or backward formulations. The latter are useful for *supervisory control* optimizations and require, of course, some a-priori information regarding the system trajectory. The advantage of backward models is the relatively low computational effort they require for their simulation. However, in this text the emphasis is laid on forward formulations.

### 2.2.3 Relevant Dynamic Phenomena

As mentioned, control-oriented models must be of small order. This is equivalent to looking for models that include only dynamic phenomena which evolve on similar time scales. As Figure 2.5 shows, in general, three types of dynamic signals must be distinguished:

b) signals with relevant dynamics;
a) signals with fast dynamics; and
c) signals with slow dynamics.

A good model will contain only the relevant dynamic effects (otherwise "stiff systems" are obtained which pose difficult problems to numerical integration routines), where "relevant dynamic effects" are defined by the variable of primary interest (e.g., the measured output of the system).

Signals of type a) are very fast compared to signals of type b) and should be modeled as variables which depend in a purely *algebraic* way on the system inputs and the relevant variables b).

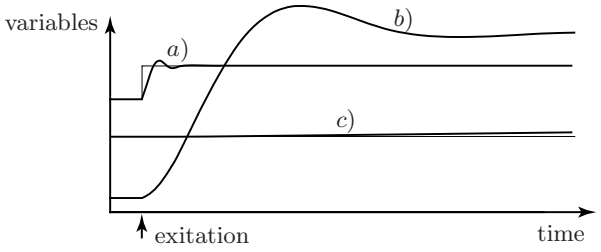



**Fig. 2.5.** Classification of variables into fast (a – idealized as algebraic), relevant (b – the dynamic variables) and slow (c – idealized as static).

Signals of type c) are very slow compared to signals of type b) and should be modeled as *constants* (which may be adapted after a longer period).

Unfortunately, there are no simple and systematic rules with which one could decide a-priori which dynamic effects can be modeled in what way. Here, experience and/or iterations will be necessary, making system modeling partially a subject of "engineering art."

Notice that the variables a) and b) together form a system of differential and algebraic equations (DAE) which is not directly solvable by standard numerical integration algorithms. Additional efforts will be necessary in this situation and more on that topic will be said below.

## 2.3 Reservoir-Based Approach

When modeling any physical system there are two main classes of objects that must be taken into account:

- *reservoirs*, for instance of thermal or kinetic energy, of mass or of information; and
- *flows*, for instance, of heat, mass, etc., flowing between the reservoirs.

The notion of a reservoir is fundamental in this text, and only systems including one or more reservoirs exhibit dynamic behavior. There is a *level variable* associated to each reservoir, which is a function of the reservoir's content (in the control literature often the term *state variable* is used for this quantity). The flows are typically driven by the differences in the reservoir levels. Several examples of that will be given below.

General guidelines to formulate a control-oriented model encompass (at least) the following seven steps:

1. define the system boundaries (what inputs, what outputs, ...);
2. identify the relevant reservoirs (for mass, energy, information, ...) and the corresponding level variables;
3. formulate the differential equations (conservation laws) for all relevant reservoirs as shown in Equation (2.6)

$$\frac{d}{dt}\,(\text{reservoir content}) = \sum \text{inflows} - \sum \text{outflows}; \tag{2.6}$$

4. formulate the (usually nonlinear) algebraic relations that express the flows between the reservoirs as functions of the level variables;
5. resolve implicit algebraic loops, if possible, and simplify the resulting mathematical relations as much as possible;
6. identify the unknown system parameters using some experiments; and
7. validate the model with experiments that have not been used to identify the system parameters.

Obviously, step 6 and 7 are only possible if the system to be modelled is available for experiments. If this is not the case, for instance because the system is not yet constructed, one has to rely on design specifications to estimate the unknown parameters.

Of course, the inverse approach is also possible and often useful: the not-yet specified system parameters $p$ are found by solving an optimization problem, where a cost function

$$J(p) = \int_{t_0}^{t_f} L(x(t))\,dt, \quad \dot{x}(t) = f(x(t), u(t), p), \quad x(t_0) = x_0 \tag{2.7}$$

is minimized by choosing the best possible parameters $p_{\text{opt}}$ (the input $u(t)$ and the initial condition $x_0$ are assumed to be fixed and known in this formulation). Appendix I introduces some of the methods that can be used to solve these problems.

*Example 2.3 (Water Tank).*

Below, the first four steps are shown in more detail by deriving a model of a cylindrical water-tank system, as depicted in Figure 2.6.

*Step 1:*

The system's input is $u(t) = \overset{*}{m}_{in}(t)$, i.e., the inflowing water massflow is assumed to be arbitrarily assignable (say, by an underlying fast massflow control loop) and the system's output is $h(t)$, i.e., the measured height of the water in the tank.

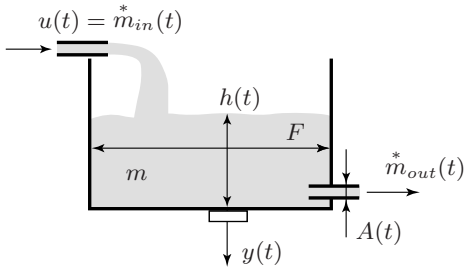



**Fig. 2.6.** Water-tank system, $m(t)$ mass of water in tank, $h(t)$ corresponding height, $F$ tank-floor area, $A =$ outflow orifice area

*Step 2:*

The only relevant reservoir is the mass $m(t)$ of water in the tank; the dynamic effects in the sensing device are fast and can, therefore, be neglected (type a) variable). The water temperature, and therefore its density, is assumed to change very slowly (type c) variable) such that it is assumed to be constant (otherwise mass and height would not be strictly proportional).

*Step 3:*

A mass balance is formulated for the tank. The resulting differential equation is

$$\frac{d}{dt}m(t) = u(t) - \overset{*}{m}_{out}(t) \tag{2.8}$$

*Step 4:*

The water massflow leaving the tank can be expressed using Bernoulli's law (i.e., assuming incompressible and frictionless flow conditions, see Section 2.4.2) as shown in Equation (2.9)

$$\overset{*}{m}_{out}(t) = A(t)\rho v(t), \quad v(t) = \sqrt{2\Delta p(t)/\rho}, \quad \Delta p(t) = \rho g h(t) \tag{2.9}$$

where $A$ is the area of the orifice, $\rho$ the density of the water, $v(t)$ the velocity of the water in the orifice, $\Delta p$ the pressure difference over the orifice and $g$ the earth's gravitation constant. Inserting Equations (2.9) in (2.8) yields the results shown in (2.10);

$$\frac{d}{dt}m(t) = \rho F \frac{d}{dt}h(t) = u(t) - A(t)\rho\sqrt{2gh(t)} \tag{2.10}$$

A block diagram showing all relevant reservoirs and flows will be called a *causality diagram* (see Figure 2.7 for the water-tank example). In such a diagram, the driving and the driven variables are clearly displayed, which permits the identification of the causalities of the analyzed system. This tool will become more important when dealing with complex systems, as will be the case in Section 2.6.

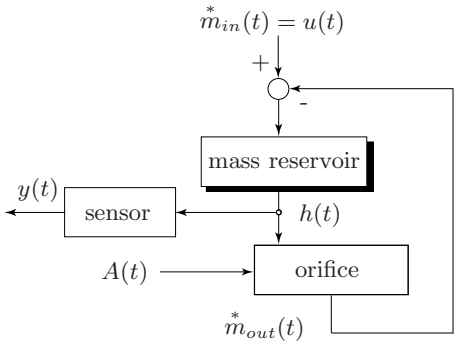



**Fig. 2.7.** Causality diagram of the water-tank system. In these diagrams, shaded blocks represent dynamical subsystems (containing reservoirs), and plain blocks represent static subsystems.

## 2.4 Basic Modeling Elements

In the next subsections, some important engineering areas will be reviewed from the point of view of this text, i.e., modeling dynamic phenomena for control purposes. It is obvious that only the simplest concepts can be discussed on this level. Readers interested in more detailed discussions can find references to dedicated textbooks in the reference section.

### 2.4.1 Mechanical Systems

Modeling mechanical systems is one of the oldest scientific disciplines. The description of general mechanical systems is rather difficult and students interested in that topic are referred to the corresponding courses. The objective of this section is to show how mechanical systems can be analyzed using the paradigms of general system dynamics and to facilitate the transition between mechanics and control systems. To be able to do that only a simple class of mechanical systems are considered below, *viz.* rigid bodies that move in a plane. Energy methods will be used in this section instead of the (equivalent) formulations such as Newton or Euler equations. This approach is often simpler when constrained and connected systems are to be analyzed.

Two levels of complexity will be distinguished in this text:

- single rigid body with one degree of freedom (DOF); and
- linked multiple rigid bodies with multiple DOF.

The central elements required to model mechanical devices are the kinetic and potential energies of these systems. For a *point mass* $m$ moving in a plane, the kinetic energy is given by

$$T(t) = \frac{1}{2} m (v_x^2(t) + v_y^2(t)) \tag{2.11}$$

where the variables $v_{...}$ designate the velocities in the two Cartesian directions $\{x, y\}$. For a rigid body of mass $m$ with non-zero planar extension two kinetic energies have to be considered:

- translational energy

$$T_t(t) = \frac{1}{2} m (v_{x,cg}^2(t) + v_{y,cg}^2(t)) \tag{2.12}$$

  where $v_{...,cg}$ are the velocities in the two Cartesian directions $\{x, y\}$ of the body's center of gravity; and
- rotational energy

$$T_r(t) = \frac{1}{2} \Theta \, \omega^2(t) \tag{2.13}$$

  where $\omega$ is the body's angular velocity vector (assumed to be always perpendicular to the plane in which the body moves) and $\Theta \in I\!R$ is its inertia around the body's center of gravity.

The angular velocity vector $\omega$ is the same for all reference points on the rigid body. In contrast, the translational velocity vector $v$ depends on the specific location chosen as a reference point. All of the following considerations are valid only when the latter is taken at the body's center of gravity, as assumed in Equation (2.12).

The potential energy can always be expressed as a function only of the body's coordinates, i.e.,

$$U(t) = U(x(t), y(t)). \tag{2.14}$$

This is all that is needed for point masses. For bodies with non-zero planar extension in gravitation fields, the mass is concentrated in the body's center of gravity. Other potential fields ("springs," etc.) can be handled similarly (see the examples below).

Using the concepts introduced above, the total mechanical energy of one body is defined as

$$E(t) = T(t) + U(t) \tag{2.15}$$

and the differential equations ("step 3") are obtained by a simple mechanical power balance

$$\frac{d}{dt} E(t) = \sum_{i=1}^{k} P_i(t) \tag{2.16}$$

where $P_i$ are the mechanical powers acting on the body.

When multiple bodies that have more degrees of freedom and that are linked by some holonomic[2] or nonholonomic constraints are considered, then a Lagrange formulation is best. This will be discussed in more detail later in this section.

### Single Rigid Body With One DOF

The simplest case is encountered when only one translation or rotation is possible. The modeling procedure is best explained using some examples.

*Example 2.4 (Cruising Car).*

The objective of this example is to derive a model that can be used to design a robust cruise controller, i.e., a control loop that keeps the speed of a vehicle at a desired setpoint despite unobservable disturbances (grade, wind, etc.) and which is robust with respect to changing system parameters (speed setpoint, vehicle mass, gear ratio, etc.) and neglected dynamic effects (engine torque delays, sensors, etc.).

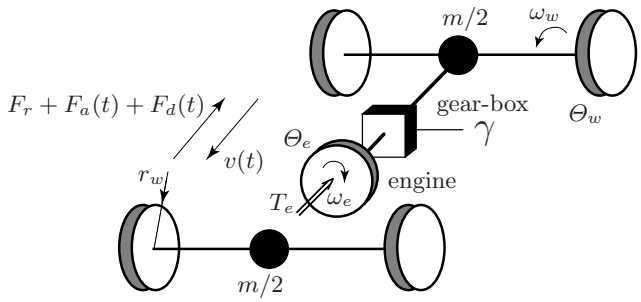



**Fig. 2.8.** Simplified vehicle structure, vehicle mass $m$ (including wheel masses) concentrated in the two axles.

Again, the first four modeling steps for the "passenger car" illustrated in Figure 2.8 are listed below.

[2] "Holonomic" means that the constraints are expressible as smooth functions of only the coordinates and possibly the time. If these functions are not smooth (e.g., hard bounds for an otherwise arbitrarily moving part), or if the constraints are imposed by restrictions on the velocities (e.g., a bicycle's wheels moving on a plane), then the problem is nonholonomic.

**Assumptions**:

1. The clutch is engaged such that the gear ratio $\gamma$ is constant.

2. No drivetrain elasticities and wheel-slip effects need be considered, i.e., the following relations are valid: $v(t) = r_w\omega_w(t)$ and $\omega_w(t) = \gamma\omega_e(t)$.

3. The vehicle must overcome rolling friction (force acting on vehicle is given by $F_r = c_r mg$) and aerodynamic drag ($F_a(t) = 1/2\rho c_w A v^2(t)$).

4. All other forces are lumped into an unknown disturbance $F_d(t)$.

5. The kinetic energy of a moving part is $1/2mv^2$ (pure translation) and $1/2\Theta\omega^2$ (pure rotation); no potential energy effects need be considered (even road).

*Step 1:*

The system's input is the engine torque $T_e$, which can be assumed to be arbitrarily controllable. In fact, the time delay caused by the engine dynamics is small compared to the typical time constants of the vehicle speed dynamics. Therefore, the engine dynamics are assumed to be a type a) phenomenon, see Figure 2.5. The system output is a signal $y(t)$ that is proportional to the vehicle speed $v(t)$.

*Step 2:*

The relevant reservoirs are the kinetic energies stored in the vehicle's translational and rotational degrees of freedom, i.e.,

$$E_{tot} = \frac{1}{2}mv^2(t) + 4\frac{1}{2}\Theta_w\omega_w^2(t) + \frac{1}{2}\Theta_e\omega_e^2(t). \tag{2.17}$$

Notice that the vehicle mass $m$ includes the mass of the engine flywheel and the wheels.

Using Assumption 2, Equation (2.17) can be simplified to

$$E_{tot} = \frac{1}{2}\left[m + \frac{4\Theta_w}{r_w^2} + \frac{\Theta_e}{\gamma^2 r_w^2}\right] v^2(t). \tag{2.18}$$

Obviously, the "level variable" is the vehicle speed $v(t)$.

*Step 3:*

The "flows" acting on the system are the mechanical powers affecting the system, i.e.,

$$P_+(t) = T_e(t)\omega_e(t), \quad \text{and} \quad P_-(t) = (F_r + F_a(t) + F_d(t))v(t) \tag{2.19}$$

The differential equation is therefore found by

$$\frac{d}{dt}E_{tot}(t) = P_+(t) - P_-(t). \tag{2.20}$$

Combining Equations (2.18), (2.19) and (2.20), an explicit form can be found

$$\frac{1}{2}\left[m + \frac{4\Theta_w}{r_w^2} + \frac{\Theta_e}{\gamma^2 r_w^2}\right] 2v(t)\frac{d}{dt}v(t) = T_e(t)v(t)\frac{1}{r_w\gamma} - (F_r + F_a(t) + F_d(t))v(t). \tag{2.21}$$

*Step 4:*

This step consists simply of using Assumption 3 and rearranging (2.21) to obtain

$$M(\gamma, m)\frac{d}{dt}v(t) = \frac{1}{r_w\gamma}T_e(t) - (c_r mg + \frac{1}{2}\rho c_w A v^2(t) + F_d(t)) \tag{2.22}$$

where

$$M(\gamma, m) = \left[m + \frac{4\Theta_w}{r_w^2} + \frac{\Theta_e}{\gamma^2 r_w^2}\right] \tag{2.23}$$

is the system's gear-ratio-dependent total inertia.

**Comments:**

The assumption that the engine torque is the *input* to the system is, of course, physically not correct:

1. the real system input is the throttle angle $\alpha_t$, as shown in Figure 2.9;
2. the *static* engine torque depends on the engine speed and on the throttle angle in a nonlinear way $T_e = f(\omega_e, \alpha_t)$, as shown in Figure 2.10; and
3. the *dynamic* engine torque reacts with some delay to changes in the throttle position mainly due to the intake manifold dynamics and to induction-to-powerstroke delays.

However, the nonlinearity can be compensated for by a static "inverse map," as shown by gray lines in Figure 2.9. In this case, $T_e(t) \approx T_{e,des}(t)$ and the dynamic effects are, as mentioned, small such that they can be dealt with later by designing a robust control system.

The transmission efficiency has been assumed to be 100%, which, of course, is again not realistic. A more realistic approach is to assume an affine input/output torque relation

$$T_{out} = \eta(\gamma)T_{in} - T_0(\gamma) \tag{2.24}$$

where the efficiency $\eta(\gamma)$ (in the order of 95% to 97% for a cog-wheel gearbox) and the idling losses $T_0(\gamma)$ (in the order of a few percent of the maximum torque) depend on the actual gear ratio $\gamma$.

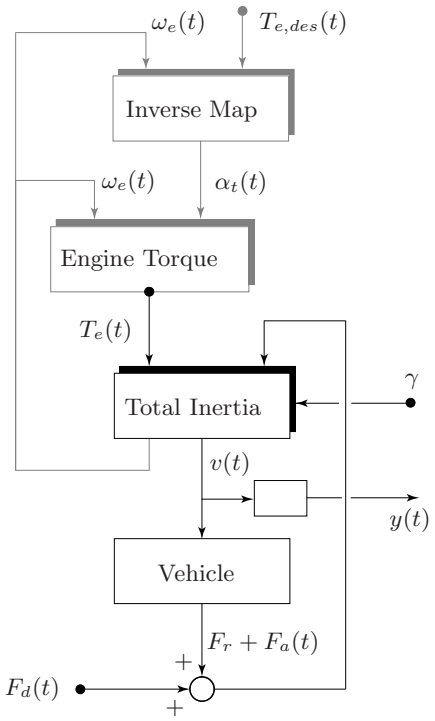



**Fig. 2.9.** Causality diagram of the longitudinal system dynamics, see "Comments" in the text for the gray part.

**Remark:** The last example might give the impression that using the energy reservoir paradigm is a "detour," and that a direct formulation such as Euler equations, is more efficient. In simple cases, this can be true. The benefits of the energy-based approach, however, will become evident when more complex systems must be modeled.

The next complication arises when both translations and rotations must be considered. However, since the body is assumed to have only one DOF, there must be only one special coordinate $q(t)$ upon which all Cartesian variables depend, i.e.,

$$T_t(t) = \frac{1}{2} m \left[ v_{x,cg}^2(q(t), \dot{q}(t)) + v_{y,cg}^2(q(t), \dot{q}(t)) \right] \tag{2.25}$$

$$T_r(t) = \frac{1}{2} \Theta \, \omega^2(q(t), \dot{q}(t)) \tag{2.26}$$

for (2.12) and (2.13), respectively, and

$$U(t) = U(x(q(t)), y(q(t))) \tag{2.27}$$

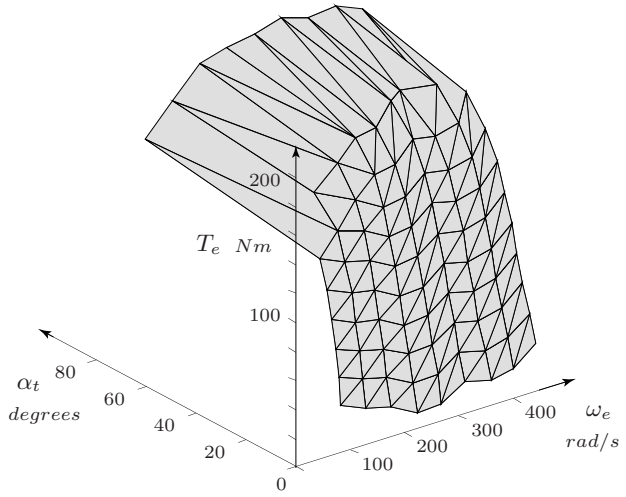



**Fig. 2.10.** Static relation between engine speed, throttle-valve opening angle and engine torque (two-liter SI engine).

for the potential energy (2.14). This approach is detailed in the following example. Notice that it is quite important to choose a suitable coordinate $q(t)$ to avoid cumbersome formulations. In some cases this choice is not obvious and some iterations might be be necessary.

*Example 2.5 (Nonlinear Pendulum).*

The example analyzed below is a rod that is in the earth's gravitational field and that is connected to a linear spring (see Figure 2.11). This inverted pendulum has one DOF and its model can be derived using the same approach used in the previous example, i.e., by using the total energy of the system as the conserved quantity. However, in addition to kinetic energies, in this case also potential energies will be present.

*Step 1:*

The pendulum is assumed to be a thin cylinder and to have a uniform density distribution. The spring slides without friction along the vertical axis. Zero friction is assumed in the pendulum's bearing as well. No external forces act on the system. Accordingly, the total energy will remain constant.

*Step 2:*

The relevant reservoirs are the kinetic and the potential energy of the pendulum and the potential energy stored in the spring. The corresponding level variables are the angular velocity $\omega = \dot{\varphi}$ and the angular position $\varphi$. To be more specific, the kinetic energy is given by

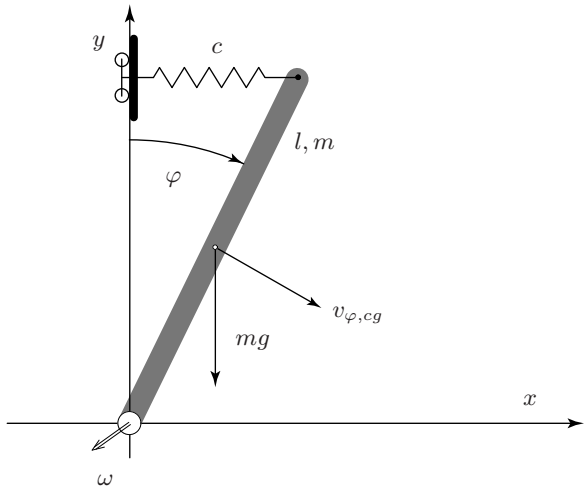



**Fig. 2.11.** Nonlinear pendulum.

$$T(t) = T_{trans}(t) + T_{rot}(t) = \frac{1}{2}m\left[\frac{l\dot{\varphi}(t)}{2}\right]^2 + \frac{1}{2}\frac{ml^2}{12}\dot{\varphi}^2(t) \tag{2.28}$$

and the potential energy

$$U(t) = U_{grav}(t) + U_{spring}(t) = mg\frac{l}{2}(\cos(\varphi) - 1) + \frac{1}{2}c(l\sin(\varphi))^2 \tag{2.29}$$

*Step 3:*
The total energy is

$$E(t) = T_{trans}(t) + T_{rot}(t) + U_{grav}(t) + U_{spring}(t) \tag{2.30}$$

and the energy conservation law yields

$$\frac{d}{dt}E(t) = 0 \tag{2.31}$$

*Step 4:*
After some straightforward algebraic manipulations, the following system equation is obtained

$$\frac{1}{3}ml^2\ddot{\varphi}(t) = \left[\frac{l}{2}mg - cl^2\cos(\varphi(t))\right]\sin(\varphi(t)) \tag{2.32}$$

### Linked Multiple Rigid Body with Multiple DOF

The general formulation for bodies in three dimensions (e.g., models of the attitude dynamics of satellites) is not straightforward. Fortunately, many technical problems are much easier to solve in particular if the motion of the system to be modeled is restricted to a plane. The problems are particularly easy when the single bodies are linked by forces acting on them. This situation is explored in the next example.

*Example 2.6 (Elastic Rotor).*

The system illustrated in Figure 2.12 can be seen as (a strongly simplified) model of the rotor of a gas turbine and generator system. The two masses at the end of the shaft model the compressor/generator inertia $\Theta_1$ and the turbine inertia $\Theta_2$. The two torques are the (driving) turbine torque $T_2$ and the (braking) compressor/generator torque $T_1$.

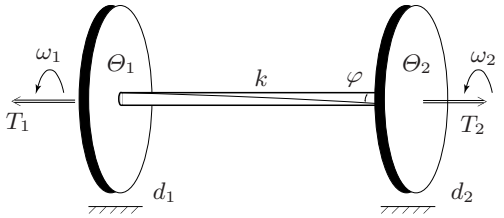



**Fig. 2.12.** Elastic rotor with a shaft of zero mass and elasticity $k$, the mass being concentrated in the inertias at either end of the shaft. The coefficients $d_1$ and $d_2$ parametrize the friction losses.

*Step 1:*
The system inputs are the torques $T_1$ and $T_2$; the system output is the rotor speed at the compressor/generator side ($\omega_1$).

*Step 2:*
The system has three reservoirs:

- the kinetic energy of the compressor/generator inertia (parameter $\Theta_1$);
- the kinetic energy of the turbine inertia (parameter $\Theta_2$); and
- the potential energy stored in the elasticity of the shaft (parameter $k$).

*Step 3:*
The conservation laws for the compressor/generator inertia are

$$\frac{d}{dt}\left(\frac{1}{2}\theta_1\omega_1^2(t)\right) = -P_{mech,1}(t) - P_{mech,2}(t) + P_{mech,3}(t) \tag{2.33}$$

and for the turbine inertia

$$\frac{d}{dt}\left(\frac{1}{2}\theta_2\omega_2^2(t)\right) = -P_{mech,4}(t) - P_{mech,5}(t) + P_{mech,6}(t) \tag{2.34}$$

The shaft's potential energy conservation law is

$$\frac{d}{dt}\left(\frac{1}{2}k\varphi^2(t)\right) = -P_{mech,3}(t) + P_{mech,4}(t) \tag{2.35}$$

*Step 4:*
The mechanical powers are given by

$$\begin{aligned} P_{mech,1} &= T_1 \cdot \omega_1, \\ P_{mech,2} &= \omega_1 \cdot d_1 \cdot \omega_1, \\ P_{mech,3} &= \omega_1 \cdot k \cdot \varphi, \\ P_{mech,4} &= \omega_2 \cdot k \cdot \varphi, \\ P_{mech,5} &= \omega_2 \cdot d_2 \cdot \omega_2, \\ P_{mech,6} &= T_2 \cdot \omega_2 \end{aligned} \tag{2.36}$$

where $P_{mech,2}$ and $P_{mech,5}$ approximate the friction losses occurring in the bearings (parameters $d_1$ and $d_2$ are the corresponding coefficients).

*Step 5:*
Equations (2.33) and (2.34) can be simplified to obtain the well-known Euler equations

$$\theta_1\frac{d}{dt}\omega_1(t) = -T_1(t) - d_1 \cdot \omega_1(t) + k \cdot \varphi(t) \tag{2.37}$$

$$\theta_2\frac{d}{dt}\omega_2(t) = T_2(t) - d_2 \cdot \omega_2(t) - k \cdot \varphi(t) \tag{2.38}$$

and the "kinematic" relation

$$\frac{d}{dt}\varphi(t) = \omega_2(t) - \omega_1(t) \tag{2.39}$$

A third-order system of ODE results. The ODE are linear and, thus, represent the simplest case.

When multiple degrees of freedom and some constraints are present in a system (e.g., several rigid bodies linked to form a mechanism, say a robot), then these constraints reduce the system's degrees of freedom and a smaller set of *generalized coordinates* can be used that describe the system's degrees of freedom. Physical insight is important for the identification of the correct

generalized coordinates. If such information is available, then a Lagrangian formulation is often the best choice.

The Lagrange function for a holonomic and scleronomic[3] mechanical system is defined as the difference between the total kinetic and the potential energy

$$L(q, \dot{q}) = T(q, \dot{q}) - U(q) \tag{2.40}$$

where the variables $q = [q_1, \ldots, q_n]$ represent the system's generalized coordinates (completely describing the $n$ degrees of freedom available) and $\dot{q} = [\dot{q}_1, \ldots, \dot{q}_n]$ are the corresponding generalized velocities. As in the unconstrained case, the energies $T_j$ and $U_j$ of the $j = 1, \ldots, m$ rigid bodies are defined by Equations (2.12), (2.13) and (2.14). The holonomic constraints are satisfied by expressing the physical coordinates $x$ and $y$ as functions of the generalized coordinates, i.e.,

$$T(t) = \sum_{j=1}^{m} T_j(x(q(t)), y(q(t)), \dot{x}(q(t), \dot{q}(t)), \dot{y}(q(t), \dot{q}(t))) \tag{2.41}$$

$$U(t) = \sum_{j=1}^{m} U_j(x(q(t)), y(q(t))) \tag{2.42}$$

According to the theory of Lagrangian mechanics, the system dynamics can be described by

$$\frac{d}{dt}\left\{\frac{\partial L}{\partial \dot{q}_k}\right\} - \frac{\partial L}{\partial q_k} = Q_k, \quad k = 1, \ldots, n \tag{2.43}$$

where $Q_k$ represents the $k^{th}$ "generalized force" (forces and torques) acting on the $k^{th}$ generalized coordinate $q_k$.

The most important advantage of the Lagrangian formulation is that the action of the binding forces (that arise due to the constraints) does not appear in the system equations. This is the consequence of the d'Alembert's *virtual energy principle* which states that the binding forces produce zero work for all admissible ("virtual") displacements of the systems.

*Example 2.7 (Nonlinear Pendulum on a Cart).*

The system to be modeled is shown in Figure 2.13. It has two rigid bodies (the cart and the pendulum), i.e., $m = 2$ and two DOF (the motion along the straight and horizontal rail and the rotation around the bearing on the cart), i.e., $n = 2$. Accordingly, the first generalized coordinate is $q_1 = y$ the distance from the origin. The second generalized coordinate is $q_2 = \varphi$ the

[3] "Scleronomic" means that the constraining functions do not vary in time.

angle of rotation of the pendulum. The generalized force associate with $q_1$ is a force acting horizontally on the cart. This force is considered to be the control input $u(t)$. The generalized force associated with $q_2$ is a torque acting on the pendulum. In this example this torque is assumed to be zero all of the time.

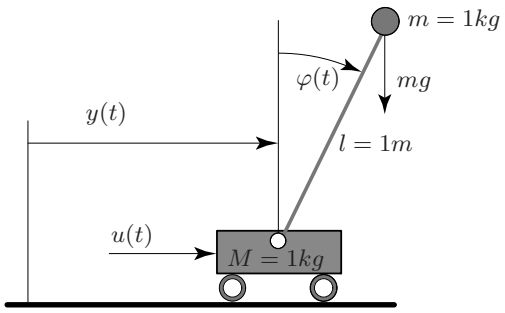



**Fig. 2.13.** Pendulum on a cart, $u(t)$ is the force acting on the cart ("input"), $y(t)$ the distance of the cart to an arbitrary but constant origin, and $\varphi(t)$ the angle of the pendulum.

To simplify the discussion, friction is neglected and the pendulum is assumed to consist of a zero-mass beam and a point mass concentrated at its upper end (see Example 2.5 for the more general case). For the (normalized) parameter values of Figure 2.13, the kinetic and potential energies of the system are given by

$$T_1(t) = \tfrac{1}{2}\dot{y}^2(t) \qquad U_1(t) = 0$$

$$T_2(t) = \tfrac{1}{2}\dot{y}^2(t) + \cos(\varphi(t))\dot{\varphi}(t)\dot{y}(t) + \tfrac{1}{2}\dot{\varphi}^2(t) \qquad U_2(t) = \cos(\varphi)g$$

where the potential energy has been normalized to zero at the bearing center.

Inserting these expressions into the Lagrange equations (2.43), the following equations are obtained

$$2\ddot{y}(t) + \ddot{\varphi}(t)\cos(\varphi(t)) - \dot{\varphi}^2(t)\sin(\varphi(t)) = u(t) \tag{2.44}$$

$$\ddot{\varphi}(t) + \cos(\varphi(t))\ddot{y}(t) - g\sin(\varphi(t)) = 0 \tag{2.45}$$

Solving these coupled equations for the two second derivatives of $y(t)$ and $\varphi(t)$ yields

$$\ddot{y}(t) = \frac{\dot{\varphi}(t)^2\sin(\varphi(t)) - g\cos(\varphi(t))\sin(\varphi(t)) + u(t)}{2 - \cos^2(\varphi(t))} \tag{2.46}$$

$$\ddot{\varphi}(t) = \frac{2g\sin(\varphi(t)) - \dot{\varphi}(t)^2\cos(\varphi(t))\sin(\varphi(t)) - \cos(\varphi(t))u(t)}{2 - \cos^2(\varphi(t))} \tag{2.47}$$

**Remark 1:** The mathematical problem encountered in the last example is generic. In fact, using the Lagrangian formalism, in general produces equations of the form

$$M(q(t))\ddot{q}(t) = f(q(t), \dot{q}(t), u(t)), \qquad q = [q_1, \ldots, q_n] \tag{2.48}$$

where $M(.)$ is an $n \times n$ "mass matrix." It can be shown that for the class of mechanical systems analyzed here, the matrix $M(.)$ is always regular (to be more precise $M(.) = M(.)^T > 0$) and, therefore, invertible. Hence, an explicit formulation of the system's dynamic equations is always possible

$$\ddot{q}(t) = M^{-1}(q(t))f(q(t), \dot{q}(t), u(t)) \tag{2.49}$$

**Remark 2:** The original energy-based approach can be generalized to what is known as a Hamiltonian formulation. The Hamiltonian of a classical mechanical system is simply the total energy as defined in (2.15). The coordinates used in the Hamiltonian formalism are the generalized coordinates $q$ (as in the Lagrangian approach) and the *generalized momentum* $p$ (i.e., not the velocities $\dot{q}$). For point masses $p = m\,\dot{q}$, a similar expression can be derived for a rigid body (translational and rotational momentum). The Hamiltonian formalism proves to be extendable to many other disciplines, especially infinite dimensional systems and quantum mechanics. More on these aspects can be found in [16].

**Remark 3:** For *nonholonomic* systems, the inclusion of the nonholonomic constraints into the kinetic and potential energies is not directly possible. Similar to the ideas used in constrained optimization (see Section 6.2), sometimes the problem can be solved by introducing additional variables ("Lagrange multipliers"). This procedure is explained for the case of $\nu < n$ *linear* differential constraints, i.e., for systems which have to satisfy the additional requirements

$$\alpha_j^T \dot{q}(t) = 0, \qquad j = 1, \ldots, \nu, \qquad \alpha_j^T = [\alpha_{j,1}, \ldots \alpha_{j,n}], \qquad \alpha_{j,k} \in I\!R \tag{2.50}$$

The Lagrangian is again given by Equation (2.40), but the differential equations are obtained by replacing Equation (2.43) by

$$\frac{d}{dt}\left\{\frac{\partial L}{\partial \dot{q}_k}\right\} - \frac{\partial L}{\partial q_k} - \sum_{j=1}^{\nu} \mu_j \alpha_{j,k} = Q_k, \qquad k = 1, \ldots, n, \tag{2.51}$$

where the nonholonomic constraints are included using the Lagrange multipliers $\mu_j$. The $n + \nu$ coupled Equations (2.50) and (2.51) must be solved for the unknowns $\ddot{q}_k$ and $\mu_j$, a process which typically also requires computing the time derivatives of the constraints (see Example 2.8).

*Example 2.8 (Ball on Wheel).*

The system to be analyzed in this example consists of a ball that must be kept on top of a wheel (see Figure 2.14).

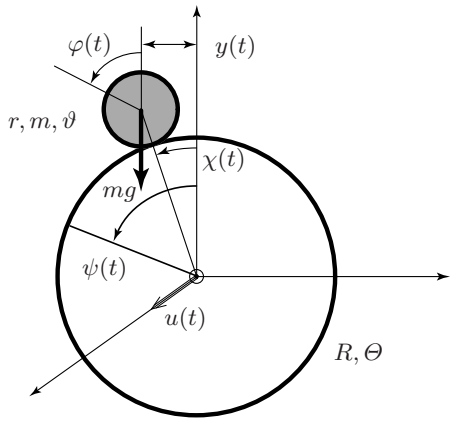



**Fig. 2.14.** Ball on a wheel system. Wheel: inertia around its center of gravity $\Theta$, radius $R$. Ball: mass $m$, inertia around its center of gravity $\vartheta$, radius $r$.

The system has three DOF

- the rotation of the wheel around its center, angle $\psi(t)$;
- the rotation of the ball around the center of the wheel, angle $\chi(t)$; and
- the rotation of the ball around its own center, angle $\varphi(t)$.

It is assumed that a "no-slip" condition is satisfied at all times, i.e., that

$$R\dot{\psi} - (R+r)\dot{\chi} + r\dot{\varphi} = 0 \tag{2.52}$$

*Step 1:*
The system input is the torque $u(t)$ to the wheel and the output the horizontal distance $y(t) = (R+r)\sin(\chi)$ of the center of the ball to the vertical reference axis.

*Step 2:*
The relevant reservoirs are the three kinetic energies

$$T(t) = \frac{1}{2}m(R+r)^2\dot{\chi}^2(t) + \frac{1}{2}\vartheta\dot{\varphi}^2(t) + \frac{1}{2}\Theta\dot{\psi}^2(t) \tag{2.53}$$

with the three level variables $\dot{\chi}$, $\dot{\varphi}$ and $\dot{\psi}$ and the single potential energy

$$U(t) = -mg(1-\cos(\chi))(R+r) \tag{2.54}$$

with the level variable $\chi$.

*Step 3:*

The formulation of the Lagrange equations in this nonholonomic case ($n = 3$, $\nu = 1$, $q_1 = \psi$, $q_2 = \chi$, $q_3 = \varphi$) is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} - \mu\alpha_k = Q_k \tag{2.55}$$

with $Q_1 = u(t)$, $Q_2 = Q_3 = 0$ and

$$\alpha_1\dot{q}_1 + \alpha_2\dot{q}_2 + \alpha_3\dot{q}_3 = 0 \tag{2.56}$$

with $\alpha_1 = R$, $\alpha_2 = -(R + r)$ and $\alpha_3 = r$.

*Step 4:*

After some straightforward steps involving Equations (2.55) and the *differentiation* of Equation (2.56), the following set of equations is obtained

$$\begin{aligned} \Theta\ddot{\psi} &= u + R\mu \\ m(R+r)^2\ddot{\chi} &= mg(R+r)\sin(\chi) - (R+r)\mu \\ \vartheta\ddot{\varphi} &= r\mu \\ R\ddot{\psi} &= (R+r)\ddot{\chi} - r\ddot{\varphi} \end{aligned} \tag{2.57}$$

This set of four equations define the four unknown variables $\{\ddot{\psi}, \ddot{\chi}, \ddot{\varphi}, \mu\}$, where $\{\ddot{\varphi}, \mu\}$ are easy to eliminate.

The remaining two equations then read

$$M \begin{bmatrix} \ddot{\psi} \\ \ddot{\chi} \end{bmatrix} = \begin{bmatrix} u \\ mg(R+r)\sin(\chi) \end{bmatrix} \tag{2.58}$$

with

$$M = \begin{bmatrix} \Theta + \vartheta\frac{R^2}{r^2} & -\vartheta\frac{R(R+r)}{r^2} \\ -\vartheta\frac{R(R+r)}{r^2} & m(R+r)^2 + \vartheta\frac{(R+r)^2}{r^2} \end{bmatrix} \tag{2.59}$$

After some steps one obtains

$$\det M = \frac{(R+r)^2\,\Gamma}{r^2} \tag{2.60}$$

where the scalar $\Gamma$ is given by

$$\Gamma = \Theta\vartheta \ + m(\vartheta R^2 + \Theta r^2) \tag{2.61}$$

Therefore, the mass matrix $M$ is regular for all physically meaningful parameter values such that its inversion is possible, yielding the final system equations

$$\begin{aligned}\ddot{\psi}(t) &= \left[(mr^2+\vartheta)u(t) + mgR\vartheta\sin(\chi(t))\right]/\Gamma \\ \ddot{\chi}(t) &= \left[\vartheta R u(t) + (\Theta r^2 + \vartheta R^2)mg\sin(\chi(t))\right]/[\Gamma(r+R)]\end{aligned} \tag{2.62}$$

This example is particularly interesting because it has some special structural features, i.e., a substantial part of the system dynamics is unobservable/uncontrollable. This will be analyzed in some detail in Section 4.4.

### 2.4.2 Hydraulic Systems

In general, hydraulic systems are described by partial differential equations (Navier-Stokes equations). However, for control purposes, simpler formulations of the specific technical device to be analyzed are necessary. For that purpose, a network of basic elements is formulated using lumped parameter building blocks (valves, ducts, compressibility nodes, etc.). Examples of such descriptions are presented below.

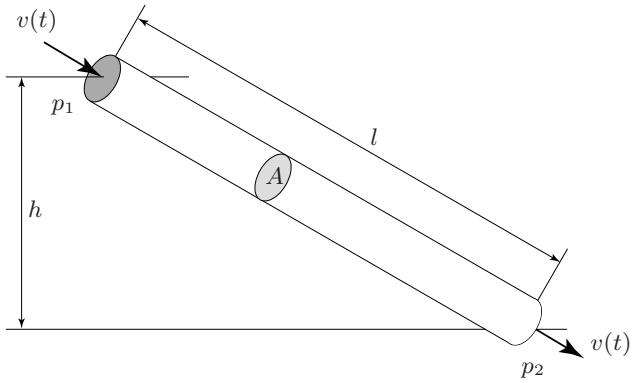



**Fig. 2.15.** Water duct in gravitational field.

#### Hydraulic Ducts

A typical element of a hydraulic network is shown in Figure 2.15. This hydraulic duct connects two nodes of known pressure $p_1$ and $p_2$. In the following, this tube is assumed to have a constant diameter (the extension to varying diameters can be made by pasting together several piecewise constant tube sections). The velocity of the fluid flowing through the tube is given by the momentum conservation law

$$m\frac{d}{dt}v(t) = \rho l A\frac{d}{dt}v(t) = A(p_1(t) - p_2(t)) + A\rho gh - F_f(t) \tag{2.63}$$

The density $\rho$ is assumed to be known and constant (incompressible fluid) and the friction force acting on the fluid may be approximated by

$$F_f(t) = A \cdot \lambda(v(t)) \cdot \frac{l}{d} \cdot \frac{\rho}{2} \cdot sgn(v(t)) \cdot v^2(t) \tag{2.64}$$

The parameter $\lambda(v)$ depends on the fluid velocity and on the material properties of the tube walls (numerical values are listed in "Moody diagrams").

Although the compressibility of a liquid fluid is very small, gas bubbles or elastic walls can lead to substantial elastic effects. These effects can be described using a lumped-parameter approach. The key idea of such a building block is illustrated in Figure 2.16. The fluid flowing into the compressibility element, in general, is not equal to the fluid leaving it, such that the upper lid of the element moves and, hence, a volume change $\Delta V(t)$ with respect to the initial (unloaded) volume $V_0$ takes place. Due to the elasticity of the compressibility element, this change in volume produces a concomitant change of the pressure.

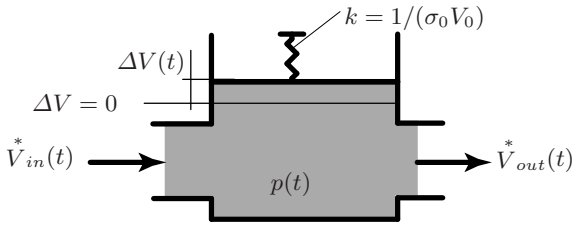



**Fig. 2.16.** Lumped-parameter element used to include compressibility into the hydraulic network.

These qualitative arguments can be made precise using the following mathematical description. For the volume, the conservation law

$$\frac{d}{dt}V(t) = \overset{*}{V}_{in}(t) - \overset{*}{V}_{out}(t) \tag{2.65}$$

must be satisfied and the compressibility effects may be approximated by

$$p(t) = \frac{1}{\sigma_0} \cdot \frac{\Delta V(t)}{V_0}, \qquad \Delta V(t) = V(t) - V_0 \tag{2.66}$$

The compressibility constant $\sigma_0$ and the nominal volume $V_0$ must be determined experimentally for the situation at hand. A combination of ducts and compressibility elements is used to approximate the behavior of long pipes.

### Hydraulic Turbines

In general, hydraulic turbines are simpler to model than gas turbines because the working fluid is essentially incompressible (cavitation effects are neglected here). A simple but useful model of a Pelton turbine is described below. Similar ideas can be formulated for Francis or Kaplan turbines.

Figure 2.17 shows the system structure[4] and the main variables. The water jet flows through a nozzle which converts potential energy (pressure) into kinetic energy. The nozzle contains a movable pin such that the massflow can be modulated for control purposes. The Pelton turbine turns at a speed which is assumed to be constant for the time being. The power transfer from the water to the turbine is accomplished by momentum exchange.

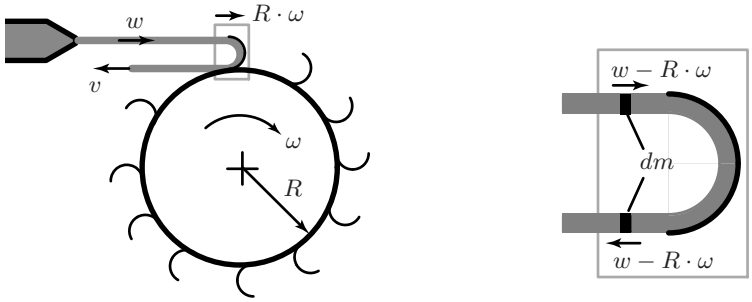



**Fig. 2.17.** Pelton turbine: definition of system variables (left) and moving momentum-exchange control volume (right).

The momentum transfer from the water jet to the Pelton turbine wheel takes place in the moving control volume shown on the right side of Figure 2.17. The momentum balance for a mass element $dm$ before and after the Pelton turbine is

$$dB = 2(w - R \cdot \omega) \cdot dm \tag{2.67}$$

where the mass element is defined by the volume flow $\overset{*}{V}$ and the density $\rho$ as

$$dm = \overset{*}{V} \cdot \rho \cdot dt \tag{2.68}$$

Therefore, the mean force acting on the Pelton turbine is given by

$$F_T = \frac{dB}{dt} = 2\,(w - R \cdot \omega) \cdot \frac{dm}{dt} = 2\,\rho \cdot (w - R \cdot \omega) \cdot \overset{*}{V} \tag{2.69}$$

[4] In reality, several 10 cups are distributed on a Pelton wheel and three to five nozzles are typically acting on the wheel. In the modeling approach shown here it is assumed that a large number of small cups is distributed on the Pelton wheel, i.e., a continuous momentum exchange is assumed.

and the resulting wheel torque is

$$T_T = 2\,\rho \cdot R \cdot (w - R \cdot \omega) \cdot \overset{*}{V} \tag{2.70}$$

The velocity of the water after leaving the cup of the Pelton turbine is equal to $v = w - 2R \cdot \omega$, i.e., if the Pelton turbine is at rest, the water is reverted without any losses (no friction or hydraulic losses have been modeled yet); and if the turbine turns at $R \cdot \omega = w$ (i.e., the tangential speed of a cup is the same speed as the speed of the water jet), the water jet is not decelerated. In both cases, no net power $T_T \cdot \omega$ is transferred from the water jet to the Pelton wheel. The maximum power transfer occurs at a wheel speed of $R \cdot \omega = 1/2w$, at which speed the water is completely decelerated to $v = 0$, i.e., all of its kinetic energy is transferred to the wheel.

### 2.4.3 Electromagnetic Systems

Electromagnetic systems often can be formulated as RLC networks, i.e., all electromagnetic phenomena are concentrated in resistances, inductances and capacitances (this corresponds to the usual "lumped-parameter" approximation).

Two classes of reservoir elements are important in this context:

- magnetic energy, stored in magnetic fields $B$; and
- electric energy, stored in electric fields $E$.

A systematic mathematical representation is given in the following table.

**Table 2.1: Linear electric elements.**

| | capacitance (electric) | inductance (magnetic) |
|---|---|---|
| energy | $W_E = \frac{1}{2}C \cdot U^2(t)$ | $W_M = \frac{1}{2}L \cdot I^2(t)$ |
| level variable | $U(t)$ (voltage) | $I(t)$ (current) |
| conservation law | $C \cdot \frac{d}{dt}U(t) = I(t)$ | $L \cdot \frac{d}{dt}I(t) = U(t)$ |

The parameters $C$ and $L$ are determined by the geometric and material properties of the corresponding elements. To confirm the equations shown in the last row, an energy balance can be used, e.g., for the capacitance

$$\frac{d}{dt}W_E(t) = P_E = U(t) \cdot I(t) \tag{2.71}$$

(the term $P_e = U(t) \cdot I(t)$ describes the total electric power flowing through the capacitance). Differentiating $W_E$ and using this energy balance equation yields the result shown in the last line of Table 2.4.3.

$$\frac{1}{2}C \cdot 2 \cdot U(t) \cdot \frac{d}{dt}U(t) = U(t) \cdot I(t) \quad \Rightarrow \quad C \cdot \frac{d}{dt}U(t) = I(t) \tag{2.72}$$

(similarly for $W_M$).

Working with RLC networks, two rules ("Kirchhoff's laws") have been developed that are useful:

1: The algebraic sum of all currents in each network node is zero.
2: The algebraic sum of all voltages following a closed network loop is zero.

These rules are equivalent to the energy balance, but more convenient to use. The following example shows the general procedure.

*Example 2.9 (Electric Oscillator).*

Electric oscillators are used in many devices, for instance as filters in signal-processing applications. The simplest possible oscillator is shown in Figure 2.18.

