Modeling and Simulation for Automatic Control

Olav Egeland and Jan Tommy Gravdahl
Norwegian University of Science and Technology
Trondheim, Norway

MARINE CYBERNETICS
Trondheim, Norway
http://www.marinecybernetics.com

For ordering see URL: http://www.marinecybernetics.com. The book can also be ordered by sending an e-mail to:

info@marinecybernetics.com

or via fax:

MARINE CYBERNETICS AS
P.O. Box 4607, NO-7451 Trondheim, Norway
fax: [+47] 72 81 00 18

ISBN 82-92356-01-0

Corrected second printing June 2003

Produced from camera-ready copy supplied by the author using *Scientific WorkPlace*.

Printed and bound by Tapir Trykkeri, Trondheim, Norway.

Preface

Modeling and simulation of dynamic processes are very important subjects in control systems design. Most processes that are encountered in practical controller design are very well described in the engineering literature, and it is important that the control engineer is able to take advantage of this information. It is a problem that several books must be used to get the relevant modeling information of a particular process, and it may take a long time to go through all the necessary material. The idea of this book is to supply the control engineer with a sufficient modeling background to design controllers for a wide range of processes. In addition, the book provides a good starting point for going into the specialist literature of different engineering disciplines. In this connection the references indicate where to start. The book also contains more material than what will normally be covered in the lectures of a typical course, so that students may return to the book at a later stage and find additional information about a particular subject. This will be more efficient than to extract the required information from a series of other books. In this sense the book will be of great value for practicing control engineers.

The development of new products and systems is often done in a team