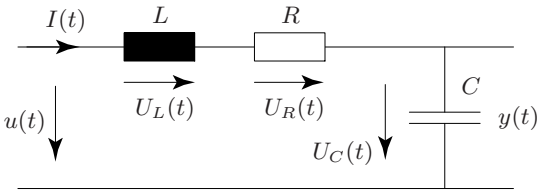



**Fig. 2.18.** RLC oscillator ($R$=resistance, $C$=capacitance, $L$=inductance).

Based on the definitions introduced in that figure, the following modeling steps are necessary.

Step 1: Input: $u(t)$, input voltage
Output: $y(t)$, output voltage

Step 2: Two relevant reservoirs, magnetic energy in $L$ and electric energy in $C$

Step 3: Apply the second Kirchhoff rule: $U_L(t) + U_R(t) + U_C(t) = u(t)$

Step 4: Apply the "C" and "L laws": $U_L(t) = L \cdot \frac{d}{dt} I(t)$, $I(t) = C \cdot \frac{d}{dt} U_C(t)$
and Ohm's law: $U_R(t) = R \cdot I(t)$

Step 5: Definitions: $y(t) = U_C(t)$, $I(t) = \frac{d}{dt} Q(t)$
Reformulation: $I(t) = C \cdot \frac{d}{dt} y(t)$, $\frac{d}{dt} I(t) = C \cdot \frac{d^2}{dt^2} y(t)$
Result: $L \cdot C \cdot \frac{d^2}{dt^2} y(t) + R \cdot C \cdot \frac{d}{dt} y(t) + y(t) = u(t)$

### 2.4.4 Electromechanical Systems

In control systems, electromechanical devices play an important role since most actuators belong to this class. The fundamental law is the Lorentz law which states that a current $I$ flowing in a conductor with length $l$ in a magnetic field $B$ will be affected by a force $F$ according to

$$F = I \cdot (l \times B) \tag{2.73}$$

(the symbol "$\times$" is the vector product in $\mathbb{R}^3$, $l$ and $B$ are to be seen as vectors in $\mathbb{R}^3$).

Closely connected to that is Faraday's induction law which states that a voltage $U$ is induced in a conductor of length $l$ that moves with velocity $v$ in a homogeneous magnetic field $B$ according to the relation

$$U = -v \cdot (l \times B) \tag{2.74}$$

Two examples of such electromechanical systems are given below.

*Example 2.10 (Loudspeaker).*

The electromagnetic part consists of a coil with diameter $D$ and $N$ windings that is immersed in a homogeneous permanent magnetic field with strength $B$. Therefore, a current that is forced through this coil will produce a force that is given by

$$F(t) = B \cdot N \cdot D \cdot \pi \cdot I(t) = \kappa \cdot I(t) \tag{2.75}$$

An equation of this form will be called the *motor law* in this text.

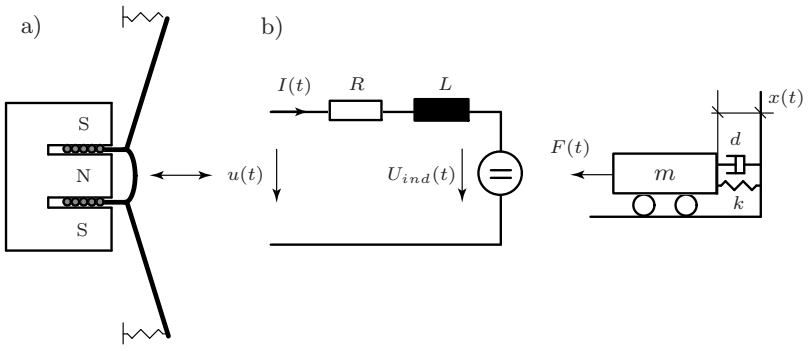



**Fig. 2.19.** Simplified loudspeaker, a) mechanical structure, b) electromechanical network.

This force couples the electromagnetic part with the mechanical part through the usual mechanical power balance. The loudspeaker's membrane can be modeled as a linear damped mass-spring system (where in reality the damping and the spring "constants" are depending on frequency and amplitude).

$$m \cdot \ddot{x}(t) = \kappa \cdot I(t) - k \cdot x(t) - d \cdot \dot{x}(t) \tag{2.76}$$

The electric part is essentially an RL network with the additional self-inductance term (2.74). The last term models the effects of the electric voltage induced in the coil when it oscillates inside the magnetic field, i.e.,

$$U_{ind}(t) = B \cdot n \cdot d \cdot \pi \cdot v(t) = \kappa \cdot v(t) \tag{2.77}$$

This term will be referred to as the *generator law*.

The electric part of the system analyzed can therefore be modeled by

$$L \cdot \frac{d}{dt} I(t) = -R \cdot I(t) - \kappa \cdot \dot{x}(t) + u(t) \tag{2.78}$$

The two coupled equations, (2.76) and (2.78), constitute a simple model of the loudspeaker system.

Most electric motors used in control loops are not linear but rotational. A classification of such motors is possible according to the commutation mechanisms used:

1. Classical DC drives have a mechanical commutation of the current in the rotor coils and constant (permanent magnet) or time-varying stator fields (external excitation).
2. Modern brushless DC drives have an electronic commutation of the stator current and permanent magnet on the rotor (i.e., no brushes).
3. AC drives have an electronic commutation of the stator current and use self-inductance to build up the rotor fields.

It turns out that the first two classes of motors can be modeled quite easily using the two laws, (2.73) and (2.74), while the third class is much more difficult to describe precisely and, thus, beyond the scope of this text.

*Example 2.11 (Permanently Excited DC Motor).*

A schematic representation of the system to be modeled is shown in Figure 2.20.

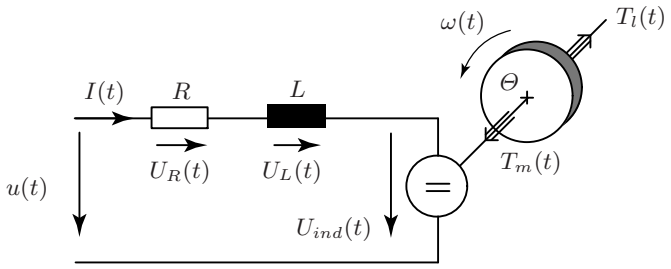



**Fig. 2.20.** Simplified DC motor structure.

*Step 1:*
The system input is the armature voltage $u(t)$ (control input) and the load torque $T_l(t)$ (disturbance). The system output is the measurement of the rotor speed $\omega(t)$. The motor is permanently excited such that the parameters $\kappa$ in the motor and generator law are constants (temperature dependencies and saturation effects are neglected). The mechanical part has linear viscous friction losses.

*Step 2:*
Two relevant reservoirs are present:

- the magnetic energy stored in the rotor coil, level variable $I(t)$;
- the kinetic energy stored in the rotor, level variable $\omega(t)$.

*Step 3:*
The two energy conservation laws yield:

$$\begin{aligned} L_A \cdot \tfrac{d}{dt} I(t) &= -R_A \cdot I(t) - U_{ind}(t) + u(t) \\ \Theta \cdot \tfrac{d}{dt} \omega(t) &= T_m(t) - T_l(t) - d \cdot \omega(t) \end{aligned} \tag{2.79}$$

*Step 4:*
The power flows can be found using the generator and the motor law

$$\begin{aligned} U_{ind}(t) \cdot I(t) &= \kappa \cdot \omega(t) \cdot I(t) \\ T_m(t) \cdot \omega(t) &= \kappa \cdot I(t) \cdot \omega(t) \end{aligned} \tag{2.80}$$

Notice that if compatible units are used and if the electromechanical conversion has no additional[5] losses, then the same constant $\kappa$ must appear in the motor and generator law.

Inserting Equation (2.80) into Equation (2.79), the following final formulation of a DC motor model is obtained

$$\begin{aligned} L_A \cdot \tfrac{d}{dt} I(t) &= -R_A \cdot I(t) - \kappa \cdot \omega(t) + u(t) \\ \Theta \cdot \tfrac{d}{dt} \omega(t) &= \kappa \cdot I(t) - T_l(t) - d \cdot \omega(t) \end{aligned} \tag{2.81}$$

### 2.4.5 Thermodynamic Systems

First the most important thermodynamic concepts are recalled, see [13] for more details. After that, three examples are given that show the general approach for incompressible and compressible fluids.

The central notion in thermal systems is the internal energy $U$ ("heat," vibration energy of the molecules, ...) of a closed system. In general, the internal energy of a body of constant mass $m$ and with homogeneous temperature $\vartheta_1$ is given by

$$U(\vartheta_1) = m \int_0^{\vartheta_1} c_v(\vartheta) d\vartheta \tag{2.82}$$

where the integration base point "0" is an arbitrarily chosen reference value (e.g., the thermodynamic zero Kelvin point).

For incompressible systems (solids and fluids), the internal energy is often well approximated by

$$U(\vartheta_1) = c \cdot m \cdot \vartheta_1, \quad c = c_v = c_p \tag{2.83}$$

i.e., the specific heat at constant pressure and at constant volume are assumed to be identical and independent of the temperature (the latter assumption is also valid for compressible gases as long as the temperature variations are not too large).

Heat energy flows associated with mass transfer are denoted as enthalpy flows and are given by

[5] The ohmic losses in the electric part and the friction losses in the mechanical part are considered by the losses in the conservation laws.

$$\overset{*}{H}(t) = c_p(\vartheta(t)) \cdot \overset{*}{m}(t) \cdot \vartheta(t) \tag{2.84}$$

where the difference

$$R = c_p(\vartheta) - c_v(\vartheta) \tag{2.85}$$

is constant for all temperatures for *ideal gases* (an assumption that is valid as long as the gas is at relatively high temperatures and low pressures).[6]

In a closed and homogeneous system, the three thermodynamic states pressure $p$, temperature $\vartheta$ and density $\rho$ are not independent but are linked through an algebraic relation of the form

$$f(p, \rho, \vartheta) = 0 \tag{2.86}$$

For gases at low pressures and high temperatures, the ideal-gas law

$$p - \rho \cdot R \cdot \vartheta = 0 \tag{2.87}$$

is often a good approximation for $f(p, \rho, \vartheta) = 0$ (the gas constant $R$ has been defined above).

Heat transfer can occur with three basic mechanisms:

- Heat conduction in a body, for the one-dimensional case (thin cylinder of cross section area $A$ and length $l$) following the basic Fourier law

$$\overset{*}{Q} = \frac{\kappa \cdot A}{l} \cdot (\vartheta_1 - \vartheta_2) \tag{2.88}$$

  where the thermal conductivity $\kappa$ is a property of the conductor.

- Heat convection between a solid body with contact area $A$ and the surrounding gas, according to the Newton law

$$\overset{*}{Q} = k \cdot A \cdot (\vartheta_1 - \vartheta_2) \tag{2.89}$$

  where the heat transfer coefficient $k$ depends on the surface and the gas flow properties.

- Heat radiation from a body whose surface $A$ is at temperature $\vartheta_1$ to its surroundings at $\vartheta_2$, given by the Stefan-Boltzmann law

$$\overset{*}{Q} = \epsilon \cdot \sigma \cdot A \cdot (\vartheta_1^4 - \vartheta_2^4) \tag{2.90}$$

  where the emissivity $\epsilon$ is a property of the radiating surface and the parameter $\sigma$ the Stefan-Boltzmann constant.

[6] In this text, the symbol $R$ denotes the *specific* gas constant. The universal gas constant $\Re = 8.314$ J/mol K is related to $R$ by $\Re = R \cdot \mathcal{M}_\mathcal{A}$, where $\mathcal{M}_\mathcal{A}$ is the molar mass of the specific gas $A$ that is analyzed. As an example, for air $\mathcal{M}_\mathcal{A}$=0.029 kg/mol, the specific gas constant is $R = 8.314/0.029 = 287$ J/kg K.

These first principles are sufficient to model several thermal dynamic systems. The following three examples give a first impression of how to proceed.

*Example 2.12 (Stirred Tank).*

For the synthesis of various products, the chemical industry uses batch reactors, which are built as shown schematically in Fig 2.21. The mixer guarantees uniform thermal conditions inside the reactor, such that a lumped-parameter approach is valid.

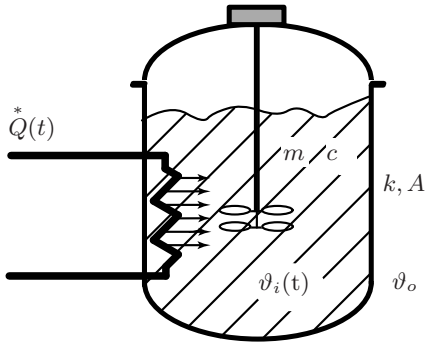



**Fig. 2.21.** Stirred reactor system.

The system parameters important in this example are:

| | | | |
|---|---|---|---|
| $\vartheta_i, \vartheta_o$ | = temperatures inside and outside of the reactor | K | |
| $m$ | = mass in the reactor | kg | |
| $c$ | = specific heat of the reactor liquid | J/(kg K) | (2.91) |
| $A$ | = active heat exchange surface of the reactor | $m^2$ | |
| $k$ | = heat transfer coefficient of the active surface | $W/(m^2\,K)$ | |

*Step 1:*
The reactor fluid has a uniform temperature $\vartheta_i$ distribution ("lumped parameter assumption"), and the temperature of the environment $\vartheta_o$ is constant.

The heat exchanger can impose an arbitrary heat flux (positive or negative) to the reactor's liquid. Heat flows through the reactor's poorly insulated wall.

The only relevant reservoir is the thermal (internal) heat stored in the reactor's liquid.

The reaction taking place inside the reactor is assumed to be neutral, i.e., no heat is generated or absorbed by the reactions inside the reactor (the general case is discussed in Section 2.4.7).

*Step 2:*
The level can be normalized to the temperature difference, i.e.,

$$\vartheta(t) = \vartheta_i(t) - \vartheta_o \tag{2.92}$$

such that the internal energy stored is given by

$$U(t) = m \cdot c \cdot \vartheta(t) \tag{2.93}$$

*Step 3:*
The energy balance yields the (only) differential equation

$$\frac{d}{dt}U(t) = m \cdot c \cdot \frac{d}{dt}\vartheta(t) = \overset{*}{Q}_{in}(t) - \overset{*}{Q}_{out}(t) \tag{2.94}$$

*Step 4:*
The heat flows are given by

$$\overset{*}{Q}_{in}(t) = u(t), \tag{2.95}$$

and

$$\overset{*}{Q}_{out}(t) = k \cdot A \cdot \vartheta(t) \tag{2.96}$$

*Step 5:*
The final formulation is then

$$m \cdot c \cdot \frac{d\vartheta(t)}{dt} = u(t) - k \cdot A \cdot \vartheta(t) \tag{2.97}$$

*Example 2.13 (Heat Exchanger I, ODE Formulation).*

Heat exchangers transfer heat between two fluids without bringing the two in contact. The most common technical solution is to separate the two fluids by a heat transfer layer, and to operate the system in "counter-flow conditions," see Figure 2.22. All combinations of fluids are possible (liquid/liquid, liquid/gas, gas/gas) and even phase changes are encountered (e.g., in boilers).

In the following, a liquid/liquid heat exchanger is considered, i.e., the fluids will be assumed to be incompressible and constant uniform massflows $\overset{*}{m}_1(t)$ and $\overset{*}{m}_2(t)$ will be assumed. The walls between the two fluids are assumed to be thin such that they do not introduce any additional relevant heat dynamics.

The heat exchanger can be modeled as a series connection of lumped parameter cells $j = 1, \ldots, j_{max}$, as shown in Figure 2.23. Each cell has the

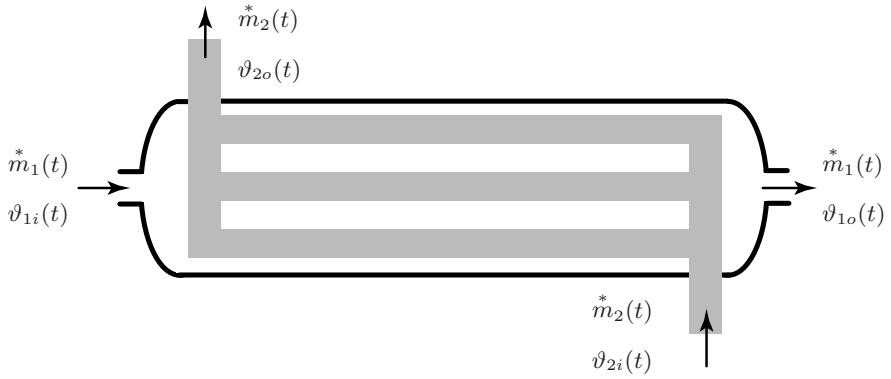



**Fig. 2.22.** Counter flow heat exchanger.

"length" $A_j$ (the active heat exchange contact area) and the heat transfer coefficient $k_j$. For the sake of simplicity, all cells are assumed to be identical (same size, heat transfer coefficient, mass, specific heat, etc.). However, the extension to nonuniform lengths or heat transfer characteristics is straightforward. It is assumed that the temperature $\vartheta_{1o,j/2o,j}$ of the fluids leaving the elements is the same as inside the elements (the usual lumped parameter assumption).

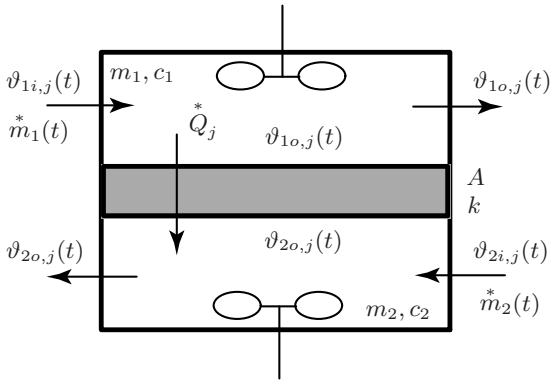



**Fig. 2.23.** Heat exchanger element, massflows and specific heat constants.

The three main equations governing a cell's behavior are the two energy balances

$$\frac{d}{dt}U_{1,j} = m_1 \cdot c_1 \cdot \frac{d\vartheta_{1o,j}(t)}{dt} = \overset{*}{m}_1 \cdot c_1 \cdot (\vartheta_{1i,j}(t) - \vartheta_{1o,j}(t)) - \overset{*}{Q}_j(t) \quad (2.98)$$

and

$$\frac{d}{dt}U_{2,j} = m_2 \cdot c_2 \cdot \frac{d\vartheta_{2o,j}(t)}{dt} = \overset{*}{m}_2 \cdot c_2 \cdot (\vartheta_{2i,j}(t) - \vartheta_{2o,j}(t)) + \overset{*}{Q}_j(t) \quad (2.99)$$

and the heat transfer equation

$$\overset{*}{Q}_j(t) = k \cdot A \cdot (\vartheta_{1o,j}(t) - \vartheta_{2o,j}(t)) \quad (2.100)$$

**Remark**: Sometimes the formulation

$$\overset{*}{Q}_j(t) = k \cdot A \cdot (\frac{\vartheta_{1i,j} + \vartheta_{1o,j}}{2} - \frac{\vartheta_{2i,j} + \vartheta_{2o,j}}{2}) \quad (2.101)$$

is used. This approach better captures the changing temperature profile inside a cell. However, it introduces algebraic loops that must be dealt with separately.

The causality diagram of one cell element is shown in Figure 2.24. Several cells can be stacked together to produce finite-element models of heat exchangers of any desired order.

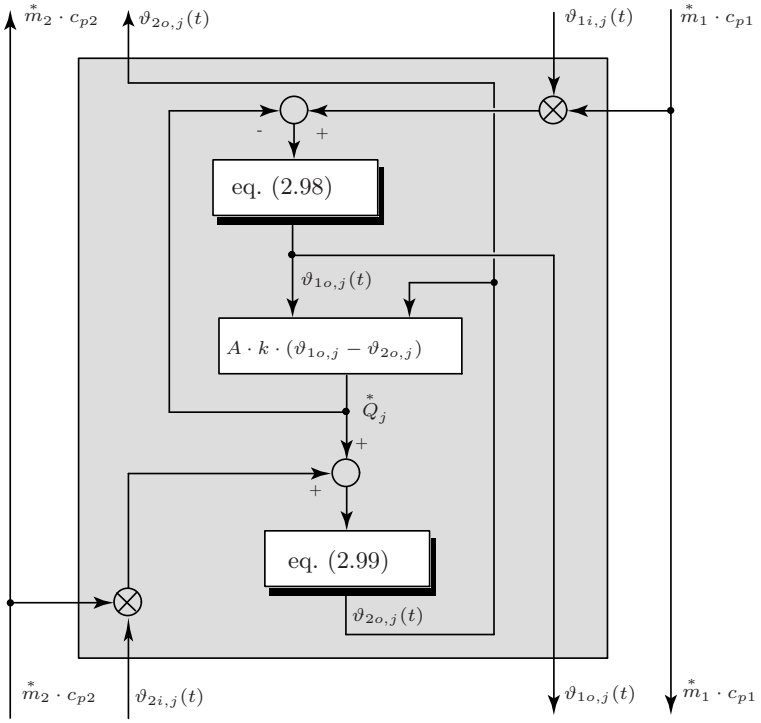



**Fig. 2.24.** Heat exchanger element causality, any desired number of elements can be stacked.

The static behavior of one cell element can be found by setting Equations (2.98) and (2.99) both to zero

$$\frac{\vartheta_{1o,j} - \vartheta_{1i,j}}{\vartheta_{1i,j} - \vartheta_{2i,j}} = -\left[1 + \frac{\overset{*}{m}_1 \cdot c_1}{k \cdot A} + \frac{\overset{*}{m}_1 \cdot c_1}{\overset{*}{m}_2 \cdot c_2}\right]^{-1} \tag{2.102}$$

and

$$\frac{\vartheta_{2o,j} - \vartheta_{2i,j}}{\vartheta_{1i,j} - \vartheta_{2i,j}} = \left[1 + \frac{\overset{*}{m}_2 \cdot c_2}{k \cdot A} + \frac{\overset{*}{m}_2 \cdot c_2}{\overset{*}{m}_1 \cdot c_1}\right]^{-1} \tag{2.103}$$

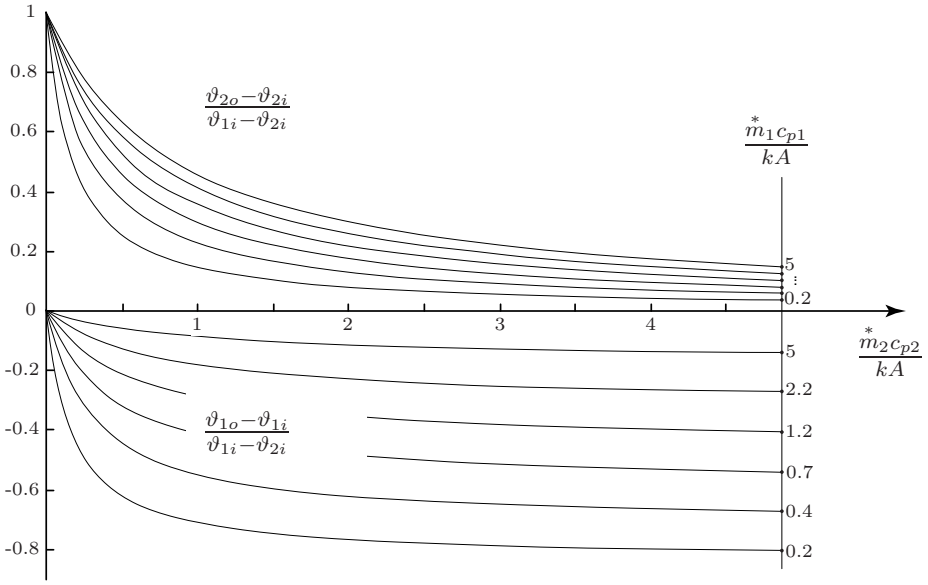



**Fig. 2.25.** Static behavior of one heat-exchanger element (see Figure 2.23) for the normalized massflows $\overset{*}{m}_1\, c_{p1}/(k\, A)$ (six different values) and $\overset{*}{m}_2\, c_{p2}/(k\, A)$ as x-axis. The temperatures $\vartheta_{2i} < \vartheta_{1i}$ are constant.

Figure 2.25 shows the static gains as functions of the normed massflows, i.e., the temperature equilibriums are nonlinear functions of the two massflows (which are assumed to be constant parameters). These relations can be applied serially for multiple cells, producing the total heat exchanger gain.

The correct number of cells can be selected only when experimental data is available. Nonlinear least squares methods will be used for the parameter identification problem (see Chapter 3). By shrinking the cell size to zero, a partial differential equation describing the heat exchanger can be obtained

(see Section 2.5). Notice that his formulation is not directly applicable to control problems and that the resulting solutions are not necessarily more precise than a lumped-parameter cell formulation.

*Example 2.14 (Gas Receiver).*

A more complex situation arises when the considered fluids are compressible and this case is discussed now using an example. The system analyzed below is illustrated in Figure 2.26. A *receiver* is a fixed volume for which the thermodynamic states (pressures, temperatures, composition, etc.) are assumed to be the same all over the volume ("lumped parameter system").

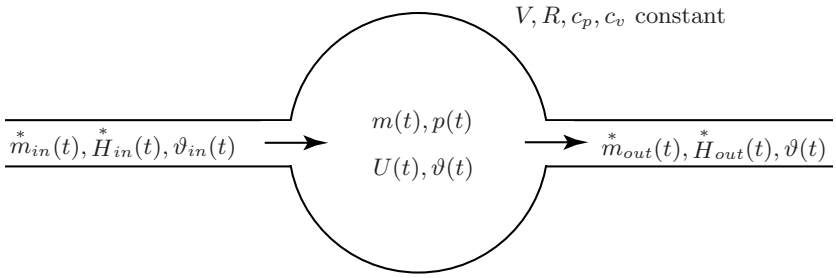



**Fig. 2.26.** Receiver of volume $V$ with compressible gas with constant specific heats $c_p$ and $c_v$, and gas constant $R$.

*Step 1:*

The inputs and outputs are the mass $\overset{*}{m}_{in/out}(t)$ and the enthalpy flows $\overset{*}{H}_{in/out}(t)$. Notice that the "lumped parameter" assumption requires that the outflowing gas has the same temperature as the gas inside the receiver.

*Step 2:*

Two reservoirs are to be considered in this case: one stores the internal energy $U(t)$ and the other the mass $m(t)$. The corresponding level variables are chosen to be the temperature $\vartheta(t)$ and the pressure $p(t)$.[7]

*Step 3:*

Assuming that no heat or mass transfer through the walls and that no substantial changes in potential or kinetic energy occur in the flow, then the following two (coupled) differential equations describe such a receiver

[7] There are other possibilities. However, engineering intuition suggests that particular choice.

$$\frac{d}{dt}U(t) = \overset{*}{H}_{in}(t) - \overset{*}{H}_{out}(t) \tag{2.104}$$

and

$$\frac{d}{dt}m(t) = \overset{*}{m}_{in}(t) - \overset{*}{m}_{out}(t) \tag{2.105}$$

*Step 4:*

Using the ideal-gas law (2.87) and the caloric relations (2.83) and (2.84) the two equations (2.104) and (2.105) can be reformulated as (the time dependencies are omitted for space reasons)

$$\frac{d}{dt}\vartheta = \frac{\vartheta \cdot R}{p \cdot V \cdot c_v} \cdot \left\{ c_p \cdot \overset{*}{m}_{in} \cdot \vartheta_{in} - c_p \cdot \overset{*}{m}_{out} \cdot \vartheta - (\overset{*}{m}_{in} - \overset{*}{m}_{out}) \cdot c_v \cdot \vartheta \right\} \tag{2.106}$$

and

$$\frac{d}{dt}p(t) = \frac{\kappa \cdot R}{V} \cdot \left\{ \overset{*}{m}_{in}(t) \cdot \vartheta_{in}(t) - \overset{*}{m}_{out}(t) \cdot \vartheta(t) \right\} \tag{2.107}$$

**Remark 1:** Notice that it is not mandatory to transform the equations (2.104) and (2.105) to the forms (2.106) and (2.107). It is, of course, also possible to keep the original reservoir formulation and to derive the system's pressure and temperature using the two equations (2.83) and (2.84). However, since only $p$ and $\vartheta$ are measurable, (2.106) and (2.107) are often preferred when describing the system dynamics.

**Remark 2:** In Example 2.14, an *adiabatic* (no heat-exchange) receiver was assumed (good insulation, large volume-to-surface ratio, rapid gas-exchange processes, etc.). The other extreme situation is to assume *isothermal* conditions, where large heat exchange areas keep the gas temperature inside the receiver constant (no insulation, small volume-to-surface ratio, long residence times of the gas inside the receiver, etc.). In this case, instead of equations (2.106) and (2.107) the dynamics of the receiver are described by

$$\frac{d}{dt}\vartheta(t) = 0 \tag{2.108}$$

and, using the ideal gas law and the isothermal assumption,

$$\frac{d}{dt}p(t) = \frac{R \cdot \vartheta}{V} \cdot \left\{ \overset{*}{m}_{in}(t) - \overset{*}{m}_{out}(t) \right\} \tag{2.109}$$

Reality is often in-between these two extreme cases and engineering intuition is necessary to decide if a mixed approach or one of the two extremes describes best the situation at hand.

### 2.4.6 Fluiddynamic Systems

#### Valves

The flow of fluids between two reservoirs is determined by "valves" whose inputs are the pressures up- and downstream. The difference between these two level variables drives the fluid in a nonlinear way through such orifices. Of course, this problem is at the heart of fluid dynamics such that a large amount of theoretical and practical knowledge exists.

For the purposes pursued in this text the two simplest formulations will suffice. The following assumptions are made:

- no friction effects in the flow (friction is accounted for using correction factors that are determined experimentally);
- no inertial effects in the flow (the mass of fluid around the valve is small compared to the mass stored in the receivers to which the valve is connected);
- completely insulated conditions (no additional energy, mass, ... enters the system); and
- all flow phenomena are "zero-dimensional," i.e., no spatial effects are considered.

Fluids often can be assumed to be *incompressible* (constant density), e.g., liquids or gases at low Mach numbers. In this case, applying Bernoulli's law, the flow through a valve can be modeled by

$$\overset{*}{m}(t) = c_d \cdot A \cdot \sqrt{2 \cdot \rho} \cdot \sqrt{p_{in}(t) - p_{out}(t) + \frac{1}{2}\rho\, v_{in}^2} \tag{2.110}$$

where the following definitions have been used:

- $\overset{*}{m}$: massflow through the valve;
- $c_d$: discharge coefficient (models flow restrictions, friction, etc.);
- $A$: open area of the valve;
- $\rho$: density of the fluid (constant);
- $v_{in}$: velocity of the fluid upstream of the valve;
- $p_{in}$: pressure of the fluid upstream of the valve; and
- $p_{out}$: pressure of the fluid downstream of the valve.

Quite often the kinetic energy term $\frac{1}{2}\rho\, v_{in}^2$ can be neglected compared to the pressure energy term $p_{in} - p_{out}$.

The most important and versatile flow-control block for *compressible* fluids is the *isenthalpic throttle*. When modeling this device, the key assumption is that the behavior of the flow may be separated into two distinct parts, i.e., first the acceleration part (pressure decreases), up to the narrowest point, during which the flow remains laminar, and second the deceleration part, where the

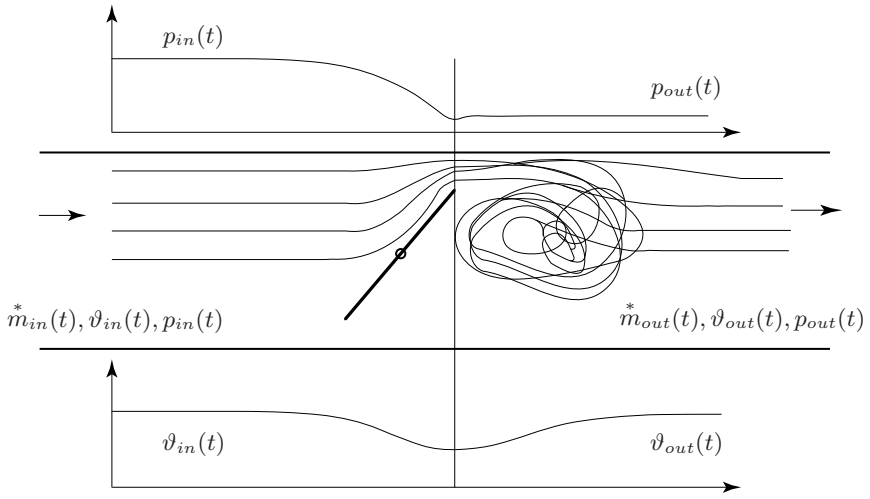



**Fig. 2.27.** Compressible flow behavior in a valve with pressure distribution (top) and temperature distribution (bottom).

flow becomes fully turbulent. In the first part, all the potential energy stored in the flow (with pressure as its level variable) is converted isentropically (i.e., without losses) into kinetic energy. In the second part, the flow becomes fully turbulent and all of its kinetic energy is dissipated into thermal energy, i.e., no pressure recuperation takes place. The consequences of these assumptions are that the pressure in the narrowest part of the valve is (approximately) equal to the downstream pressure and that the temperature of the flow before and after the valve is (approximately) the same (hence, its name, see Figure 2.27).

Using the first law of thermodynamics and the isentropic expansion relations for a perfect gas [13], the following equation is obtained

$$\overset{*}{m}(t) = c_d \cdot A(t) \cdot \frac{p_{in}(t)}{\sqrt{R \cdot \vartheta_{in}(t)}} \cdot \Psi(p_{in}(t), p_{out}(t)) \tag{2.111}$$

where the function $\Psi(.)$ is given by the expression

$$\Psi(p_{in}(t), p_{out}(t)) = \begin{cases} \sqrt{\kappa \cdot \left(\frac{2}{\kappa+1}\right)^{\frac{\kappa+1}{\kappa-1}}} & \text{for } p_{out} < p_{cr} \\ \left(\frac{p_{out}}{p_{in}}\right)^{1/\kappa} \cdot \sqrt{\frac{2\,\kappa}{\kappa-1} \cdot \left[1 - \left(\frac{p_{out}}{p_{in}}\right)^{\frac{\kappa-1}{\kappa}}\right]} & \text{for } p_{out} \geq p_{cr} \end{cases} \tag{2.112}$$

and where the critical pressure is defined by

$$p_{cr} = \left[\frac{2}{\kappa+1}\right]^{\frac{\kappa}{\kappa-1}} p_{in} \tag{2.113}$$

At this pressure level, the flow in the narrowest part reaches sonic conditions. Since the throttle geometry is assumed to have an unfavorable form (i.e., no convergent-divergent "Laval" shape, see [19]), the flow is "choked" at this velocity and no further speed increase can take place. The form of (2.112) is shown in Figure 2.28 for $\kappa = c_p/c_v \approx 1.4$, which is realistic for air.

Figure 2.28 also shows an approximate formulation for $\Psi$, which is indicated in Equation (2.114). Figure 2.28 shows that the two curves are quite similar (they are almost identical for gases with $\kappa = 1.55$) such that the simpler formulation will usually be acceptable, especially since the errors introduced by imprecisely known discharge coefficients $c_d$ will often be larger.

$$\Psi(p_{in}(t), p_{out}(t)) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } p_{out} < 0.5\, p_{in} \\ \sqrt{\frac{2\, p_{out}}{p_{in}} \cdot \left[1 - \frac{p_{out}}{p_{in}}\right]} & \text{for } p_{out} \geq 0.5\, p_{in} \end{cases} \tag{2.114}$$

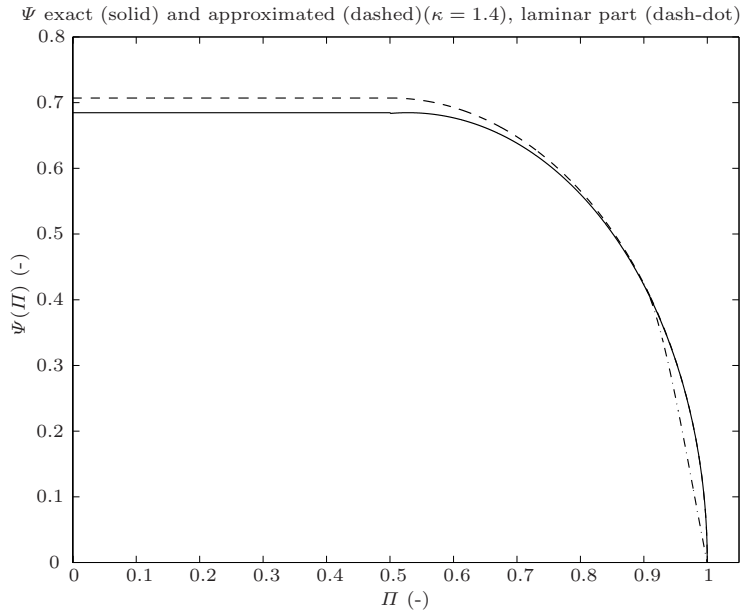



**Fig. 2.28.** Isentropic and approximated flow characteristics for subsonic conditions; pressure ratio $\Pi = p_{out}/p_{in}$.

The curves shown in Figure 2.28 indicate that the nonlinear characteristic has a singularity (in both formulations) at $p_{in} = p_{out}$ (inserting this relation in the receiver Equation (2.105) yields a system that does not satisfy the Lipschitz conditions [20]). To overcome this problem, a laminar flow condition can be assumed for very small pressure ratios. This (physically quite reasonable) assumption yields a differentiable formulation at zero pressure difference and thus eliminates the problems mentioned. A pragmatic approach to model such a behavior is to assume that there exists a threshold

$$\Pi_{tr} < 1 \tag{2.115}$$

below which the function $\Psi(\Pi)$ is defined by Equation (2.114). If larger pressure ratios occur, then a cubic approximation $\widetilde{\Psi}$ of $\Psi$ of the form

$$\widetilde{\Psi}(\Pi) = a \cdot (\Pi - 1)^3 + b \cdot (\Pi - 1) \tag{2.116}$$

with

$$a = \frac{\Psi'_{tr} \cdot (\Pi_{tr} - 1) - \Psi_{tr}}{2\,(\Pi_{tr} - 1)^3}, \qquad b = \Psi' - 3\,a \cdot (\Pi_{tr} - 1)^2 \tag{2.117}$$

is used ($\Psi_{tr}$ is the value of $\Psi$ and $\Psi'_{tr}$ is the value of $\partial\Psi/\partial\Pi$ at the threshold $\Pi_{tr}$). The special form of Equation (2.116) guarantees a symmetric behavior around the critical pressure ratio and a smooth transition at the departure point. Figure 2.28 shows an example of this curve (dash-dotted curve, for reasons of clarity an unrealistically large threshold $\Pi_{tr} = 0.9$ has been chosen).

### Gas Turbines

For control system applications, the fluid-dynamic behavior of gas turbines often can be simplified to be a static element with the input variables pressure $p_3$ before and pressure $p_4$ after the turbine, the temperature $\vartheta_3$ at turbine intake and the turbine rotor speed $\omega_t$ ("causes").

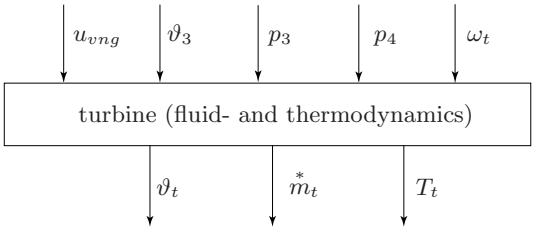



**Fig. 2.29.** Causality diagram of a gas turbine.

The turbine massflow $\overset{*}{m}_t$, the outlet temperature $\vartheta_4$, and the shaft torque $T_t$ are the outputs ("effects"). If the turbine has a variable-nozzle geometry an additional control input $u_{vnt}$ is important; otherwise the turbine radius $r_t$, and therefore the open area, are assumed to be constant.

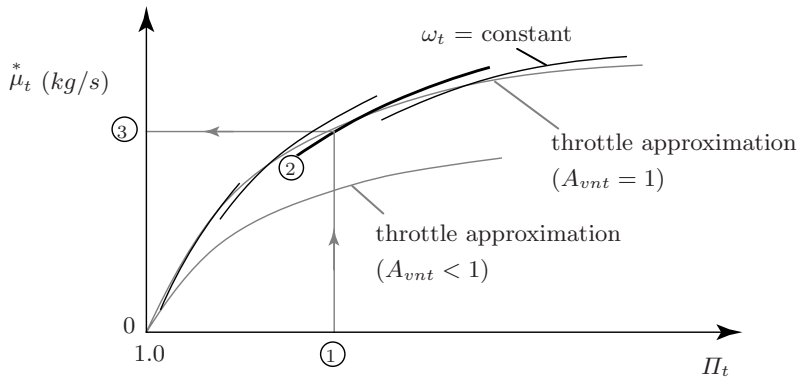



**Fig. 2.30.** Gas turbine massflow behavior.

As a first approximation, the flow behavior of gas turbines may be assumed to be similar to that of an isenthalpic valve, as discussed above (see Figure 2.30). Turbine maps also are in use. In both representations, all variables are referenced to a nominal case $\vartheta_{3,ref}, p_{3,ref}$

$$\begin{aligned} \Pi_t &= \frac{p_3}{p_4} \\ \mu_t &= \overset{*}{m}_t \cdot \sqrt{\vartheta_3/\vartheta_{3,ref}}/(p_3/p_{3,ref}) \\ c_u &= \frac{r_t \cdot \omega_t}{c_{us}} \\ c_{us} &= \sqrt{2\, c_p \cdot \vartheta_3 \cdot [1 - \Pi_t^{(1-\kappa)/\kappa}]} \end{aligned} \tag{2.118}$$

The turbine efficiency $\eta_t$ is more difficult to model. Typically, measurements must be taken to get precise indications for that quantity. Figure 2.31 shows results for a small automotive-type turbine. Once the efficiency is known, the torque $T_t$ that is delivered by the turbine to the mechanical shaft can be computed as follows

$$T_t = \frac{\eta_t \cdot \overset{*}{m}_t \cdot c_p \cdot \vartheta_3}{\omega_t} \cdot \left[1 - \Pi_t^{(1-\kappa)/\kappa}\right] \tag{2.119}$$

where $c_p$ is the specific heat at constant pressure and $\kappa$ the ratio $\kappa = c_p/c_v$ of the gases flowing through the turbine.

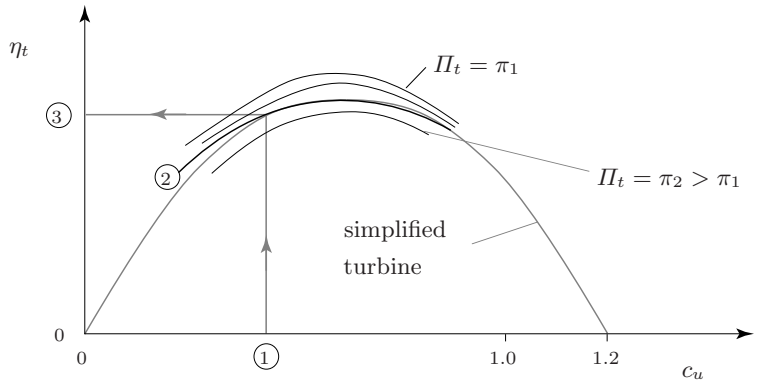



**Fig. 2.31.** Gas turbine efficiencies.

Finally, using a first-law argument and assuming perfect gases, the turbine outlet temperature $\vartheta_4$ can be estimated using the equation

$$\vartheta_4 = \vartheta_3 \cdot \left[1 - \eta_t \cdot \left(1 - \Pi_t^{(1-\kappa)/\kappa}\right)\right] \tag{2.120}$$

### Compressors

Compressors are – as turbines – simplified by modeling them as static elements that have the causality structure shown in Figure 2.32. The pressure before the compressor, $p_1$, and after the compressor, $p_2$, the temperature $\vartheta_1$ of the inflowing gases, and the compressor rotor speed $\omega_c$ are the causes.

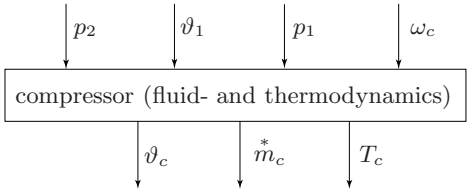



**Fig. 2.32.** Causality diagram of a gas compressor.

The effects are the compressor massflow $\overset{*}{m}_c$, the temperature of that fluid after the compressor $\vartheta_2$, and the torque absorbed by the compressor, $T_c$.

For the estimation of both the compressor massflow and the compressor efficiency, measured maps are necessary. Figure 2.33 shows the idealized struc-

ture of such a map. Using the known pressure ratio $\Pi_c = p_2/p_1$ (point 1 in Figure 2.33) and the known speed $\omega_c$ (point 2) of the compressor, the massflow can be interpolated from the map data (point 3). Similarly, the compressor efficiency $\eta_c$ (point 4) is found in that map.

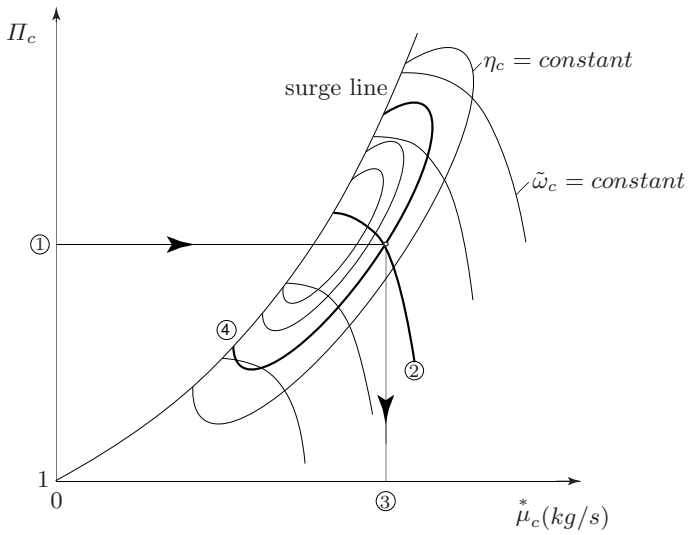



**Fig. 2.33.** Mass flow and efficiency map of a gas compressor.

Once the efficiency is known, the torque absorbed by the compressor can be found using the equation

$$T_c = \frac{\overset{*}{m}_c \cdot c_p \cdot \vartheta_1}{\eta_c \cdot \omega_c} \cdot \left[\Pi_c^{(\kappa-1)/\kappa} - 1\right] \tag{2.121}$$

where $c_p$ is the specific heat at constant pressure and $\kappa$ the ratio $\kappa = c_p/c_v$ of the gases flowing through the compressor.

Finally, using a first-law argument and assuming perfect gases, the compressor outlet temperature $\vartheta_2$ can be estimated using the equation

$$\vartheta_2 = \vartheta_1 \cdot \left[1 + \frac{1}{\eta_c} \cdot \left(\Pi_c^{(\kappa-1)/\kappa} - 1\right)\right] \tag{2.122}$$

### 2.4.7 Chemical Systems

Chemically active systems have many similarities to thermal systems, with one additional feature: the mass fraction of the species present in the system change (in a closed system, total mass is constant). Typically, this is associated with a change of the amount of internal energy. The kinetic characteristics will determine the rate ("speed") of the reactions, while the equilibrium state can be determined by thermodynamic arguments only. Here, just the most elementary notions will be presented. More on this topic can be found in [18].

Chemical reactions typically involve two reactants ("binary reactions," the probability that three species react directly is much smaller), such that only the following reaction type will be discussed here

$$\alpha \cdot A + \beta \cdot B \leftrightarrow \gamma \cdot C + \delta \cdot D \tag{2.123}$$

where the integer coefficients $\{\alpha, \beta, \gamma, \delta\}$ describe the stoichiometry of the reaction and the double arrow in (2.123) indicates that the reaction can evolve in both directions.

Chemical reactions are best described on a molecular basis, i.e., instead of the species mass their molar numbers $n_A$, $n_B$, etc., are often used.[8] Closely associated to that is the notion of concentration $[A] = n_A/V$, i.e., the number of molecules $n_A$ in a given volume $V$. The rate of formation from the left to the right for the binary reaction (2.123) is given by

$$\frac{d^-}{dt}[A] = -\alpha \cdot r^- \cdot [A]^\alpha \cdot [B]^\beta \tag{2.124}$$

i.e., the probability that the reaction $\rightarrow$ takes place is proportional to the probability that the necessary number of molecules $A$ and $B$ are in contact. Of course, the inverse reaction $\leftarrow$ also is possible, i.e.,

$$\frac{d^+}{dt}[A] = \alpha \cdot r^+ \cdot [C]^\gamma \cdot [D]^\delta \tag{2.125}$$

such that the total rate of formation of the species $A$ in reaction (2.123) (in $mol/(s\ m^3)$) is given by

$$\frac{d}{dt}[A] = \alpha \cdot \left(r^+ \cdot [C]^\gamma \cdot [D]^\delta - r^- \cdot [A]^\alpha \cdot [B]^\beta\right) \tag{2.126}$$

The rate of the reaction (its kinetics) is given by the rate "constants" $r^+$, and $r^-$ that depend on the pressure, certain constants and – most importantly

[8] One mol of $A$ (i.e., $n_A = 1$ mol) is $6.022\ldots \cdot 10^{23}$ molecules of that species, and the connection to the mass is given by the molar "mass" $\mathcal{M}_\mathcal{A}$, which is equivalent to the mass of 1 mol of $A$.

– on the temperature of the reacting phase. Usually an Arrhenius model (see Figure 2.34) is used to describe these dependencies

$$r^+ = k^+(\vartheta, p, \ldots) \cdot e^{-E^+/(\Re \cdot \vartheta)} \tag{2.127}$$

where $\Re$ is the universal gas constant (8.314 J/mol K). The variable $k^+$ is the pre-exponential factor which can be a moderate function of pressure and temperature, and $E^+$ is the activation energy.

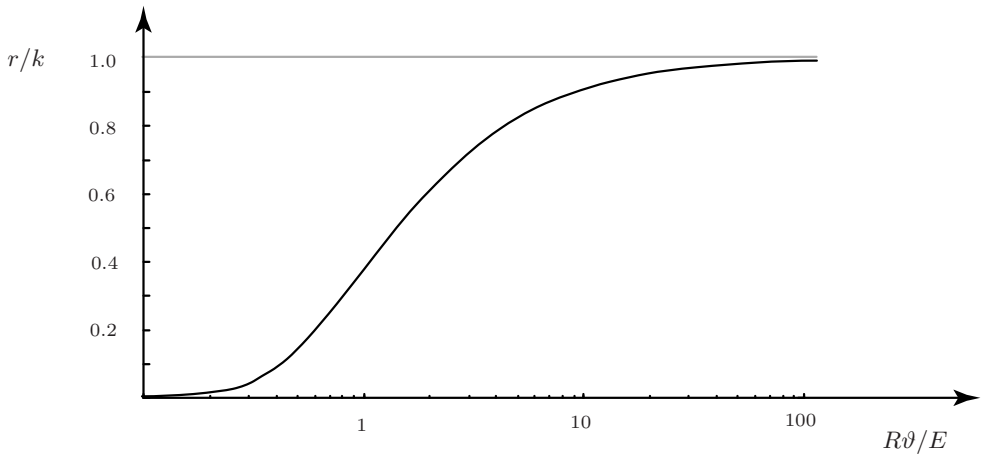



**Fig. 2.34.** Arrhenius function.

The Boltzmann term $e^{-E^+/(\Re\vartheta)}$ indicates the fraction of all collisions that have sufficient energy to start the reaction (2.123). Similarly,

$$r^- = k^-(\vartheta, p, \ldots) \cdot e^{-E^-/(\Re \cdot \vartheta)} \tag{2.128}$$

is the reaction kinetic for the backward reaction. In most cases the four parameters

$$\{k^+, k^-, E^+, E^-\}$$

must be determined experimentally.[9]

Using these preliminaries, species conservation laws can be expressed. Chemical reactions can produce or consume internal energy. A reaction, therefore, represents a heat flow and that effect must be included in the energy conservation laws. These points will be illustrated in the following example.

[9] For very simple reactions "ab initio" quantum-mechanical computations can be feasible.

*Example 2.15 (Continuously Stirred Tank Reactor).*

The system to be modeled is a continuously stirred tank reactor (CSTR, see Figure 2.35) in which the following reaction takes place

$$A + B \rightarrow C \tag{2.129}$$

The molecule $A$ is assumed to be the limiting species, i.e., the amount of $B$ in the inflow and in the CSTR is always much larger than necessary for the reaction (2.129) to take place (the concentration $[B]$, therefore, can be assumed to remain constant). The species $C$ is the product and it is continuously removed from the tank. The dissociation $A + B \leftarrow C$ is assumed to be negligible. Moreover, it is assumed that the specific heat $c$, the density $\rho$, and the mass $m$ of the fluid in the CSRT are constant (a fast level control loop is present). Also, the CSTR is perfectly insulated, such that the only heat exchange with the environment occurs through the heat exchanger. Finally, the usual lumped-parameter assumptions are made, for instance that the concentration $[C_o(t)]$ of the species $C$ in the outflow is the same as in the CSTR, i.e., $[C_o(t)] = [C(t)]$.

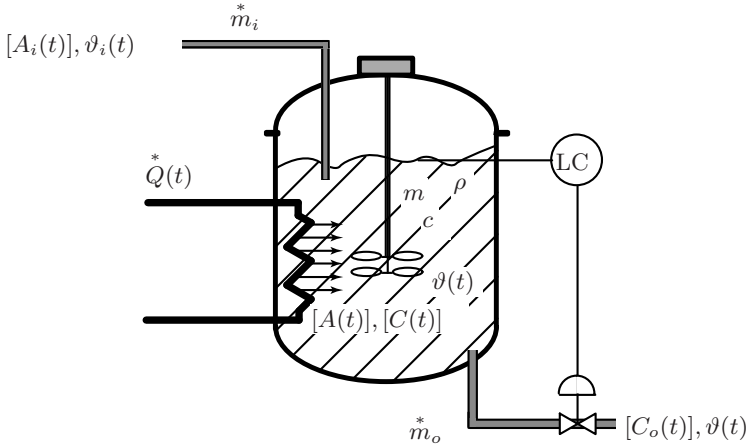



**Fig. 2.35.** Continuously stirred chemical reactor.

*Step 1:*

The system inputs consist of one control signal $\overset{*}{Q}(t)$, i.e., the rate of heat transferred by the heat exchanger and two disturbances, i.e., the reactant concentration $[A_i(t)]$ and the temperature $\vartheta_i(t)$ in the incoming flow $\overset{*}{m}_i$. The measured system outputs are the concentration of the species $[C]$ and the temperature $\vartheta(t)$ of the outgoing flow $\overset{*}{m}_o$.

Notice that since the mass in the CSTR is assumed to be constant the massflow rates satisfy the relation

$$\overset{*}{m}_i = \overset{*}{m}_o = \overset{*}{m} \tag{2.130}$$

and the constant density assumption leads to

$$\overset{*}{V}_i = \overset{*}{V}_o = \overset{*}{V} = \overset{*}{m}/\rho \tag{2.131}$$

*Step 2:*
Using the assumptions mentioned, three[10] reservoirs must be modeled:

- $n_A$, i.e., the amount of species $A$ in the CSTR, level variable $[A]$;
- $n_C$, i.e., the amount of species $C$ in the CSTR, level variable $[C]$; and
- $U$, i.e., the internal energy in the CSTR, level variable $\vartheta$.

*Step 3:*
For species $A$, the conservation laws yield

$$\frac{d}{dt}n_A(t) = \overset{*}{V} \cdot [A_i(t)] - \overset{*}{V} \cdot [A(t)] - V \cdot k^- \cdot [B] \cdot e^{-E/(\mathcal{R}\cdot\vartheta(t))} \cdot [A(t)] \tag{2.132}$$

for species $C$

$$\frac{d}{dt}n_C(t) = -\overset{*}{V} \cdot [C(t)] + V \cdot k^- \cdot [B] \cdot e^{-E/(\mathcal{R}\cdot\vartheta(t))} \cdot [A(t)] \tag{2.133}$$

and for the CSTR energy

$$\frac{d}{dt}U(\vartheta(t), n_A(t), n_C(t)]) = \overset{*}{H}_i(\vartheta_i(t)) - \overset{*}{H}_o(\vartheta(t)) + \overset{*}{Q}(t) \tag{2.134}$$

where the enthalphy flows are given by

$$\overset{*}{H}_i(\vartheta_i(t)) = \overset{*}{m} \cdot c \cdot \vartheta_i(t) \quad \text{and} \quad \overset{*}{H}_o(\vartheta(t)) = \overset{*}{m} \cdot c \cdot \vartheta(t)$$

respectively.

*Step 4:*
In chemical systems, the additional point that must be considered is the fact that the internal energy is a function of the temperature *and* the chemical composition and, therefore,

$$\begin{aligned} dU(\vartheta, n_A, n_B, n_C) &= \frac{\partial U}{\partial \vartheta} \cdot d\vartheta + \frac{\partial U}{\partial n_A} \cdot dn_A + \frac{\partial U}{\partial n_B} \cdot dn_B + \frac{\partial U}{\partial n_C} \cdot dn_C \\ &= \rho \cdot V \cdot c \cdot d\vartheta + H_A \cdot dn_A + H_B \cdot dn_B + H_C \cdot dn_C \end{aligned}$$

[10] As mentioned above, the concentration $[B]$ is assumed to be constant.

where $H_A$, $H_B$, and $H_C$ are the enthalpies of formation for the corresponding species.[11] Notice that in Equation (2.135) neither the specific heat $c$ nor the enthalpies of formation $H_x$ are assumed to depend on the temperature $\vartheta$.

Since in reaction (2.129) the stoichiometry is $\{1, 1, 1\}$ in each reaction the changes in the species must be $-dn_A = -dn_B = +dn_C$. Accordingly, the total enthalpy release is $H_0 = H_A + H_B - H_C$ (one molecule of $A$ and one molecule of $B$ disappear, i.e., their enthalpy of formation is released, and one molecule $C$ is created, which consumes the corresponding enthalpy of formation).

*Step 5:*

Inserting Equation (2.135) in the conservation laws (2.132) to (2.134), and simplifying the resulting expressions, the following ODEs are obtained

$$\begin{aligned}
\tau \tfrac{d}{dt}[A(t)] &= [A_i(t)] - \left(1 + \tau \cdot k \cdot e^{-E/(\mathcal{R}\cdot\vartheta(t))}\right) \cdot [A(t)] \\
\tau \tfrac{d}{dt}[C(t)] &= -[C(t)] + \tau \cdot k \cdot e^{-E/(\mathcal{R}\cdot\vartheta(t))} \cdot [A(t)] \\
\tau \tfrac{d}{dt}\vartheta(t) &= \vartheta_i(t) - \vartheta(t) + \tfrac{1}{\rho\, c} \tfrac{\overset{*}{Q}(t)}{\overset{*}{V}} + \tfrac{\tau}{\rho\, c} \cdot H_0 \cdot k \cdot e^{-E/(\mathcal{R}\cdot\vartheta(t))} \cdot [A(t)]
\end{aligned} \tag{2.135}$$

where the residence time $\tau$, the overall reaction rate $k$ and the reaction enthalpy $H_0$ are defined by

$$\tau := V/\overset{*}{V}, \qquad k := k^{-} \cdot [B], \qquad H_0 = H_A + H_B - H_C \tag{2.136}$$

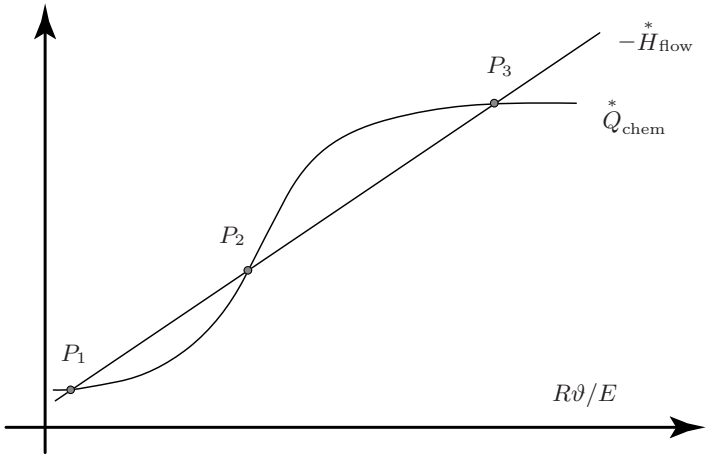



**Fig. 2.36.** Steady-state points of the CSTR reaction as a function of temperature.

[11] Following the usual definition, the enthalpy of formation $H_x$ of a species $x$ is positive if energy is needed to form $x$.

Assuming no control action $\overset{*}{Q} = 0$, a fixed reactant intake temperature $\vartheta_i$, and a constant concentration $[A_i]$, the steady-state conditions of the CSTR can be derived by setting the time derivatives in Equation (2.135) to zero and solving the coupled system of algebraic equation for the CSTR temperature

$$\overset{*}{H}_{\text{flow}}(\vartheta) + \overset{*}{Q}_{\text{chem}}(\vartheta) = 0 \tag{2.137}$$

where (assuming an exothermal reaction, i.e., $\vartheta > \vartheta_i$)

$$\overset{*}{H}_{\text{flow}}(\vartheta) = \overset{*}{m} \cdot c \cdot (\vartheta_i - \vartheta) \tag{2.138}$$

is the heat removed by the massflow and

$$\overset{*}{Q}_{\text{chem}}(\vartheta) = H_0 \cdot \frac{V \cdot k \cdot e^{-E/(\mathcal{R}\cdot\vartheta)}}{1 + \tau \cdot k \cdot e^{-E/(\mathcal{R}\cdot\vartheta)}} \cdot [A_i] \tag{2.139}$$

is the heat produced by the chemical reaction. This implicit equation for the equilibrium temperature $\vartheta$ cannot be solved analytically. However, plotting the two heat flows, the typical situation shown in Figure 2.36 is encountered, i.e., three equilibrium points exist. The equilibrium $P_2$ is unstable, while $P_1$ and $P_3$ are potentially stable. However, often the point $P_2$ is the desired operating point (best yield). Accordingly, an appropriate feedback control system that stabilises the CSTR dynamics at that point must be added.

## 2.5 Distributed Parameter Systems

Many physical systems are best described by partial differential equations (PDE, for instance, fluids, electromagnetic fields, etc.). The underlying physical laws (Navier-Stokes, Maxwell, etc.) are relatively simple to formulate but extremely difficult to solve, especially for non-trivial boundary conditions.

Typically, such PDE are obtained by analyzing an infinitesimally small "lumped parameter" cell. This is shown in the following example of a heat exchanger that continues the derivations started in Section 2.13.

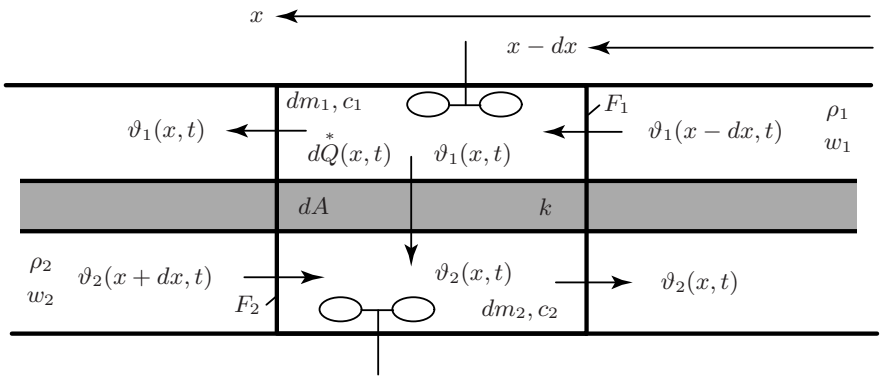



**Fig. 2.37.** Infinitesimally small heat exchanger element.

*Example 2.16 (Heat Exchanger II, PDE Formulation).*

The elementary cell of the PDE model is shown in Figure 2.37. An internal energy balance for the upper part yields

$$\frac{\partial}{\partial t}\left[dm_1 c_1 \vartheta_1(x,t)\right] = \overset{*}{m}_1 c_1 \left[\vartheta_1(x-dx,t) - \vartheta_1(x,t)\right] - kdA\left[\vartheta_1(x,t) - \vartheta_2(x,t)\right] \tag{2.140}$$

Introducing the fluid flow speeds $w_1$ and $w_2$, the fluid densities $\rho_1$ and $\rho_2$, the flow cross sections $F_1$ and $F_2$, the total heat exchanger length $L$ (all these parameters are assumed to be constant) and using a series expansion of $\vartheta(x,t)$ in $x$, Equation (2.140) can be approximated as follows

$$F_1 dx \rho_1 c_1 \frac{\partial \vartheta_1(x,t)}{\partial t} \approx -F_1 w_1 \rho_1 c_1 \frac{\partial \vartheta_1(x,t)}{\partial x} dx - k\frac{A}{L} dx \left[\vartheta_1(x,t) - \vartheta_2(x,t)\right] \tag{2.141}$$

The final formulation is obtained collecting terms and dividing by $dx$

$$\frac{\partial \vartheta_1(x,t)}{\partial t} = -w_1 \frac{\partial \vartheta_1(x,t)}{\partial x} - \frac{1}{\tau_1}\left[\vartheta_1(x,t) - \vartheta_2(x,t)\right] \tag{2.142}$$

with $\tau_1 = \frac{Lc_1F_1\rho_1}{kA}$. The lower part can be analyzed in the same way yielding

$$\frac{\partial \vartheta_2(x,t)}{\partial t} = w_2 \frac{\partial \vartheta_2(x,t)}{\partial x} + \frac{1}{\tau_2}\left[\vartheta_1(x,t) - \vartheta_2(x,t)\right] \tag{2.143}$$

with $\tau_2 = \frac{Lc_2F_2\rho_2}{kA}$.

Systems described by PDEs such as (2.142) or (2.143) are of infinite dimension, i.e., there is no finite number of ODEs that can exactly reproduce the behavior modeled by (2.142) or (2.143). To clarify this observation, the heat exchanger equations are analyzed for the simplified situation in which $k = 0$ (no heat transfer). In this case the lumped-parameter approach of Example 2.13 yields for one cell[12] the following equation (see Equation (2.98)

$$m \cdot c \cdot \frac{d\vartheta_{o,j}(t)}{dt} = \overset{*}{m} \cdot c \cdot (\vartheta_{i,j}(t) - \vartheta_{o,j}(t)) \tag{2.144}$$

and the heat-exchanger behavior is modeled by stacking together $n$ such first-order systems. Defining

$$m = \rho \cdot F \cdot \frac{L}{n}, \quad \overset{*}{m} = \rho \cdot F \cdot w$$

with $m$ representing the mass in one cell, $F$ the constant cross section area and $L$ the total length of the heat-exchanger tube, and $w$ the velocity of the flow, Equation (2.144) can be simplified to

$$\tau \cdot \frac{d\vartheta_{o,j}(t)}{dt} = \vartheta_{i,j}(t) - \vartheta_{o,j}(t), \quad \tau = \frac{L}{n \cdot w}, \quad j = 1, \ldots, n \tag{2.145}$$

The PDE (2.142) for the simplified case $k = 0$ reads

$$\frac{\partial \vartheta(x,t)}{\partial t} = -w \cdot \frac{\partial \vartheta(x,t)}{\partial x} \tag{2.146}$$

This hyperbolic PDE is known as the (simplified) advection equation, and it is known to be hard to solve with numerical PDE solvers. Fortunately, a closed-form solution is possible using the "Ansatz"

$$\vartheta(x,t) = f(t - \frac{x}{w}) \tag{2.147}$$

which is based on physical intuition (the flow is described by a "plug flow" approach). It is easy to see that the function $f(.)$ satisfies the PDE (2.146)

$$\frac{\partial f(t - \frac{x}{w})}{\partial t} = \frac{\partial f(t - \frac{x}{w})}{\partial (t - \frac{x}{w})} \cdot \frac{\partial (t - \frac{x}{w})}{\partial t} = g(t - \frac{x}{w}) \cdot 1$$

[12] Since with $k = 0$ the two streams are isolated, no distinction is necessary between the hot and the cold flow.

and,

$$\frac{\partial f(t-\frac{x}{w})}{\partial x} = \frac{\partial f(t-\frac{x}{w})}{\partial (t-\frac{x}{w})} \cdot \frac{\partial (t-\frac{x}{w})}{\partial x} = g(t-\frac{x}{w}) \cdot \frac{-1}{w}$$

respectively. Obviously, Equation (2.146) is satisfied because

$$g(t-\frac{x}{w}) = -w \cdot \frac{-1}{w} \cdot g(t-\frac{x}{w})$$

Therefore, taking $x$ at the input ($x = 0$) and at the output ($x = L$), the input/output dynamics $u(t) = \vartheta(t,0) \rightarrow y(t) = \vartheta(t,L)$ is described by

$$y(t) = u(t-\frac{L}{w}) = u(t-T) \tag{2.148}$$

which is, of course, a simple delay element with delay $T = \frac{L}{w}$.

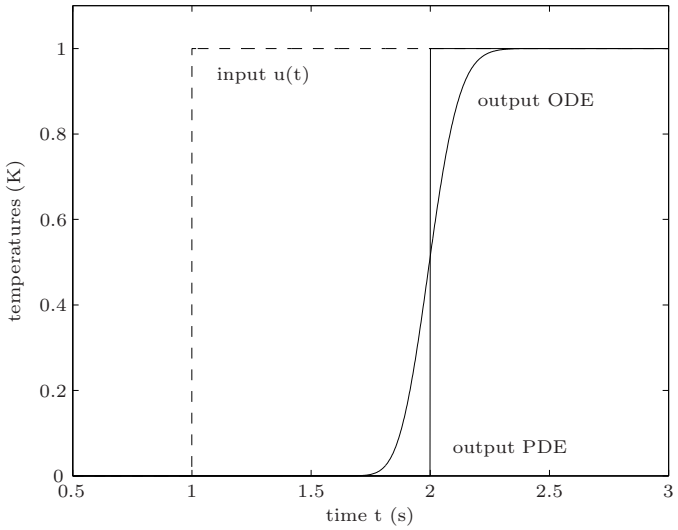



**Fig. 2.38.** Numerical solution obtained with a series connection of $n = 100$ elements (2.145) and exact solution of the simplified heat-exchanger system with $\vartheta(0,t) = u(t) = h(t)$. Parameter values: $L = 1\ m$, $v = 1\ m/s$

Figure 2.38 shows the comparison of the solutions obtained with Equation (2.148) and with 100 instances of Equation (2.145) connected in series. The input signal used is $u(t) = h(t)$. Despite the rather high order, the ODE approximation is not able to capture the essential dynamics of the system, i.e., its "shock-wave" behavior and this is true for any $n < \infty$.

## 2.6 Building Larger Models

Models of large systems can be built only by managing the complexity through modularization. The general approach introduced at the beginning of this chapter, which introduced the ideas of reservoirs and flows, follows this guideline by combining the three main ideas of modularization:

- **encapsulation**: each reservoir represents an "atom" module and all dynamic effects are enclosed in the reservoir's level;
- **interfacing**: the flows constitute the natural interfaces between the modules; and
- **superstructures**: can be built directly from "atoms" since the interfaces are physically meaningful flows.

In the process of developing larger models it is recommended to first identify all reservoirs before any flow-modeling steps are made. Reservoirs constitute *breaking points* in the multiple feedforward and feedback loops present in all dynamic systems, i.e., the level variable of a reservoir is a quantity that is decoupled from the rest of the model by the reservoir's integrating behavior. In particular, implicit loops are broken by that.

After all relevant reservoirs have been identified the system structure has to be mapped by defining the flows between the single reservoir modules. The main representation tools for complex structures are the causality diagrams that have been informally introduced above. Causality essentially means identifying the driving level variables and the resulting flows. Only after these system-theoretic preparations engineering physics becomes important.

Encoding the models using appropriate software and numerical simulation should not be started before the mentioned steps have been made. While this top-down approach may seem slower at the beginning, experience has shown countless times that it leads to the correct results much faster.[13] The best way to learn this technique is to solve non-trivial examples. The following case study and – even more so – the exercises are first steps in that direction.

[13] "Two hours of simulation can easily save you five minutes of thinking!" *Anonymous frustrated engineer*

## 2.7 Case Study: Water-Propelled Rocket

In this case study, a water-propelled "rocket" (WPR) is analyzed and its operation parameters are chosen such that the top height the WPR reaches is maximized. The main objective of this case study is to illustrate the steps necessary to model a system that includes more than one class of dynamic elements. In this example, mechanical, thermodynamic and fluid dynamic subsystems will be included. In addition, the definitions of *hybrid systems* and *state events* are introduced.

### System Modeling

*Problem Definition*

The relevant system parameters and coordinates are illustrated in Figure 2.39. The WPR (say a mineral water bottle with an appropriate nozzle) of mass $m_r$ and volume $V_r$, is filled at start with a certain mass of water $m_w(0)$ that occupies the volume $m_w(0)/\rho$, where $\rho$ is the density of the water. The remaining volume $V(0) = V_r - m_w(0)/\rho$ is filled with air at pressure $p(0)$ and temperature $\vartheta(0)$.

At $t = 0$, the nozzle is opened and the resulting water jet produces the thrust needed to lift the WPR. After all of the water has been ejected at $t = t_1$, the dynamic behavior of the system changes and the pressurized air produces thrust. When the air pressure in the WPR becomes equal to the ambient pressure at $t = t_2$, the WPR completes its trajectory in a ballistic mode until it hits the ground at $t = t_3$.

Generally, a system that changes its dynamic behavior depending on discrete events is referred to as a hybrid system. Such a change can either be triggered by a certain condition for the time $t$, for the system input $u$, or – as in the case of the WPR – by a condition for one or more states $x_i(t)$ of the system being fulfilled ("state events"). Here, the switching conditions to be fulfilled are $m_w(t_1) = 0$, $p(t_2) = p_a$, and $h(t_3) = 0$. In the following, the water-thrust phase, the air-thrust phase, and the ballistic flight phase are modeled separately.

The trajectory of the WPR is described by a vertical velocity $v(t)$ and a vertical position $h(t)$. The velocity of the water or air, respectively, ejected through the nozzle is $w(t)$. This velocity is relative to the WPR and is positive when flowing in the direction indicated in Figure 2.39.

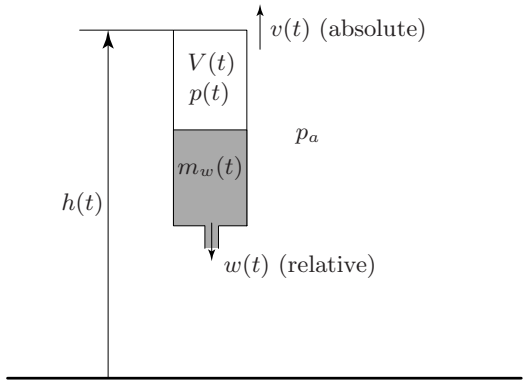



**Fig. 2.39.** Illustration of some of the system parameters and coordinates.

The following assumptions are adopted:

1. Only the vertical motion is modeled, i.e., the WPR is assumed to move only upward or downward.
2. The only two forces acting on the WPR are gravity[14] and thrust. In particular, aerodynamic friction is neglected.[15]
3. The expansion of the pressurized air inside the WPR is isentropic, i.e., no heat transfer or friction need to be considered.[16]
4. Compared to the water mass and the mass of the empty rocket, the air mass is neglected. The air mass flow is only considered for the thrust in the second phase of the flight (air-thrust phase).
5. The flow of the water through the nozzle can be modeled using Bernoulli's law, i.e., that flow is assumed to be incompressible and without friction. The flow of the air can be modeled using Equation (2.111) where the function $\Psi(p(t), p_a)$ is approximated using Equation (2.114). The gravity and the acceleration of the WPR do not influence these behaviors.[17]

[14] The acceleration due to gravity $g = 9.81$ m/s$^2$ is assumed to be constant as the heights reached by the WPR are in the order of 100 m, only.

[15] The top speeds achieved during the trajectory are on the order of 15 m/s. The corresponding aerodynamic forces are approximately one order of magnitude smaller than the gravitational force acting on the empty WPR.

[16] The expansion takes place in a few tenths of a second. In that short time span no substantial heat exchange can take place.

[17] The maximum accelerations $a_{max}$ are on the order of 100 m/s$^2$. The pressure changes at the nozzle induced by these accelerations are on the order of $\Delta p \approx \rho \cdot a_{max} \cdot l$ where $\rho$ is the density of the water or air and $l$ the length of the water on top of the nozzle. Inserting a typical value of $l \approx 0.1$ m yields a pressure change of less than 0.1 bar. Obviously, the influence of the hydrostatic pressure $\Delta p = \rho\, g\, l$ is one order of magnitude smaller.

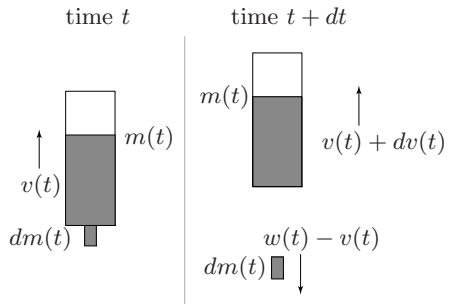



**Fig. 2.40.** Illustration of the momentum balance equations.

*Model Equations for the Water-Thrust Phase* $0 < t \leq t_1$

The equations of the mechanical part of the systems are derived using a momentum balance as illustrated in Figure 2.40. Neglecting all second-order terms, the change in momentum is given by the expression

$$\begin{aligned} dB(t) &= B(t + dt) - B(t) \\ &= [m(t)(v(t) + dv(t)) - dm(t)(w(t) - v(t))] - [(m(t) + dm(t))v(t)] \\ &= m(t) \cdot dv(t) - dm(t) \cdot w(t) \end{aligned}$$

According to the momentum conservation law, this difference must be equal to the external force $F_e$ acting on the system times the infinitesimal time interval $dt$, i.e.,

$$dB(t) = F_e(t) \cdot dt = -g \cdot m(t) \cdot dt \tag{2.149}$$

Combining the last two equations yields

$$m(t) \cdot dv(t) = dm(t) \cdot w(t) - g \cdot m(t) \cdot dt \tag{2.150}$$

The infinitesimal mass element $dm$ is given by the expression[18]

$$dm(t) = \rho \cdot F \cdot w(t) \cdot dt \tag{2.151}$$

and inserting this expression into Equation (2.150) yields, after dividing all terms by $dt$, the final equation describing the vertical motion of the WPR ($T(t)$ is the "thrust" produced by the out-flowing water)

$$m(t) \cdot \frac{d}{dt} v(t) = -g \cdot m(t) + T(t), \qquad T(t) = \rho \cdot F \cdot w^2(t) \tag{2.152}$$

$$\frac{d}{dt} h(t) = v(t) \tag{2.153}$$

where $m(t) = m_r + m_w(t)$.

[18] The nozzle area $F$ is not the geometric nozzle area; it includes the fluid dynamic restriction effects often represented by a discharge coefficient $c_d$.

The mass of the water in the WPR changes according to the equation

$$\frac{d}{dt}m_w(t) = -\rho \cdot F \cdot w(t) \tag{2.154}$$

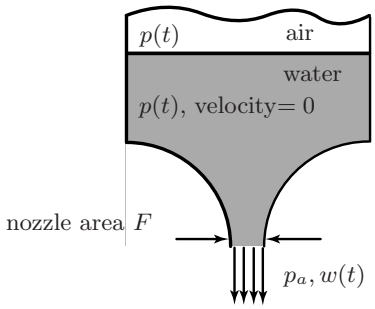



**Fig. 2.41.** Illustration of the definitions used in equations (2.154) and (2.155).

The velocity $w(t)$ of the water through the nozzle is defined by the pressure difference over the nozzle $p(t) - p_a$ according to Bernoulli's law

$$\frac{1}{2} \cdot \rho \cdot w^2(t) + p_a = p(t) \tag{2.155}$$

where $p_a$ is the ambient pressure. As mentioned above, in this equation the increase of pressure due to gravity and acceleration is neglected. Also note that it is assumed that the water leaving the nozzle has expanded to the ambient pressure $p_a$. Solving Equation (2.155) for $w(t)$ yields the following equation

$$w(t) = \sqrt{\frac{2}{\rho}} \cdot \sqrt{p(t) - p_a} \tag{2.156}$$

The last missing element for the water-thrust phase is the description of the pressure $p(t)$ inside the WPR. Starting at $p(0)$, the pressure $p(t)$ will decrease due to the increasing volume $V(t)$ available for the air. This volume is linked to the mass of water left in the WPR

$$V(t) = V_r - \frac{m_w(t)}{\rho} \tag{2.157}$$

Assuming isentropic conditions, the pressure inside the air plenum can be derived using the equation[19]

$$p(t) = \left(\frac{V(0)}{V(t)}\right)^{\kappa} \cdot p(0) \tag{2.158}$$

[19] Remember: In this phase, the mass of air in the WPR remains constant.

where $\kappa = c_p/c_v$ is the isentropic exponent of air and $p(0)$ the pressure at the start of the WPR. At the time $t = t_1$ the condition $m_w(t_1) = 0$ is reached ("water burn out") and the air-thrust phase begins.

Although not needed for $t < t_1$, the temperature of the air can be calculated using the following *algebraic* equation (the variable $m_w(t)$ is defined by Equation (2.154))

$$\vartheta(t) = [(V_r - m_w(0)/\rho)/(V_r - m_w(t)/\rho)]^{\kappa-1} \cdot \vartheta(0) \tag{2.159}$$

Notice that $m_w(0)$ (the amount of water in the WPR), $p(0)$ (the pressure of the air in the WPR), and $\vartheta(0)$ (the temperature of the air in the WPR) at $t = 0$ are the initial conditions that must be defined a priori.

*Model Equations for the Air-Thrust Phase $t_1 < t \leq t_2$*

In the water-thrust phase only three reservoirs were considered (one for the mass of water, and two for the vertical motion of the rocket). In the air-thrust phase, obviously, the water mass is zero and this reservoir can be discarded. However, since the mass of air in the WPR is not constant anymore, its dynamics and the dynamics of the air's internal energy must be modeled now to describe fully the state of the WPR.[20]

In the air-thrust phase, the rocket is modeled as an adiabatic gas receiver (see Example 2.14). The level variable of the internal energy of the air is its temperature $\vartheta(t)$. Since no heat transfer takes place and no enthalpy flows into the rocket one obtains by simplifying Equation (2.106) the following equation for this variable (pro memoria: $c_p - c_v = R$)

$$\frac{d}{dt}\vartheta(t) = \frac{\vartheta^2(t) \cdot R^2}{p(t) \cdot V_r \cdot c_v} \cdot \overset{*}{m}_{out}(t) \tag{2.160}$$

As mentioned above, the out-flowing air mass is modeled using the ideas introduced in Section 2.4.6. More specifically, Equation (2.111) and the simplification (2.114) can be used to compute the air mass flow exiting the rocket.

The level variable of the air mass is the pressure $p(t)$. Since no air mass flows into the rocket, the dynamics of the pressure are described by (see Equation (2.107))

$$\frac{d}{dt}p(t) = -\frac{\kappa \cdot R}{V_r} \cdot \overset{*}{m}_{out}(t) \cdot \vartheta(t) \tag{2.161}$$

[20] During the water-thrust phase these two reservoirs were, of course, present as well, but since the mass of air $m_{air}(t)$ remained constant and since its internal energy, with level variable $\vartheta(t)$, was directly proportional to the mass of the water (see Equation (2.159)), there was no need to model them separately.

where the temperature of the air $\vartheta(t)$ is known form Equation (2.160).

The mass of air $m_{air}(t)$ is much smaller than the (constant) rocket mass $m_r$. Therefore, following the same ideas for the momentum exchange introduced above, the dynamics of the vertical motion are described now by

$$m_r \cdot \frac{d}{dt} v(t) = -g \cdot m_r + T_{air}(t) \tag{2.162}$$

$$\frac{d}{dt} h(t) = v(t) \tag{2.163}$$

where the thrust $T_{air}(t)$ produced by the outflowing air can be modeled as

$$T_{air}(t) = \rho_{air}(t) \cdot F \cdot w^2(t) = \overset{*}{m}{}^2_{out}(t)/(\rho_{air}(t) \cdot F) \tag{2.164}$$

This formulation avoids computing explicitly the velocity $w(t)$ of the outflowing air (the mass flow is already known). Obviously, the density of the air $\rho_{air}(t)$ is defined by

$$\rho_{air}(t) = p(t)/(R \cdot \vartheta(t)) \tag{2.165}$$

where the pressure $p(t)$ id defined by Equation (2.161) and the temperature $\vartheta(t)$ by Equation (2.160).

To start the simulation, the initial conditions $v(t_1)$, $h(t_1)$, $\vartheta(t_1)$, and $p(t_1)$ must be known. These values can all be taken over from the final values obtained in the water-thrust phase (see Equations (2.152) (2.158), and (2.159)).

*Model Equations for the Ballistic Phase $t_2 < t \leq t_3$*

In the last phase no thrust is present anymore and the WPR follows a ballistic trajectory. Therefore, neglecting all aerodynamic effects, the dynamics of the WPR are

$$m_r \cdot \frac{d}{dt} v(t) = -g \cdot m_r \tag{2.166}$$

$$\frac{d}{dt} h(t) = v(t) \tag{2.167}$$

with the initial conditions $v(t_2)$ and $h(t_2)$ being equal to the values of the corresponding variables at the end of the air-thurst phase.

### Model Simulation

The equations describing the complete system have to be encoded in an appropriate simulation environment. In this case study Matlab/Simulink is used for that purpose. For the water-thrust phase, the top level of the corresponding Simulink program is shown in Figure 2.42. This program is in the "cause/effect" form that is recommended in the model development process. For the air-thrust phase, the approach described above yields a similar structure.

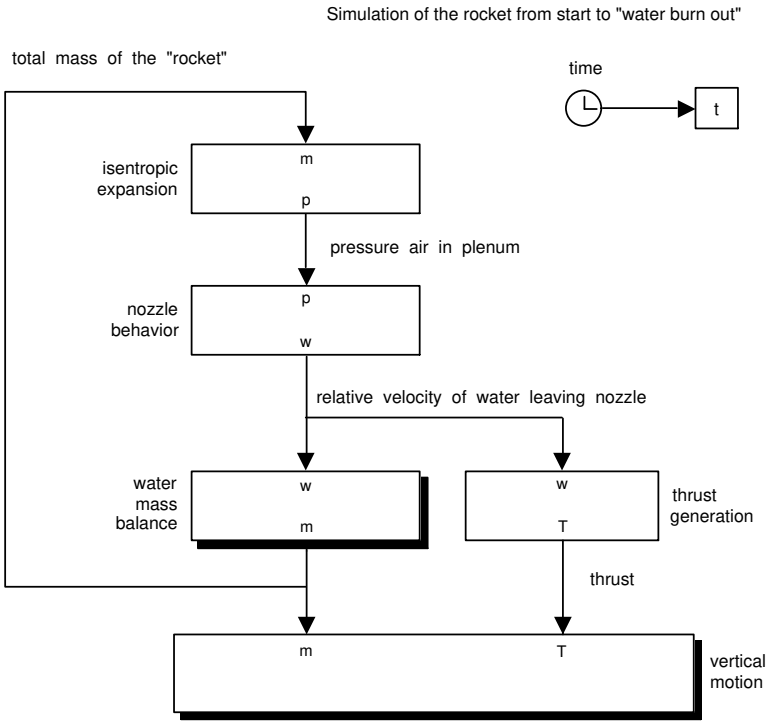



**Fig. 2.42.** Top level of the Simulink program used to simulate the water-thrust phase of the WPR system.

The contents of the five blocks that are shown in Figure 2.42 are displayed in Figure 2.46. All parameters are defined in a separate .m Matlab file shown in Figure 2.45. This .m file contains also all commands needed to compute the results shown in Figure 2.44.

First, the simulation for the water-thrust phase is run. When there is no water left, the simulation is stopped and the simulation time, the pressure and the temperature of the air in the WPR, and its velocity and height are stored and passed on to the second simulation, which models the phase where the air thrust propels the rocket. Once the pressure inside the bottle reaches ambient conditions, the same procedure is repeated and the model describing the purely ballistic flight becomes active.

Such a switching between two modes is referred to as a *state event* because it is triggered by a condition on the system's state variables (for instance, the water mass $m_w(t)$ in the WPR reaching 0). If a simple simulation of the system behavior is the only objective it is often possible to handle state events by checking after every integration step if the condition is met and, depending on the outcome, to continue the simulation in the previous or in the new mode.

However, if accurate results are needed,[21] a more precise state-event detection is mandatory. Figure 2.43 illustrates the main idea of an iteration that can be used to detect a state event with a predefined accuracy. After the event has been detected, i.e., when $m_w(t) < 0$, the simulation is interrupted and restarted with a reduced integration step. This process is repeated until the desired detection accuracy is achieved.

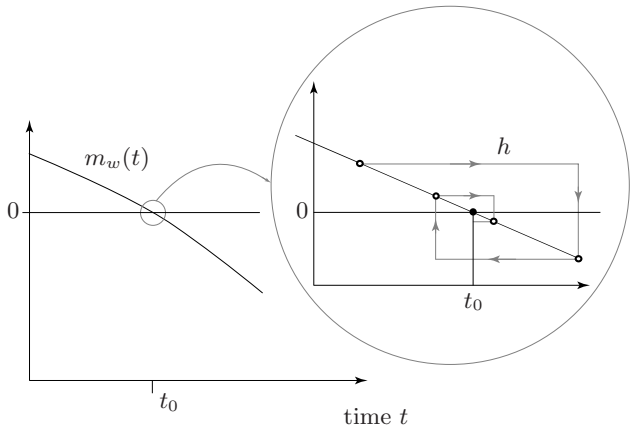



**Fig. 2.43.** Iterative state-event detection.

[21] A typical example is the case where such a simulation is part of a numerical optimization problem which is solved using gradient-based methods. In such a situation, the errors induced by the not perfect state-event detection must be substantially smaller than the variations caused by the finite differences used to approximate the gradients.

The results of the simulation are shown in the top row of Figure 2.44. Notice that the "air-thrust phase" has a small but not negligible influence on tthe WPR trajectory. These results indicate that a water content at start of approximately 40% of the total volume maximizes the top height that the WPR reaches during its flight. Experiments have been carried out to verify the quality of the WPR model. The results of these measurements, indicated by crosses in Figure 2.44, show a reasonable agreement between the observed and the the simulated maximum heights.

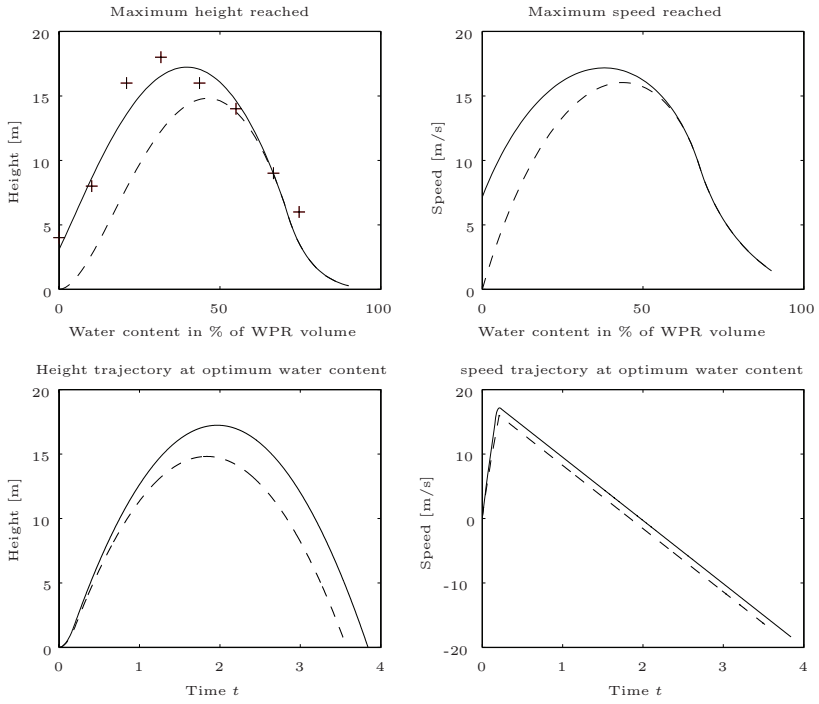



**Fig. 2.44.** Top two plots: maximum height and maximum speed reached during the flight of the WPR for varying initial water levels. The crosses in the top left diagram indicate measured values. Bottom two plots: height and speed trajectories for a WPR with optimal water content at start. In all plots the dashed curves are the results when neglecting the thrust contribution of the air after "water burn out."

Qualitatively, this result can be explained by the following arguments:

- If the WPR is filled with little water, a large amount of energy can be stored in the pressurized air. However, this energy will not produce much thrust because of the lack of sufficient propellant.
- On the other hand, if too much water is used, little energy can be stored in the pressurized air. Moreover, the WPR will be heavier and lifting that additional mass will require a large amount of the energy stored in the pressurized air.

Accordingly, the optimum has to lie in between the two extreme values $m_w(0) = 0$ and $m_w(0) = V_r \cdot \rho$. As the results shown in Figure 2.44 (top row) demonstrate, the optimum is a bit closer to the lower value, because in this case two positive effects (higher energy content and lighter WPR) stand against one negative effect (less propellant).

The trajectory of the WPR for the case in which the water content at start is chosen at its optimum value is shown in Figure 2.44 in the bottom two plots. At $t_1 = 0.22$ s the "water burn-out" is reached and shortly after that the "air burn-out as well. At that point in time the WPR starts its ballistic flight. At $t = 1.95$ s its velocity $v(t)$ is zero and the WPR reaches its maximum height. The velocity then reverses its sign and the WPR impacts at $t = 3.84$ s on the ground.

Of course, several other parameters can be varied. For instance, increasing the initial pressure $p(0)$ or reducing the mass of the empty WPR $m_r$ will always improve the performance. Interestingly, the nozzle area $F$ also should be chosen as large as possible. In fact, for larger $F$, the thrust at start is increased and the "burn" phase is shortened. By that, the WPR avoids lifting the propellant mass and, therefore, uses the energy stored in the pressurized air to reach higher "burn-out" speeds.

The complete set of files required to run the simulations shown in this section can be found at

`http://www.idsc.ethz.ch/education/lectures/system-modeling`

```
% parameters
pa      = 0.97e5;        % ambient pressure Pa
p0      = 5.1e5;         % pressure in WPR at start Pa
rho     = 1000;          % density water kg/m^3
rho_air = 1.2;           % density air kg/m^3
F       = (0.0045^2)*pi; % area nozzle m^2
Vl      = 0.0006;        % volumen WPR m^3
ml      = 0.195;         % mass empty WPR kg
g       = 9.81;          % gravitational acceleration m/s^2
kappa   = 1.35;          % kappa of air -
c_d     = 0.9;           % flow coefficient (compressible)
theta_0 = 300;           % temperature air at beginning K
R       = 287;           % ideal gas constant J/(kg*K)
% initial conditions
v0  =  0.0;              % initial velocity
h0  =  0.0;              % initial height
% load nozzle function (psi) and measured data
psi_param;               % load Psi function parameters
load measured_data       % measured heights
% optimization parameter = mass of water in WPR at start
res1=[];
res2=[];
mw_vec=0.0*Vl*rho:0.01*Vl*rho:0.9*Vl*rho;
for mw0 =  mw_vec;       % mass of water in WPR at start
     V0 = Vl-mw0/rho; % air volume in WPR at start
     m_air_0 = p0*V0/(R*theta_0);
% simulation of system until "water burn out"
     sim('rocketv2');water
     p1=p(end);
     m_air1=m_air_0;
     theta1=p1*Vl/(R*m_air1);
% simulation of system until "air burn out"
     sim('rocketv_noWater_v2');
% remaining flight ballistic
     t=[t_water;t_air];
     h=[h_water;h_air];
     v=[v_water;v_air];
     a=[a_water;a_air];
     nnn=max(size(t));   % find water burn out
     t1=t(nnn);          % time at burn out
     v1=v(nnn);          % velocity at burn out
     h1=h(nnn);          % height at burn out
     t2=v1/g;            % time until WPR reaches v=0
     h2=h1+0.5*v1^2/g;   % maximum height of rocket
     t3=sqrt(2*h2/g);    % time to reach h=0
     maxv=max(v);        % maximum velocity
% store result
     res1=[res1;h2];
     res2=[res2;maxv];
end;
```

**Fig. 2.45.** Matlab .m file including the definition of all system parameters and the commands required to compute the result shown in the top row of Figure 2.44.

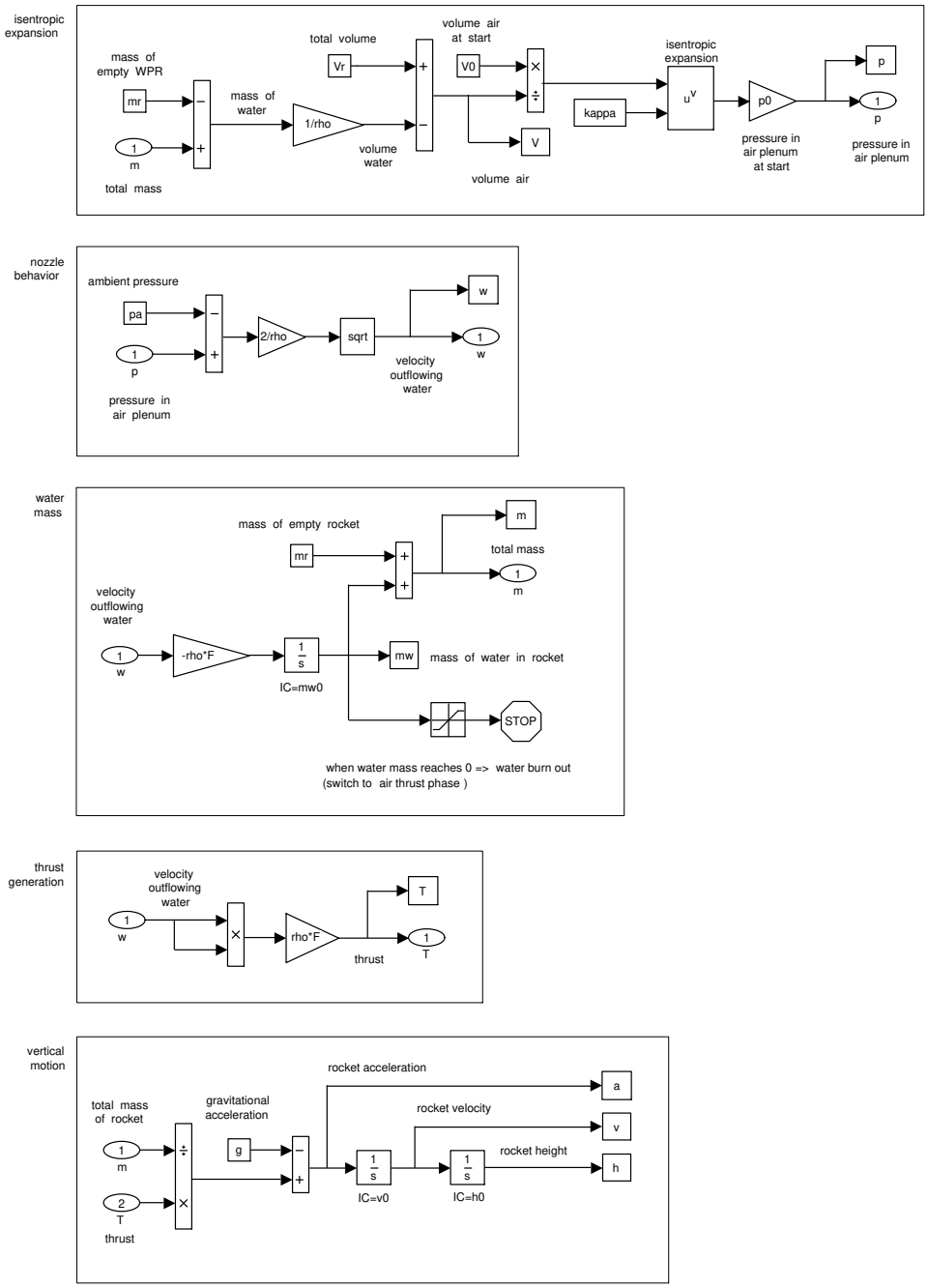



**Fig. 2.46.** Details of the building blocks shown in Figure 2.42.

# 3

# Model Parametrization

The problem of experimentally determining the system parameters will be addressed in this chapter in some detail. The starting point is the section about planning experiments, which shows what excitation and measurement values are best to efficiently identify the system parameters.

The classical approach used for *static* and *linear* systems is then introduced. These ideas may be applied to linear *dynamic* systems as well. However, since these methods are presented in other courses offered at ETH, this subject is not discussed here.

Nonlinear methods, which rely on numerical optimizations methods, will be necessary to identify the physical parameters of more general systems. For instance, the problem of identifying the parameters of a nonlinear dynamic system may be formulated as such a numerical optimization problem. These aspects are discussed in the third section.

Optimization methods form the core of all parameter identification approaches. Appendix I summarizes the most important aspects of closed-form and numerical parameter optimizations. Much more could be said on these two topics. The interested reader is referred to the references mentioned in the appendix.

## 3.1 Planning Experiments

Designing experiments is a difficult problem requiring significant experience with such tasks. Only a few parts of this process can be accomplished using formal methods [15]. One important example of such a formalism is the choice of the correct input signals. These signals must be chosen such that all relevant dynamic and static effects inside the plant are excited with the correct amount of input energy. Other points to be taken into consideration are:

- Measurements for linear or nonlinear model identification.
- Noise at input and output.
- Safety issues.

The data obtained experimentally may be used for two purposes:

1. To identify unknown system structures and system parameters (see next two sections).
2. To validate the results of the system modeling and parameter identification process.

Both purposes are usually accomplished by comparing the output signals acquired in the experiments with those computed in the simulations, using in both cases the same input signals. It is of fundamental importance *not* to use the same data set for purpose 1 and purpose 2. In fact, if the same data set is used for the parameter identification *and* the model validation, then a good agreement between measurements and simulation outputs is not difficult to achieve. This approach is essentially nothing more than a (complicated) way of interpolation. The true quality of a model may only be assessed by comparing the prediction of that model with measurement data that have not been used in the model parametrization. The model has a certain credibility only if these two outputs are sufficiently similar and only in this case the model is useful for the objectives listed in Section 1.1.

## 3.2 Least Squares Methods for Linear Systems

### 3.2.1 Introduction

The theory of least squares estimation is used to fit the parameters of a linear and static model assumed to be able to explain the observed input/output measurements. Since the model never predicts the measurements exactly, it is assumed that an additional error source acts on the measured output (see Figure 3.1). This error may be assumed to be a deterministic or a stochastic variable. Both formulations are equivalent, as long as these errors are completely unpredictable and not correlated with the inputs.

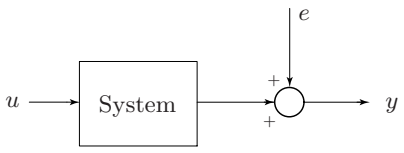



**Fig. 3.1.** Elementary least-squares model structure.

In this section, it is assumed that the output of the system shown in Figure 3.1 may be approximated by a *linear*[1] equation of the form

$$y(k) = h^T(u(k)) \cdot \pi + e(k) \tag{3.1}$$

where $k \in [1, \ldots, r]$ is an integer that indicates the number of the independent measurements and $u(k) \in I\!R^m$ is the k-th input vector, $y(k) \in I\!R$ the k-th output[2] of the system, and $e(k) \in I\!R$ the k-th measurement error. The vector $h(.) \in I\!R^q$ is the "regressor" and depends on the input $u(.)$ in a nonlinear, but algebraic way that is assumed to be known exactly. For the sake of brevity and simple notation, the dependency on $u$ will be dropped below. The vector $\pi \in I\!R^q$ contains all "parameters" of the system, i.e., all constants that are related to physical properties of the system (mass, elasticity, specific heat, ...) and that are not known at the outset.. Typically, there will be many more measurements than unknown parameters, i.e., it is assumed that $r >> q$.

The objective of this section is to derive methods with which it is possible to estimate the unknown parameters $\pi$ using the measurements of the inputs and outputs and the model (3.1). The following definitions are introduced to simplify the notation:

$$\begin{aligned} \widetilde{y} &= [y(1), y(2), \ldots, y(r)]^T,\ \widetilde{y} \in I\!R^r \\ \widetilde{e} &= [e(1), e(2), \ldots, e(r)]^T,\ \widetilde{e} \in I\!R^r \\ H &= [h(1), h(2), \ldots, h(r)]^T,\ H \in I\!R^{r \times q} \end{aligned} \tag{3.2}$$

The error between all the observed and expected outputs is then given by

$$\widetilde{e} = \widetilde{y} - H \cdot \pi, \qquad \widetilde{e} \in I\!R^r \tag{3.3}$$

Now, that set of parameters $\pi_{LS}$ is sought which requires the "smallest" error to explain the differences between measurement result and prediction. To be more precise, $\pi_{LS}$ is that set of parameters $\pi$ that minimizes the norm

$$\epsilon = \widetilde{e}^T \cdot W \cdot \widetilde{e} \tag{3.4}$$

Here, a symmetric and positive definite matrix $W \in I\!R^{r \times r}$ is used to include all a priori information available. If all measurements are equally reliable, then $W = I$ is an obvious choice. If not all measurements are equally reliable, then a $W = \mathrm{diag}\{w_i\}$ may be used where the scalars $w_i$ reflect the relative quality of the corresponding measurement. Non-diagonal matrices $W$ might be used to penalize unlikely combinations of errors.

[1] The nonlinear counterpart will be discussed in Section 3.3.

[2] "Multivariable" versions of this formulation where $y(k) \in I\!R^p$ may be handled easily by interpreting the $p$ outputs as originating from $p$ single-output systems.

Notice that the minimization of the scalar $\epsilon$, as defined in (3.4) with respect to the unknown parameters $\pi$, makes sense only if the model (3.1) is a good approximation of the real system behavior. In other words, if the regressor does not reflect the system behavior sufficiently well, then the error (3.3) must be large. Looking for parameters that minimize it thus will not produce any meaningful parameter estimations.

### 3.2.2 Solution of the Least Squares Problem

The solution of the least squares (LS) minimization problem is straightforward. Inserting the definition (3.3) into the error criterion (3.4) yields

$$\epsilon = \widetilde{y}^T \cdot W \cdot \widetilde{y} - 2 \cdot \widetilde{y}^T \cdot W \cdot H \cdot \pi + \pi^T \cdot H^T \cdot W \cdot H \cdot \pi \tag{3.5}$$

This is a quadratic form of the unknown parameters $\pi$. As shown in Section 6.1, for such quadratic objective functions the sufficient conditions for a unique global minimum are

$$\frac{\partial \epsilon}{\partial \pi} = -2 \cdot H^T \cdot W \cdot \widetilde{y} + 2 \cdot H^T \cdot W \cdot H \cdot \pi = 0 \tag{3.6}$$

and

$$\frac{\partial^2 \epsilon}{\partial \pi^2} = 2 \cdot H^T \cdot W \cdot H > 0 \tag{3.7}$$

Therefore, the explicit solution has the form

$$\pi_{LS} = \left[H^T \cdot W \cdot H\right]^{-1} H^T \cdot W \cdot \widetilde{y} \tag{3.8}$$

and the condition (3.7) is equivalent to requiring a regressor $H$ to have full column rank ($W$ is *chosen* to be positive definite). The condition $\text{rank}\{H\} = q$ formalizes the fact that the number of parameters has to be non-redundant, i.e., that all $q$ parameters $\pi_i$ are required to explain the data.

The solution (3.8) may be efficiently computed with modern CACSD tools. The operator $(.)^\dagger$ with

$$M^\dagger = (M^T \cdot M)^{-1} \cdot M^T, \quad M \in \mathbb{R}^{r \times q},\ r > q,\ \text{rank}\{M\} = q \tag{3.9}$$

is a special case of the more general Moore-Penrose pseudo inversion. It can be shown that if the error $e$ is an uncorrelated white noise signal with mean value zero and variance $\sigma$, then the expected value of the parameter estimation $\pi_{LS}$ is equal to its true value,[3] and the covariance matrix of the parameter estimation is equal to $\sigma^2 \cdot (H^T \cdot W \cdot H)^{-1}$.

[3] Of course, this holds only if the model (3.1) perfectly describes the true system behavior.

For the case $W = I$, the solution (3.8) also permits a rather interesting geometric interpretation. As Figure 3.2 shows, for the special case of $q = 2$ and $r = 3$, the solution obtained by the least-squares approach is that set of parameters $[\pi_{LS,1}, \pi_{LS,2}]^T$ that can explain the observed data with minimum-norm error. Since the model can only predict data that is included in the subspace spanned by the columns of $H = [\widetilde{h}_1, \widetilde{h}_2]$, the minimum-error requirement is equivalent to the condition that the error must be orthogonal to that subspace.

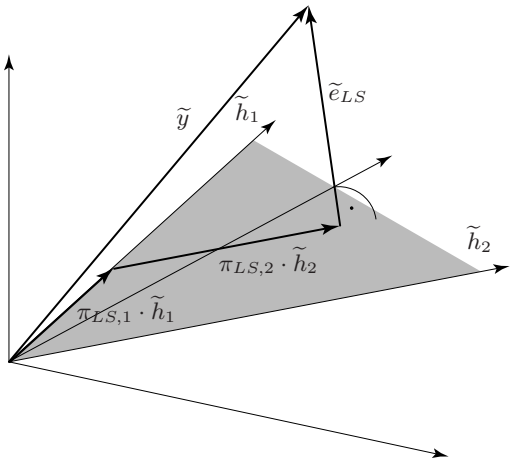



**Fig. 3.2.** Geometric interpretation of the least-squares solution as that parameter set that produces a vector in the subspace spanned by the columns $\widetilde{h}_i$ of $H$ which has the smallest possible distance to the observed data ($W = I$).

### 3.2.3 Iterative Solution of the LS Problem

Up to now, a batch-like approach has been assumed, i.e., first all measurements are taken, then this data is organized according to the definitions (3.2), and, finally, the solution (3.8) is computed. Of course, in this computation the matrix inversion part is the most time-consuming step. Now, assuming that $r$ measurements have been taken and that a solution (3.8) has been computed with this data set, it would be numerically very inefficient to repeat the full matrix inversion procedure when one additional measurement result becomes available. Instead, an *iterative* solution of the form

$$\pi_{LS}(r+1) = \pi_{LS}(r) + \delta(r+1) \cdot [y(r+1) - h^T(r+1) \cdot \pi_{LS}(r)], \qquad (3.10)$$

initialized by

$$\pi_{LS}(0) = E\{\pi\} \tag{3.11}$$

would be much more efficient, especially if the computations of the update of the correction direction $\delta(.)$ do not involve any matrix inversion steps. Notice that the form of this recursive formulation is not arbitrary. First, equation (3.10) shows that changes in $\pi$ are made only if there is a prediction error (the term in square brackets). Second, if there is a prediction error, the changes in $\pi$ are linearly dependent in magnitude on that quantity, but the direction in which the changes are made does not depend on the error.

Such a recursive formulation of the least-squares problem is indeed possible and will be presented below. The cornerstone of the recursive formulation is a slightly simplified version of the Matrix Inversion Lemma.

**Lemma**: Suppose $M \in \mathbb{R}^{n\times n}$ is a regular matrix ($\det(M) \neq 0$) and $v \in \mathbb{R}^n$ is a column vector which satisfies the condition $1 + v^T \cdot M^{-1} \cdot v \neq 0$. In this case

$$[M + v \cdot v^T]^{-1} = M^{-1} - \frac{1}{1 + v^T \cdot M^{-1} \cdot v} \cdot M^{-1} \cdot v \cdot v^T \cdot M^{-1} \tag{3.12}$$

The proof by inspection of this lemma is straightforward (multiply from the left with $M + v \cdot v^T$). The key point of this lemma is that no additional matrix inversion is needed on the right-hand side of equation (3.12). In other words, if the inverse of a matrix $M$ is known and if a rank-one matrix $v \cdot v^T$ is added to that matrix, the inversion of the new matrix $M + v \cdot v^T$ may be carried out very efficiently. This result is now used to derive the recursive LS estimation algorithm.

The starting point for this is a reformulation of the solution (3.8) of the LS problem as

$$\pi_{LS}(r) = \left[\sum_{k=1}^{r} h(k) \cdot h^T(k)\right]^{-1} \cdot \sum_{k=1}^{r} h(k) \cdot y(k) \tag{3.13}$$

Notice that $W = I$ has been used in the last equation (it will be shown below how to include an error weighting in a recursive formulation). To simplify the notation, a matrix $\Omega$ defined by

$$\Omega(r) = \left[\sum_{k=1}^{r} h(k) \cdot h^T(k)\right]^{-1} \tag{3.14}$$

is introduced. Using the matrix inversion lemma, $\Omega(r+1)$ may be written as

$$\Omega(r+1) = \Omega(r) - \frac{1}{1 + c(r+1)} \cdot \Omega(r) \cdot h(r+1) \cdot h^T(r+1) \cdot \Omega(r) \tag{3.15}$$

where $c(r+1) = h^T(r+1) \cdot \Omega(r) \cdot h(r+1)$ has been used to simplify the notation.

Combining (3.15) and equation (3.13) (with $r+1$ instead of $r$) yields

$$\begin{aligned}\pi_{LS}(r+1) = \pi_{LS}(r) &+ \Omega(r)\cdot h(r+1)\cdot y(r+1)\\ &-\tfrac{1}{1+c(r+1)}\cdot\Omega(r)\cdot h(r+1)\cdot h^T(r+1)\cdot\pi_{LS}(r)\\ &-\tfrac{1}{1+c(r+1)}\cdot\Omega(r)\cdot h(r+1)\cdot h^T(r+1)\cdot\Omega(r)\cdot h(r+1)\cdot y(r+1)\end{aligned} \tag{3.16}$$

Rearranging terms, using the fact that $c$ is a scalar (and hence may be moved to a more convenient place), and that $1 - c/(1+c) = 1/(1+c)$, this equation may be simplified to the following form

$$\pi_{LS}(r+1) = \pi_{LS}(r) + \frac{1}{1+c(r+1)}\cdot\Omega(r)\cdot h(r+1)\cdot[y(r+1) - h^T(r+1)\cdot\pi_{LS}(r)] \tag{3.17}$$

which, together with the "update" equation for the gain matrix $\Omega$ (3.15), corresponds to the form as described in equation (3.10).

### 3.2.4 Some Extensions

#### Exponential Forgetting

The definition of a weighted error, as introduced in equation (3.3) for the standard case, is often reformulated for the recursive case in the following way

$$\epsilon(r) = \sum_{k=1}^{r} \lambda^{r-k}\cdot[y(k) - h^T(k)\cdot\pi_{LS}(r)]^2, \quad \lambda < 1 \tag{3.18}$$

This formulation introduces an "exponential forgetting" in the recursive parameter estimation. The implicit assumption is that the older they are, past errors should have a smaller influence on the result of the parameter estimation. Such an exponential-forgetting formulation can cope with slowly[4] varying parameters as well.

Repeating the analysis shown above, the following recursive LS algorithm with exponential forgetting is obtained

$$\pi_{LS}(r+1) = \pi_{LS}(r) + \frac{1}{\lambda+c(r+1)}\cdot\Omega(r)\cdot h(r+1)\cdot[y(r+1) - h^T(r+1)\cdot\pi_{LS}(r)] \tag{3.19}$$

The update equation in this case has the form

$$\Omega(r+1) = \frac{1}{\lambda}\cdot\Omega(r)\cdot\left[I - \frac{1}{\lambda+c(r+1)}\cdot h(r+1)\cdot h^T(r+1)\cdot\Omega(r)\right] \tag{3.20}$$

[4] Slowly relative to the convergence speed of the LS estimation.

**Simplified Algorithms**

Many other identification algorithms have been proposed in the literature. All trade-off convergence rate for reduced computational burden. A typical example of these alternatives is Kaczmarz's projection algorithm whose key idea is the observation that each new prediction error

$$e(r+1) = y(r+1) - h^T(r+1) \cdot \pi(r) \tag{3.21}$$

contains new information on the parameters $\pi$ only in the direction of $h(r+1)$. Assuming $\pi(r)$ to be close to the "true" parameter value, that parameter $\pi(r+1)$ is sought, which requires the smallest possible change $\pi(r+1)-\pi(r)$, in order to be able to explain the new observation. This problem formulation yields the following constrained minimization problem

$$J(\pi) = \frac{1}{2} \cdot (\pi(r+1)-\pi(r))^T \cdot (\pi(r+1)-\pi(r)) + \mu \cdot [y(r+1)-h^T(r+1) \cdot \pi(r+1)] \tag{3.22}$$

The necessary conditions for the minimum solution are

$$\begin{aligned} \tfrac{\partial J}{\partial \pi(r+1)} &= \pi(r+1) - \pi(r) - \mu \cdot h(r+1) = 0 \\ \tfrac{\partial J}{\partial \mu} \quad &= y(r+1) - h^T(r+1) \cdot \pi(r+1) = 0 \end{aligned} \tag{3.23}$$

Solving this set of linear equations for $\pi(r+1)$ and $\mu$, the following solution is obtained

$$\pi(r+1) = \pi(r) + \frac{h(r+1)}{h^T(r+1) \cdot h(r+1)} \cdot [y(r+1) - h^T(r+1) \cdot \pi(r)] \tag{3.24}$$

Usually this solution is modified as

$$\pi(r+1) = \pi(r) + \frac{\gamma \cdot h(r+1)}{\lambda + h^T(r+1) \cdot h(r+1)} \cdot [y(r+1) - h^T(r+1) \cdot \pi(r)] \tag{3.25}$$

where the two real parameters $0 < \gamma < 2$ and $0 < \lambda < 1$ may be used to achieve desired convergence and forgetting properties.

Kaczmarz's projection algorithm requires substantially less computational effort than the classical LS solution (3.17). It is, however, also known to converge much slower than the classical LS algorithm. Which of the many variants is best will, therefore, depend on problem-specific aspects such as computational resources available, convergence speed requirements, etc.

## 3.3 Nonlinear LS Methods

### 3.3.1 Problem Formulation

In this section it is assumed that a dynamic system may be approximated by a model which is formulated as a set of nonlinear ODE

$$\frac{d}{dt}\hat{x}(t) = f(\hat{x}(t), u(t), \hat{\pi}), \quad \hat{y}(t) = h(\hat{x}(t), u(t), \hat{\pi}) \tag{3.26}$$

Here, $\hat{x}(t) \in I\!R^n$ is the state, $u(t) \in I\!R^m$ the input, and $\hat{y}(t) \in I\!R^p$ the output of the model at the time $t$. For the sake of simplicity, in the following it is assumed that $m = p = 1$, but the generalization to multivariable situations is straightforward.

As in the last section, the vector $\pi \in I\!R^q$ contains all parameters of the system, i.e., all constants that are related to physical properties of the system (mass, elasticities, specific heat ...). The objective of this section is to present methods with which it is possible to obtain estimates $\hat{\pi}$ of the parameters $\pi$ using some information on the real system that has been modeled by (3.26).

For this purpose it is assumed that a set of input/output measurements are performed on the *real* system

$$t = [t_1, t_2, \ldots, t_r]^T, \quad u = [u_1, u_2, \ldots, u_r]^T, \quad y = [y_1, y_2, \ldots, y_r]^T \tag{3.27}$$

where the abbreviation $u_i = u(t_i)$ and $y_i = y(t_i)$ is used in the following.

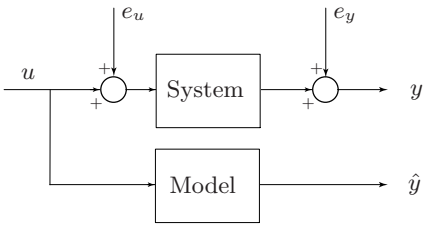



**Fig. 3.3.** Problem set-up, $u$ is a deterministic signal, $e_u$ and $e_y$ are two zero-mean uncorrelated white-noise signals. The block "model" can be any mathematical description, e.g. equation (3.26).

As shown in Figure 3.3, in most cases the input signal of the model will be chosen to be identical to the expected value of the measured input (the issue of choosing suitable inputs $u$ was discussed in Section 3.1). The output of the model $\hat{y}$ will not be equal to the measured output $y$ because of the noise signals $e_u$ and $e_y$ (both assumed to be zero mean uncorrelated white noise signals), because of system/model mismatches, because of differences

between the system's $x(0)$ and the model's $\hat{x}(0)$ initial conditions, and because of incorrect parameter values $\hat{\pi} \neq \pi$. In the following it will be assumed that only the latter effect is important. This is a substantial simplification and models obtained in this way must be used with care, especially for designing feedback controllers.

With these preparations and using an initial estimation $\hat{\pi}(0)$ of the unknown parameter values $\pi$, the model can be used to predict the output $\hat{y}$ by numerically solving the ODE (3.26). With this, an error performance index can be defined

$$\epsilon = \sum_{i=1}^{r} \rho_i \cdot (y_i(\pi) - \hat{y}_i(\hat{\pi}))^2 \tag{3.28}$$

that must be minimized by choosing appropriate parameter values $\pi_{LS}$. The weights $\rho_i \in I\!R+$ express some a priori information available on the system behavior.[5]

### 3.3.2 Nonlinear Programming Solution

In order to transform this problem into a nonlinear programming problem, a specific input $u$ is chosen and the computation of the objective function (3.28), even though it includes the solution of a system of ODE, is interpreted as just some complicated nonlinear equation $L : I\!R^q \to I\!R_+$ that relates a specific set of parameters $\hat{\pi}$ to a specific value of the error $\epsilon$. This function $L(\hat{\pi})$ is then the quantity that must be minimized. Figure 3.4 gives an overview of the structure of this numerical algorithm.

The problem of finding the minimizing parameters $\hat{\pi}_{LS}$ is, in general, a nonlinear problem, although there are situations where a nonlinear system can lead to a linear parameter estimation problem ("l.i.p." systems – linear in the parameters, see Example 3.1). Accordingly, the problem will, in general, be nonconvex, i.e., many local minima will exist and no guarantee to find the global minimum can be given.

*Example 3.1 (Car on a Plane – 2).* In Example 2.2 of Chapter 2, the equation describing the motion of a car on a plane was derived to be

$$m\frac{d}{dt}v(t) = -\{k_0 + k_1 \cdot v(t)^2\} + F(t) \tag{3.29}$$

First, it is assumed that $y = \dot{v}$, $u = F/m$, $\pi_1 = -k_0/m$, and $\pi_2 = -k_1/m$. The differential equation is nonlinear, the unknown parameters, however, influence

[5] For instance, if $u$ is a step and if only the transients are important, $\rho_i$ could tend toward zero when the model output approaches its steady-state value.

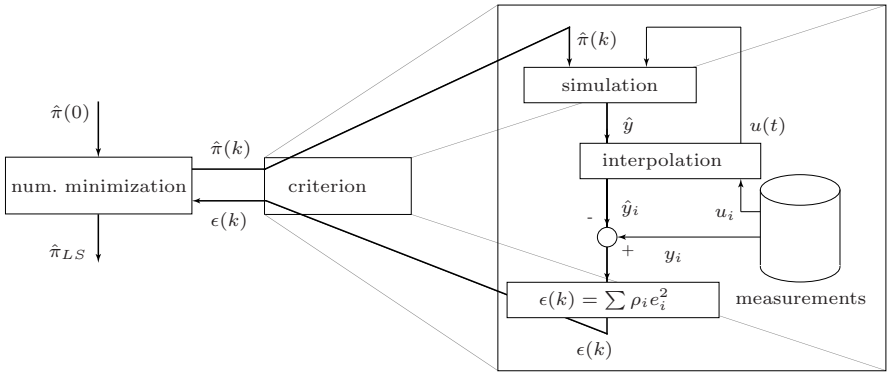



**Fig. 3.4.** Parameter identification as a nonlinear programming problem.

the output linearly. Therefore, the problem may be solved using a least-squares approach.

A typical method to estimate the unknown parameters $\pi$ is to accelerate the vehicle to a desired speed, then to reduce the propulsion force $F$ to zero (open the clutch), to measure the deceleration $y_i$ of the vehicle every $\tau$ seconds, and to store all these measurement results in one vector $\widetilde{y} = [y_1, \ldots, y_r]$.

The vehicle deceleration can then be computed using equation (3.29)

$$\begin{bmatrix} \dot{v}_1 \\ \ldots \\ \dot{v}_r \end{bmatrix} = \begin{bmatrix} 1 & v_1^2 \\ \ldots & \ldots \\ 1 & v_r^2 \end{bmatrix} \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = H \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} \tag{3.30}$$

**Remark 1:** The parameters $\pi_i$ are not simple physical parameters but represent aggregated quantities (for instance, in the case of $\pi_1$, rolling friction divided by vehicle mass). This is a typical situation in parameter estimation problems. It is rare that all individual physical parameters can be estimated directly.

**Remark 2:** The regressor $H$ in equation (3.30) depends on the velocity of the vehicle. In this problem formulation, the velocity was considered as a known input and the acceleration as the (only) output. This example shows that the physical causalities in some cases do not define what the inputs and outputs to a specific least-squares estimation problem are.

**Remark 3:** If instead of the acceleration $\dot{v}$ only the velocity $v$ can be measured, the problem is not "l.i.p." anymore. In fact, even though in this simple case the ODE (3.29) is solvable for $u = 0$

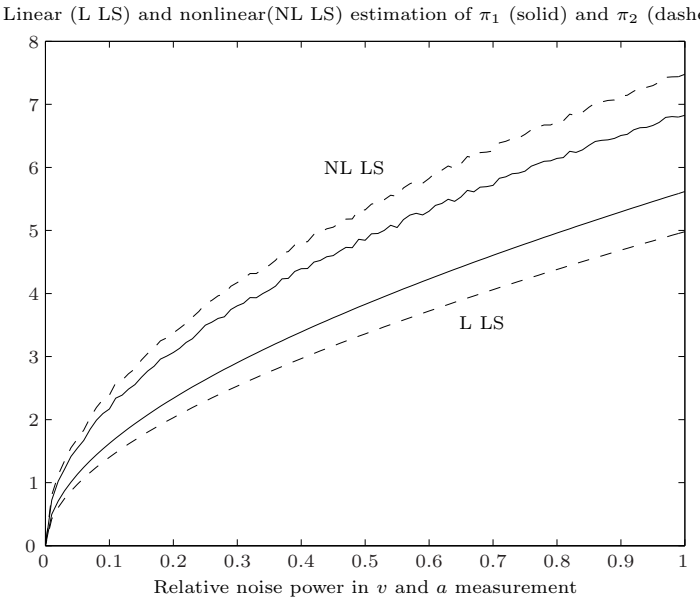



**Fig. 3.5.** Relative errors between estimated and exact values of $\pi_1$ and $\pi_2$.

$$v(t) = \sqrt{\frac{\pi_1}{\pi_2}} \cdot \tan\left\{\arctan\{\sqrt{\frac{\pi_2}{\pi_1}} \cdot v(0)\} - \sqrt{\pi_1 \cdot \pi_2} \cdot t\right\} \tag{3.31}$$

the error equation $\widetilde{v} - v$ is now nonlinear in the unknown parameters $\pi_1,\ \pi_2$ and a closed-form LS solution no longer can be obtained.

**Remark 4:** Figure 3.5 shows the results of a linear and of a nonlinear least-squares estimation of the parameters $\pi_1$ and $\pi_2$ as explained above. Notice that with decreasing noise signal power ($\sigma^2$) the estimation of the parameters improves. The estimation errors are proportional to the standard deviation $\sigma$ of the noise signals $n_u$ and $n_y$, i.e., they are proportional to the root of the noise signal power.

**Remark 5:** As Figure 3.5 shows, the result of the linear least-squares estimation is better than those of its nonlinear counterpart. This is not always the case, as many other effects influence the quality of the outcome of these two approaches (sampling times, problem duration, etc.).

**Remark 6:** In practice, such a nonlinear least-squares identification is only feasible for a very limited number of unknown parameters. For larger models, which include too many unknown parameters, a modular approach has to be used. As shown in Figure 3.6, such an approach requires intermediate output signals. The parameters of the complete model are identified sequentially by

first concentrating on the first module (using input $u$ and output $y_1$). Once the parameters of this model are identified, the subsequent modules are identified.

Two situations may arise:

- The measurable output of the module $i-1$ is equal to the input to module $i$. In this (simpler) case, a "local identification" of module $i$ using the input $y_{i-1}$ and the output $y_i$ is possible.
- The measurable output of the module $i-1$ is not the input to module $i$. In this case, the results of the identification of the previous modules must be used to identify the parameters of module $i$ (in Figure 3.6 the input would be $y_1$ and the output $y_3 = y$).

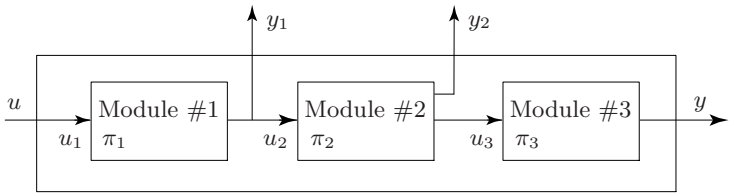



**Fig. 3.6.** Substructuring of an identification problem using intermediate outputs.

**Remark 7:** Another aspect that must be considered carefully is the problem of choosing suitable initial conditions $\pi_1$ when starting the nonlinear least-squares estimation. Obviously, the error function (3.28) will have (many) local minima and, in general, it will not be a convex function of the unknown parameters $\pi$. A good estimation of the true value of these parameters will, therefore, be necessary before starting the identification process.

**Remark 8:** A last word of caution is in order with respect to normalization. Numerical problems may arise in the identification part if incompatible physical units are used in the model formulation. Before starting the parameter identification, a normalization, as introduced in Section 4.2.1, is strongly recommended.

# 4

# Analysis of Linear Systems

**Remark:** For those students who followed the courses "Control Systems I and II," most of the material presented in this chapter will be a repetition. However, since several points are discussed in more detail and since additional material is introduced, it is not recommended to skip these classes.

## 4.1 Introduction

The models obtained in the last chapter are usually nonlinear and have physical ("non-normalized") states $z$, inputs $v$, and outputs $w$

$$\frac{d}{dt}z(t) = f(z(t), v(t), t), \qquad w(t) = g(z(t), v(t), t) \tag{4.1}$$

Unfortunately, nonlinear systems are difficult to analyze and systems with non-normalized variables are prone to numerical problems. Moreover, the single state variables are hard to compare to one another. For these reasons, a systematic normalization procedure is introduced below and the models are linearized around a chosen trajectory. In this form, several fundamental questions can be answered:

- Assuming known input signals and initial conditions, what output signals are to be expected?
- Which points in the state space may be reached by appropriate input signals?
- Which points in the state space may be observed by analyzing the associated output signals?
- What parts of a dynamic system are relevant for the input/output characteristics?

In addition to this analysis of linear systems, a few ideas and methods applicable to nonlinear systems will be presented as well.

## 4.2 Normalization and Linearization

### 4.2.1 Normalization

Normalization can be done in several ways. Here, a "setpoint-based" approach is followed, i.e., it is assumed that the system usually operates around a constant[1] point such that for all level variables $z_i(t)$ with $i = 1, \ldots, n$ there is a known and physically meaningful reference value $z_{i,0}$. Reference inputs $v_{j,0}$ $(j = 1, \ldots, m)$ and nominal outputs $w_{k,0}$ $(j = 1, \ldots, p)$ are associated to each of these setpoints $z_{i,0}$.

The normalization then consists of replacing $z_i(t)$, $v_j(t)$ and $w_k(t)$ by their normalized counterparts $x_i(t)$, $u_j(t)$ and $y_k(t)$ through the transformation

$$z_i(t) = z_{i,0}\, x_i(t), \quad v_i(t) = v_{i,0}\, u_i(t), \quad w_i(t) = w_{i,0}\, y_i(t) \tag{4.2}$$

Of course, the new variables $x_i(t)$, $u_j(t)$ and $y_k(t)$ will have no physical units.

Since $z_{i,0}$ is constant the time derivative of the normalized level variable will be

$$\frac{d}{dt} z_i(t) = z_{i,0} \frac{d}{dt} x_i(t) \tag{4.3}$$

The normalization may be expressed in vector notation as

$$z = T \cdot x, \quad T = \mathrm{diag}\{z_{1,0}, \ldots, z_{n,0}\} \tag{4.4}$$

It will be shown later that such a "similarity transformation" does not change the systems characteristics (stability, input-output behavior, etc.).

The unit of (4.3) is $1/s$. Therefore, a time normalization is proposed sometimes

$$t = \tau\, q \tag{4.5}$$

where $\tau$ is the nominal time constant of the system. However, since the models obtained with the approach introduced in the last chapter contain only "relevant dynamic phenomena" (i.e., all states are evolving on similar time scales) this additional transformation is not applied in this chapter.

*Example 4.1 (Water Tank, Normalization).*

This example is a continuation of Example 2.3. The ODE (2.10) is normalized using the constant nominal value $h_0$ for the water level and $\overset{*}{m}_{in,0}$ for the input, i.e.,

[1] In this text, it will be assumed that $z_{i,0}$ is a constant. The extension to the case in which the reference "point" is a trajectory rather than a constant is not difficult, but leads to time-varying systems.

$$h(t) = h_0\, x(t), \qquad \overset{*}{m}_{in}(t) = \overset{*}{m}_{in,0}\, u(t) \tag{4.6}$$

such that the new variables $x(t)$ and $u(t)$ will be dimensionless and have a magnitude around 1. Using the definitions (4.6), equation (2.10) can be rewritten as

$$\rho F h_0 \frac{d}{dt} x(t) = \overset{*}{m}_{in,0}\, u(t) - A\,\rho\,\sqrt{2gh_0}\,\sqrt{x(t)} \tag{4.7}$$

Notice that if the setpoint is chosen to be an equilibrium point (which is usually the case), the quantities of in-flowing and out-flowing water have to be equal in this situation

$$\overset{*}{m}_{in,0} = A\,\rho\,\sqrt{2gh_0}. \tag{4.8}$$

The example shows one very important point: if the setpoints are chosen to be equilibrium points, then they may no longer be chosen independently. In the case of Example 4.1, this resulted in a simple equality for the input and output flows. In the general case of (4.1) this can pose far more difficult problems.

*Example 4.2 (Cruising Car II, Normalization).*

This example follows up on the problem introduced in Example 2.4. Here the difficulty of a zero setpoint value will be encountered. In fact, the setpoint is $\{v_0, T_{e,0}, 0\}$, i.e., the nominal disturbance force $F_d(t)$ is assumed to be zero[2].

Of course, $F_{d,0} = 0$ cannot be used in the transformation (4.2). Therefore the disturbance $F_d$ is normalized using another variable that has a physical interpretation; in this case, twice the aerodynamic force[3]. Accordingly, the normalized system variables are chosen to be

$$v(t) = v_0\, x(t), \qquad T_e(t) = T_{e,0}\, u_1(t), \qquad F_d(t) = \rho\, c_w\, A\, v_0^2\, u_2(t). \tag{4.9}$$

The normalized version of equation (2.22) therefore has the form

$$M(\gamma, m) v_0 \frac{d}{dt} x(t) = \frac{T_{e,0}}{r_w \gamma} u_1(t) - (c_r m g + \frac{1}{2}\rho c_w A v_0^2 x^2(t) + \rho c_w A v_0^2 u_2(t)). \tag{4.10}$$

The setpoint $\{v_0, T_{e,0}, 0\}$ is again assumed to be an equilibrium point such that

$$\frac{T_{e,0}}{r_w \gamma} = c_r m g + \frac{1}{2}\rho c_w A v_0^2. \tag{4.11}$$

[2] This corresponds to an even road and no wind. It would not make sense to assume the car to be running uphill most of the time.

[3] Other choices are, of course, possible; however, this special choice will lead to particularly simple system equations.

### 4.2.2 Linearization

Linearization is the main prerequisite for system analysis and controller synthesis. In fact, a closed and complete theory for controller synthesis is available only for linear systems.

The system after normalization has the form

$$\frac{d}{dt}x(t) = f_0(x(t), u(t), t),\ \ y(t) = g_0(x(t), u(t), t) \tag{4.12}$$

with $x(t) \in I\!R^n, u \in I\!R^m, y \in I\!R^p$.

Linearization is the technique which permits one to make assertions about the system's behavior in a "small" neighborhood (a ball of radius $r$)

$$B_r := \{x \in I\!R^n \mid\ ||x - x_e||^2 + ||u - u_e||^2 \leq r\} \tag{4.13}$$

around a chosen equilibrium point $\{x_e, u_e\}$, for which $f_0(x_e, u_e, t) = 0$ holds true. Often (but not always), the equilibrium point coincides with the reference point used to normalize the system equations. Notice that the system (4.12) can have many different isolated equilibria and even equilibrium "subspaces" (manifolds). Also notice that many authors use a translation of the coordinate system to move the equilibrium point to the origin, i.e.,

$$\widetilde{x}(t) = x(t) - x_e,\ \ \widetilde{u}(t) = u(t) - x_e,\ \ \widetilde{y}(t) = y(t) - g_0(x_e, u_e, t) \tag{4.14}$$

such that in the new coordinates the system dynamics are described by the ODEs

$$\frac{d}{dt}\widetilde{x}(t) = \widetilde{f}_0(\widetilde{x}(t), \widetilde{u}(t), t),\ \ \widetilde{y}(t) = \widetilde{g}_0(\widetilde{x}(t), \widetilde{u}(t), t) \tag{4.15}$$

that satisfy $\widetilde{f}_0(0, 0, t) = 0$. Only small deviations from its *normalized* setpoints are analyzed in order to linearize the system, and the following new variables are introduced

$$\begin{aligned} x_i(t) &= x_e + \delta x_i(t) \ \text{ with } \ |\delta x_i| \ll 1, \\ u_i(t) &= u_e + \delta u_i(t) \ \text{ with } \ |\delta u_i| \ll 1, \\ y_i(t) &= y_e + \delta y_i(t) \ \text{ with } \ |\delta y_i| \ll 1 \end{aligned} \tag{4.16}$$

Using these definitions to expand equation (4.12) into a Taylor series and neglecting all terms of second and higher order, a set of linear differential equations is obtained

$$\frac{d}{dt}\delta x(t) = \frac{\partial f_0}{\partial x}|_{x_e,u_e}\delta x(t) + \frac{\partial f_0}{\partial u}|_{x_e,u_e}\delta u(t) \tag{4.17}$$

The associated set of linear output equations has the form

$$\delta y(t) = \frac{\partial g_0}{\partial x}|_{x_e,u_e}\delta x(t) + \frac{\partial g_0}{\partial u}|_{x_e,u_e}\delta u(t) \tag{4.18}$$

*Example 4.3 (Water Tank III, Linearization).*

This example follows up on the problem introduced in Examples 2.3 and 4.1. Using the definitions (4.16) to expand equation (4.7) into a Taylor series and neglecting all terms of second or higher order, the following linear differential equation is obtained

$$\rho F h_0 \frac{d}{dt}\delta x(t) = \overset{*}{m}_{in,0}(1+\delta u(t)) - A\rho\sqrt{2gh_0}(1+\frac{1}{2}\delta x(t)) \tag{4.19}$$

which using (4.8) can be simplified to

$$\rho F h_0 \frac{d}{dt}\delta x(t) = \overset{*}{m}_{in,0}\delta u(t) - A\rho\sqrt{2gh_0}\frac{1}{2}\delta x(t) \tag{4.20}$$

or, by rearranging terms and again using (4.8)

$$\tau\frac{d}{dt}\delta x(t) = 2\delta u(t) - \delta x(t) \tag{4.21}$$

where the "time constant" $\tau$ is defined by

$$\tau = \frac{F}{A}\sqrt{\frac{2h_0}{g}}. \tag{4.22}$$

This equation shows that the system becomes slower when the ratio $F/A$ or the setpoint $h_n$ are increased. The "gain" of the system is 2 in its normalized form (a consequence of the Bernoulli law) and equal to $2h_0/\overset{*}{m}_{in,0}$ in its non-normalized representation.

*Example 4.4 (Cruise Control III, Linearization).*

Linearizing equation (4.10) using the small deviations[4]

$$\begin{aligned} x(t) &= 1+\delta x(t), \quad |\delta x| \ll 1 \\ u_1(t) &= 1+\delta u_1(t), \, |\delta u_1| \ll 1 \\ u_2(t) &= 0+\delta u_2(t), \, |\delta u_2| \ll 1 \end{aligned} \tag{4.23}$$

yields

$$M(\gamma,m)v_0\frac{d}{dt}\delta x = \frac{T_{e,0}}{r_w\gamma}(1+\delta u_1) - \left[c_r mg + \frac{1}{2}\rho c_w A v_0^2\left[1+2\delta x+\delta x^2+2\delta u_2\right]\right] \tag{4.24}$$

Neglecting the second-order terms and inserting (4.11) yields the desired linear differential equation

[4] Remember: $u_2$ is a disturbance and assumed to be zero in equilibrium conditions.

$$M(\gamma, m) \cdot v_0 \cdot \frac{d}{dt}\delta x(t) \approx \frac{T_{e,0}}{r_w \cdot \gamma} \cdot \delta u_1(t) - \rho \cdot c_w \cdot A \cdot v_0^2 \cdot (\delta x(t) + \delta u_2(t)) \quad (4.25)$$

This equation can be put into its normal form by dividing all terms by the factor $\rho \cdot c_w \cdot A \cdot v_n^2$ and using equations (4.11) and (2.23)

$$\tau(\gamma,\, m,\, v_0) \cdot \frac{d}{dt}\delta x(t) = k(m, v_0) \cdot \delta u_1(t) - \delta x(t) - \delta u_2(t) \quad (4.26)$$

where

$$\tau(\gamma,\, m,\, v_0) = \frac{1}{\rho \cdot c_w \cdot A \cdot v_0} \left( m + \frac{\Theta_e}{\gamma^2 \cdot r_w^2} \right) \quad (4.27)$$

and

$$k(m,\, v_0) = \frac{1}{2} + \frac{c_r \cdot m \cdot g}{\rho \cdot c_w \cdot A \cdot v_0^2} \quad (4.28)$$

Both $\tau$ and $k$ are functions of the physical parameters. Since for a given vehicle only the mass $m$, the gear ratio $\gamma$ and the vehicle speed setpoint $v_n$ can vary substantially, the dependency of the control-oriented parameters $\tau$ and $k$ on these physical parameters is emphasized by using the notation ...$(\gamma,\, m,\, v_n)$.

In general, the linearized system equations thus obtained will have the form

$$\begin{aligned} \tfrac{d}{dt}\delta x(t) &= A\delta x(t) + B\delta u(t) \\ \delta y(t) &= C\delta x(t) + D\delta u(t) \end{aligned} \quad (4.29)$$

For the sake of simplicity, the prefix $\delta$ will often be omitted below. This system respresentation forms the basis of "modern" control design methods and will be used widely in the following parts.

Notice that the system (4.29) may be described using (infinitely) many other coordinate systems. The change of coordinates is given by the similarity transformation

$$x = T\widetilde{x}, \qquad T \in I\!R^{n \times n}, \quad det(T) \neq 0 \quad (4.30)$$

where the columns of $T$ are the unit vectors of the new coordinate frame expressed in the old coordinate system. In the new coordinates, the system is described by

$$\begin{aligned} \tfrac{d}{dt}\widetilde{x}(t) &= T^{-1}AT\widetilde{x}(t) + T^{-1}Bu(t) \\ y(t) &= CT\widetilde{x}(t) + Du(t) \end{aligned} \quad (4.31)$$

The fundamental system properties (IO-behavior, stability, controllability, etc.) are independent of the coordinates chosen. However, if (4.29) is derived by physical arguments (as shown above) there are strong reasons for sticking to these "natural" coordinates.

## 4.3 Solution of Linear ODE

One of the most important problems in the analysis of dynamic systems is the prediction of the state $x$ and the output $y$ of a linear system $\{A, B, C, D\}$ when its initial condition $x(0)$ and the control signal $u$ are known ("initial value problem").

The emphasis here is on arbitrary inputs $u$ (so Laplace transform or other closed-form methods are not applicable). The existence of a solution is guaranteed for sufficiently reasonable inputs (i.e., piecewise continuous signals with countable discontinuities, see [20]).

The actual computations will be a combination of a large number of "preprocessing" computations and a relatively small number of on-line computations. Matrix exponentials lie at the core of this problem

$$e^{At} = I + \frac{1}{1!}At + \frac{1}{2!}(At)^2 + \frac{1}{3!}(At)^3 + \cdots + \frac{1}{n!}(At)^n + \cdots \tag{4.32}$$

where $A \in \mathbb{R}^{n\times n}$ is a real square matrix. As in the scalar case, the matrix exponential also satisfies

$$\frac{de^{At}}{dt} = A\,e^{At} = e^{At}\,A \tag{4.33}$$

where in equation (4.33) it is emphasized that $e^{At}$ and $A$ commute. However, there are differences to the scalar case, e.g., it is true that

$$e^a \cdot e^b = e^{a+b} \tag{4.34}$$

while in the matrix case in general

$$e^A \cdot e^B \neq e^{A+B} \tag{4.35}$$

Only if $A$ and $B$ commute ($AB = BA$), then

$$e^A \cdot e^B = e^{A+B} \tag{4.36}$$

and since obviously $At$ and $A\tau$ do commute for arbitrary $t, \tau \in \mathbb{R}$, it follows that

$$e^{At}e^{-At} = e^{A(t-t)} = e^0 = I \tag{4.37}$$

and therefore

$$(e^{At})^{-1} = e^{-At} \tag{4.38}$$

In "Regelungstechnik I" the "explicit" solution to the initial value problem

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \tag{4.39}$$

was shown to be

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}B\,u(\tau)\,d\tau \tag{4.40}$$

The output signal is then given by

$$y(t) = C\,e^{At}x_0 + \int_0^t C\,e^{A(t-\tau)}B\,u(\tau)\,d\tau + D\,u(t) \tag{4.41}$$

Since the system is assumed to be time-invariant, $t_0$ may be assumed (without loss of generality) to be equal to zero.

The matrix $e^{At}$ is often referred to as the "transition matrix" $\Phi_t^A$. With this notation, the "flow" of a system can be concisely captured, e.g.

$$\Phi_{t_1}^A\,\Phi_{t_2}^A = \Phi_{t_1+t_2}^A \text{ und } \left[\Phi_t^A\right]^{-1} = \Phi_{-t}^A \tag{4.42}$$

The second term in equation (4.40) is sometimes called the "convolution operator"

$$\sigma(t) * u(t) = \int_0^t \sigma(t-\rho)\,u(\rho)\,d\rho \tag{4.43}$$

and the function

$$\sigma(t) = C\,e^{At}B \tag{4.44}$$

is the system's impulse response.

The explicit computation of the matrix exponential can be carried out following the definition (4.32). However, this approach is not recommended since it may cause numerical problems. In all modern CACSD packages much better algorithms are available [5] such that even large-scale problems can be solved efficiently and reliably.

## 4.4 Properties of Linear Systems

### 4.4.1 Jordan Forms and Stability of Linear Systems

Stability is the most important concept in the analysis of dynamic systems. There are several definitions of stability and only the most important ones will be presented below. Stability is always connected to a metric, i.e., the size of vectors is important. The symbol $||x||$ will be used to denote this operation, and any norm will be acceptable, for instance the Euclidean length

$$x \in I\!R^n, \qquad ||x||^2 := \sum_{i=1}^{n} x_i^2 \tag{4.45}$$

For nonlinear and time-varying systems, the definition of stability is mathematically not straightforward and will be discussed in more detail in Section 5.2. In the linear and time-invariant case rather intuitive definitions may be used. Moreover, in this case stability is a global concept, i.e., if the equilibrium point $x = 0$ is stable, then this is true for all finite initial conditions $x(0)$ in which the system may be at $t = 0$.

A pragmatic definition of stability valid for a linear time-invariant system

$$\frac{d}{dt}x(t) = A \cdot x(t), \quad x(0) = x_0, \quad 0 < ||x_0|| < \infty \tag{4.46}$$

distinguishes three possible cases. The system (4.46) is defined to be

- asymptotically stable if $\lim_{t\to\infty} ||x(t)|| = 0$;
- stable if $||x(t)|| < \infty \;\; \forall\, t \in [0, \infty]$; and
- unstable if $\lim_{t\to\infty} ||x(t)|| = \infty$,

where the norm is the Euclidean length (4.45). It will be shown that the question of stability of this linear system is intimately related to the eigenvalues of the matrix $A$. Therefore, the general eigenstructure of a matrix $A$ must be derived as a first step.

### Jordan Forms

The matrix $A$ is associated to the linear function that maps elements $x$ of a real vector space $\mathbb{R}^n$ into elements $y$ in the same space

$$A \; : \mathbb{R}^n \;\; \rightarrow \;\; \mathbb{R}^n, \tag{4.47}$$

according to the relation

$$y \;=\; A\, x, \quad x, y \;\in \mathbb{R}^n, \quad A \in \mathbb{R}^{n\times n} \tag{4.48}$$

An eigenvector $v_i$ of $A$ is a vector which is mapped, modulo a length change, onto itself, i.e.,

$$A\; v_i \;=\; \lambda_i\; v_i \tag{4.49}$$

The scaling parameter $\lambda_i$ will be defined below. Equation 4.49 is a homogenous equation for $v_i$ and it has a non-trivial solution iff the matrix

$$(\lambda_i\, I - A)\, v_i = 0 \tag{4.50}$$

is singular, i.e., has a rank smaller than $n$. Accordingly, the eigenvalues $\lambda_i$ must be chosen such that

$$\det(\lambda_i\, I - A) = 0 \tag{4.51}$$

Even for real matrices $A \in \mathbb{R}^{n \times n}$ the eigenvalues $\lambda_i$ and the eigenvectors $v_i$ are, in general, complex entities. However, in this case they always arise in complex conjugate pairs, such that the linear combination of these pairs is again real.

If $n$ *linearly independent* eigenvectors exist, then these eigenvectors can be used to form a similarity transformation

$$V = [v_1, \ldots, v_n] \tag{4.52}$$

which – by definition – diagonalizes $A$, i.e.,

$$A\,V \;=\; V\,\Lambda \quad \Rightarrow \quad V^{-1}\,A\,V \;=\; \Lambda \tag{4.53}$$

where

$$\Lambda \;=\; \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ \ldots & & & \ldots \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} \tag{4.54}$$

If all eigenvalues are distinct, i.e., $\lambda_i \neq \lambda_j$ for all $i \neq j$, then $n$ independent eigenvectors always exist [3]. If there are multiple eigenvalues, then the situation is more complex. Two scalars[5] are important in that discussion:

- the multiplicity $r_i$ of the corresponding eigenvalue $\lambda_i$; and
- the rank loss $\rho_i$ of the matrix $\lambda_i I - A$ associated with the eigenvalue $\lambda_i$.

When all eigenvalues are distinct, of course $r_i = \rho_i = 1$ for all $i = 1, \ldots, n$ and the matrix $A$ can be diagonalized by the similarity transformation (4.52).

In the general case, three distinct situations arise:

1 the rank loss $\rho_i = 1$, i.e., only one independent eigenvector exists for the $r_i > 1$ identical eigenvalues $\lambda_i$;
2 the rank loss $\rho_i < r_i$, i.e., less than the multiplicity of $\lambda_i$ independent eigenvectors exist; or
3 the rank loss $\rho_i = r_i$, i.e., sufficient independent eigenvectors exist to diagonalize that part of $A$ that belongs to $\lambda_i$.

The third case is similar to the regular case. If all multiple eigenvalues of $A$ satisfy that condition, then $A$ can be diagonalized.[6]

In the first and in the second case, $A$ is not diagonalizable but can be brought to what is known as a "Jordan Form." In this form, the matrix $A$ is similar to a block diagonal matrix, each block being associated to one eigenvalue. The form of a single block is almost diagonal but can have 1 on the upper diagonal. The exact structure depends on the rank loss. In the first case, where $\rho_i = 1$, the Jordan block is full, i.e.,

[5] Sometimes $r_i$ is denoted as the *algebraic* and $\rho_i$ as the *geometric* multiplicity.
[6] In this case the matrix $A$ is said to be semi-simple.

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_i & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & 0 & \lambda_i \end{bmatrix} \tag{4.55}$$

Matrices that satisfy the condition $\rho_i = 1$ for all multiple eigenvalues are called "cyclic." In the second case, where $1 < \rho_i < r_i$, the i-th Jordan block is not full but contains $r_i - \rho_i$ elements 1 in the upper semi-diagonal. Matrices of this type are neither diagonal nor cyclic and will be denoted below as "mixed." Figure 4.1 gives an overview of the possible cases.

For cyclic or mixed matrices the transformation (4.52) cannot be formed because there are fewer than $n$ linearly independent eigenvectors. In this situation, *generalized* eigenvectors are required to obtain a transformation $x = \widetilde{V} \cdot z$ with $\det(\widetilde{V}) \neq 0$ that transforms the system matrix $A$ of (4.46) into a Jordan form [3]. The following recursion defines the necessary generalized eigenvectors that are associated to a multiple eigenvalue $\lambda_i$ with $\rho_i = 1$

$$\begin{aligned} (\lambda_i I - A) \cdot v_i &= 0 \\ (\lambda_i I - A) \cdot w_{i,1} &= v_i \\ \dots &\quad \dots \\ (\lambda_i I - A) \cdot w_{i,r_i-\rho_i} &= w_{i,r_i-\rho_i-1} \end{aligned}$$

In the case where $1 < \rho_i < r_i$ the iteration can be aborted once sufficient generalized eigenvectors are found to form a complete transformation $\widetilde{V}$.

### Stability of Linear Systems

The question of stability of a linear system can be answered by combining the expression (4.40) for the general solution of a linear system with the coordinate transformation (4.52). Starting with the system

$$\frac{d}{dt}x(t) = Ax(t), \qquad x(0) = x_0 \neq 0 \tag{4.56}$$

(the input $u(t) = 0$ for stability analysis) and applying the eigentransformation $x = Vz$ (4.52), the Jordan system

$$\frac{d}{dt}z(t) = V^{-1}AVz(t), \qquad z(0) = V^{-1}x_0 \neq 0 \tag{4.57}$$

is obtained. Notice that the system (4.57) has the same stability properties as the system (4.56) because

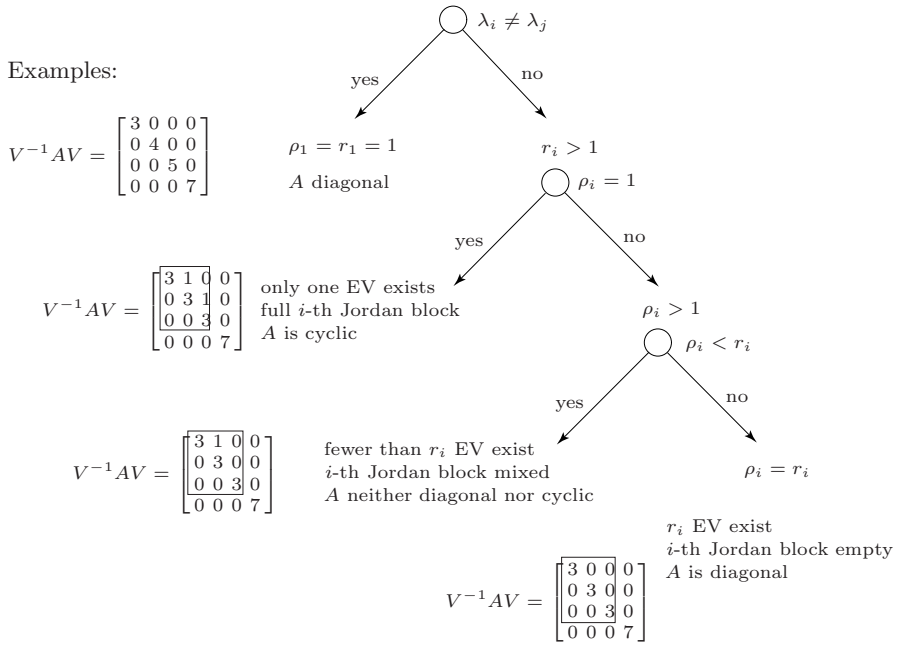



**Fig. 4.1.** Classification of matrix eigenstructures.

$$||x||^2 = x^T x = z^T V^T V z \quad \Rightarrow \quad \sigma_{min}(T)||z|| \leq ||x|| \leq \sigma_{max}(T)||z|| \tag{4.58}$$

Therefore, if $||z|| \to 0$, then also $||x|| \to 0$, and if $||z(t)|| < \infty \; \forall \; t$, then also $||x(t)|| < \infty \; \forall \; t$ (the constants $\sigma_i$ are the singular values of the matrix $V$ and – since $\det(V) \neq 0$ – they cannot be zero).

Using the general solution (4.40) the following equation is obtained

$$Vz(t) = e^{At}Tz(0) \Rightarrow z(t) = V^{-1}e^{At}Vz(0) \tag{4.59}$$

and using the definition (4.32), $z(t)$ can be written as

$$z(t) = V^{-1}\left\{I + \frac{1}{1!}At + \frac{1}{2!}(At)^2 + \frac{1}{3!}(At)^3 + \cdots\right\} V\, z(0) \tag{4.60}$$

Multiplying the factors $V^{-1}$ and $V$ inside the parentheses and adding the identity $I = V^{-1}V$ between each pair $A$ $A$, the last equation can be written as

$$\begin{aligned} z(t) = \big\{I + \tfrac{1}{1!}V^{-1}AV\,t + \tfrac{1}{2!}V^{-1}AVV^{-1}AV\,t^2 \\ + \tfrac{1}{3!}V^{-1}AVV^{-1}AVV^{-1}AV\,t^3 + \cdots\big\}\, z(0) \end{aligned} \tag{4.61}$$

$$\Rightarrow \quad z(t) = \left\{I + \frac{1}{1!}\Lambda\, t + \frac{1}{2!}\Lambda^2 t^2 + \frac{1}{3!}\Lambda^3 t^3 + \cdots\right\} z(0) = e^{\Lambda t} z(0) \tag{4.62}$$

The case in which the matrix $\Lambda$ is diagonal is analyzed first. In this case, for each element of the vector $z(t)$, the following relation is valid

$$z_i(t) = e^{\lambda_i t} z_i(0) = e^{\sigma_i t} \left[\cos(\omega_i t) + j \sin(\omega_i t)\right] z_i(0) \quad i = 1, \cdots, n \tag{4.63}$$

where $\sigma_i$ is the real and $\omega_i$ the imaginary part of the $i$-th eigenvalue $\lambda_i$. Three situations can therefore be encountered:

1. all eigenvalues have negative real parts $\sigma_i < 0$;
2. some eigenvalues have zero real parts $\sigma_i = 0$, but none has a positive real part; or
3. At least one eigenvalue has a positive real part $\sigma_i > 0$.

In the first case, obviously $z_i(\infty) = 0$ for all $i = 1, \cdots, n$ and therefore also $||x(\infty)|| = 0$ (see (4.58)). The system is therefore (exponentially) asymptotically stable.

In the third case, one or more $z_i(t)$ go to $\infty$, and, in general, all elements of $x(t)$ also will tend to $\infty$, i.e., the system is unstable.

The limiting case between these two extremes is that some eigenvalues have zero real parts. Using the series expansion (4.63) it is clear that the $z_i(t)$ associated with the zero real part eigenvalues do not tend to 0 with time going to infinity, but remain finite in magnitude. The system is therefore stable, but not asymptotically stable.

It must be emphasized that the last result is only valid for systems (4.56) with matrices $A$ which are similar to a diagonal matrix. For cyclic or mixed matrices $A$, this is no longer true. In fact, systems with multiple eigenvalues on the imaginary axis can be stable only if the corresponding Jordan blocks are diagonal.[7] Systems with mixed or cyclic Jordan blocks associated to multiple eigenvalues on the imaginary axis will always have some state variables growing without bounds as $t \to \infty$. The two most important examples of matrices $A$ that have cyclic structures and multiple eigenvalues on the imaginary axis are discussed below.

### Two Examples of Critical Cyclic Systems

The first example of a critical cyclic, and hence unstable dynamic system is the "series double integrator structure"

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{4.64}$$

[7] All other eigenvalues are, of course, assumed to have negative real parts.

These equations model several physical phenomena, for instance a mass point acceleration in free space. Obviously, the system has two eigenvalues $\lambda_1 = \lambda_2 = 0$ and since $r_1 = 2$, $\rho_1 = 1$ its structure is cyclic.

In this simple case the system description is already in Jordan form and it is easy to verify that the system is unstable using the general solution (4.40) for $u = 0$ and the definition of the exponential function (4.32)

$$x(t) = \left\{ I + \frac{1}{1!} A\, t + \frac{1}{2!} A^2\, t^2 + \cdots \right\} x(0) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} x(0) \tag{4.65}$$

because $A^2 = A^3 = \ldots = 0$. Accordingly, there is no $x_2(0)$ with $|x_2(0)| \neq 0$ sufficiently small to prevent $|x_1(t)|$ to grow without bounds, i.e., the system is Lyapunov unstable. A physical interpretation of this observation is that if the mass point has a non-zero initial velocity $x_2(0) \neq 0$ the position $x_1(t)$ will grow linearly with time and tend to infinity for $t \to \infty$.

The second example analyzed in this section has the form

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & c & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \tag{4.66}$$

Figure 4.2 shows a block-diagram representation of the dynamic system associated to this matrix $A$. The value of the parameter $c$ can be either 1 or 0. In both cases, the eigenvalues of this $A$ are $\lambda_1 = \lambda_2 = +j$ and $\lambda_3 = \lambda_4 = -j$, i.e., the system is critical and since there are multiple eigenvalues on the imaginary axis it can be Lyapunov stable or unstable, depending on the structural properties of $A$.

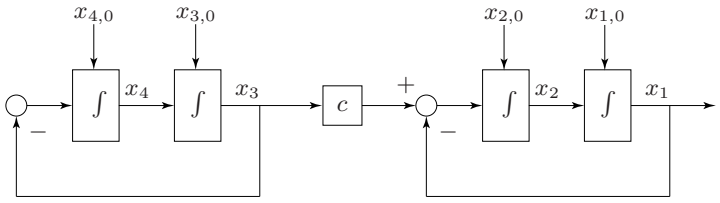



**Fig. 4.2.** Block diagram of the system of Example (4.66).

In the case that $c = 0$, the structural properties are $r_1 = r_2 = 2$ and $\rho_1 = \rho_2 = 2$, i.e. the matrix $A$ is semi-simple and can be transformed into a diagonal matrix. This corresponds to two independent undamped oscillators that each have the same eigenfrequency $\omega = 1$. The amplitude of these oscillations is constant and, therefore, the system is Lyapunov stable.

In case $c = 1$, the structural properties are $r_1 = r_2 = 2$ and $\rho_1 = \rho_2 = 1$, i.e. the matrix $A$ is cyclic and, therefore, the corresponding system is unstable. The objective of this example is to give an intuitive justification for this.

The starting point is the Jordan form of $A$

$$J = \begin{bmatrix} j & 1 & 0 & 0 \\ 0 & j & 0 & 0 \\ 0 & 0 & -j & 1 \\ 0 & 0 & 0 & -j \end{bmatrix} \tag{4.67}$$

The transformation $\widetilde{V}$ that transforms $A$ into $J = \widetilde{V}^{-1} A \widetilde{V}$ is defined by

$$\widetilde{V} = [v_1, w_1, v_2, w_2] \tag{4.68}$$

where the eigenvectors $v_1$ and $v_2$ are given by

$$v_1 = \begin{bmatrix} 1 \\ j \\ 0 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 1 \\ -j \\ 0 \\ 0 \end{bmatrix} \tag{4.69}$$

The vectors $w_1$ and $w_2$ are the generalized eigenvectors defined by

$$(\lambda_1 I - A)\, w_1 = v_1, \quad (\lambda_2 I - A)\, w_2 = v_2 \tag{4.70}$$

and their values are

$$w_1 = \begin{bmatrix} 0 \\ 1 \\ j\,2 \\ -2 \end{bmatrix}, \quad w_2 = \begin{bmatrix} 0 \\ 1 \\ -j\,2 \\ -2 \end{bmatrix} \tag{4.71}$$

The transformation $\widetilde{V}$ transforms the physical coordinates $x$ into the modal coordinates $z$ according to the rule $\widetilde{V} z = x$. In these modal coordinates $z$ two *independent* subsystems can be analyzed

$$\frac{d}{dt}\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} j & 1 \\ 0 & j \end{bmatrix}\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix}, \quad \frac{d}{dt}\begin{bmatrix} z_3(t) \\ z_4(t) \end{bmatrix} = \begin{bmatrix} -j & 1 \\ 0 & -j \end{bmatrix}\begin{bmatrix} z_3(t) \\ z_4(t) \end{bmatrix} \tag{4.72}$$

Notice that these two subsystems each are similar to the "double series integrator" system (4.64), the difference being that instead of $\lambda_1 = \lambda_2 = 0$ now the eigenvalues are either $\lambda_1 = \lambda_2 = j$ or $\lambda_3 = \lambda_4 = -j$. The solutions (4.40) of these linear differential equations are

$$\begin{aligned} z_1(t) &= e^{j\,t}\, z_1(0) + t\, e^{j\,t}\, z_2(0) \\ z_2(t) &= e^{j\,t}\, z_2(0) \\ z_3(t) &= e^{-j\,t}\, z_3(0) + t\, e^{-j\,t}\, z_4(0) \\ z_4(t) &= e^{-j\,t}\, z_4(0) \end{aligned}$$

which, using Euler's equation, can be written as

$$
\begin{aligned}
z_1(t) &= (\cos(t) + j\ \sin(t))\, z_1(0) + t\,(\cos(t) + j\ \sin(t))\, z_2(0) \\
z_2(t) &= (\cos(t) + j\ \sin(t))\, z_2(0) \\
z_3(t) &= (\cos(t) - j\ \sin(t))\, z_3(0) + t\,(\cos(t) - j\ \sin(t))\, z_4(0) \\
z_4(t) &= (\cos(t) - j\ \sin(t))\, z_4(0)
\end{aligned}
$$

Figure 4.3 shows the time behavior of $z_1(t)$ and $z_2(t)$ for the chosen initial conditions. For $z_3(t)$ and $z_4(t)$ very similar results are obtained, only the rotation in the complex plane is clockwise.

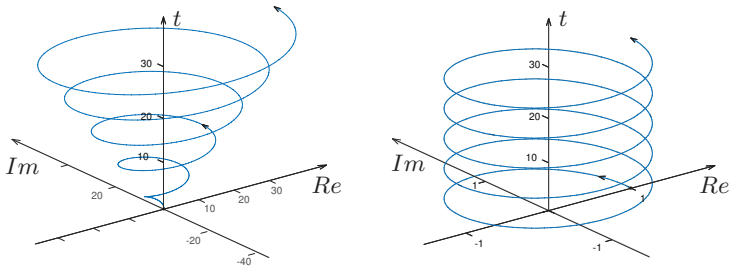



**Fig. 4.3.** Left $z_1(t)$, right $z_2(t)$; initial conditions $z_1(0) = 0 + j\,0$ and $z_2(0) = 1 + j\,0$.

The variables $z_i(t)$ are complex, but that is just a mathematical technicality. The main point that becomes clearly visible in Figure 4.3 is that two first-order systems with identical eigenvalues (in the previous example that eigenvalue was 0, in this example it is either $j$ or $-j$), when connected in series, lead to an ever increasing amplitude of the second system. For the simple structure chosen in this example, this behavior can be inferred immediately, for instance, from Figure 4.2. In more complex cases, however, this resonance effects can only be discovered by the analysis shown above.

For the sake of completeness, Figure 4.4 shows the transients of the state variables $x_1(t)$ and $x_3(t)$ defined in Figure 4.2. To obtain these variables, one simply uses the eigentransformation introduced in equation (4.68) and the solution $z(t)$ derived above. The output of the first oscillator is a harmonic signal whose frequency coincides with the resonant frequency of the second oscillator. Clearly, this leads to an ever-increasing amplitude of that oscillation. Since this requires a constant influx of energy into the system, it is clear that no passive mechanical system can exist that exhibits this behavior. An external power source is needed for that.

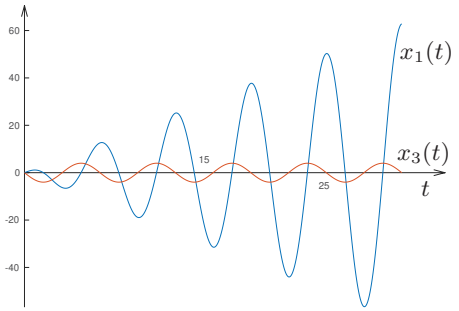



**Fig. 4.4.** Time-domain behavior of $x_1(t)$ and $x_3(t)$ as defined in Figure 4.2.

### 4.4.2 Reachability and Observability of Linear Systems

#### Introduction

Two fundamental questions concerning the linear system

$$\frac{d}{dt}x(t) \;=\; A\cdot x(t) \;+\; B\cdot u(t), \quad y(t) \;=\; C\cdot x(t) \;+\; D\cdot u(t) \tag{4.73}$$

are:

- What points $x_0 \in I\!R^n$ can be reached in finite time $T$ by finite energy inputs $u$ when $x(0) = 0$?
- What initial conditions $x(0) \in I\!R^n$ can be reconstructed using only the output $y(t)$?

It can be shown that the set of reachable points $\mathcal{R}$ and the set of observable points $\mathcal{O}$ form a linear vector (sub)space in $I\!R^n$. The state space $I\!R^n$ can therefore be subdivided into four subspaces, as shown in Figure 4.5. The arrows show influence relations between the different subspaces. The rules used to derive that structure are:

- The input $u$ may only act on the reachable subspaces.
- The output $y$ may only be influenced by the observable subspaces.
- The not reachable subspaces may not be influenced by a subspace that is influenced by the input $u$.
- The not observable subspaces may not influence a subspace that itself influences $y$.

Accordingly, there must be a choice of coordinates $z$ in which the system (4.73) is described in a simple structure, as shown in Figure 4.6. Notice that the system matrices relevant for the input-output behavior ("minimal realization") are the quadruple $\{A_{33}, B_3, C_3, D\}$, only. In other words, the transfer function $u \to y$ is given by

$$P(s) = C_3(sI - A_{33})^{-1}B_3 + D \tag{4.74}$$

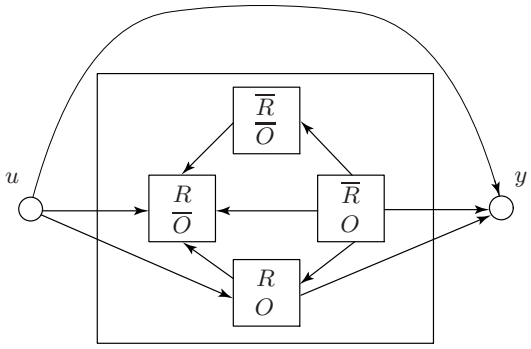



**Fig. 4.5.** Structure of state space with reachable and observable subspaces.

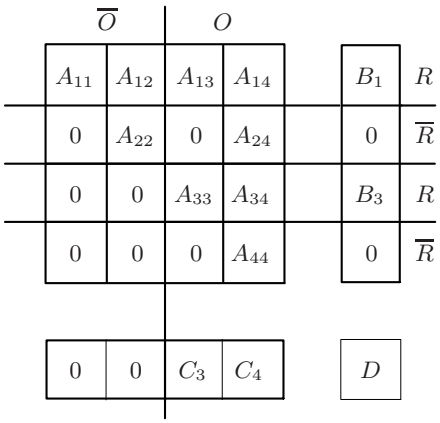



**Fig. 4.6.** Canonical coordinate system.

In the following, explicit conditions for the reachability and observability questions will be derived.

### Cayley-Hamilton Theorem

Before being able to find these conditions, an important result of matrix algebra must be introduced. As mentioned above, the eigenvalues of a matrix $A$ coincide with the roots of the "characteristic polynomial" $\Pi(s)$ which is defined by

$$\Pi(s) = \det(sI - A) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0 \tag{4.75}$$

Since the computation of a matrix determinant involves only multiplications and additions of the matrix elements, the expression (4.75) is indeed a polynomial in $s$. The equivalence between eigenvalues of $A$ and roots of $\Pi(s)$ is

visible from the definition (4.49) of the eigenproblem. In fact, the equation

$$(\lambda_i I - A)v_i = 0 \tag{4.76}$$

has only nontrivial solutions for those $\lambda_i$ for which the matrix $\lambda_i I - A$ has less than full rank. This is equivalent to the requirement that $\det(sI - A)$ is equal to zero. With these preliminaries, it is now possible to state the Cayley-Hamilton theorem:

*Each square matrix $A$ satisfies its own characteristic polynomial,*[8] *i.e.,*

$$A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I = 0 \tag{4.77}$$

The most important consequence of this theorem is that the n-th power of $A$ does not introduce any new linearly independent directions in the sequence of matrices (4.77) that have not been introduced in the other terms up to the (n-1)-th power of $A$.

### Reachability Conditions

Using the general solution (4.40), the set of all states that can be reached at time $\tau$ starting at $x(0) = 0$ can be written as

$$x(\tau) = e^{A\tau} \int_0^\tau e^{-A\sigma} B\, u(\sigma)\, d\sigma \tag{4.78}$$

The matrix $e^{A\tau}$ is regular for each instant $\tau < \infty$[9]. Therefore, instead of analyzing $x(\tau)$ the unique state $x^*(\tau) = e^{-A\tau}x(\tau)$ will be considered first

$$x^*(\tau) = \int_0^\tau e^{-A\sigma} B\, u(\sigma)\, d\sigma \tag{4.79}$$

Using the definition (4.32) for the matrix exponential, the last equation can be rewritten as

$$x^*(\tau) = \int_0^\tau \left\{ I - \frac{1}{1!}A\sigma + \frac{1}{2!}(A\sigma)^2 - \frac{1}{3!}(A\sigma)^3 + \cdots \right\} B\, u(\sigma)\, d\sigma \tag{4.80}$$

and from that

$$x^*(\tau) = B \int_0^\tau u(\sigma)\, d\sigma - AB \int_0^\tau \frac{1}{1!}\sigma u(\sigma)\, d\sigma + A^2 B \int_0^\tau \frac{1}{2!}\sigma^2 u(\sigma)\, d\sigma \cdots \tag{4.81}$$

[8] Pro memoria: $\det(\lambda I - A) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_1\lambda + a_0$ with the real constants $a_i$ is the chracteristic polynomial of $A$.

[9] This can be proven using equation (4.42) and the fact that $\det(A \cdot B) = \det(A) \cdot \det(B)$.

By choosing a specific control signal $u$, one can independently choose the terms

$$v_i = \int_0^\tau \frac{(-1)^i}{i!}\sigma^i u(\sigma)\, d\sigma, \qquad i = 0, 1, \ldots \tag{4.82}$$

The reachable states $x^*$ are then defined by the linear equation

$$x^*(\tau) = \begin{bmatrix} B & AB & A^2B & A^3B & \cdots \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ \vdots \end{bmatrix} = R\,v \tag{4.83}$$

Equation (4.83) can have a solution $v$ if, and only if, the vector $x^*$ is within the range of the matrix $R$. Obviously, the system (4.73) is *completely reachable* (all points in $\mathbb{R}^n$ can be reached from $x(0) = 0$) if, and only if, the matrix $R$ has full rank. If $R$ spans only a subspace of $\mathbb{R}^n$ then only the points $x^*$ contained in that subspace may be reached. Moreover, all states $x(\tau) = e^{A\tau}x^*(\tau)$ must be in the same subspace of $\mathbb{R}^n$ as $x^*(\tau)$ is[10] such that these assertions are valid for the original problem as well.

Notice that there is a potential problem with the application of this result. In fact, the matrix $R$ has $n$ rows but infinitely many columns, therefore the task of identifying the range of $R$ seems to be a difficult one. Fortunately, the Cayley-Hamilton theorem is available, such that the search for a new direction in $\mathbb{R}^n$ can be aborted after the n-th power of $A$, i.e., the truncated matrix

$$\mathcal{R}_n = \begin{bmatrix} B & AB & A^2B & A^3B & \cdots & A^{n-1}B \end{bmatrix} \tag{4.84}$$

fully defines the reachable set for $x^*$.

Sometimes the dual notion of controllable states is used, which identifies those initial conditions $x(0) \neq 0$ that can be forced to the origin in finite time by a suitable control signal $u(t)$. For linear continuous-time systems, the set of reachable and controllable states is identical. Especially if a system is completely reachable, then it is completely controllable as well.

It has to be emphasized that there is a fundamental difference between controllability and reachability in the following sense. While a completely controllable system may be brought to the origin and *kept there* (simply by setting $u(t) = 0$ for $t > T$), starting at any initial condition $x(0)$, a completely reachable system can only be brought to a certain point $x(T)$ in the state space at time $t = T$. However, in general, it cannot be forced to stay there for $t > T$ by any control action.

[10] Multiplying $R$ by $e^{A\tau}$ does not add any new directions as definition (4.42) shows, and again using the Cayley-Hamilton theorem, the assertion follows.

**Observability Conditions**

In the observability problem, the question is analyzed whether it is possible to reconstruct the initial condition $x(0)$ of the system (4.73) by analyzing only the output signal $y(t)$.

Since only the initial condition is of interest here, it is possible to let $u(t) = 0$ for all times $t$. Under this condition, the output and its derivatives are defined by

$$y(t) = Cx(t)\ \tfrac{d}{dt}y(t) = CAx(t)\ \tfrac{d^2}{dt^2}y(t) = CA^2x(t)\ \tfrac{d^3}{dt^3}y(t) = CA^3x(t)\ \text{etc.} \tag{4.85}$$

Evaluating this equation for $t = 0$ yields the linear equation

$$\begin{bmatrix} y(t) \\ \frac{d}{dt}y(t) \\ \frac{d}{dt^2}y(t) \\ \vdots \end{bmatrix}_{(t=0)} = w(0) = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \cdot x(0) = O \cdot x(0) \tag{4.86}$$

By "measuring" the vector $w(0)$ (assume that this can be done by a set of differentiators) the unknown initial condition $x(0)$ can therefore be reconstructed uniquely if, and only if, the kernel of the matrix $O$ is empty. In fact, if this is not true, then there are two different initial conditions, $x_1(0)$ and $x_2(0)$, which produce the same output $w(0)$

$$w(0) = O \cdot x_1(0) \tag{4.87}$$

$$w(0) = O \cdot x_2(0) \tag{4.88}$$

Of course, in this case the difference $\Delta x(0) = x_1(0) - x_2(0)$ satisfies the equation

$$0 = O \cdot (x_1 - x_2) = O \cdot \Delta x(0) \quad \Rightarrow \quad \Delta x \in \mathrm{Ker}\,\{O\} \tag{4.89}$$

The initial condition $x(0)$ is uniquely determined by (4.86) if, and only if, the rank of $O$ is full (i.e., the kernel of $O$ contains only the element 0)

Again, thanks to the Cayley-Hamilton theorem, only the first $n$ blocks of the matrix $O$ must be analyzed. The truncated observability matrix

$$\mathcal{O}_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{4.90}$$

is therefore sufficient to decide whether or not the system (4.73) is completely observable.

*Example 4.5 (An Academic System).* The system

$$A = \begin{bmatrix} -0.7 & 0.2 & -0.1 \\ -1.5 & -2.0 & 0.5 \\ 0.9 & 0.6 & -1.3 \end{bmatrix} \; b = \begin{bmatrix} -3 \\ 5 \\ 1 \end{bmatrix} \; c^T = \begin{bmatrix} -2 \\ 0 \\ 4 \end{bmatrix}$$

with $\mathcal{R}_n$ and $\mathcal{O}_n$ defined by

$$\mathcal{R}_n = \begin{bmatrix} -3 & 3 & -3 \\ 5 & -5 & 5 \\ 1 & -1 & 1 \end{bmatrix} \; \mathcal{O}_n = \begin{bmatrix} -2 & 0 & 4 \\ 5 & 2 & -5 \\ -11 & -6 & 7 \end{bmatrix}$$

is neither completely controllable nor completely observable. On the one hand only the states on the one-dimensional submanifold $\mu[-3, 5, 1]^T$ ($\mu \in \mathbb{R}$) can be reached starting in the origin by any control signal $u(t)$. On the other hand, by analyzing the output signal $y(t)$, it is not possible to distinguish any initial conditions that differ by a vector $\mu[0.5963\cdots, -0.7454\cdots, 0.2981\cdots]^T$.

*Example 4.6 (Ball-on-Wheel).*

The system analyzed in this example is the plant (2.8) that was introduced in Section 2. Linearizing the system equations around the equilibrium point $\psi = \dot{\psi} = \chi = \dot{\chi} = 0$ yields the state-space system description

$$\delta\dot{x}(t) = A \cdot \delta x(t) + b \cdot \delta u(t), \quad \delta y(t) = c \cdot \delta x(t) \tag{4.91}$$

with

$$\delta x(t) = \begin{bmatrix} \delta\psi(t) \\ \delta\dot{\psi}(t) \\ \delta\chi(t) \\ \delta\dot{\chi}(t) \end{bmatrix} \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & a_1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & a_2 & 0 \end{bmatrix} \quad b = \begin{bmatrix} 0 \\ b_1 \\ 0 \\ b_2 \end{bmatrix} \quad c^T = \begin{bmatrix} 0 \\ 0 \\ c_1 \\ 0 \end{bmatrix} \tag{4.92}$$

and

$$a_1 = \frac{mgR\vartheta}{\Gamma}, \; a_2 = \frac{mg(R^2\vartheta + r^2\Theta)}{(R+r)\Gamma}$$

$$b_1 = \frac{mr^2 + \vartheta}{\Gamma}, \; b_2 = \frac{R\vartheta}{(R+r)\Gamma}, \; c_1 = R + r$$

To simplify the notation, the following numerical values are used for the system parameters

$$\Theta = 5\,\vartheta, \; R = 3\,r, \; \vartheta = r^2 m, \; g = 10\,\mathrm{m}^2/\mathrm{s}, \; r = 0.2\,\mathrm{m}, \; m = 1\,\mathrm{kg} \tag{4.93}$$

With these preparations it is easy to compute the controllability and observability matrices. The first has the form

$$\mathcal{R}_n = \begin{bmatrix} 0 & \frac{50}{19} & 0 & \frac{75}{76} \\ \frac{50}{19} & 0 & \frac{75}{76} & 0 \\ 0 & \frac{5625}{722} & 0 & \frac{13125}{1444} \\ \frac{5625}{722} & 0 & \frac{13125}{1444} & 0 \end{bmatrix}, \tag{4.94}$$

The determinant of $\mathcal{R}_n$ is not zero ($\det(\mathcal{R}_n) = 263.444$), i.e., the system is completely controllable. However, the system is not completely observable. In fact, the determinant of the observability matrix

$$\mathcal{O}_n = \begin{bmatrix} 0 & 0 & \frac{4}{5} & 0 \\ 0 & 0 & 0 & \frac{4}{5} \\ 0 & 0 & \frac{140}{19} & 0 \\ 0 & 0 & 0 & \frac{140}{19} \end{bmatrix} \tag{4.95}$$

is equal to zero. Moreover, for this special case it is easy to see that the two state variables $\psi$ and $\dot{\psi}$ cannot be estimated using the output signal $y$ as the only information. This is obvious for the "ball-on-wheel" system: the angle and the velocity of rotation of the main wheel is irrelevant for the dynamics of the ball. Only the acceleration of the wheel has an influence.

The linearized system is unstable. Its eigenvalues are

$$\lambda_1 = \lambda_2 = 0, \ \lambda_3 = -\lambda_4 = 3.03488 \tag{4.96}$$

According to the Lyapunov principle, the nonlinear system must be unstable as well. The question is now, can the system be stabilized by an appropriate feedback action? As it will be shown below, this is not possible. However, this does not mean that the system cannot meet the desired objective, i.e., to balance the ball at the equilibrium position.

Since two state variables are not observable, the transfer function from $u \rightarrow y$ must be of second order. Some straightforward computations (use Mathematica) confirm that, using the numerical values listed in (4.93), the transfer function is

$$P(s) = \frac{15}{19\, s^2 - 175} \tag{4.97}$$

This shows that the two observable states are those whose dynamics are described by the two eigenvalues $\lambda_{3,4}$. Accordingly, this part of the system can be stabilized by an appropriate feedback control system. The remaining part, described by the two eigenvalues $\lambda_{1,2}$, cannot be influenced by feedback action. This subsystem is critical, i.e., its eigenvalues have zero real parts. Moreover, this subsystem has a double eigenvalue at zero. The linearized system can, therefore, be stable (but not asymptotically stable) or unstable, depending on

the structure of the matrix $A$. Since the system was shown to be completely controllable, the matrix $A$ must be similar to a Jordan matrix,[11] i.e., the unobservable subsystem must be unstable. Accordingly, in this particular case there is no possibility to stabilize the system in an arbitrary reference point and for arbitrary initial conditions. In other words, the ball can be kept on top of the wheel, but, in general, the wheel speed and the wheel angle cannot be controlled to asymptotically reach zero.

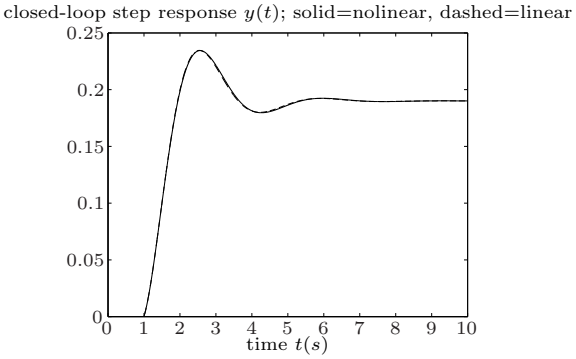



**Fig. 4.7.** Reference step responses of the closed-loop "ball-on-wheel" system; all initial conditions equal to zero.

For illustration purposes, the following feedback controller

$$C(s) = \frac{k_2\, s + k_1}{\tau\, s + 1} = \frac{2.2417\, s + 17.05}{0.01\, s + 1} \tag{4.98}$$

is used. The closed-loop system response to a reference step $r = 0 \rightarrow 0.06$ is shown in Figure 4.7. Figure 4.8 shows the behavior of the "hidden state variables," i.e., $\chi(t)$ and $\dot{\chi}(t)$. To stabilize the ball at an off-axis position on the wheel, the latter must continue to accelerate and thus the angle $\chi(t)$ must diverge (eventually, the speed of the wheel will reach its saturation limits).

If the objective is to stabilize the ball at the unstable equilibrium $\chi_0 = \dot{\chi}_0 = 0$, the system follows the trajectories illustrated in Figure 4.9. The ball can be stabilized, but even in this case the state variables of the unobservable subsystem diverge.

[11] Analyzing the structure of the matrix $A$ in (4.92) the validity of this assertion is obvious. In fact, that matrix $A$ is block triangular and the first diagonal block, which represents the unobservable subsystem, is already in Jordan form.

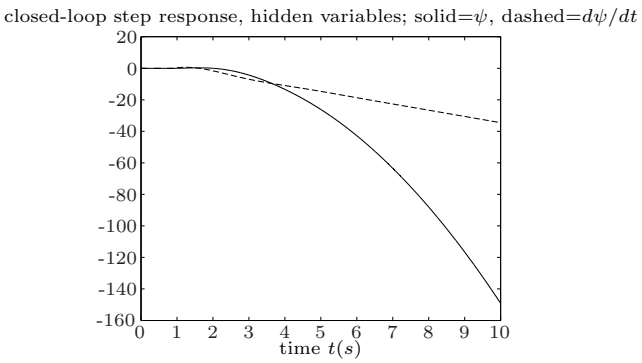



**Fig. 4.8.** Behavior of the "hidden state variables" for the same reference step as illustrated in Figure 4.7.

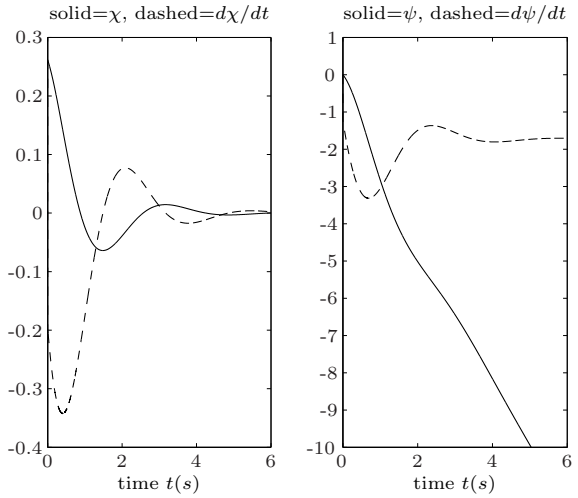



**Fig. 4.9.** Behavior of the closed-loop system for an initial condition of $\chi(0) = \pi/12\ rad$, $\dot{\chi}(0) = \pi/12\ rad/s$ and reference values $\chi_r = \dot{\chi}_r = 0$ (equilibrium).

## 4.5 Balanced Realization and Order Reduction

In the last section, the questions of *structural* controllability and observability were answered, i.e., using the matrices $R$ and $O$, the question was answered whether the system has any non-reachable or non-observable subspaces. The answer is a clear "yes" or "no"; however, this does not give any information whether a specific subspace is "well reachable" or "barely reachable." Closely linked to that question is the problem of system order reduction, i.e., the process of identifying those states that do not substantially influence the input/output behavior of the system and the algorithms with which these "not important" states can be eliminated.

*Example 4.7 (Poorly Controllable System).* The system pair

$$\{A, b\} = \left\{ \begin{bmatrix} -1 & \varepsilon & 0 \\ 0 & 0 & 1 \\ 0 & -a_0 & -a_1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

satisfies the controllability conditions for arbitrary $\varepsilon \neq 0$ (the controllability matrix $R$ is non-singular for all $\varepsilon \neq 0$). However, it is easy to see from the special structure of $A$ that for small $|\varepsilon|$ the state $x_1$ is very difficult to influence, i.e., very large control efforts will be required to achieve the desired state trajectory.

The quantitative controllability of a system can be analyzed using its (infinite) Gramian matrix

$$W_R = \int_0^\infty e^{A\sigma} B\, B^T e^{A^T\sigma}\, d\sigma \tag{4.99}$$

The closer $W_R$ is to a singular matrix, the less controllable the corresponding system will be. Similarly, the infinite observability Gramian

$$W_O = \int_0^\infty e^{A^T\sigma} C^T\, C\, e^{A\sigma}\, d\sigma \tag{4.100}$$

provides an indication as to how well-observable the system is.

The Gramians (4.99) and (4.100) exist iff the system $\{A, B, C, D\}$ is asymptotically stable. The computation of these two matrices appears to be difficult (integration over an infinite time span). The following result, however, simplifies those steps considerably.

*If the system is Hurwitz ($A$ is asymptotically stable), the two Gramian matrices represent the (unique) solution of the two Lyapunov equations*

$$AW_R + W_R A^T = -BB^T \tag{4.101}$$

*and*

$$A^T W_O + W_O A = -C^T C \tag{4.102}$$

Reliable algorithms exist for the solution of these two linear matrix equations and all modern CASD tools include corresponding software (the corresponding Matlab command is `lyap`).

With the last example in mind, one might be tempted to simply delete those system parts that do not significantly contribute to the controllability or observability Gramian matrices. However, this can lead to wrong results because certain state variables might be poorly controllable, but very well

observable such that their total contribution to the IO behavior remains notable.[12]

The correct approach is to look for a coordinate transformation $T \cdot x_b = x$, which transforms the original system into a system whose controllability and observability Gramians are equal and diagonal, i.e.,

$$W_{R,b} = W_{O,b} = \mathrm{diag}(\sigma_i), \qquad i = 1, \ldots, n \tag{4.103}$$

Since the Gramians are by construction symmetric and positive definite, the real scalars $\sigma_i$ are all positive.

This transformation can be found using the Gramians introduced above. The details are

$$T = T_R T_O \tag{4.104}$$

where $T_R$ is found using an eigendecomposition[13] of the controllability Gramian

$$W_R = V_R \Lambda_R^2 V_R^T \quad \rightarrow \quad T_R = V_R \Lambda_R \tag{4.105}$$

and $T_O$ follows from the observability Gramian using the following transformation and eigendecomposition

$$\widetilde{W}_O = T_R^T W_O T_R = V_O \Lambda_O^2 V_O^T \quad \rightarrow \quad T_O = V_O \Lambda_O^{-1/2} \tag{4.106}$$

The Gramians of the "balanced realization" $\{T^{-1}AT, T^{-1}B, CT, D\}$ will have the following form (up to permutations in order to arrange the diagonal elements in descending order)

$$W_{R,b} = W_{O,b} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \sigma_n \end{bmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0 \tag{4.107}$$

Notice that a proper normalization (see Section 4.2.1) is important for this procedure to work correctly, otherwise the magnitudes of the state variables cannot be compared among each other.

Assuming that the last $\nu$ elements $\sigma_i$ with $i = \nu, \ldots, n$ are substantially smaller than the other first $n - \nu$ elements $\sigma_j$ with $j = 1, \ldots \nu - 1$ than the contribution of the last $\nu$ balanced modes to the system's IO behavior may be neglected. Therefore, the following system order reduction algorithms can

[12] If the system analyzed in Example 4.7 has a measurement vector $c = [1/\varepsilon,\ 0,\ 0]$, its transfer function is $P(s) = 1/[(s+1)(s^2 + a_1\, s + a_0)]$ for all $\varepsilon \neq 0$.

[13] Recall that a symmetric matrix always has $n$ linearly independent eigenvectors which are orthogonal to each other. Moreover, the eigenvalues of a positive definite matrix $W_R > 0$ are always real and positive.

be applied. As a first step, after balancing (4.104), the system is partitioned as follows

$$\begin{aligned} \tfrac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t) \\ y(t) \quad &= [C_1 \;\; C_2] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + Du(t) \end{aligned} \tag{4.108}$$

where $x_1 \in \mathbb{R}^{n-\nu}$ and $x_2 \in \mathbb{R}^{\nu}$. Since the contribution of the last $\nu$ states is small, the system can be reduced by simply omitting the corresponding elements, i.e.,

$$\begin{aligned} \tfrac{d}{dt}x_1(t) &= A_{11}x_1(t) + B_1u(t) \\ y(t) \quad &= C_1x_1(t) + Du(t) \end{aligned} \tag{4.109}$$

Typically, this will yield good agreement in the frequency domain but, in general the DC gains of the original system (4.108) and the reduced-order system (4.109) will be different.

If this is to be avoided, a "singular perturbation" approach is better suited where the dynamics of the last $\nu$ states is neglected, but not their DC contributions. The details of that approach are

$$\frac{d}{dt}x_2(t) \approx 0 \quad \Rightarrow x_2(t) \approx -A_{2,2}^{-1}[A_{2,1}x_1(t) + B_2u(t)] \tag{4.110}$$

and

$$\begin{aligned} \tfrac{d}{dt}x_1(t) &\approx [A_{11} - A_{1,2}A_{2,2}^{-1}A_{2,1}]x_1(t) + [B_1 - A_{1,2}A_{2,2}^{-1}B_2]u(t) \\ y(t) \quad &\approx [C_1 - C_2A_{2,2}^{-1}A_{2,1}]x_1(t) + [D - C_2A_{2,2}^{-1}B_2]u(t) \end{aligned} \tag{4.111}$$

This approach is always feasible if the original system was asymptotically stable, and this assumption is not restrictive since the existence of the Gramians (4.99) and (4.100) requires the same assumption to be satisfied.

*Example 4.8 (An Academic System).* This example shows the results of the balancing and order reduction procedure using Matlab for a SISO system of order $n = 7$. The system matrices (with no physical background) are available at the lecture's Web site. The resulting system responses are displayed in Figure 4.10.

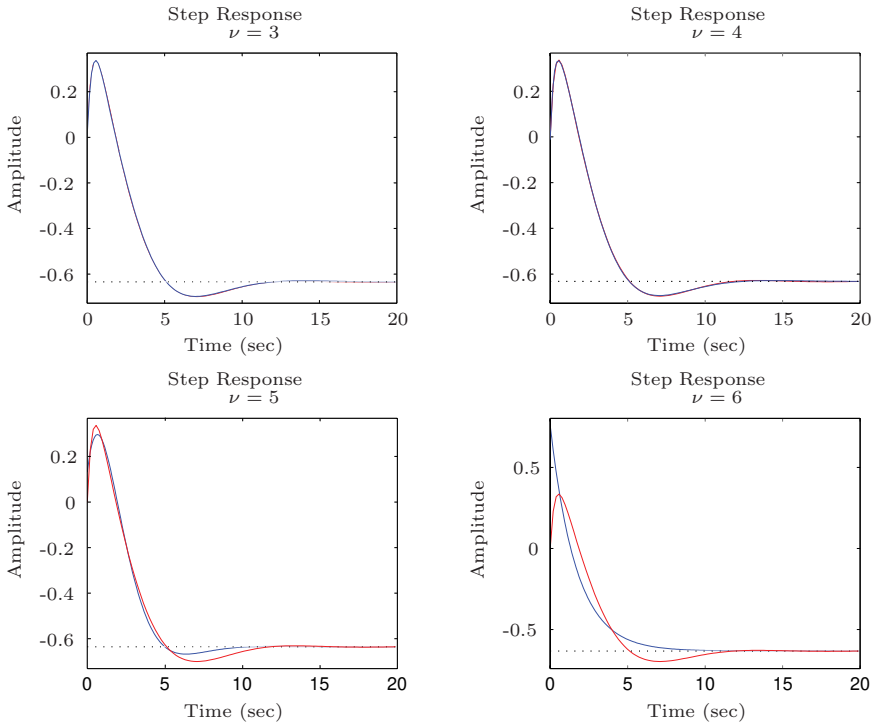



**Fig. 4.10.** Step responses for increasing $\nu$, red=original and blue=reduced-order outputs.

## 4.6 Zero Dynamics

The connection between the dynamic behavior of a system (4.114) and its poles was discussed in detail in Section 4.4.1. In this subsection, the influence of the zeros on the dynamic properties of the system will be clarified. It was mentioned in RT I that non-minimumphase zeros limit the achievable bandwidth on any closed-loop system. A precise justification for this, which was not given in RT I, will be shown below using a special coordinate transformation. To simplify this discussion, the plant to be analyzed is assumed to be a SISO system. Similar, but mathematically more complex results hold for MIMO and even nonlinear systems [8].

In the SISO case, the transfer function matrix

$$P(s) = C \cdot [s\, I - A]^{-1} \cdot B + D \tag{4.112}$$

is a scalar rational function which can always be written in the following form

$$P(s) = k \frac{s^{n-r} + b_{n-r-1} s^{n-r-1} + \ldots + b_1 s + b_0}{s^n + a_{n-1} s^{n-1} + a_{n-2} s^{n-21} + \ldots + a_2 s^2 + a_1 s + a_0} \tag{4.113}$$

The order $n$ is defined as the highest power of $s$ in the denominator polynomial and the relative degree $r$ is defined as the difference between the highest power in the denominator polynomial and the highest power in the numerator polynomial. This parameter will be shown to have an important role in the discussion of system zeros.

It is easy to verify that the following choice of coordinates produces a state-space description which, after its Laplace transformation, coincides with the transfer function (4.113)

$$\begin{aligned}\tfrac{d}{dt}x(t) &= \begin{bmatrix} 0 & 1 & 0 & \dots\dots & 0 \\ 0 & 0 & 1 & \dots\dots & 0 \\ \cdot & \cdot & \cdot & \cdot\ \cdot & \cdot \\ 0 & 0 & 0 & \dots\dots & 1 \\ -a_0 & -a_1 & -a_2 & \dots\dots & -a_{n-1} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ 0 \\ k \end{bmatrix} u(t) \\ y(t) &= \begin{bmatrix} b_0 & \dots & b_{n-r-1} & 1 & 0 & \dots & 0 \end{bmatrix} x(t)\end{aligned} \tag{4.114}$$

As usual, there are (infinitely) many other possible choices of coordinate systems to realize the transfer function (4.113). However, this special form ("controller canonical form with gain") has the minimum number of parameters and, moreover, it is easily constructed once (4.113) is known. Of course, the coordinates chosen in this approach have no physical interpretation, i.e., the state variables $x(t)$ may not be identified with the storage variables of the system.

The starting point for the following discussions is the system's relative degree $r$, which is equal to the number of differentiations required to have the input $u(t)$ explicitly appear in the "output" $y^{(r)}(t)$. This interpretation can be easily verified using the special choice of coordinates introduced above

$$\begin{aligned} y(t) &= cx(t), \\ \dot{y}(t) &= c\dot{x}(t) = cAx(t) + cbu(t) = cAx(t), \\ \ddot{y}(t) &= \tfrac{d}{dt}\dot{y}(t) = cA^2x(t) + cAbu(t) = cA^2x(t), \\ &\dots \\ y^{(r-1)}(t) &= \tfrac{d}{dt}y^{(r-2)}(t) = cA^{r-1}x(t) + cA^{r-2}bu(t) = cA^{r-1}x(t), \\ y^{(r)}(t) &= \tfrac{d}{dt}y^{(r-1)}(t) = cA^rx(t) + cA^{r-1}bu(t) = cA^rx(t) + ku(t) \end{aligned} \tag{4.115}$$

where, of course, $r \leq n$.

The *zero dynamics* of the system is now defined to be the behavior of (4.114) for those special inputs $u^*(t)$ and initial conditions $x^*$ for which

the output $y(t)$ is identical to zero for a finite interval.[14] This is an interesting problem setting because in all reference tracking problems the controller tries to force the error to zero. If the plant in such a situation has internal dynamics that are unstable, but not visible at the system's output, problems are to be expected.

For $y(t)$ to be zero for all times it is necessary that all of its derivatives are zero as well. The following coordinate transformation $z = \Phi^{-1} \cdot x$ is introduced to facilitate the subsequent simplifications that are made using this observation

$$\begin{aligned}
z_1 &= y = cx &&= [b_0 x_1 + b_1 x_2 + \ldots + b_{n-r-1} x_{n-r} + x_{n-r+1}] \\
z_2 &= \dot{y} = cAx &&= [b_0 x_2 + b_1 x_3 + \ldots + b_{n-r-1} x_{n-r+1} + x_{n-r+2}] \\
&\ldots \\
z_r &= y^{(r-1)} = cA^{r-1}x &&= [b_0 x_r + b_1 x_{r+1} + \ldots + b_{n-r-1} x_{n-1} + x_n]
\end{aligned} \tag{4.116}$$

The remaining $n - r$ coordinates must be chosen such that the transformation $\Phi^{-1}$ is regular and that their derivatives do not depend on the input $u$ either. Obviously, the choice

$$\begin{aligned}
z_{r+1} &= x_1 \\
z_{r+2} &= x_2 \\
&\ldots \\
z_n &= x_{n-r}
\end{aligned} \tag{4.117}$$

satisfies both requirements.

To simplify the following discussion the vector $z$ is partitioned into two subvectors

$$z = \begin{bmatrix} \xi \\ \eta \end{bmatrix}, \quad \xi = \begin{bmatrix} z_1 \\ \ldots \\ z_r \end{bmatrix}, \quad \eta = \begin{bmatrix} z_{r+1} \\ \ldots \\ z_n \end{bmatrix} \tag{4.118}$$

In the new coordinates the system (4.113) has the form

$$\frac{d}{dt}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \left[\begin{array}{ccccc|ccccc} 0 & 1 & 0 & \ldots & 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & 0 & 1 & 0 & \ldots & \ldots & \ldots & 0 \\ - & - & r^T & - & - & - & - & s^T & - & - \\ \hline 0 & \ldots & \ldots & \ldots & 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 1 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 & 0 & \ldots & \ldots & 0 & 1 \\ - & - & p^T & - & - & - & - & q^T & - & - \end{array}\right] \cdot \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \left[\begin{array}{c} 0 \\ \ldots \\ 0 \\ k \\ \hline 0 \\ \ldots \\ \ldots \\ 0 \end{array}\right] \cdot u \tag{4.119}$$

[14] Of course, the trivial solution $u^* = 0$, and $x^* = 0$ is not of interest here.

and, obviously, $y = \xi_1$. Figure 4.11 illustrates the structure of the system description introduced above. The precise form of the vectors $r^T$ and $s^T$ is not important for the following considerations. The vectors $p^T$ and $q^T$ are shown below.

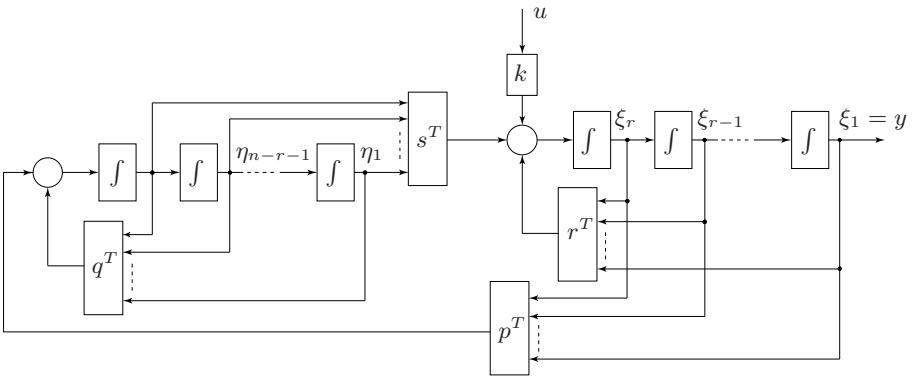



**Fig. 4.11.** System structure illustrating the concept of the system's zero dynamics.

In order to have an identically vanishing output, it is therefore necessary and sufficient to choose the following initial conditions and control signal

$$\xi^*(0) = 0, \qquad u^*(t) = -\frac{1}{k} s^T \eta^*(t) \tag{4.120}$$

The initial condition $\eta_0^* \neq 0$ may be chosen arbitrarily. If the system is initialized and controlled according to (4.120), the output $y(t)$ and the state variables $\xi(t)$ will be zero for all times $t > 0$. The trajectories of state variables $\eta(t)$ will be governed by the equations

$$\frac{d}{dt}\eta^*(t) = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ - & - & q^T & - & - \end{bmatrix} \cdot \eta^*(t) = Q \cdot \eta^*(t), \qquad \eta^*(0) = \eta_0^* \tag{4.121}$$

and these equations define the *zero dynamics* of the system (4.114). If the matrix $Q$ is asymptotically stable, then the system is minimum phase. This notion is usually introduced in connection with the system's zeros, describing the fact that they all have negative real parts. These two definitions are consistent, which can be seen from the definition of the vector $q^T$

$$q^T = [-b_0, \ -b_1, \ \dots, \ -b_{n-r-2}, \ -b_{n-r-1}] \tag{4.122}$$

This equation can be derived using the definition of $z_n$ (4.117), the original system equation (4.114), the definition of $z_1$ (4.117), again the coordinate transformation (4.117), and the definition (4.118)

$$\begin{aligned}\dot{z}_n &= \dot{x}_{n-r} \\ &= x_{n-r+1} \\ &= z_1 - b_0 x_1 \ldots - b_{n-r-1} x_{n-r} \\ &= z_1 - b_0 z_{r+1} \ldots - b_{n-r-1} z_n \\ &= p^T \xi + q^T \eta\end{aligned} \tag{4.123}$$

where $p^T = [1, 0, \ldots, 0]$. Therefore, the eigenvalues of $Q$ coincide with the (transmission) zeros of (4.114) and the roots of the numerator (4.113). If one or more of the eigenvalues of $Q$ have a positive real part, then the system has non-minimumphase zeros and its zero dynamics are unstable. In this case, the input $u(t)$ may not be chosen such that the output $y(t)$ is (almost) zero before the states $\eta$ associated with the zero dynamics are (almost) zero. In fact, if this condition is violated the internal state variables $\eta(t)$ can diverge without $y(t)$ being affected by that. Of course, this situation must be avoided and this corresponds to the constraint that the bandwidth of the closed-loop system must be substantially smaller than the "slowest" non-minimumphase zero.

*Example 4.9 (Zero Dynamics).* The following example illustrates the ideas introduced above for a SISO system of order $n = 4$ and relative degree $r = 2$.

$$\tfrac{d}{dt}x(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 \end{bmatrix} \cdot x(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ k \end{bmatrix} \cdot u(t), \quad k \neq 0 \tag{4.124}$$

$$y(t) = \begin{bmatrix} b_0 & b_1 & 1 & 0 \end{bmatrix} \cdot x(t) + \begin{bmatrix} 0 \end{bmatrix} \cdot u(t)$$

The system's relative degree $r$ is indeed equal to 2 as can bee seen from

$$\begin{aligned} y(t) &= b_0 x_1(t) + b_1 x_2(t) + x_3(t) \\ \dot{y}(t) &= b_0 x_2(t) + b_1 x_3(t) + x_4(t) \\ \ddot{y}(t) &= -a_0 x_1(t) - a_1 x_2(t) + (b_0 - a_2) x_3(t) + (b_1 - a_3) x_4(t) + k\, u(t) \end{aligned} \tag{4.125}$$

For this example the coordinate transformation $z = \Phi^{-1} \cdot x$ has the form

$$\begin{aligned} z_1 &= y = b_0 x_1 + b_1 x_2 + x_3 \\ z_2 &= \dot{y} = b_0 x_2 + b_1 x_3 + x_4 \end{aligned} \tag{4.126}$$

and

$$\begin{aligned} z_3 &= x_1 \\ z_4 &= x_2 \end{aligned} \tag{4.127}$$

Therefore

$$\Phi^{-1} = \begin{bmatrix} b_0 & b_1 & 1 & 0 \\ 0 & b_0 & b_1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \text{ and } \Phi = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -b_0 & -b_1 \\ -b_1 & 1 & b_0\, b_1 & b_1^2 - b_0 \end{bmatrix}$$

Notice that, by construction, $\det(\Phi) = \det(\Phi^{-1}) = 1$. Partitioning

$$\xi = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad \eta = \begin{bmatrix} z_3 \\ z_4 \end{bmatrix} \tag{4.128}$$

in the new coordinates the system

$$\frac{d}{dt} z(t) = \Phi^{-1} A\, \Phi\ z(t) + \Phi^{-1} b\, u(t), \quad y(t) = c\, \Phi\ z(t)$$

has the form

$$\frac{d}{dt} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \eta_1(t) \\ \eta_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ r_1 & r_2 & s_1 & s_2 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & -b_0 & -b_1 \end{bmatrix} \cdot \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \\ \eta_1(t) \\ \eta_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ k \\ 0 \\ 0 \end{bmatrix} \cdot u(t) \tag{4.129}$$

and, obviously, $y(t) = \xi_1(t)$. The coefficients $r_1, r_2, s_1, s_2$ are listed below

$$r_1 = b_0 - a_2 - b_1(b_1 - a_3)$$

$$r_2 = b_1 - a_3$$

$$s_1 = b_0\, b_1(b_1 - a_3) - a_0 - b_0(b_0 - a_2)$$

$$s_2 = (b_1 - a_3)(b_1^2 - b_0) - a_1 - (b_0 - a_2)b_1$$

Figure 4.11 illustrates the structure of this system in the new coordinates.

Choosing the following initial conditions

$$\xi_1^*(0) = \xi_2^*(0) = 0 \tag{4.130}$$

and control signal

$$u^*(t) = -\frac{1}{k} \left[ s_1 \eta_1^*(t) + s_2 \eta_2^*(t) \right] \tag{4.131}$$

yields a zero output $y(t) = 0$ for all $t \geq 0$. The initial conditions $\eta_1^*(0) \neq 0$ and $\eta_2^*(0) \neq 0$ may be chosen arbitrarily. The trajectories of state variables $\eta_1(t)$ and $\eta_2(t)$, in this case, are defined by the equations

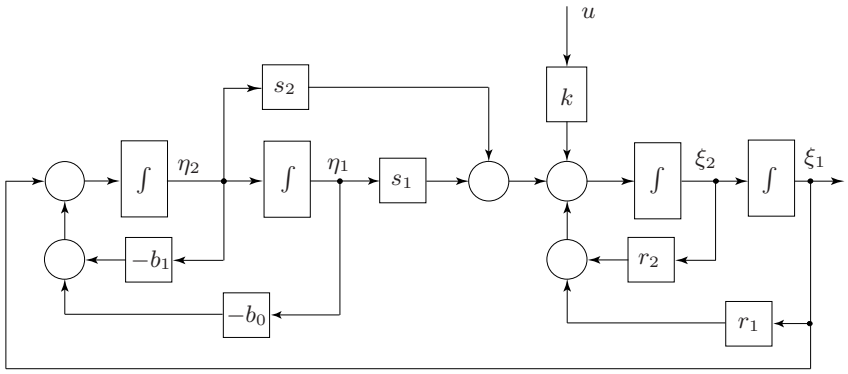



**Fig. 4.12.** System structure of the example's zero dynamics.

$$\frac{d}{dt}\eta^*(t) = \begin{bmatrix} 0 & 1 \\ -b_0 & -b_1 \end{bmatrix} \cdot \eta^*(t) \tag{4.132}$$

Notice that the main result, i.e., the fact that for $y(t) = 0$ for all $t > 0$ the system dynamics are governed by the last equation can be directly inferred from the definitions

$$\eta_1 = x_1, \quad \eta_2 = x_2$$

with

$$\tfrac{d}{dt}\eta_1(t) = \tfrac{d}{dt}x_1(t) = x_2(t) = \eta_2(t)$$
$$\tfrac{d}{dt}\eta_2(t) = \tfrac{d}{dt}x_2(t) = x_3(t)$$

and the output equation of the original system

$$y(t) = b_0\, x_1(t) + b_1\, x_2(t) + x_3(t)$$

In fact, if $y(t) = 0$ then

$$x_3(t) = -b_0\, x_1(t) - b_1\, x_2(t) = -b_0\, \eta_1(t) - b_1\, \eta_2(t)$$

and therefore

$$\tfrac{d}{dt}\eta_1(t) = \eta_2(t)$$
$$\tfrac{d}{dt}\eta_2(t) = -b_0\, \eta_1(t) - b_1\, \eta_2(t)$$

## 4.7 Case Study: Geostationary Satellite

### Introduction

In this case study, the dynamic behavior of a *geostationary satellite* is analyzed. Geostationary satellites orbit around the earth in its equatorial plane with exactly the sidereal[15] rotation period of the earth. An observer on earth will therefore always see this satellite at the same position in the sky. Obviously, this property is very useful in many applications, most importantly for communication purposes.[16]

In this case study, the equations that describe the motion of the satellite will be derived and a nominal periodic orbit will be obtained with these equations. The equations will then be linearized around this orbit. The linear equations, which describe the behavior of the satellite close to its nominal orbit, will be used to understand the basic properties and limitations of the satellite dynamics. The results obtained in this way are important for the subsequent step in which an orbit-stabilizing controller is designed. However, this design will not be discussed in this case study.

### System Modeling

*Problem Definition*

The relevant system parameters and coordinates are illustrated in Figure 4.13. The satellite has two degrees of freedom. A convenient choice to quantify them is to use a radius $r$ and an azimuth angle $\varphi$ as coordinates.

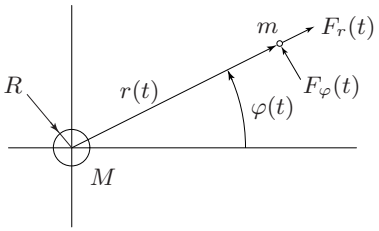



**Fig. 4.13.** Illustration of the system parameters and coordinates in the equatorial plane.

[15] The sidereal day is measured with the stars as a reference frame. The sidereal day is slightly shorter than the solar day, i.e., a sidereal day is only 23 hours and 56 minutes and 4.1 seconds long.

[16] Interestingly, this idea was first suggested in 1947 by Arthur C. Clarke, who became later famous for his novel "2001 – A Space Odyssey," written when working with Stanley Kubrick on the film of the same name.

The following assumptions are adopted:

1. No other celestial bodies are considered, i.e., only the gravitational forces of the earth-satellite system must be considered.[17]
2. The mass $M$ of the earth is much larger than the mass $m$ of the satellite. Therefore, the center of gravity of the two bodies may be placed at the center of gravity of the earth, i.e., the earth is assumed to define an inertial reference frame.
3. The satellite always remains in the equatorial plane, i.e., the two variables $r$ and $\varphi$ completely describe its position in space.
4. The attitude of the satellite is kept constant by an on-board control system that is not part of this analysis.[18] Therefore, it may be assumed that two mutually independent control forces can be applied in the radial and tangential directions.

### Nonlinear Model

*Lagrange Function*

With the simplifications mentioned, it is possible to formulate the dynamics of the satellite orbiting the earth using Lagrange's method

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{r}}\right] - \frac{\partial L}{\partial r} = F_r \tag{4.133}$$

and

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\varphi}}\right] - \frac{\partial L}{\partial \varphi} = F_\varphi \cdot r \tag{4.134}$$

Notice that the second degree of freedom is a rotation. Accordingly, the associated generalized force must be a torque.

The Lagrange function $L = T - V$ is the difference of the kinetic energy

$$T = \frac{1}{2} \cdot m \cdot \dot{r}^2 + \frac{1}{2} \cdot m \cdot (r \cdot \dot{\varphi})^2 \tag{4.135}$$

and the potential energy $V$. This variable, which only depends on $r$, can be computed using Newton's law of gravitation

$$V = \int_R^r F(\rho)\, d\rho = \int_R^r G \cdot \frac{M \cdot m}{\rho^2}\, d\rho = G \cdot M \cdot m \cdot (\frac{1}{R} - \frac{1}{r}), \quad r > R \tag{4.136}$$

[17] The gravitational forces of the sun and the moon and other forces are treated as disturbances that will be compensated for by an appropriate orbit-stabilizing control system.

[18] This problem is rather difficult to solve, see e.g. [7].

where $G = 6.673 \ldots 10^{-11}$ m$^3$/(s$^2$ kg) is a constant.[19] The mass of the earth is approximately $M = 5.974 \ldots 10^{24}$ kg and its radius is $R = 6.367 \ldots 10^6$ m. The mass of the satellite will be not relevant as long as it is much smaller than the mass of the earth.

As Figure 4.14 shows, the potential energy is 0 on the surface of the earth (this is an arbitrary but convenient choice) and reaches 90% of its maximum value

$$V_\infty = \frac{GMm}{R} \tag{4.137}$$

at a distance of only 10 times the radius of the earth. Assuming a rapid acceleration at the start from standstill to a speed $v_0$, the energy balance

$$\frac{1}{2} \cdot m \cdot v_0^2 = G \cdot M \cdot m \cdot (\frac{1}{R} - \frac{1}{r}) \tag{4.138}$$

permits the computation of the kinetic energy that a satellite must have to reach a certain orbit height $r - R$. The *escape velocity* $v_\infty$, i.e., that velocity that is required to completely leave the gravitation field of the earth, can be found using the last equation by choosing $r = \infty$

$$v_\infty = \sqrt{\frac{2\,G \cdot M}{R}} \approx 1.12 \ldots 10^4 \; m/s \tag{4.139}$$

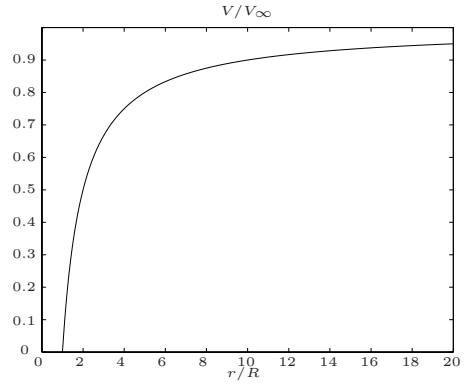



**Fig. 4.14.** Potential energy as a function of the normalized orbit radius.

[19] The derivations shown in this case study assume a non-relativistic setting. Below the orbital speed will be shown to be much smaller than the speed of light, such that this assumption is well justified.

*System Equations*

Inserting the equations (4.135) and (4.136) into the Lagrange equations (4.133) and (4.134) yields, after the following intermediate steps

$$\begin{aligned} \tfrac{\partial L}{\partial \dot{r}} &= m \cdot \dot{r} & \tfrac{\partial L}{\partial \dot{\varphi}} &= m \cdot r^2 \cdot \dot{\varphi} \\ \tfrac{d}{dt}(\tfrac{\partial L}{\partial \dot{r}}) &= m \cdot \ddot{r} & \tfrac{d}{dt}(\tfrac{\partial L}{\partial \dot{\varphi}}) &= m \cdot r^2 \cdot \ddot{\varphi} + 2\, m \cdot r \cdot \dot{\varphi} \cdot \dot{r} \\ \tfrac{\partial L}{\partial r} &= m \cdot r \cdot \dot{\varphi}^2 - G \cdot M \cdot m \cdot \tfrac{1}{r^2} & \tfrac{\partial L}{\partial \varphi} &= 0 \end{aligned} \tag{4.140}$$

the first result

$$m \cdot \ddot{r} = m \cdot r \cdot \dot{\varphi}^2 - G \cdot M \cdot m \cdot \frac{1}{r^2} + F_r \tag{4.141}$$

$$m \cdot r^2 \cdot \ddot{\varphi} = -2\, m \cdot r \cdot \dot{\varphi} \cdot \dot{r} + F_\varphi \cdot r \tag{4.142}$$

These equations are simplified by introducing the control accelerations

$$u_r = F_r/m, \quad \text{or} \quad u_\varphi = F_\varphi/m, \ \text{respectively} \tag{4.143}$$

by dividing both sides of the last two equations by the satellite mass $m$ and the second equation also by the satellite radius squared

$$\ddot{r} = r \cdot \dot{\varphi}^2 - G \cdot M \cdot \frac{1}{r^2} + u_r \tag{4.144}$$

$$\ddot{\varphi} = -2\, \dot{\varphi} \cdot \dot{r} \cdot \frac{1}{r} + u_\varphi \cdot \frac{1}{r} \tag{4.145}$$

*Circular Orbit*

The two equations, (4.144) and (4.145), are the starting points for the derivation of the reference solution. For geostationary conditions, the periodic reference orbit must be chosen to be circular with constant angular velocity, i.e.,

$$\begin{aligned} u_r = 0, \quad \ddot{r} = 0, \quad \dot{r} = 0, \quad r = r_0 \\ u_\varphi = 0, \quad \ddot{\varphi} = 0, \quad \dot{\varphi} = \omega_0, \quad \varphi = \omega_0 \cdot t \end{aligned} \tag{4.146}$$

where $\omega_0$ is the sidereal angular velocity

$$\omega_0 = 7.29 \ldots 10^{-5}\ rad/s \tag{4.147}$$

Inserting this desired trajectory in (4.144) yields the only non-trivial condition

$$0 = r_0 \cdot \omega_0^2 - G \cdot M \cdot \frac{1}{r_0^2} \tag{4.148}$$

which permits the computation of the radius of the geostationary orbit

$$r_0 = \left(\frac{G \cdot M}{\omega_0^2}\right)^{1/3} \approx 4.217 \ldots 10^7 \ m \tag{4.149}$$

or approximately 6.2 times the radius of the earth. The resulting tangential speed is approximately $r_0\,\omega_0 \approx 3.07 \ldots 10^3 \ m/s$. As mentioned above, no relativistic effects need to be considered at that value. As Figure 4.14 illustrates, the energy required to reach the geostationary orbit radius is more than 80% of the escape energy. In addition to that the satellite must be accelerated to the tangential speed $r_0\,\omega_0$. For typical geostationary satellites with a mass between $2 \cdot 10^3$ and $3 \cdot 10^3$ kg, however, the corresponding kinetic energy is one order of magnitude smaller.

*State-Space Formulation*

The system description obtained so far consists of two coupled ordinary second-order differential equations (ODE). This form is not suitable for the subsequent analysis steps. A first-order ODE formulation, which is more useful, can be obtained as follows. First, the following definitions are introduced

$$\begin{aligned} x_1(t) = r, \quad x_2(t) = \dot{r}, \quad u_1(t) = u_r \\ x_3(t) = \varphi, \quad x_4(t) = \dot{\varphi}, \quad u_2(t) = u_\varphi \end{aligned} \tag{4.150}$$

Second, the following vectors are formed

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix}, \quad u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \tag{4.151}$$

Using these expressions, the system dynamics can be described with a first-order vector ODE

$$\frac{d}{dt}x(t) = f(x(t), u(t)) \tag{4.152}$$

where

$$f(x(t)) = \begin{bmatrix} x_2(t) \\ x_1 \cdot x_4^2(t) - G \cdot M / x_1^2(t) + u_1(t) \\ x_4(t) \\ -2\,x_2(t) \cdot x_4(t)/x_1(t) + u_2(t)/x_1(t) \end{bmatrix} \tag{4.153}$$

To complete the system description, the measured variables must be defined. It is assumed that the normed[20] radius $r/r_0$ and the azimuth $\varphi$ are the only available signals. In state space form, these measurements can be modeled by

$$y(t) = h(x(t)) \tag{4.154}$$

where

$$h(x(t)) = \begin{bmatrix} x_1(t)/r_0 \\ x_3(t) \end{bmatrix} \tag{4.155}$$

### Model Linearization

The model (4.152) is now linearized around the nominal orbit

$$x_0(t) = \begin{bmatrix} r_0 \\ 0 \\ \omega_0 \cdot t \\ \omega_0 \end{bmatrix}, \quad u_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{4.156}$$

Notice that in this case the nominal trajectory is *not* an equilibrium point but a periodic solution of the system equations. However, introducing the notation

$$x(t) = x_0(t) + \delta x(t), \quad u(t) = u_0 + \delta u(t), \quad y(t) = y_0(t) + \delta y(t) \tag{4.157}$$

to describe the small deviations around that orbit, a procedure similar to the one used in the case of a fixed equilibrium point, can be used

$$\begin{aligned} \frac{d}{dt}x_0(t) + \frac{d}{dt}\delta x(t) &= f(x_0(t) + \delta x(t), u_0 + \delta u(t)) \\ &\approx f(x_0(t), u_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0(t),u_0} \cdot \delta x(t) + \left.\frac{\partial f}{\partial u}\right|_{x_0(t),u_0} \cdot \delta u(t) \end{aligned} \tag{4.158}$$

Since, by construction,

$$\frac{d}{dt}x_0(t) = f(x_0(t), u_0) \tag{4.159}$$

the linearized system is described by the equation

[20] The radius $r$ is normed by the geostationary radius $r_0$ (4.149) to facilitate a later control system design. With this choice both channels have an acceleration as input and a non-dimensional signal as output.

$$\frac{d}{dt}\delta x(t) = \left.\frac{\partial f}{\partial x}\right|_{x_0(t),u_0} \cdot \delta x(t) + \left.\frac{\partial f}{\partial u}\right|_{x_0(t),u_0} \cdot \delta u(t) \tag{4.160}$$

The explicit computation of the Jacobian matrices yields the following results

$$\frac{\partial f}{\partial x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ x_4^2 + 2\,G \cdot M/x_1^3 & 0 & 0 & 2\,x_1 \cdot x_4 \\ 0 & 0 & 0 & 1 \\ 2\,x_2 \cdot x_4/x_1^2 - u_2/x_1^2 & -2\,x_4/x_1 & 0 & -2\,x_2/x_1 \end{bmatrix}, \quad \frac{\partial f}{\partial u} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/x_1 \end{bmatrix} \tag{4.161}$$

Evaluating the Jacobians along the periodic nominal solution (4.156) yields the result

$$\left.\frac{\partial f}{\partial x}\right|_{x_0(t),u_0} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\,\omega_0^2 & 0 & 0 & 2\,r_0 \cdot \omega_0 \\ 0 & 0 & 0 & 1 \\ 0 & -2\,\omega_0/r_0 & 0 & 0 \end{bmatrix}, \quad \left.\frac{\partial f}{\partial u}\right|_{x_0(t),u_0} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/r_0 \end{bmatrix} \tag{4.162}$$

Note that the definition of the orbital radius $r_0$ (4.149) has been used to simplify the element $(2,1)$ of the first Jacobian. Surprisingly, the resulting system matrices are time-*in*variant! In general, this is not the case when nonlinear systems are linearized around a periodic solution. However, in this rather simple example, constant system matrices result. Also note that the measurement equation (4.155) is already linear so no additional steps are necessary for that part.

### System Analysis

The system that is considered in this section was derived in the last section. For convenience, the standard notation

$$\dot{x}(t) = A \cdot x(t) + B \cdot u(t), \quad y(t) = C \cdot x(t) \tag{4.163}$$

will be adopted below, where the deviations $\delta x$, $\delta u$ and $\delta y$ will be denoted by $x$, $u$ and $y$, respectively. The matrices $A$ and $B$ have been defined above

$$A = \left.\frac{\partial f}{\partial x}\right|_{x_0(t),u_0}, \quad B = \left.\frac{\partial f}{\partial u}\right|_{x_0(t),u_0} \tag{4.164}$$

and the matrix $C$ follows from (4.155)

$$C = \begin{bmatrix} 1/r_0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.165}$$

For the linearized system $\{A, B, C, 0\}$, defined above, the following questions will be discussed:

1. What are the stability properties of the linearized system?
2. Is the linearized system controllable and observable?
3. Which actuators and sensors are most important?
4. What are the salient dynamic properties of the linearized system?

*System Stability*

The stability properties of the linearized system are analyzed by computing the eigenvalues of the matrix $A$. The eigenvalues are the roots of the characteristic equation of $A$, i.e.,

$$\det(s \cdot I - A) = \det \begin{bmatrix} s & -1 & 0 & 0 \\ -3\,\omega_0^2 & s & 0 & -2\,r_0 \cdot \omega_0 \\ 0 & 0 & s & -1 \\ 0 & 2\,\omega_0/r_0 & 0 & s \end{bmatrix} \tag{4.166}$$

The determinant of this matrix is best found by using the element $(3, 3)$ as the first expansion factor

$$\det(s \cdot I - A) = s \cdot \det \begin{bmatrix} s & -1 & 0 \\ -3\,\omega_0^2 & s & -2\,r_0 \cdot \omega_0 \\ 0 & 2\,\omega_0/r_0 & s \end{bmatrix} \tag{4.167}$$

$$= s \cdot \left[ s \cdot (s^2 + 4\,\omega_0^2) - (-1) \cdot (-3\,\omega_0^2 \cdot s) \right] \tag{4.168}$$

$$= s^2 \cdot (s^2 + \omega_0^2) \tag{4.169}$$

The roots of this characteristic equation are $\{0, 0, +j\,\omega_0, -j\,\omega_0\}$. As it is often the case in celestial mechanics, the linearized system only has eigenvalues on the imaginary axis. The eigenvalue pair $\pm j\,\omega_0$ shows that the linearized system includes oscillatory modes whose eigenfrequency coincides with the angular speed of the satellite. The double root in the origin indicates that the linearized system could be Lyapunov stable or unstable. One way to determine the stability properties of the system is to compute the rank of the matrix

$$M = (s \cdot I - A)|_{s=0} \tag{4.170}$$

which turns out to be three, i.e., there is only one eigenvector associated with the double eigenvalue $s = 0$. Therefore, the matrix $A$ is cyclic and the linearized system is unstable.

*Controllability and Observability*

Again, the starting point for this analysis is the linearized system description

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\,\omega_0^2 & 0 & 0 & 2\,r_0 \cdot \omega_0 \\ 0 & 0 & 0 & 1 \\ 0 & -2\,\omega_0/r_0 & 0 & 0 \end{bmatrix} \cdot x + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/r_0 \end{bmatrix} \cdot u \tag{4.171}$$

and

$$y = \begin{bmatrix} 1/r_0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot x \tag{4.172}$$

With these matrices, the controllability matrix $\mathcal{R}$ is found to be

$$\mathcal{R} = [B, A \cdot B, A^2 \cdot B, A^3 \cdot B] = \begin{bmatrix} 0 & 0 & 1 & 0 & \dots \\ 1 & 0 & 0 & 2\,\omega_0 & \dots \\ 0 & 0 & 0 & 1/r_0 & \dots \\ 0 & 1/r_0 & -2\,\omega_0/r_0 & 0 & \dots \end{bmatrix} \tag{4.173}$$

Obviously, this linearized system is completely controllable. In fact, the first four columns already are linearly independent (the determinant of that submatrix is $-1/r_0^2$).

If the radial thruster fails, the matrix $B$ is reduced to $b_2 = [0,\ 0,\ 0,\ 1/r_0]^T$. The corresponding controllability matrix has the form

$$\mathcal{R}_2 = \begin{bmatrix} 0 & 0 & 2\,\omega_0 & 0 \\ 0 & 2\,\omega_0 & 0 & -2\,\omega_0^3 \\ 0 & 1/r_0 & 0 & -4\,\omega_0/r_0 \\ 1/r_0 & 0 & -4\,\omega_0^2/r_0 & 0 \end{bmatrix} \tag{4.174}$$

The determinant of this matrix is $-12\,\omega_0^4/r_0^2$, i.e., not zero on the reference orbit. Accordingly, the satellite remains completely controllable even if the radial thruster fails.

However, if the tangential thruster fails, the input matrix $B$ is reduced to $b_1 = [0,\ 1,\ 0,\ 0]^T$ and the controllability matrix

$$\mathcal{R}_1 = \begin{bmatrix} 0 & 1 & 0 & -\omega_0^2 \\ 1 & 0 & -\omega_0^2 & 0 \\ 0 & 0 & -2\,\omega_0/r_0 & 0 \\ 0 & -2\,\omega_0/r_0 & 0 & 2\,\omega_0^3/r_0 \end{bmatrix} \tag{4.175}$$

in this case, has only three independent columns (the fourth column is equal to $-\omega_0^2$ times the second column), i.e., the linearized system no longer is completely controllable.

The linearized system with no sensor faults is completely observable, as can easily be seen analyzing the first four rows of its observability matrix

$$\mathcal{O} = \begin{bmatrix} C \\ A \cdot C \\ A^2 \cdot C \\ A^3 \cdot C \end{bmatrix} = \begin{bmatrix} 1/r_0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1/r_0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{4.176}$$

If the radial sensor fails, the measurement matrix is reduced to $c_2 = [0,\ 0,\ 1,\ 0]$ and the resulting observability matrix $\mathcal{O}_2$ has the form

$$\mathcal{O}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -2\,\omega_0/r_0 & 0 & 0 \\ -6\,\omega_0^3/r_0 & 0 & 0 & -4\,\omega_0^2 \end{bmatrix} \tag{4.177}$$

This matrix is regular (its determinant is $-12\,\omega_0^4/r_0^2$). Accordingly, the linearized system remains completely observable even if the radial sensor fails.

However, if the tangential sensor fails, the measurement matrix is reduced to $c_1 = [1/r_0,\ 0,\ 0,\ 0]$ and the resulting observability matrix $\mathcal{O}_1$

$$\mathcal{O}_1 = \begin{bmatrix} 1/r_0 & 0 & 0 & 0 \\ 0 & 1/r_0 & 0 & 0 \\ 3\,\omega_0^2/r_0 & 0 & 0 & 2\,\omega_0 \\ 0 & -\omega_0^2/r_0 & 0 & 0 \end{bmatrix} \tag{4.178}$$

in this case is singular (the third column is zero). Accordingly, the linearized system is not completely observable anymore.

Note that the analysis presented in this section has immediate consequences for the design of the satellite. The main result is that the tangential actuator and sensor are more important for the operation of the satellite than their corresponding radial counterparts. In fact, if the latter fail, the tangential counterpart can still guarantee system stability in a reconfigured control scheme.

### Transfer Function

The system has two inputs and two outputs and, therefore, its transfer function $P(s)$ has four scalar transfer function elements. Note that since the internal description $\{A, B, C, 0\}$ is known, the derivation of the transfer function matrix is not difficult

$$P(s) = \begin{bmatrix} P_{11}(s) & P_{12}(s) \\ P_{21}(s) & P_{22}(s) \end{bmatrix} = C \cdot (s \cdot I - A)^{-1} \cdot B = \frac{C \cdot \mathrm{Adj}(s \cdot I - A) \cdot B}{\det(s \cdot I - A)} \quad (4.179)$$

The determinant was derived in the last section. The adjoint of a matrix $M$ is defined by

$$[\mathrm{Adj}(M)]_{j,i} = -1^{i+j} \cdot \det(M_{i,j}) \quad (4.180)$$

where $M_{i,j}$ is obtained from $M$ by deleting the $i$-th row and $j$-th column. Analyzing the structure of the matrices $B$ and $C$ it becomes clear that only four of the 16 elements of $\mathrm{Adj}(s \cdot I - A)$ are relevant, *viz.* the elements $\{(1,2), (1,4), (3,2), (3,4)\}$.

The details of the computations of these four elements are as follows:

$$[\mathrm{Adj}(M)]_{1,2} = -\det \begin{bmatrix} -1 & 0 & 0 \\ 0 & s & -1 \\ 2\,\omega_0/r_0 & 0 & s \end{bmatrix} = s^2 \quad (4.181)$$

$$[\mathrm{Adj}(M)]_{1,4} = -\det \begin{bmatrix} -1 & 0 & 0 \\ s & 0 & -2\,\omega_0 r_0 \\ 0 & s & -1 \end{bmatrix} = 2\,\omega_0 \cdot r_0 \cdot s \quad (4.182)$$

$$[\mathrm{Adj}(M)]_{3,2} = -\det \begin{bmatrix} s & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 2\,\omega_0/r_0 & s \end{bmatrix} = -s \cdot 2\,\omega_0/r_0 \quad (4.183)$$

$$[\mathrm{Adj}(M)]_{3,4} = -\det \begin{bmatrix} s & -1 & 0 \\ -3\,\omega_0^2 & s & -2\,\omega_0 \cdot r_0 \\ 0 & 0 & -1 \end{bmatrix} = s^2 - 3\,\omega_0^2 \quad (4.184)$$

With these four elements, the matrix $C \cdot \mathrm{Adj}(s \cdot I - A) \cdot B$ can be computed

$$\begin{bmatrix} s^2/r_0 & 2\,\omega_0 \cdot s/r_0 \\ -2\,\omega_0/r_0 \cdot s & (s^2 - 3\,\omega_0^2)/r_0 \end{bmatrix} \quad (4.185)$$

Finally, after canceling all common terms, the transfer function has the form

$$P(s) = \begin{bmatrix} \dfrac{1}{r_0 \cdot (s^2 + \omega_0^2)} & \dfrac{2\,\omega_0}{r_0 \cdot s \cdot (s^2 + \omega_0^2)} \\ \dfrac{-2\,\omega_0}{r_0 \cdot s \cdot (s^2 + \omega_0^2)} & \dfrac{s^2 - 3\,\omega_0^2}{r_0 \cdot s^2 \cdot (s^2 + \omega_0^2)} \end{bmatrix}$$

As mentioned in Section 4.7, the transfer function of the linearized system shows that the linearized system is unstable.

The single transfer functions show interesting properties. As was seen above, the linearized system is completely controllable and observable with only the tangential thruster and sensor working. Accordingly, the transfer function $P_{2,2}(s)$ is the only one that has no pole/zero cancellations. If either the tangential sensor or the tangential thruster fail, one pole/zero cancellation takes place. If they both fail, then two pole/zero cancellations take place, as shown in the element $P_{1,1}(s)$.

Note that even if the linearized system *is* completely controllable and observable, and hence stabilizable, with only the tangential thruster and sensor working, the associated control design problem is rather difficult. In fact, the corresponding SISO transfer function $P_{22}(s)$ has a non-minimumphase zero at $\sqrt{3} \cdot \omega_0$, which limits the attainable crossover frequency to approximately $0.85\,\omega_0$.

Now the advantages of a MIMO control approach become obvious. In fact, the MIMO system $\{A, B, C, 0\}$, defined in (4.171) and (4.172), has no finite transmission zeros. This can be seen by computing the determinant of the matrix

$$Z(s) = \begin{bmatrix} (s \cdot I - A) & -B \\ C & D \end{bmatrix} \quad (4.186)$$

which turns out to be

$$\det Z(s) = \frac{1}{r_0^2} \quad (4.187)$$

Since no finite $s$ can make this determinant equal to zero, no finite zero exists. In other words, the MIMO system $P(s)$ is minimumphase. Compared to the SISO design using only $P_{22}(s)$, the design of a suitable controller using both inputs and outputs will yield a better closed-loop system performance.

### Simulations

In this section two simulations are shown. The main objective is to convey a first impression of the differences between the linear system behavior, as analyzed above, and the nonlinear system behavior. The system is assumed to be in its reference state for all $t < 0$. At that moment a test signal

$$d(t) = \begin{cases} d_0 & 0 \leq t < 1 \\ 0 & 1 \leq t \end{cases} \tag{4.188}$$

is applied on one of the two control inputs (one of the thrusters is briefly fired). In order to clearly show the differences between the linear and the nonlinear system behavior, a first simulation is performed with $d_0$ rather small and a second simulation is run with $d_0$ very large.

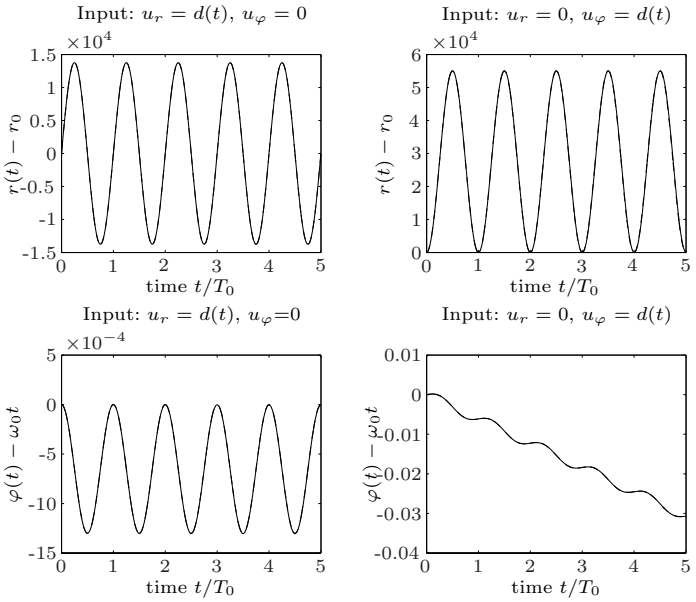



**Fig. 4.15.** System behavior for $d_0 = 1\ m/s^2$, orbit period $T_0 = 2\,\pi/\omega_0$. Solid: nonlinear system response; dashed: linearized system response.

Figure 4.15 shows that, for small inputs $u_r$ and $u_\varphi$, the linearized and the original nonlinear system behave almost in the same way. Note the unstable behavior of the azimuth error $\varphi(t) - \omega_0 \cdot t$ of the satellite (lower right plot). A closer inspection of that plot also shows the non-minimumphase nature of the transfer function $P_{22}(s)$. In fact, for small $t/T_0$, the deviation $\varphi(t) - \omega_0 \cdot t$ reaches positive values while the "gain" of the system is negative, as witnessed by the steady decrease for larger $t$.

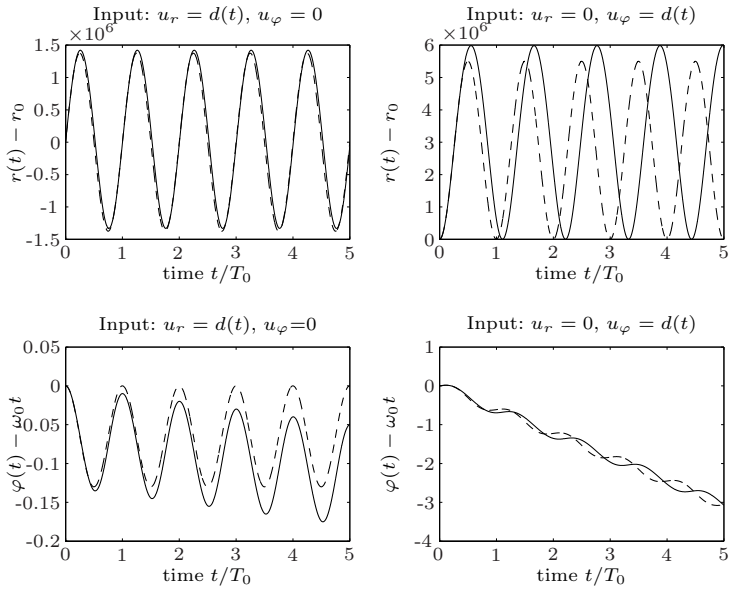



**Fig. 4.16.** System behavior for $d_0 = 100\ m/s^2$, orbit period $T_0 = 2\,\pi/\omega_0$. Solid: nonlinear system response; dashed: linearized system response.

For large inputs, the nonlinear system behavior clearly differs from its linearized counterpart.[21] The most interesting result is visible in the plot in the lower left corner. The nonlinear behavior of the azimuth of the satellite appears[22] to be *unstable* even when only the input $u_r$ is active. This behavior is *not* predicted by the linearized system, as the transfer function $P_{21}(s)$ shows. However, $P_{21}(s)$ has poles with zero real parts and the Lyapunov principle does not hold for such critical systems. The satellite system is an example of a system whose higher-than-linear terms force it into an unstable behavior.

[21] Of course, the input $d_0 = 100\ m/s^2$ is unrealistic. It has been chosen to highlight the differences between the linearized and the nonlinear system.

[22] Note that a simulation can never *prove* that a system is unstable.

# 5

# Analysis of Nonlinear Systems

## 5.1 Some Definitions

Linear systems have the following three special sets: (1) one isolated equilibrium point, (2) entire equilibrium subspaces, and (3) periodic orbits with one fixed frequency but arbitrary amplitude. For nonlinear systems those sets are important as well. However, nonlinear systems can have more complicated *limit sets*.[1] For instance, nonlinear systems can have *isolated* periodic orbits ("limit cycles") to which all trajectories converge that start sufficiently close to that orbit. Section 5.2.6 shows an example of such a system.

To analyze the stability of such periodic orbits a formulation proposed by Floquet may be used [6]. In this approach the question of stability of a periodic orbit is translated to the problem of stability of a discrete time system. As Figure 5.1 shows, the periodic orbit $x(t) \in I\!R^n$ is "asymptotically stable" (under some mild additional assumptions) if the intersections $w(k)$ with an $n-1$-dimensional hyperplane do converge to a fixed point $w_0$.

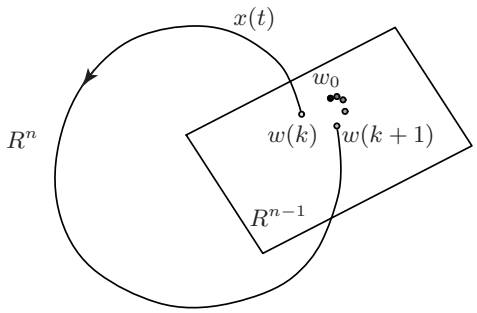



**Fig. 5.1.** Definitions for Floquet stability.

[1] The precise definition of these objects is deferred to Section 5.3.2.

Nonlinear systems can have even more complicated limit sets. If these sets are attractive, i.e., if all trajectories converge to them provided that they start sufficiently close, then they are referred to as "strange attractors." The system trajectories in such a strange attractor are not periodic, but "almost periodic" in the sense that trajectories do never leave that bounded set, if they start in it. Section 5.3.3 shows an example of such an object. Closely related to strange attractors are *chaotic* systems. A few remarks on that vast topic are made in Section 5.3.4.

Before discussing these more complex limit sets, the stability of isolated equilibrium points will be analyzed in the next section. The equilibrium points of nonlinear systems have several somewhat unexpected properties:

- nonlinear systems can have (infinitely) many isolated equilibrium points;
- such an equilibrium point can have a finite region of attraction only (as discussed in the Example 5.4);
- an equilibrium point can be non-exponentially asymptotically stable (this phenomenon is discussed in Example 5.1); and
- if an equilibrium point is unstable the state of the system can "escape to infinity" (its norm can become infinitely large) in finite time (Example 5.2 illustrates that effect).

## 5.2 Stability of Nonlinear Systems

### 5.2.1 Definition Lyapunov Stability

For isolated equilibrium points Lyapunov stability [6] is the most useful stability concept. Lyapunov stability always is connected to a constant equilibrium point $x_e$ of a system

$$\frac{d}{dt}x(t) = f(x(t), t), \quad x(t_0) = x_0 \neq 0 \tag{5.1}$$

where $f(x_e, t) = 0 \; \forall \, t > t_0$.[2]

[2] Notice that if a constant equilibrium point $x_e$ exists, one can always assume $x_e = 0$. If this is not the case, a coordinate transformation $\tilde{x} = x - x_e$ yields a system description that satisfies this condition.

The point $x_e = 0$ is *stable in the sense of Lyapunov* at $t = t_0$ if for any $R > 0$ there exists a $r(R, t_0)$ such that if $||x_0|| < r(R, t_0)$ then $||x(t)|| < R$ for all times $t > t_0$. In other words, a system is Lyapunov stable if for any bound $R$ one chooses, there is a bound on the initial condition $||x_0|| < r(R, t_0) \leq R$ such that the trajectory $x(t)$ is bounded by $||x(t)|| < R$ for all times. If $r$ does not depend on $t_0$ then the system is *uniformly* Lyapunov stable (ULS).

The same point $x_e$ is *asymptotically stable* if it is ULS and *attractive*, i.e.

$$\lim_{t \to \infty} x(t) = x_e = 0 \tag{5.2}$$

It is *exponentially* asymptotically stable if there exist constant scalars $a > 0$ and $b > 0$ such that

$$||x(t)|| \leq a \cdot e^{-b \cdot (t - t_0)} \cdot ||x_0|| \tag{5.3}$$

In general, only an exponentially asymptotically stable equilibrium is acceptable for technical applications (this form of stability is "robust" with respect to modeling errors). Examples of non-exponential asymptotically stable systems are analyzed in Section 5.1.

### 5.2.2 First and Second-Order Systems

First-order time-invariant nonlinear systems

$$\frac{d}{dt}x(t) = f(x(t)), \quad t, x \in I\!R \tag{5.4}$$

are the easiest to analyze as the variables $t$ and $x$ can be separated into a standard integration problem

$$\int \frac{dx}{f(x)} = \int dt = t + c \tag{5.5}$$

that can be solved explicitly in many cases (the constant $c$ is used to satisfy the initial conditions). The following example illustrates this approach.

*Example 5.1 (Non-Exponential Asymptotic Stability).*

If a linear system is asymptotically stable then it is always exponentially asymptotically stable. This is not true for nonlinear systems. The first-order nonlinear system

$$\frac{d}{dt}x(t) = -x^3(t), \qquad x(0) = x_0 \neq 0 \tag{5.6}$$

is an example of a critical system (see below) which turns out to be asymptotically stable, but which converges slower than exponentially to its equilibrium

point. In this very simple case it is possible to find the explicit solution of the ODE (5.6)

$$\int \frac{dx}{-x^3} = \int dt \quad \Rightarrow \quad \frac{1}{2\,x^2} = t + c$$

and, therefore

$$x(t) = x_0 \cdot (2 \cdot t \cdot x_0^2 + 1)^{-1/2}$$

This solution approaches the equilibrium slower than exponentially. In fact, inserting the explicit solution of the ODE (5.6) in the condition (5.3)

$$||x(t)|| = ||x_0|| \cdot (2 \cdot t \cdot x_0^2 + 1)^{-1/2} \leq a \cdot e^{-b\,t} \cdot ||x_0||$$

yields the inequality

$$1 \leq a \cdot e^{-b\,t} \cdot (2 \cdot t \cdot x_0^2 + 1)^{1/2}$$

For $\lim t \to \infty$ the right-hand side of this inequality tends to 0 for all possible $0 < a, b < \infty$, and this contradiction shows that the equilibrium is approached slower than exponentially.

Historically, stability theory started with two-dimensional time-invariant nonlinear systems, because their solutions can be plotted on a plane $x_1(t)$, $x_2(t)$ (for which the term "phase plane" was coined). Such systems can always be described by two scalar ODE

$$\begin{aligned} \tfrac{d}{dt}x_1(t) &= f_1(x_1, x_2), \qquad x_1(0) = x_{1,0} \\ \tfrac{d}{dt}x_2(t) &= f_2(x_1, x_2), \qquad x_2(0) = x_{2,0} \end{aligned} \tag{5.7}$$

Unfortunately, many important effects that are observed in general dynamic systems (for instance, deterministic chaos) are not possible in "smooth" (differentiable) two-dimensional continuous-time systems (in "flatland," the whole space is always divided into two distinct regions by any closed curve). Nevertheless, the analysis of two-dimensional systems is worthwhile because this simple situation already illustrates some of the main points.

The starting point is to develop (5.7) into a Taylor series (assuming that this is possible) around an isolated equilibrium (which is assumed to be in the origin)

$$\begin{aligned} \tfrac{d}{dt}x_1(t) &= a_{11}x_1(t) + a_{12}x_2(t) + \widetilde{f}_1(x_1, x_2), \qquad x_1(0) = x_{1,0} \\ \tfrac{d}{dt}x_2(t) &= a_{21}x_1(t) + a_{22}x_2(t) + \widetilde{f}_2(x_1, x_2), \qquad x_2(0) = x_{2,0} \end{aligned} \tag{5.8}$$

with

$$a_{i,j} = \frac{\partial f_i}{\partial x_j}, \quad i, j = 1, 2 \qquad \lim_{||x|| \to 0} \frac{\widetilde{f}_i(x_1, x_2)}{||x||} = 0 \tag{5.9}$$

The four coefficients, of course, form the system matrix

$$A_{2\times 2} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{5.10}$$

and the linearized system is then given by

$$\frac{d}{dt}\delta x(t) = A \cdot \delta x(t), \qquad \delta x(t) = [\delta x_1(t), \delta x_2(t)]^T \tag{5.11}$$

The following important result is now available (also known as the Lyapunov Principle):

*Excluding the case where the matrix A has eigenvalues with zero real part, the local behavior of the nonlinear system (5.7) and the behavior of the linear system (5.11) are topologically equivalent (i.e., they have the same main "geometric characteristics").*

| eigenvalues | linearized system | nonlinear system |
|---|---|---|
| $\lambda_1 \in \mathcal{C}_-, \lambda_2 \in \mathcal{C}_-$ | Stable Focus | Stable Focus |
| $\lambda_1 \in I\!R_-, \lambda_2 \in I\!R_-$ | Stable Node | Stable Node |
| $\lambda_1 \in I\!R_+, \lambda_2 \in I\!R_-$ | Saddle | Saddle |
| $\lambda_1 \in I\!R_+, \lambda_2 \in I\!R_+$ | Unstable Node | Unstable Node |
| $\lambda_1 \in \mathcal{C}_+, \lambda_2 \in \mathcal{C}_+$ | Unstable Focus | Unstable Focus |
| $Re(\lambda_{1,2}) = 0$ | Center | ? ? ? |

Notice that since (5.11) has real coefficients, the complex eigenvalues always have the same real part. Figure 5.2 shows the plots of the different behaviors. The first five "topologies" coincide for the nonlinear system and its linearized counterpart. The sixth plot (lower right-hand corner) is valid only for a strictly linear system.

Fortunately, the Lyapunov Principle is valid also for all higher-order time-invariant nonlinear systems

$$\frac{d}{dt}x(t) = f(x(t)), \quad x(t_0) = x_0 \neq 0 \tag{5.12}$$

*As long as the linearization of a dynamic system (5.12) has no eigenvalues on the imaginary axis, its local stability properties are fully understood once the eigenvalues of the linearization are known. Particularly, if the linearization of a nonlinear system (5.12) around an isolated equilibrium point $x_e$ is asymptotically stable (unstable) then this equilibrium is an asymptotically stable (unstable) equilibrium of the nonlinear system as well.*

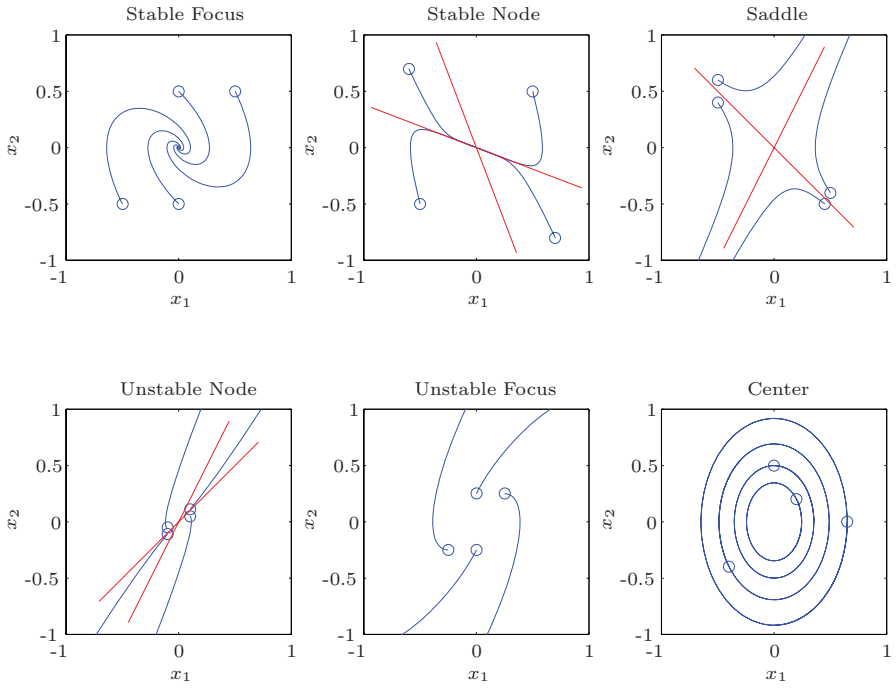



**Fig. 5.2.** Classification of the system behaviors, the dots indicate the initial points. For the second, third, and fourth case, real eigenvectors exist as indicated in the plots.

Notice that this result says nothing about the "large signal behavior," i.e., stability in the sense of Lyapunov is a strictly local concept (and local can be a very small neighborhood of the equilibrium, see next section). The critical case is when the linearization has eigenvalues with zero real part, i.e., $\lambda_{1,2} = \pm j\omega$. The nonlinear system's behavior, then, is not defined by its first-order approximation, but by the higher-order terms $\widetilde{f}(x)$.

*Example 5.2 (Critical Nonlinear System).*

The system to be analyzed is given by

$$\begin{aligned} \tfrac{d}{dt}x_1 &= -x_1 + x_2 \\ \tfrac{d}{dt}x_2 &= x_2^3 \end{aligned} \tag{5.13}$$

Obviously, this system has only one isolated equilibrium at $x_{e,1} = x_{e,2} = 0$.

The system linearization has one eigenvalue at $-1$ and one at 0. The solution of the linear system is stable (although the origin is not asymptotically

stable) while the nonlinear system is unstable (it has even *finite* escape times). This can be inferred from the explicit solution for $x_2(t)$, which is easily found to be

$$x_2(t) = \frac{x_{2,0}}{\sqrt{1 - 2\,t\,x_{2,0}^2}}$$

Obviously, with $\lim t \to 1/(2\,x_{2,0}^2)$ this state variable will "escape to infinity."

A general theory that is able to analyze high-order critical systems is known as "center manifold theory" (see [8] for an introduction to center-manifold theory). Its key idea is to separate the problem in an asymptotical convergence to a central manifold, to which the critical part of the dynamics are confined to, and then to analyze the (lower-dimensional) system dynamics on this submanifold using nonlinear stability methods (see below).

Sometimes "first integrals" can be found directly. Typically, these first integrals are conserved quantities (like Hamiltonian energy, momentum, etc.). These integrals help to understand the system behavior "at large" and are often also applicable in critical cases.

*Example 5.3 (Nonlinear Pendulum – Stability Analysis).*

Linearizing Example 2.5 at $\varphi = 0$ (upright position) yields

$$\frac{1}{3}ml^2\ddot{\varphi}(t) = \left[\frac{l}{2}mg - cl^2\right]\varphi(t) \tag{5.14}$$

The eigenvalues of the system are

$$s_{1,2} = \pm j\sqrt{3\left(\frac{c}{m} - \frac{g}{2l}\right)} \tag{5.15}$$

Hence, the linearized system is stable, but not asymptotically stable (critical case) for all parameters that satisfy the condition

$$mg < 2cl \tag{5.16}$$

According to the Lyapunov principle, the linearized system description does not suffice to decide whether the equilibrium is stable or unstable. In this example, this question can be answered using a first integral. This first integral (the conserved quantity) is the Hamiltonian (total energy) of the system defined by

$$H = \frac{1}{6}\ m\ l^2\ \dot{\varphi}^2 + \frac{1}{2}\ m\ g\ l\ (\cos(\varphi) - 1) + \frac{1}{2}\ c\ l^2 \sin^2(\varphi) \tag{5.17}$$

Solving this equation with different constant values for $H$ yields the orbits shown in Figure 5.3. For sufficiently small initial conditions (total energy) the

system is indeed Lyapunov stable: if the system is started at zero velocity $\dot{\varphi}(0) = 0$ and sufficiently small angle $\varphi < \varphi_\sigma$, it oscillates around its upper equilibrium. The limit is given by

$$\varphi_\sigma = \arccos\left\{\frac{m\,g}{2\,c\,l}\right\} \tag{5.18}$$

For initial conditions $\dot{\varphi}(0) = 0$ and $\varphi(0) > \varphi_\sigma$, the system oscillates around the lower equilibrium $\varphi_0 = \pi$ and never reaches the upper equilibrium $\varphi_0 = 0$.

Both these situations correspond to a Lyapunov stable system behavior and there is no possibility to "inject" more energy in the system if $\dot{\varphi}(0) = 0$. However, for sufficiently large initial conditions $|\dot{\varphi}(0)| > 0$ the system becomes unstable (the rotation angle $\varphi(t)$ growths without bounds as the pendulum swings through both equilibrium points). The limit between the stable and the unstable region is shown by the red trajectories in Figure 5.3.

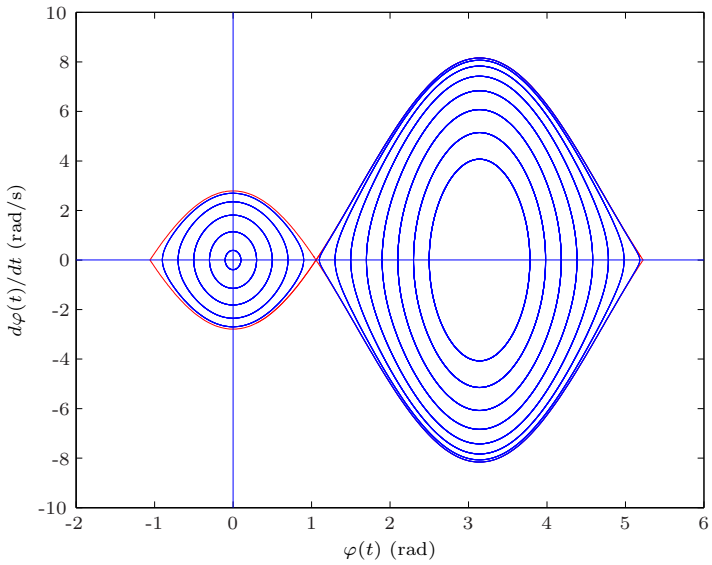



**Fig. 5.3.** Orbits of the pendulum system for different initial conditions $\varphi(0)$, separation point at $\varphi_\sigma$, upper equilibrium at $\varphi = 0$, lower equilibrium at $\varphi = \pi$.

The last example shows how powerful energy methods can be for second-order systems. Unfortunately, for higher-order systems the conclusions that can be derived by energy conservation laws are much less complete.[3]

[3] For instance, it is still an open question whether our solar system is Floquet stable or whether chaotic behavior is possible for sufficiently large observation periods!

### 5.2.3 A Glimpse on Lyapunov Theory

As mentioned in the previous sections, the *local* stability properties of the isolated equilibrium point $x_e = 0$ of a time-invariant nonlinear system

$$\frac{d}{dt}x(t) = f(x(t)), \quad x(0) \neq 0 \tag{5.19}$$

are fully described by the first-order approximation $A$ of $f(.)$

$$A = \left.\frac{\partial f}{\partial x}\right|_{x=0} \tag{5.20}$$

provided $A$ has no eigenvalues with zero real part. If this is not true or if one is interested in non-local results or in more general nonlinear systems, then Lyapunov's direct method, as briefly[4] introduced below, can be useful.

The starting point is the time-varying nonlinear system (5.1) and the definition of Lyapunov stability given at the beginning of this chapter. In addition, a few new definitions are required.

- A scalar function $\alpha(p)$ with $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is a *strictly increasing* function if $\alpha(0) = 0$ and $\alpha(p) > \alpha(q)$ if $p > q$.
- A function $V : \mathbb{R}^n \to \mathbb{R}$ is a candidate global Lyapunov function if
  - the function is positive definite, i.e., $V(x,t) > 0 \,\forall\, x \neq 0$, and $V(0,t) = 0$ $\forall\, t$; and
  - there are two strictly increasing functions $\alpha$ and $\beta$ of the norm of the state variables $||x(t)||$ that satisfy the inequalities $\beta(||x||) \leq V(x,t) \leq \alpha(||x||)$.

With these preparations, it now possible to state the two main Lyapunov theorems.

**Theorem 1** The system (5.1) is uniformly globally *stable* in the sense of Lyapunov if there is a global Lyapunov function candidate $V(x,t)$ for which the following inequality holds true

$$\frac{d}{dt}V(x,t) = \frac{\partial V(x,t)}{\partial t} + \frac{\partial V(x,t)}{\partial x}f(x,t) \leq 0 \;\; \forall\, x(t) \neq 0, \; \forall\, t \tag{5.21}$$

**Theorem 2** The system (5.1) is uniformly globally *asymptotically* stable if there is a global Lyapunov function candidate $V(x,t)$ such that $-\frac{d}{dt}V(x,t)$ is a positive definite function.

[4] The concepts mentioned are really the most basic ideas. Many extensions exist and no proofs are given here. Interested readers are referred to [20].

The main idea used for the proofs of these theorems is that if a quantity that depends in a positive-definite way on the norm of the state variables is non-increasing (strictly decreasing) in time, then the norm of the state variables may not grow (must asymptotically vanish).

Obviously, all results can be reformulated in a local way by asking all conditions to be true only in a neighborhood of the equilibrium point. In general, even if one looks only for local results, it is very difficult to find a suitable function $V(x,t)$. Sometimes the knowledge of the physical structure of the system can be used to derive Lyapunov functions. In fact, Lyapunov functions can be interpreted as generalized energy functions.

Note that the Lyapunov theorems provide sufficient but not necessary conditions, i.e., if a Lyapunov function is found, then the system properties stability or asymptotic stability are proven. If no Lyapunov function is found, i.e., if a chosen Lyapunov function candidate turns out not to satisfy the conditions of the Lyapunov theorems, then no conclusions may be drawn. In this case the system can still be stable or asymptotically stable. In particular, the conditions of the second theorem are usually difficult to satisfy. In this case an extension of Lyapunov's method proposed by LaSalle is often useful. Interested reades are referred to [6].

A universal Lyapunov function can be found in all cases only for linear systems

$$\frac{d}{dt}x(t) = A\,x(t), \qquad x(0) \neq 0 \tag{5.22}$$

using the "Ansatz"

$$V(x) = x^T P x \tag{5.23}$$

The symmetric and positive definite matrix $P$ is the solution of the Lyapunov equation

$$PA + A^T P = -Q \tag{5.24}$$

For arbitrary $Q = Q^T > 0$, a solution to this equation exists if, and only if, the matrix $A$ is a Hurwitz matrix. This result provides no new information, since it is already well known that a matrix $A$ whose eigenvalues all have negative real parts characterizes an asymptotically stable system and vice versa. However, if the matrix $A$ represents the linearization of a nonlinear system, then the Lyapunov function (5.23) can sometimes be useful to generate a Lyapunov function for the nonlinear system or to estimate its region of attraction.

*Example 5.4 (Lyapunov Analysis of a Second-Order System).*

The system to be analyzed is given by

$$\begin{aligned} \tfrac{d}{dt}x_1 &= x_1(x_1^2 + x_2^2 - 1) - x_2 \\ \tfrac{d}{dt}x_2 &= x_1 + x_2(x_1^2 + x_2^2 - 1) \end{aligned} \tag{5.25}$$

Obviously, this system has only one isolated equilibrium at $x_1 = x_2 = 0$. Linearizing the system (5.25) in the equilibrium yields the following system matrix

$$A = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} \tag{5.26}$$

Since the eigenvalues of the matrix are stable ($\lambda_{1,2} = -1 \pm j$), the system (5.25) is locally asymptotically stable. From the special form of (5.25), it is clear[5] that this property cannot be global. Therefore, the system must have a finite region of attraction. The size of this region can be very small. The following Lyapunov analysis yields an estimation of the region of attraction.

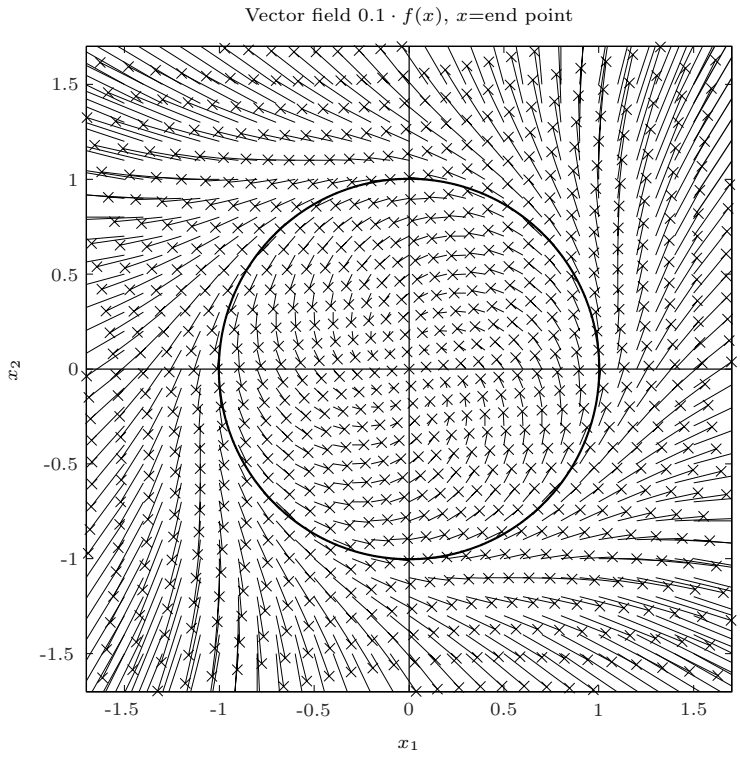



**Fig. 5.4.** Vector field of the system (5.25). To improve the clarity of the picture,, the vector field $f(x)$ has been scaled by a factor of 0.1. The symbol x indicates the end point of the vector $f(x)$ that is "attached" to the corresponding $x$.

[5] For large $||x||$, the nonlinear terms dominate the linear terms such that the system dynamics are approximated by $\dot{x}_1 \approx x_1||x||$ and $\dot{x}_2 \approx x_2||x||$. This describes a radially divergent behavior.

The chosen candidate Lyapunov function is defined by

$$V(x) = x_1^2 + x_2^2 \tag{5.27}$$

Clearly, this function satisfyies all conditions that a Lyapunov candidate must satisfy. The time derivative of $V$ along a trajectory of (5.25) is given by

$$\dot{V}(t) = 2(x_1^2 + x_2^2)(x_1^2 + x_2^2 - 1) \tag{5.28}$$

This shows that at least the region defined by $||x||^2 = x_1^2 + x_2^2 < 1$ must be included in the region of attraction. As Figure 5.4 shows, this is indeed the complete region of attraction.[6]

### 5.2.4 Circle Criterion

The circle and the Popov (see next section) stability criteria are derived using the Lyapunov theory introduced in the last section. Here, only the most important results will be presented. The proofs and several extensions can be found in [10].

The system analyzed in this section is illustrated in Figure 5.5. It consists of a linear and time-invariant SISO dynamic part described by its loop gain $L(s)$ and a "memory-less" time-varying nonlinearity $\phi(t, y)$ that is inserted in the feedback path. Quite a few feedback control problems can be reformulated to be compatible with this structure.

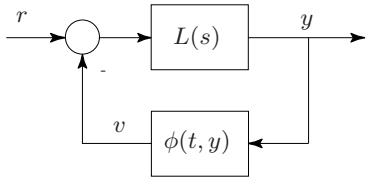



**Fig. 5.5.** Nonlinear system structure analyzed using the circle criterion.

The nonlinearity $\phi(t, y)$ is assumed to be a sector-bounded function that for all times $t$ satisfies the inequalities

$$\alpha \cdot y < \phi(t, y) \cdot y < \beta \cdot y, \quad \alpha, \beta \in I\!R, \quad 0 < \alpha < \beta \tag{5.29}$$

These inequalities are illustrated on the left in Figure 5.6. Notice that no restrictions on the rate of change of $\phi(t, y)$ are imposed.

[6] In general, it is very difficult to compute the exact region of attraction. One only succeeds in such simple and "rigged" examples as analyzed here.

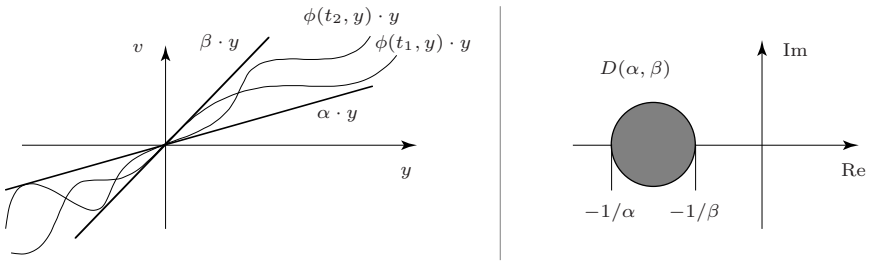



**Fig. 5.6.** Left: illustration of the sector nonlinearity $\phi(t, y)$. Right: definition of the circle $D(\alpha, \beta)$.

**Circle Criterion**: Assume that $L(s)$ is a strictly proper transfer function that has $n_+$ unstable poles and $n_0$ poles on the imaginary axis. Assume the nonlinearity $\phi(t, y)$ is sector bounded by (5.29). Then the closed-loop system of Figure 5.5 is asymptotically stable if (i) the Nyquist curve $L(j\,\omega)$ does not enter the disk $D(\alpha, \beta)$ and (ii) the Nyquist curve $L(j\,\omega)$ encircles $n_+ + n_0/2$ times[7] the disk $D(\alpha, \beta)$.

**Remarks:** Obviously, the circle criterion is a generalization of the Nyquist criterion in which $\alpha = \beta = 1$ is assumed. As for the Nyquist criterion, the circle criterion is not only valid for real-rational systems, but also for the case where $L(s)$ includes some non-rational parts (time delays). Although the circle criterion is formulated as a sufficient condition for stability, it can be shown to be necessary[8] as well. This proof, however, requires other tools than Lyapunov theory, which only yields sufficient results. Since stability is guaranteed for arbitrarily fast variations, slow variations will yield conservative stability conditions.

*Example 5.5 (Circle-Criterion Analysis of the Mathieu Equation).*

The system to be analyzed is given by

$$\begin{aligned} \tfrac{d}{dt} x_1 &= x_2 \\ \tfrac{d}{dt} x_2 &= -2\,\mu\, x_2 - (\mu^2 + a^2)\, x_1 + u(t) \end{aligned} \tag{5.30}$$

and

$$u(t) = -q\, \cos(\omega_0\, t)\, y(t), \quad y(t) = x_1(t) \tag{5.31}$$

Following the notation introduced in Figure 5.5, the loop gain is given by

$$L(s) = \frac{1}{s^2 + 2\,\mu\, s + (a^2 + \mu^2)} \tag{5.32}$$

[7] As in the linear case, counter-clockwise encirclements are counted positive and $\omega$ goes from $-\infty$ to $+\infty$.

[8] This point needs some attention. The following example will clarify it.

and the gain $\phi(t, y)$

$$\phi(t, y(t)) = q \cos(\omega_0 t) \tag{5.33}$$

Notice that in this example the gain $\phi(t, y(t))$ does not depend on $y(t)$, but only on $t$. In other words, the Mathieu system is a linear but time-varying system. The sector bounds $\alpha$ and $\beta$ in this particular case do not satisfy the condition introduced in (5.29). In this example

$$\alpha = -q < 0 < \beta = q \tag{5.34}$$

Accordingly, positive feedbacks must be considered. Moreover, it is clear that the loop gain $L(s)$ may not have any unstable poles to permit the closed-loop system to be asymptotically stable for sufficiently small $q$.

The circle criterion must be modified for this case: the condition now is that the loop gain $P(j\,\omega)$ must be completely contained in the circle defined by $-1/\alpha = 1/q$ and $-1/\beta = -1/q$ (see Figure 5.7). Since the maximum magnitude of $P(j\,\omega)$ is equal to $1/(2\mu\,a)$, a sufficient condition for the closed-loop system to remain stable is that

$$q < 2\,\mu\,a \tag{5.35}$$

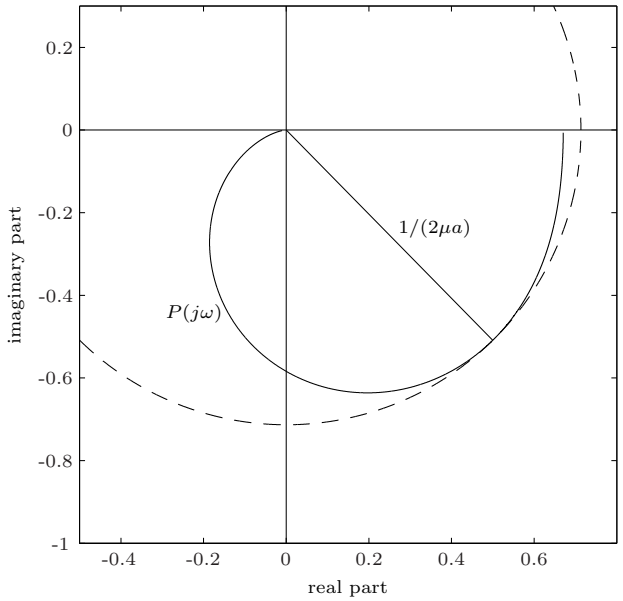



**Fig. 5.7.** Limiting circle and frequency response $P(j\,\omega)$ of the Mathieu example.

Note that this result is true for any $\omega_0$. Simulations show that the "worst" frequency, i.e., that frequency at which the smallest $q$ destabilizes the system, is given by $\omega_0 = 2\,\omega^*$ where $\omega^* = \sqrt{a^2 - \mu^2}$ is that frequency at which $|P(j\,\omega^*)| = 1/(2\,\mu\,a)$.

These simulations also show that the system is able to cope with $q > 2\,\mu\,a$, in particular if the frequency $\omega_0$ is not close to its critical value. It was stated above that the circle criterion yields sufficient and necessary conditions. However, necessity is only relevant for the worst possible nonlinear gain $\phi(t, y(t))$ that lies in the sector. In the case of the Mathieu system, the gain is limited to a special class, *viz.* $\phi(t, y(t)) = q\,\cos(\omega_0\,t)$. It would be a coincidence if this type of gain were the worst possible choice. Accordingly, for the Mathieu problem the circle criterion will be sufficient but no longer necessary.

### 5.2.5 Popov Criterion

Compared to the circle criterion, the Popov criterion is applicable to a smaller class of systems (see below). However, for these systems the Popov criterion yields powerful results and an interesting interpretation in the complex plane is possible as well.

The class of systems to which the Popov criterion may be applied is defined by the additional restrictions that are imposed on the system illustrated in Figure 5.5:

- the loop gain $L(s)$ may not have unstable poles; and
- the nonlinear gain $\phi(.)$ must be time invariant.

For such systems, the following theorem can be proven [10].

**Popov Criterion:** Assume that $L(s)$ and $\phi(.)$ satisfy the conditions mentioned above (the parameters $\alpha$ and $\beta$ are the limits shown in Figure 5.6). Then the closed-loop system illustrated in Figure 5.5 is asymptotically stable if the following condition can be satisfied for some real-positive scalar $r$

$$\text{Real}\,\{(1 + r\,j\,\omega)\,L(j\omega)\} + \frac{1}{\beta - \alpha} + \frac{\alpha}{\alpha + \beta}\,|L(j\,\omega)|^2 > 0, \quad \forall\,\omega \tag{5.36}$$

The Popov criterion yields "global" results, i.e., as long as the conditions under which this theorem holds true are satisfied, the asymptotic stability is guaranteed for arbitrary initial conditions of the linear system $L(s)$.

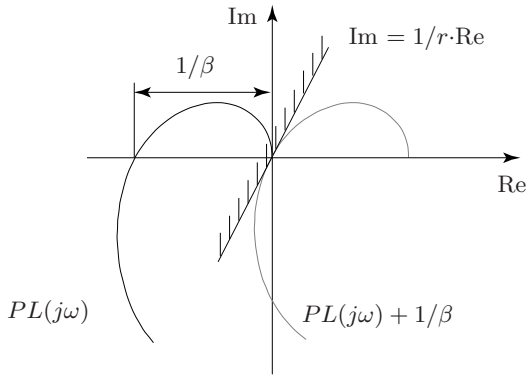



**Fig. 5.8.** Illustration of the application of Popov plots, case $\alpha = 0$.

For the special case in which $\alpha = 0$, the Popov criterion has an interesting geometric interpretation in the complex plane. In fact, by plotting the Popov plot $PL(j\,\omega)$ of the open-loop gain $L(j\,\omega)$ defined by

$$PL(j\,\omega) = \text{Real}\{L(j\omega)\} + j\,\omega\,\text{Imag}\{L(j\omega)\} \tag{5.37}$$

the maximum permissible sector slope $\beta$ can be inferred from the distance $1/\beta$ by which the Popov plot $PL(j\,\omega)$ has to be shifted such that a line exists with slope $1/r$ passing through the origin, which bounds the shifted Popov plot as shown in Figure 5.8. This is easy to see by rewriting (5.36) as follows

$$\text{Real}\{L(j\omega)\} - r\,\omega\,\text{Imag}\{L(j\omega)\} + \frac{1}{\beta} > 0 \tag{5.38}$$

and, therefore

$$\omega\,\text{Imag}\{L(j\omega)\} < \frac{1}{r} \cdot \left[\text{Real}\{L(j\omega)\} + \frac{1}{\beta}\right] \tag{5.39}$$

or

$$\text{Imag}\{PL(j\omega)\} < \frac{1}{r} \cdot \text{Real}\{PL(j\omega) + \frac{1}{\beta}\} \tag{5.40}$$

As the following example shows, this interpretation can lead to quite useful results.

In the case in which $\alpha > 0$, no simple geometric interpretation exists in the complex plane. In this situation, the expression on the left side of Equation (5.36) must be plotted for all positive frequencies $\omega$ and the condition (5.36) must be checked.

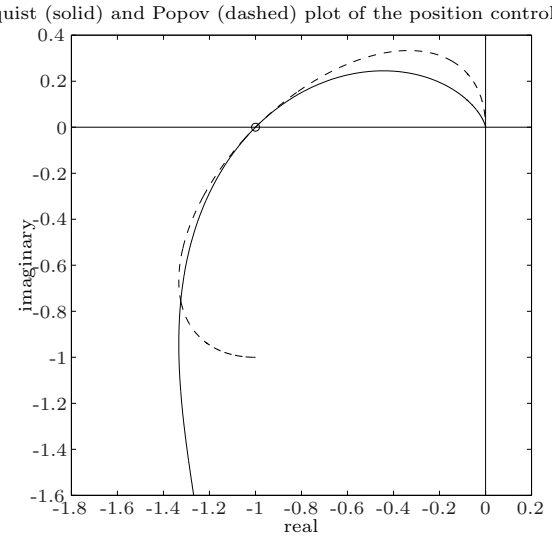



**Fig. 5.9.** Popov plot of the example 5.6.

*Example 5.6 (Linear System with Input Saturation).*

The system analyzed in this example is a position feedback loop, i.e., the position $x$ of a mechanical system of mass $m$ is to be controlled to a desired reference position $r$ using an electromechanical actuator and a P controller with gain $k_p$. The latter has a limited control authority. According to Section 2.4.4, the system is described by

$$\frac{d}{dt}x(t) = v(t)$$

$$\frac{d}{dt}v(t) = \frac{\kappa}{m}I(t)$$

$$\frac{d}{dt}I(t) = -\frac{R}{L}I(t) - \frac{\kappa}{L}v(t) + \frac{1}{L}U(t)$$

$$U(t) = \mathrm{sat}\left(k_p[r(t) - x(t)]\right)$$

Assuming, for the sake of a simple notation, that $L = R = m = \kappa = 1$, the transfer function of this system has the form

$$L(s) = \frac{1}{s}\,\frac{1}{s^2 + s + 1} \tag{5.41}$$

The Nyquist and the Popov plots of this system are shown in Figure 5.9 on the same graph. According to the Popov criterion the nonlinearity must lie

in the sector $\{0, 1\}$ in order for the closed-loop system to remain stable. Interestingly, the Nyquist plot shows that the maximum gain $k_p$ for the linear feedback is 1 as well. Therefore, for this plant, the feedback may be an arbitrary nonlinear function as long as it lies in the sector that is defined by the maximum allowable linear feedback. Notice that the Popov criterion only guarantees stability. As discussed in the courses "Regelungstechnik I&II," reset windup effects, which are to be expected in this case, can lead to very badly damped transients.

### 5.2.6 Describing Functions

#### Introduction

The describing function method is used to analyze the stability properties of a special class of nonlinear SISO systems. This class consists of all closed-loop systems that contain a *linear dynamic system*, usually described by its transfer function $L(s)$, and a *static nonlinear system*, denoted by $\phi(.)$, as in the previous sections. Figure 5.10 shows the block diagram of the system that will be analyzed below. The dynamic part must be an asymptotically stable low-pass system and the nonlinearity $\phi(.)$ is assumed to be an odd function, i.e., $\phi(-e) = -\phi(e)$ and time invariant.

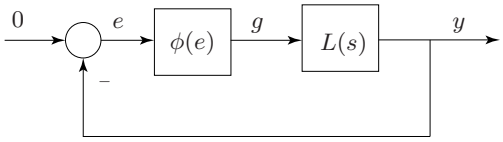



**Fig. 5.10.** Block diagram of the system analyzed with the describing function method.

Of course, the system shown in Figure 5.10 is very similar to the systems analyzed in the previous sections of this chapter (circle and Popov criterion). However, the main objective of the describing function approach is to predict the emergence or absence of *limit cycles*. Limit cycles are sustained periodic oscillations of the closed-loop system. In most cases, reasonable estimates can be given of the frequency and the amplitude of these oscillations. However, since the describing function method relies on approximations of the system behavior, all results are approximative and must be confirmed by simulations or experimentally.

### Some Remarks on Linear Systems

Before entering into the discussion of describing functions, a qualitative and intuitive re-interpretation of the well-known Nyquist stability result is given. This discussion will facilitate the introduction of the main ideas of describing functions in the subsequent sections.

In the linear case $\phi(e) = e$, the closed-loop system shown in Figure 5.10 is on its stability boundary when

$$e(t) = a \cdot \sin(\omega \cdot t), \quad y(t) = -a \cdot \sin(\omega \cdot t) = a \cdot \sin(\omega \cdot t - \pi) \tag{5.42}$$

In this case the oscillation is self-sustaining. Notice that the magnitude $a$ of the amplitude of the oscillation is not relevant in the linear case. The only conditions are that the amplitude of $e(t)$ and $y(t)$ are the same and that the phase of $y(t)$ lags by $-\pi$ the phase of $e(t)$. Of course, these two conditions are exactly the Nyquist result, which predicts marginal stability when

$$|L(j\,\omega)| = 1, \quad \text{and} \quad \arg(L(j\,\omega)) = -\pi \tag{5.43}$$

These two real-valued equations can be written as one complex-valued equation

$$L(j\,\omega) = 1 \cdot e^{-j\,\pi} = -1, \quad \text{or} \quad 1 + L(j\,\omega) = 0 \tag{5.44}$$

which is, of course, nothing more than the well-known condition that guarantees some system poles to be on the imaginary axis and, hence, the system to be marginally stable.

### Main Idea

In the nonlinear case, similar ideas can be used to assess the stability properties. Assuming the signal $e(t)$ to be the same as in the linear case, the output of the nonlinear, but static element $\phi(.)$ will be a *periodic* function as well. Therefore, this output can be written as a Fourier series.

$$g(t) = \phi(a \cdot \sin(\omega \cdot t)) = k_0(a) + \sum_{i=1}^{\infty} k_i(a) \cdot \sin(i \cdot \omega \cdot t + \varphi_i(a)) \tag{5.45}$$

The assumption that $\phi(.)$ is an odd function implies that $k_0(a) = 0$, and the assumption that $L(s)$ is a low-pass element justifies the simplification

$$g(t) \approx k_1(a) \cdot \sin(\omega \cdot t + \varphi_1(a)) \tag{5.46}$$

because any higher-frequency components in $g(t)$ will be filtered out by $L(s)$ and only the dominant component with the lowest frequency will have a strong influence on $y(t)$. Of course, this is never completely true and, in particular

for systems $L(s)$ with sharp resonances close to $i \cdot \omega$ ($i = 2, 3, \ldots$), the results shown below might not be correct.

However, assuming (5.46) to be exact, the following assertions become possible. The effect of the nonlinearity $\phi(.)$ is to change the amplitude and the phase of the signal entering this block. This can be represented compactly using a complex gain $DF(a)$ that is defined as

$$DF(a) = \frac{k_1(a) \cdot e^{j \cdot \varphi_1(a)}}{a} \tag{5.47}$$

The variable $DF(a)$ is the *describing function* of the nonlinearity $\phi(.)$. It describes the action of $\phi(.)$ on a special class of inputs $e(t)$ (harmonic functions), omitting all higher-order components in the output $g(t)$. In general, this action is a change in amplitude and a shift in phase, therefore, it can be described compactly using a complex gain. Note that this gain does not depend on the frequency $\omega$ ($\phi(.)$ is static), but on the amplitude $a$ ($\phi(.)$ is nonlinear) of the input $e(t) = a \cdot \sin(\omega\, t)$. Despite that, the describing function $DF(a)$ is similar to the transfer function $L(j\,\omega)$ in its effect: it transforms harmonic signals (approximately) to harmonic signals with a different amplitude and phase. Accordingly, the two operators can be concatenated and their action can be represented in the same Nyquist diagram.

Using the same reasoning introduced in the last section for the purely linear case, the closed-loop system shown in Figure 5.10 is marginally stable when the following two conditions are satisfied

$$k_1(a) \cdot |L(j\,\omega)| = a \cdot |DF(a)| \cdot |L(j\,\omega)| = a \tag{5.48}$$

and

$$\varphi_1(a) + \arg(L(j\,\omega)) = \arg(DF(a)) + \arg(L(j\,\omega)) = -\pi \tag{5.49}$$

Of course, these equations can again be written as one complex-valued equation that generalizes (5.42) to the nonlinear case analyzed in this section as follows

$$1 + DF(a) \cdot L(j\,\omega) = 0 \tag{5.50}$$

A limit cycle, therefore, can occur if there is a frequency $\omega^*$ and an amplitude $a^*$ for which equation (5.50) holds. This equation is highly nonlinear, and a closed-form solution is, in general, not possible. However, a graphical solution can be found as follows. First, the negative inverse of $DF(a)$ is plotted in a Nyquist plane for all possible amplitudes $a$. If this curve intersects[9] with the curve $L(j\,\omega)$, then there is a solution to equation (5.50) and the corresponding amplitude and phase are (approximately) the amplitude and phase of the limit cycle that arises in the closed-loop system. This result can be

[9] There may be several intersections.

interpreted as a generalization of the linear Nyquist theorem. In fact, if $\phi(.)$ is the identity function $\phi(e) = e$, its describing function is $DF(a) = 1$ for all gains $a$. Accordingly, the curve $-DF(a)$ is a "generalization" of the point $-1$, which, according to the Nyquist theorem, may not be part of $L(j\omega)$ to avoid sustained harmonic oscillations (the limit cycles of linear systems).

Three cases can be distinguished in general (see Figure 5.11):

- The curve $L(j\,\omega)$ neither intersects nor encircles the curve $-DF^{-1}(a)$. In this case, the closed-loop system (probably) will be asymptotically stable and no limit cycles will appear (remember: the loop gain $L(s)$ was assumed to have no unstable poles).
- The curve $L(j\,\omega)$ does not intersect the curve $-DF^{-1}(a)$, but encircles it. In this case, the closed-loop system (probably) will be unstable.
- The curve $L(j\,\omega)$ intersects the curve $-DF^{-1}(a)$. In this case, the closed-loop system can produce a limit cycle (see below).

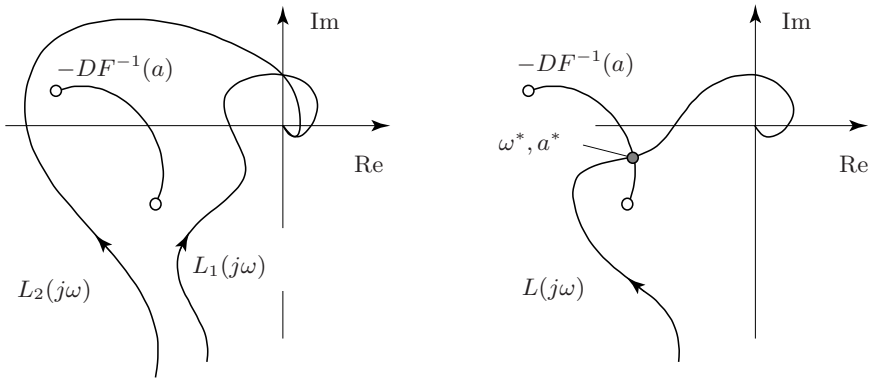



**Fig. 5.11.** Left plot: $L_1(s)$ yields an asymptotically stable closed-loop system, while $L_2(s)$ yields an unstable closed-loop system. Right plot: $L(s)$ yields a limit cycle with amplitude $a^*$ and frequency $\omega^*$.

Even if a limit cycle can appear, it is not clear whether this periodic solution is stable, i.e., whether the output $y(t)$ remains close to a periodic orbit even when small disturbances act on the closed-loop system. Figure 5.12 illustrates how the stability of the limit cycle can be analyzed graphically.

The curves in the left plot indicate that the limit cycle is stable. This can be seen using the following "Gedankenexperiment:"

- Assume that the system is on a limit-cycle with $\omega = \omega^*$ and $a = a^*$ (grey dot in Figure 5.12, left).

- At $t = t_0$ a disturbance acts briefly on the system such that the amplitude of the oscillation increases to $a^+ > a^*$ (upper black dot in Figure 5.12, left).
- Since the curve $L(j\omega)$ does not encircle $DF^{-1}(a^+)$, the closed-loop system is stable and the amplitude of the oscillation decreases until it again reaches the value $a = a^*$.

If the disturbance reduces the amplitude to $a^- < a^*$, then the loop gain $L(j\omega)$ encircles the critical point (lower black dot in Figure 5.12, left). In this case, the closed-loop system is unstable and the amplitude of the oscillation increases until it again reaches the value $a^*$.

The opposite situation is encountered in the right plot in Figure 5.12. In this case, a disturbance increasing the amplitude yields an unstable situation such that the amplitude further increases and vice versa. Accordingly, the closed-loop system will never exhibit a limit cycle characterized by the frequency $\omega^*$ and the amplitude $a^*$ because arbitrarily small disturbances will lead to ever-increasing or decreasing amplitudes.

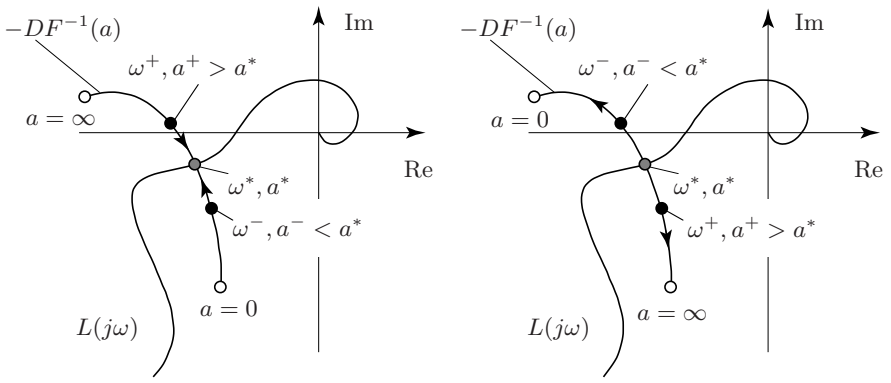



**Fig. 5.12.** Describing function and loop gain combinations indicating a stable (left plot) and unstable (right plot) limit cycle.

The subsequent section will show how the function $DF(a)$ can be computed and illustrate the describing function method with an example. More details on the theory of limit cycles and several examples of describing functions can be found in [10].

**Example: Two-Point Control System**

A two-point controller is probably the cheapest control system used in practice. Such "on-off controllers" are found in residential heating systems ("thermostats"), in cheap electric power converters and in many other applications. The control system consists of a plant $P(s)$ and a controller with the characteristics illustrated in Figure 5.13. This controller acts as a relay with hysteresis, i.e., if the error $e(t_0)$ is smaller than $e^-$ and increases monotonously with $t > t_0$, the output of the controller switches from $u^-$ to $u^+$ when $e(t) > e^+$ and vice versa. For the sake of simplicity, in this example it is assumed that $e^+ = u^+ = +1$, $e^- = u^- = -1$, that the plant $P(s)$ is a low-pass system, and that the loop gain is defined by

$$L(s) = \frac{2}{s \cdot (s+1)^2} \tag{5.51}$$

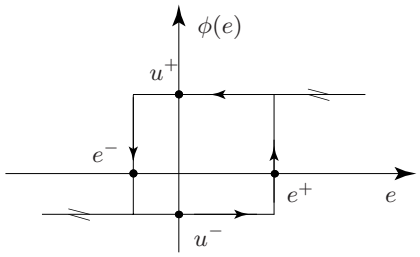



**Fig. 5.13.** Two-point controller as a relay with hysteresis.

The explicit calculation of the describing function uses Fourier theory. In particular, in equation (5.46) the variables $k_1(a)$ and $\varphi_1(a)$ are defined by

$$\begin{aligned} k_1(a) \cdot \sin(\varphi_1(a)) &= \alpha(a) = \tfrac{1}{\pi} \textstyle\int_0^{2\,\pi} \phi(a \cdot \sin(x)) \cdot \cos(x)\, dx \\ k_1(a) \cdot \cos(\varphi_1(a)) &= \beta(a) = \tfrac{1}{\pi} \textstyle\int_0^{2\,\pi} \phi(a \cdot \sin(x)) \cdot \sin(x)\, dx \end{aligned} \tag{5.52}$$

Once $\alpha(a)$ and $\beta(a)$ are known, the describing function is easily found by

$$DF(a) = \frac{\beta(a) + j\,\alpha(a)}{a} \tag{5.53}$$

As shown in Figure 5.14, the variable $\beta(a)$ in (5.52) can be found using the simpler expression

$$\beta(a) = \frac{2}{\pi} \int_{x_1}^{\pi + x_1} \sin(x) dx = \frac{2}{\pi} \left[ -\cos(x) \right]_{x_1}^{\pi + x_1} \tag{5.54}$$

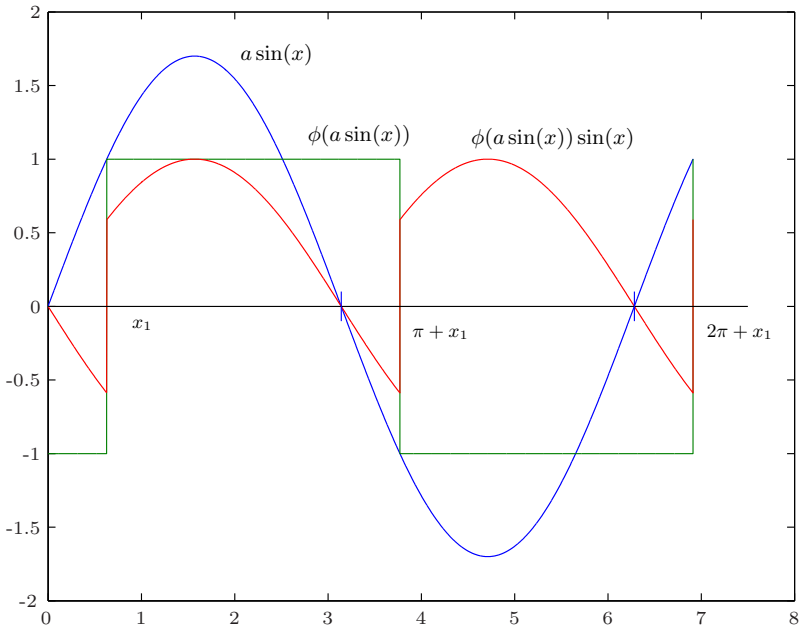



**Fig. 5.14.** Signals relevant for the computation of the variable $\beta(a)$.

The point $x_1$, at which the two-point controller switches its output from $-1$ to $+1$, is defined by

$$a \cdot \sin(x_1) = 1 \quad \Rightarrow \quad x_1 = \arcsin(\frac{1}{a}) \tag{5.55}$$

Of course, $x_1$ is defined only if $a > 1$. In this example, this is a *necessary* condition for the emergence of limit cycles. Equation (5.54) now reads

$$\beta(a) = \frac{2}{\pi}\left[\cos(\arcsin(\frac{1}{a})) - \cos(\pi + \arcsin(\frac{1}{a}))\right] \tag{5.56}$$

Using the well-known relations

$$\begin{aligned}\cos(\pi + \arcsin(\tfrac{1}{a})) &= \cos(\pi)\cos(\arcsin(\tfrac{1}{a})) - \sin(\pi)\sin(\arcsin(\tfrac{1}{a})) \\ &= -\cos(\arcsin(\tfrac{1}{a}))\end{aligned}$$

and

$$\sin^2(x) + \cos^2(x) = 1 \quad \Rightarrow \quad \cos(x) = \sqrt{1 - \sin^2(x)}$$

the variable $\beta(a)$ is found to be

$$\beta(a) = \frac{4}{\pi}\sqrt{1 - \frac{1}{a^2}} = \frac{4}{\pi \cdot a}\sqrt{a^2 - 1} \tag{5.57}$$

The variable $\alpha(a)$ can be computed using very similar intermediate steps. The result is

$$\alpha(a) = -\frac{4}{\pi \cdot a} \tag{5.58}$$

Therefore, the describing function of the two-point controller is defined by

$$DF(a) = \frac{4}{\pi \cdot a^2}(\sqrt{a^2 - 1} - j) \tag{5.59}$$

It is not difficult to see that

$$-DF^{-1}(a) = -\frac{\pi}{4} \cdot (\sqrt{a^2 - 1} + j) \tag{5.60}$$

Figure 5.15 shows the Nyquist plot of the loop gain (5.51) and the Nyquist plot of the function (5.60). In this example, the curves $L(j\,\omega)$ and $-DF^{-1}(a)$ intersect. Therefore, a limit cycle is likely to appear. Moreover, this limit cycle will be stable, i.e., persist even when disturbances act on the system (increasing values of the amplitude $a$ yield points on $-DF^{-1}(a)$ on the left side of $L(j\,\omega)$). The expected amplitude $a^*$ and frequency $\omega^*$ of the steady-state signal $e_\infty(t) = a^* \sin(\omega^* t)$ are listed in this figure as well.

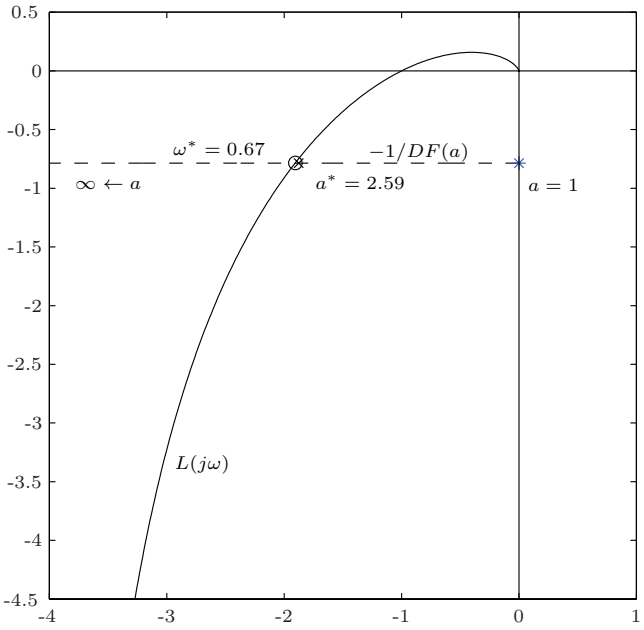



**Fig. 5.15.** Nyquist plots of the loop gain $L(j\omega)$ (5.51) and of the describing function (5.60).

As mentioned, the describing function method is based on approximations of the system behavior. It is therefore not surprising that the signal $e(t)$ does not converge exactly to $e_\infty(t)$. In fact, as Figure 5.16 shows, $e(t)$ converges after a few transient oscillations to a harmonic signal, but the amplitude and period of that signal are slightly larger than the expected values.

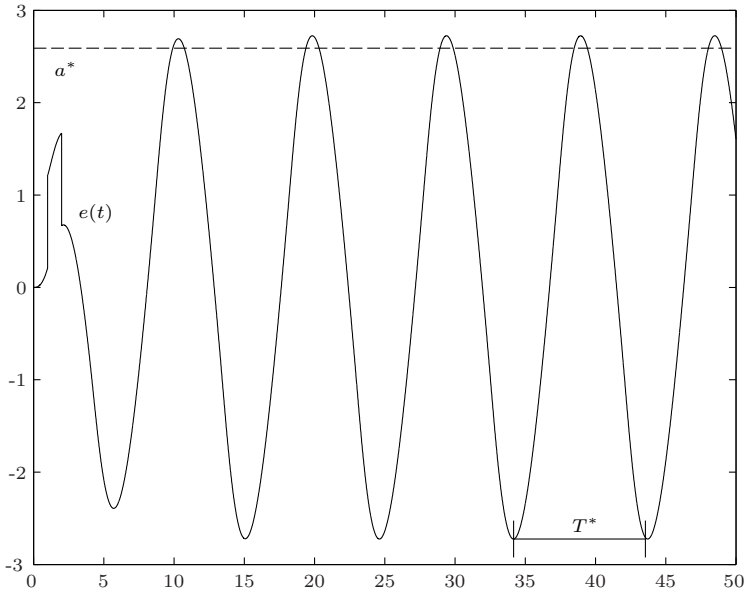



**Fig. 5.16.** Signal $e(t)$ (see Figure 5.10); amplitude $a^*$ and period $T^* = 2\pi/\omega^*$.

## 5.3 Some Notions of Chaos Theory

### 5.3.1 Introduction

This section introduces some elementary notions of chaos theory. Two examples of chaotic behavior are analyzed in more detail (the Rössler system and the "logistics equation").

The adjective "chaotic" is often used to describe a complex and not easily understandable behavior of a dynamic system. It is, however, not trivial to precisely define what a "chaotic system" is, and it is even more difficult to decide whether a given system actually is chaotic or not.

This section provides an introduction to the most elementary ideas of chaos theory and analyzes in more detail two of the most simple examples of chaotic systems. Readers interested in this field may find much more information in the references given in [2].

Some of the key ideas relevant to chaos theory are:

- period doubling;
- self similarity;
- sensitivity to initial conditions; and
- strange attractors.

All of these concepts and objects will be defined in more detail in the two examples analyzed below. Before that, some general remarks will be made.

### 5.3.2 Poincaré-Bendixson Theorem

The Poincaré-Bendixson theorem permits one to draw conclusions that are valid for general second-order continuous-time systems. Such systems are described by two coupled first-order differential equations

$$\begin{aligned} \tfrac{d}{dt}x_1(t) &= f_1(x_1(t), x_2(t)) \\ \tfrac{d}{dt}x_2(t) &= f_2(x_1(t), x_2(t)) \end{aligned} \tag{5.61}$$

where the functions $f_i(.)$ are assumed to be sufficiently smooth such that the existence and the uniqueness of the solutions of (5.61) are guaranteed.

Many results are available for such systems. The analysis of the qualitative behavior of their solutions is greatly simplified because tools of planar geometry may be applied to its phase plane (see Section 5.2.2). One of the most important results is the Poincaré-Bendixson Theorem. Two definitions are needed before this theorem can be formulated.

**Limit Set:** A point $x_\infty \in \mathbb{R}^n$ is called a limit point if there is a solution of the system

$$\frac{d}{dt}x(t) = f(x(t)), \quad x(0) \neq 0$$

which passes infinitely many times arbitrarily close to $x_\infty$. In the limit, the sequence defined by the approach times $t_i$ and the minimum-distance approach points $x_i$ will converge to $x_\infty$, i.e., $\lim_{i\to\infty} x_i = x_\infty$. The limit set of $x_0$ is the set of all limit points of the solution that starts at $x(0) = x_0$.

Notice that equilibrium points and periodic trajectories form limit sets (the "limit cycles" discussed in the last section are a typical example), but that there are limit sets that are more complex than those two simple cases. These limit sets are referred to as "strange attractors" and their most important properties will be introduced in the next subsection).

**Bounded and Closed Regions:** A subset $\Omega$ of $\mathbb{R}^n$ is a bounded and closed region if $\Omega$ is finite and if the boundary $\partial\Omega$ of $\Omega$ also is part of $\Omega$. For instance, the set $\Omega_1 = \{x \in \mathbb{R}^2 \mid |x_1| \leq 1\}$ is not bounded and the set $\Omega_2 = \{x \in \mathbb{R}^2 \mid |x| < 1\}$ is not closed.

**Poincaré-Bendixson Theorem**

If $L$ is a limit set of the system (5.61) that is completely contained in a closed and bounded region $\Omega$, then $L$ is either an equilibrium point or a periodic solution of the system (5.61).

The proof of this theorem is not easy, and interested readers are referred to [6]. What the Poincaré-Bendixson Theorem really says is that "in time-invariant sufficiently smooth two-dimensional continuous-time systems chaos is not possible." Chaos is only possible in continuous-time systems of order three and higher, and – surprisingly – in discrete-time systems of any dimension (one-dimensional discrete-time systems will be shown to exhibit surprisingly complex behavior, see Section 5.3.4).

### 5.3.3 Continuous-Time Systems of Orders Three and Higher

Linear systems of any order cannot produce chaotic behavior because all trajectories are either globally asymptotically stable, periodic or unstable. Chaos is therefore implicitly defined as being more complex than any of these classes. Such complex behavior may be found only in nonlinear systems. As shown in the last section, in the continuous-time case, the systems must be of orders three or higher.

One of the simplest known continuous-time systems with chaotic behavior is the Rössler system defined by

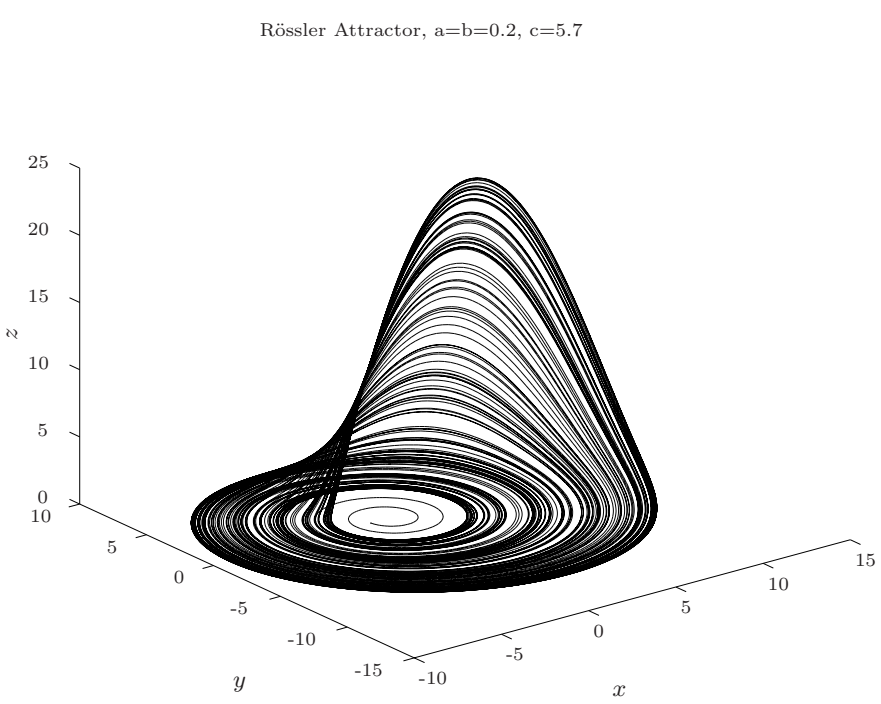



**Fig. 5.17.** Strange attractor of the Rössler system.

$$\begin{aligned} \tfrac{d}{dt}x(t) &= -y(t) - z(t) \\ \tfrac{d}{dt}y(t) &= x(t) + a \cdot y(t) \\ \tfrac{d}{dt}z(t) &= b + x(t) \cdot z(t) - c \cdot z(t) \end{aligned} \tag{5.62}$$

Parameter values leading to a chaotic behavior are, for instance, $\{a, b, c\} = \{0.2, 0.2, 5.7\}$. This chaotic behavior is characterized by the typical signs, especially the sensitivity to initial conditions.[10] As Figure 5.18 shows, arbitrarily small differences in the initial conditions are amplified even if the system does not diverge.

The most important new limit sets encountered in systems similar to (5.62) are "strange attractors." In fact, as shown in Figure 5.17, the Rössler system has a limit set that is neither an equilibrium point nor a periodic solution. In other words, the system trajectories do not diverge to infinity, but do not reach equilibria or periodic solutions either.

[10] The better-known Lorenz system, which is slightly more complex, is associated with weather forecast models. The sensitivity to initial conditions is paraphrased for this system with the famous metaphor that "The beat of a butterfly wing over the Azores can lead to a thunderstorm in Zurich."

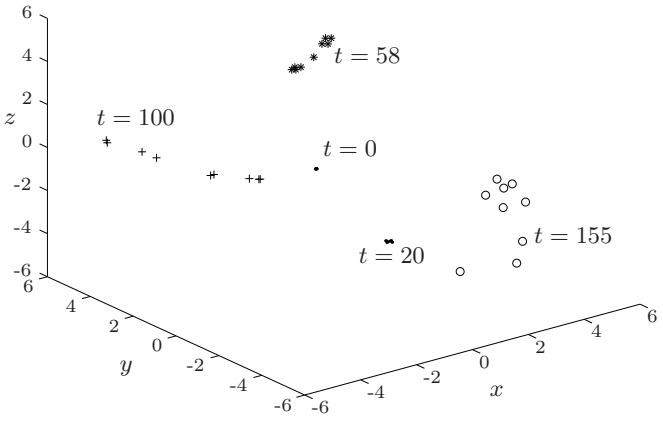



**Fig. 5.18.** Sensitivity to initial conditions: $x(0) \in \{(-1, 0, 0) + B_{r=0.01}\}$.

### 5.3.4 The Logistics Equation

The system analyzed in this section is defined by

$$x_{k+1} = f(x_k) = \mu \cdot x_k \cdot (1 - x_k) \tag{5.63}$$

and it is assumed that $x_k$ and $\mu$ are both real scalars. The parameter $\mu$ is assumed to be in the interval $\mu \in [1, 4]$. In this case, for the state variable $x_k$ the interval $x_k \in [0, 1]$ is invariant.[11]

The system (5.63) obviously has two equilibria

$$x_{0,1} = 0 \qquad \text{and} \qquad x_{0,2} = 1 - 1/\mu \tag{5.64}$$

which are both obtained by solving the equation

$$x_0 = \mu \cdot x_0 \cdot (1 - x_0) \tag{5.65}$$

Therefore, for the chosen parameter range, the nontrivial equilibrium will be in the interval $x_0 \in [0, 0.75]$. Figure 5.19 shows a graphical illustration of these observations for the choice $\mu = 2.5$.

It is well-known that for a first-order discrete-time system $x_{k+1} = f(x_k)$, an equilibrium point $x_0$ is asymptotically stable if the gradient of $f$ at the equilibrium has a magnitude smaller than 1 ("Lyapunov's Principle for discrete-time systems").

[11] The interval $[0, 1]$ is mapped by (5.63) onto the interval $[0, \mu/4]$.

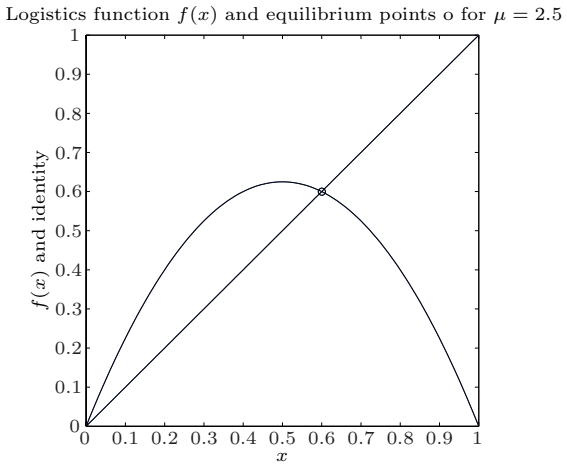



**Fig. 5.19.** Functions $f(x)$ as defined in (5.63) and identity. The intersections satisfy equation (5.65) and indicate the equilibrium points

Since the derivatives of (5.63) at the equilibriums (5.64) are given by

$$\left.\frac{d}{dx}f(x)\right|_{x=x_0} = \begin{cases} \mu & \text{for } x_0 = 0 \\ 2-\mu & \text{for } x_0 = 1 - 1/\mu \end{cases} \tag{5.66}$$

Therefore, the first equilibrium point $x_0 = 0$ is unstable for all $\mu \in (1, 4]$, while the second is asymptotically stable for all $\mu \in (1, 3)$. Figure 5.20 shows an example of a stable nontrivial equilibrium point ($\mu = 2.5$) and Figure 5.21 shows an example of an unstable nontrivial equilibrium point ($\mu = 3.05$).

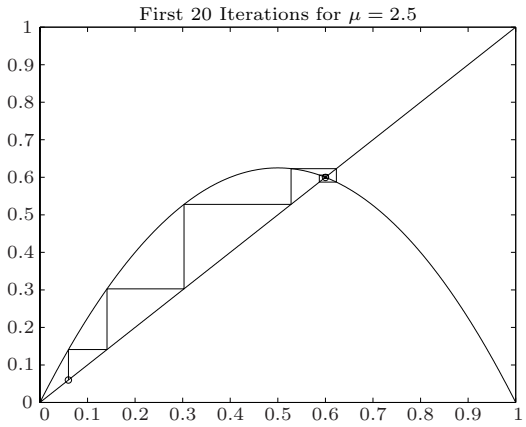



**Fig. 5.20.** First 20 iterations of the function (5.63), with $x_0 = 0.05$ and $\mu = 2.5$.

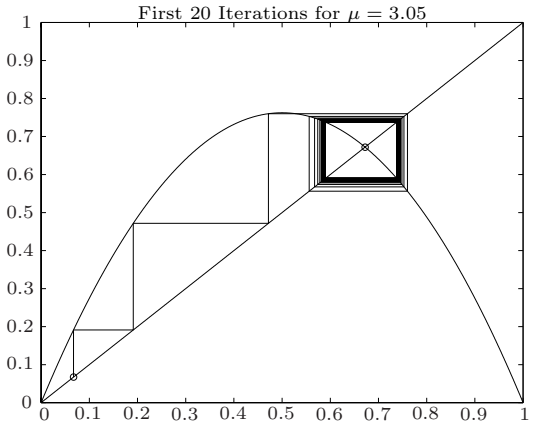



**Fig. 5.21.** First 20 iterations of the function (5.63), with $x_0 = 0.05$ and $\mu = 3.05$.

As Figure 5.21 shows, a periodic orbit emerges around the unstable equilibrium when $\mu$ is chosen slightly larger than the stability limit $\mu_1 = 3$, i.e., for $k \to \infty$, the equation

$$x_{k+2} = x_k \tag{5.67}$$

is satisfied in this case. This periodic orbit is characterized by the fact that the two-fold application of the mapping (5.63)

$$f(f(x)) = \mu^2 \cdot x \cdot (1 - x) \cdot (1 - \mu \cdot x \cdot (1 - x)) \tag{5.68}$$

produces an asymptotically stable behavior for some of its equilibrium points. It is now easy to show (see Lemma 1 below) that two of the four equilibriums of (5.68) are given by the expression (5.65).

**Lemma 1:** For any given mapping $f : \mathbb{R} \to \mathbb{R}$, denote the $n$-fold repeated iteration

$$f(f(\ldots f(f(x))\ldots)) \tag{5.69}$$

by

$$f_n : \mathbb{R} \to \mathbb{R} \tag{5.70}$$

where $n$ is a cardinal number indicating the number of repeated applications of $f$.

Then the set of all equilibria of the mapping $f_{2m}$ includes the equilibria of the mapping $f_m$ for all $m = 1, 2, \ldots$

**Proof:** Assume that $x_0$ is an equilibrium of $f_m$, i.e., it satisfies the condition

$$x_0 = f_m(x_0) \tag{5.71}$$

Then

$$\widetilde{x}_0 = f_{2m}(x_0) = f_m(f_m(x_0)) = f_m(x_0) = x_0 \tag{5.72}$$

which proves that $x_0$ is an equilibrium of $f_{2m}$ as well.

**Lemma 2:** If the gradient of a mapping $f_m$ at some equilibrium $x_0$ has a magnitude larger than one, then the gradient of the mapping $f_{2m}$ at the same equilibrium $x_0$ will also be larger than one.

**Proof:** According to the chain rule the gradient of $f_{2m}$ is given by

$$\frac{d}{dx} f_{2m} = \frac{d}{dx} f_m \cdot \frac{d}{dx} f_m \tag{5.73}$$

From this, the assertion of this Lemma follows immediately.

According to the result of Lemma 1, two of the four equilibria of the double iteration (5.68) are given by (5.64). The other two equilibrium points may be found by solving the second-order polynomial equation

$$[\mu^2 \cdot x \cdot (1-x) \cdot (1-\mu \cdot x \cdot (1-x)) - x]/[(x-0) \cdot (x-(1-1/\mu))] = 0 \tag{5.74}$$

The two solutions of this equation are given by

$$\begin{aligned} x_{0,3} &= \left(1 + 1/\mu + \sqrt{1 - 2/\mu - 3/\mu^2)}\right)/2 \\ x_{0,4} &= \left(1 + 1/\mu - \sqrt{1 - 2/\mu - 3/\mu^2)}\right)/2 \end{aligned} \tag{5.75}$$

According to the result of Lemma 2, it is clear that the two equilibria (5.64) of $f_1 = f$ – which are unstable for $\mu > 3$ – also will be unstable for $f_2$. The other two equilibria (5.75) may be stable or unstable, depending on the value of $\mu$. To analyze this in more detail, the derivative of $f_2$ is formed

$$\frac{d}{dx} f_2(x) = \mu^2 \cdot \left[1 - 2 \cdot (1+\mu) \cdot x + 6 \cdot \mu \cdot x^2 - 4 \cdot \mu \cdot x^3\right] \tag{5.76}$$

Unfortunately, inserting $x_{0,3}$ or $x_{0,4}$ into this expression yields a polynomial equation for $\mu$ of order larger than five and a closed-form solution is no longer possible. Numerical procedures must now be followed.

The result of such an analysis will be that for $\mu < \mu_2$ the gradient (5.3.4) at the two equilibrium points $x_{0,3}$ and $x_{0,4}$ will have a magnitude less than 1 where

$$\mu_2 \approx 3.4494897428\ldots \tag{5.77}$$

In other words, for all parameters $\mu \in (3, 3.4494897428\ldots)$ the periodic orbit defined by $f_2$ of (5.68) will be stable, i.e., the state $x_k$ will oscillate between the two values defined by (5.75).

For parameters $\mu > \mu_2$, another periodic orbit can be observed where the state variable follows a four-point orbit (see Figure 5.22). This "periodic

frequency doubling" phenomenon repeats itself in increasingly smaller increments of $\mu$. Using numerical tools, it can be determined that this series is given by

$$\begin{aligned}
\mu_0 &= 1 \\
\mu_1 &= 3 \\
\mu_2 &\approx 3.449489\ldots \\
\mu_3 &\approx 3.544090\ldots \\
\mu_4 &\approx 3.564407\ldots \\
\mu_5 &\approx 3.568759\ldots \\
\ldots &\approx \ldots
\end{aligned}$$

and reaches its limit at $\mu_\infty \approx 3.5699456\ldots$ where the "periodic" orbit reaches a period equal to infinity.

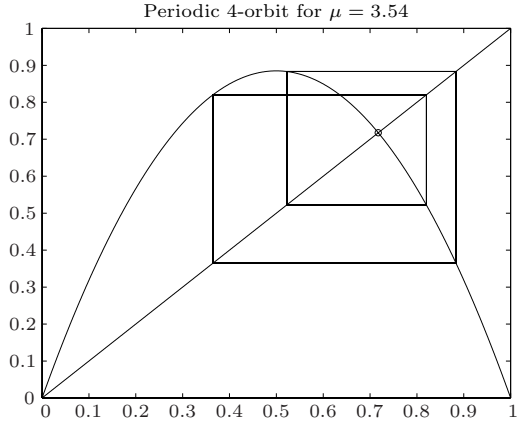



**Fig. 5.22.** Periodic four-point orbit for $\mu = 3.54$

For parameter values $\mu > \mu_\infty$, there are periodic orbits with infinite period. Therefore, an extremely complex system behavior emerges. Interestingly, there are parameter values $\mu_\infty < \mu < 4$ for which simple periodic orbits can be observed again. This is shown in Figure 5.23 where the limit set for different values of $\mu$ is shown (for instance, for $\mu = 3.626$, a periodic orbit with a periodicity of 6 arises).

Another typical feature of such "bifurcation diagrams" is shown in Figure 5.24, i.e., the bifurcation behavior exhibits a clear self-similarity in some parts of the parameter interval. Self-similarity is a phenomenon that is often encountered in fractal geometry, which many people think is better suited to

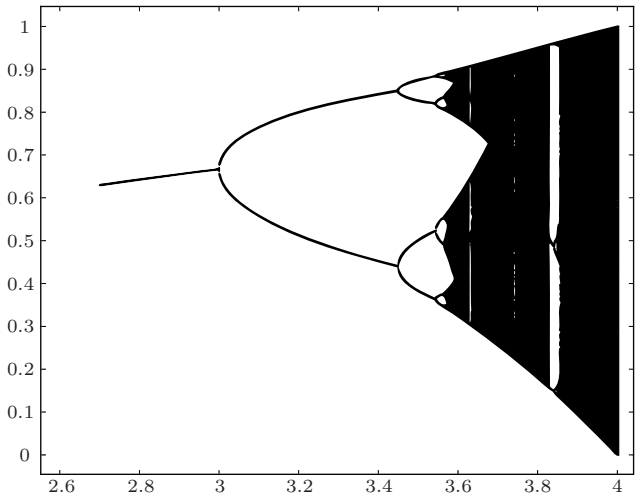



**Fig. 5.23.** Complete bifurcation diagram.

describe some shapes found in nature. This, and similar observations, have made many believe that chaos theory and fractal geometry may explain otherwise incomprehensible phenomena.

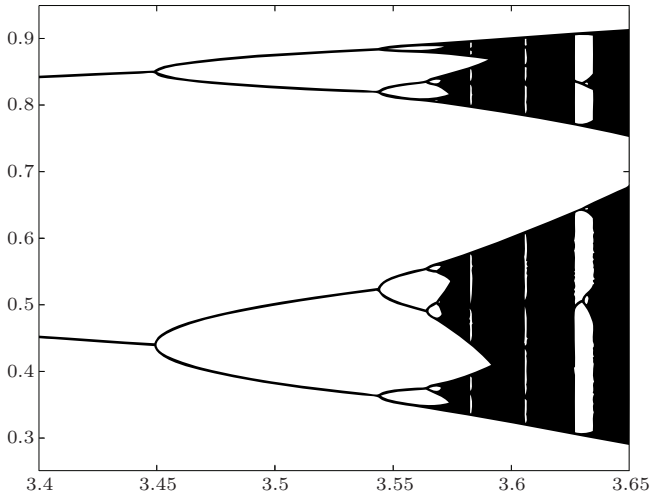



**Fig. 5.24.** One detailed part of the bifurcation diagram.

# 6

# Appendix I: Parameter Optimization

## 6.1 Problems without Constraints

In order to better understand the remarks on numerical optimization algorithms made in the next section, the most important facts of the theory of closed-form parameter optimization are repeated here.

Let $\pi = [\pi_1, \ldots, \pi_m]^T \in I\!R^m$ be a vector of arbitrary parameters and $L : I\!R^m \to I\!R_+$ a sufficiently differentiable function (the "performance index") that must be minimized.[1] Sufficient conditions for a point $\pi^o$ to be a local minimum are

$$\left.\frac{\partial L(\pi)}{\partial \pi}\right|_{\pi=\pi^o} = 0, \quad \text{and} \quad \left.\frac{\partial^2 L(\pi)}{\partial \pi^2}\right|_{\pi=\pi^o} > 0 \tag{6.1}$$

i.e., the gradient of the performance index must vanish at the minimum ($\pi^o$ is a stationary point), and the Hessian matrix of the performance index must be positive definite (in the neighborhood of $\pi^o$, $L(\cdot)$ increases everywhere). The condition 6.1 is globally sufficient only for specific cases, for instance, if it is known that the function $L(\cdot)$ is globally convex.

Necessary, but not sufficient, conditions for a local minimum are

$$\left.\frac{\partial L(\pi)}{\partial \pi}\right|_{\pi=\pi^o} = 0, \quad \text{and} \quad \left.\frac{\partial^2 L(\pi)}{\partial \pi^2}\right|_{\pi=\pi^o} \geq 0 \tag{6.2}$$

To establish whether a minimum exists, additional information is needed in this case.

[1] A maximization problem can be obtained from a minimization problem by simply multiplying the performance index by $-1$

*Example 6.1 (Quadratic Form).*

Performance index

$$L(\pi) = \frac{1}{2}\pi^T M \pi, \quad \pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 1 \\ 1 & \mu \end{bmatrix}, \quad \mu \in I\!R$$

Minimization

$$\frac{\partial L}{\partial \pi} = M\pi, \quad \frac{\partial^2 L}{\partial \pi^2} = M$$

i.e., if $M$ is non-singular ($\mu \neq 1$), then only one possible minimum $\pi^o = [0,0]^T$ exists. For $\mu = 1$, all the points on the line $\lambda \cdot [1,-1]^T$, $\lambda \in I\!R$ are minima. For $\mu > 1$, $M$ is positive definite and $\pi^o$ is a (global) minimum. For $\mu < 1$, $M$ is indefinite and $\pi^o$ is not a minimum but a saddle point (see Figure 6.1).

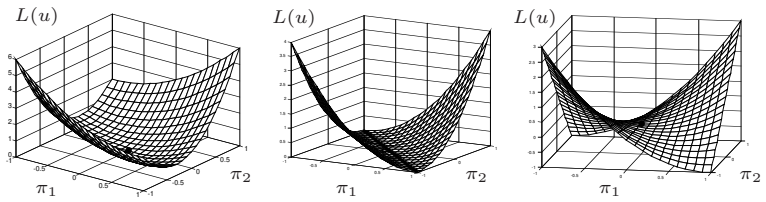



**Fig. 6.1.** $L(u)$ of Example 1, left $\mu > 1$, middle $\mu = 1$, right $\mu < 1$

*Example 6.2 (Critical Case).*

Performance index

$$L(\pi) = (\pi_1 - \pi_2^2)(\pi_1 - 3\pi_2^2) \tag{6.3}$$

Minimization

$$\frac{\partial L}{\partial \pi} = \begin{bmatrix} 2\pi_1 - 4\pi_2^2 \\ -8\pi_1\pi_2 + 12\pi_2^3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \pi^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \left.\frac{\partial^2 L}{\partial \pi^2}\right|_{\pi=\pi^0} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

i.e.,$L_\pi(\pi) = 0$ has a triple but isolated solution $\pi_1 = \pi_2 = 0$, but this solution is *not* a local minimum, as it is shown in Figure 6.2.[2] This is a typical case where the linearization is not sufficient to decide wether a stationary point is a minimum or not.

[2] The expression $L_\pi$ denotes the partial derivative of $L$ with respect to $\pi$.

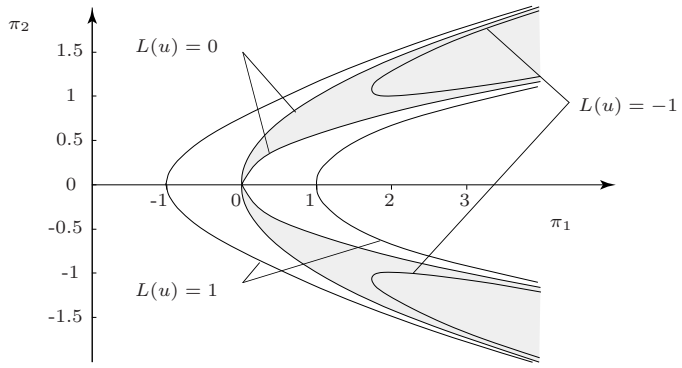



**Fig. 6.2.** Contour plot of Example (6.3)

## 6.2 Minimization with Equality Constraints

Let $\pi = [\pi_1, \ldots, \pi_m]^T \in I\!R^m$ be an $m$-dimensional vector of "control variables," and $x = [x_1, \ldots, x_n]^T \in I\!R^n$ an $n$-dimensional vector of "state variables"[3]. The functions $L : I\!R^{m+n} \to I\!R_+$ and $f : I\!R^{m+n} \to I\!R^n$ are sufficiently differentiable functions of these quantities. The optimization problem consists in finding those $\pi^o, x^o$ which minimize the performance index $L$ and, at the same time, satisfy the constraint $f(\pi^o, x^o) = 0$. With the adoption of a new variable $z = [\pi, x]^T$, the following expressions can be written in a more compact way.

In the case where $n = 1$, the "solution" of this problem can be immediately given with the help of Figure 6.3. An optimal point $z^o$ is a point where the gradient on an "iso-level curve" of the performance index and the gradient on the subset defined by the constraint are co-linear, i.e.,

$$\left.\frac{\partial L}{\partial z}\right|_{z=z^o} + \lambda \left.\frac{\partial f}{\partial z}\right|_{z=z^o} = 0 \tag{6.4}$$

where $\lambda$ is a new free scalar parameter. These $m+1$ equations, together with the constraint $f(z^o) = 0$, can be used to find the $m+2$ unknown quantities $z^o = [\pi^o, x^o]$ and $\lambda^o$.

[3] The distinction between parameter and state variables is often a matter of convenience.

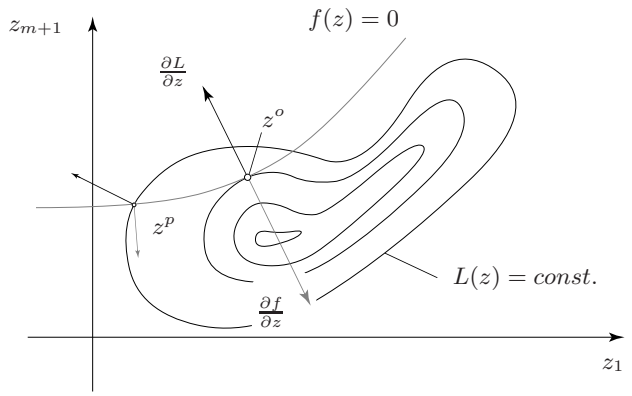



**Fig. 6.3.** Constrained optimization problem with one constraint

In the general case, the solution of the minimization problem satisfies the requirement that in the optimal point, any *arbitrary* variation *dz that satisfies the constraints*, must not cause any variation of the performance index, i.e., for

$$df(\pi, x) = f_\pi d\pi + f_x dx = 0 \tag{6.5}$$

it has to be

$$dL(\pi, x) = L_\pi d\pi + L_x dx = 0 \tag{6.6}$$

Therefore, based on equation (6.5), the variation $dx$ depends on the variation $d\pi$ through (6.7)

$$dx = -f_x^{-1} f_\pi d\pi \tag{6.7}$$

(the $n \times n$ matrix $f_x$ must be nonsingular in the optimal point, otherwise the problem is not well posed). Inserting (6.7) into (6.6) yields

$$dL(\pi, x) = [L_\pi - L_x f_x^{-1} f_\pi] d\pi = 0 \tag{6.8}$$

Since this has to be valid for any arbitrary variation $d\pi$, the following necessary condition results

$$L_\pi - L_x f_x^{-1} f_\pi = 0 \tag{6.9}$$

Sufficient conditions for a local minimum must take into account any second-order variations. Expanding the performance index in a Taylor series around the optimal point $x^o, \pi^o$, up to the second-order variations yields

$$\begin{aligned} dL \approx & [L_x(x^o, \pi^o), L_\pi(x^o, \pi^o)] \begin{bmatrix} dx \\ d\pi \end{bmatrix} \\ & + \tfrac{1}{2} [dx^T, d\pi^T] \begin{bmatrix} L_{xx}(x^o, \pi^o) & L_{x\pi}(x^o, \pi^o) \\ L_{\pi x}(x^o, \pi^o) & L_{\pi\pi}(x^o, \pi^o) \end{bmatrix} \begin{bmatrix} dx \\ d\pi \end{bmatrix} \end{aligned} \tag{6.10}$$

The linear term vanishes for (6.6). Moreover, $dx$ cannot be arbitrarily selected, but it must be related to $d\pi$ such that the constraint $f(x+dx, \pi+d\pi)=0$ is satisfied, i.e., with (6.7). Equation (6.10) then becomes

$$dL \approx \frac{1}{2} d\pi^T [-f_\pi^T (f_x^{-1})^T, I] \begin{bmatrix} L_{xx}(x^o,\pi^o) & L_{x\pi}(x^o,\pi^o) \\ L_{\pi x}(x^o,\pi^o) & L_{\pi\pi}(x^o,\pi^o) \end{bmatrix} \begin{bmatrix} -f_x^{-1} f_\pi \\ I \end{bmatrix} d\pi \tag{6.11}$$

This equation must be satisfied for an arbitrary $d\pi$.

The sufficient condition for the point $x^o, \pi^o$ being optimal, therefore, is

$$\left.\frac{\partial^2 L}{\partial \pi^2}\right|_{x^o,\pi^o,f(x,\pi)=0} > 0 \tag{6.12}$$

where

$$\left.\frac{\partial^2 L}{\partial \pi^2}\right|_{\dots} = (L_{\pi\pi} - f_\pi^T (f_x^{-1})^T L_{x\pi} - L_{\pi x} f_x^{-1} f_\pi + f_\pi^T (f_x^{-1}) L_{xx} f_x^{-1} f_\pi)\big|_{x^o,\pi^o} \tag{6.13}$$

That is to say, the Hessian matrix of the performance index in the point $x^o, \pi^o$, and for variations that satisfy the constraints, must be positive definite. If the Hessian matrix defined in equation (6.12) is only semidefinite, $x^o, \pi^o$ may be a minimum, but this must be verified using higher-order terms.

## 6.3 Numerical Algorithms

A closed-form solution is possible only in very simple cases. In most cases, one must rely on numerical methods. Generally, gradient-based or other methods can be applied. The latter approaches (genetic algorithms, Monte-Carlo methods, etc.) are not discussed here. Interested readers are referred to the appropriate courses in the INFK and ITET departments of ETH and to [12].

Two different gradient-based approaches are possible:

- "semi-analytical" methods where the performance index $L(\cdot)$ and its gradient are available in a closed form; and
- "fully numerical" methods where only the performance index is known, and its gradient must be approximated numerically using finite differences.

The semi-analytical methods generally converge much faster than the fully numerical methods, and the former are less sensitive to rounding effects.

The numerical search can be either of first- or second-order. The first-order numerical methods use the idea of the "steepest descent" and all first-order algorithms have approximately the following structure:

1. Guess an initial value for $\pi(1)$ (the argument denotes the iteration index).
2. Evaluate the gradient $\frac{\partial L}{\partial \pi}\Big|_{\pi=\pi(i)}$ using either semi-analytic or numerical methods.
3. Determine the new iteration point according to the rule

$$\pi(i+1) = \pi(i) - h(i)\left.\frac{\partial L}{\partial \pi}\right|_{\pi=\pi(i)}$$

4. Check whether the difference $|L(\pi(i+1)) - L(\pi(i))|$ is smaller than a predetermined threshold $\epsilon$. In this case, the algorithm ends, otherwise it is repeated starting at point 2.

Crucial for the convergence is the choice of the relaxation factor $h(i)$. If it is chosen too large (see Figure 6.4), the algorithm may overshoot and even become unstable. One possibility of choosing $h(i)$ in an optimal way consists of solving a one-dimensional minimization problem with constraints along the gradient, i.e., $\pi(i+1)$ is the point that generates the smallest performance index and at the same time lies in the direction of the gradient.

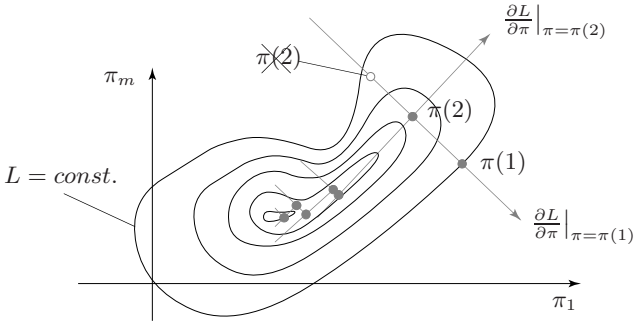



**Fig. 6.4.** Sequence of numerical first-order optimization procedure.

Figure 6.4 also shows the problem that arises with narrow "valleys:" a slow convergence is, in this case, unavoidable. The choice of $\pi(1)$ is very important as well. If, on the basis of simplifying considerations, a guess of the optimal point is possible, this additional effort is often worthwhile. Moreover, during the optimization, monitoring the parameter changes is necessary in order to keep the parameters from drifting to unrealistic solutions.

Besides the first-order methods (which use only the first-order derivatives), second-order methods are known as well. They often converge faster than first-order methods in the neighborhood of the optimum, but are inefficient far away from it. The basic idea is to approximate the performance index through a quadratic form, i.e.,

$$L(\pi) \approx L(\pi(i)) + \left.\frac{\partial L}{\partial \pi}\right|_{\pi=\pi(i)} (\pi - \pi(i)) + \frac{1}{2}(\pi - \pi(i))^T \left.\frac{\partial^2 L}{\partial \pi^2}\right|_{\pi=\pi(i)} (\pi - \pi(i)) \tag{6.14}$$

and then to use the solution of the approximate problem (which is solvable in closed form) as a new iteration value, i.e.,

$$\pi(i+1) = \pi(i) - \left[\left.\frac{\partial^2 L}{\partial \pi^2}\right|_{\pi=\pi(i)}\right]^{-1} \left[\left.\frac{\partial L}{\partial \pi}\right|_{\pi=\pi(i)}\right]^T \tag{6.15}$$

Optimization algorithms are available in program libraries, for instance, in the Matlab™ optimization toolbox. Using this software is recommended rather than attempt to "re-invent the wheel."

# References


1. Bryson A. E. and Y. C. Ho, *Applied Optimal Control*, Halsted Press, 1975.
2. Devaney R. L., *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley, 1989.
3. Gantmacher F. R., *Theory of Matrices*, Springer Verlag, 1970.
4. Glad T. and L. Ljung, *Control Theory*, Taylor & Francis, 2000.
5. Golub G. H. and C. F. Van Loan *Matrix Computations*, Johns Hopkins University Press, 1989.
6. Hahn, W., *Stability of Motion*, Springer Verlag, Berlin, 1967.
7. Hughes P. C., *Spacecraft Attitude Dynamics*, Wiley, 1986.
8. Isidori A., *Nonlinear Control Systems*, 2nd. ed., Springer Verlag, New York, 1989.
9. Kailath T., *Linear Systems*, Prentice Hall, 1980.
10. Khalil H. K., *Nonlinear Systems*, MacMillan, 1996.
11. Ljung L. and T. Glad, *Modeling of Dynamic Systems*, Prentice Hall, 1994.
12. Mitchell M., *An Introduction to Genetic Algorithms*, The MIT Press, 1998.
13. Moran M. and H. Shapiro, *Fundamentals of Engineering Thermodynamics*, Wiley, 1998.
14. Rizzoni G., *Principles and Applications of Electrical Engineering*, McGraw-Hill, 2000.
15. Santner T. J., Williams B. and Notz W., *The Design and Analysis of Computer Experiments*, Springer Verlag, 2003.
16. Scheck F., *Mechanik*, Springer Verlag, 1988.
17. Sontag E. D., *Mathematical Control Theory*, Springer Verlag, 1990.
18. Stephanopoulous G., *Chemical Process Control*, Prentice Hall, 1984.
19. Tritton D. J., *Physical Fluid Dynamics*, Calendon Press, Oxford, 1988.
20. Vidyasagar M., *Nonlinear Systems*, Springer Verlag, 1990.
21. Zhou K., *Robust and Optimal Control*, Prentice Hall, 1995.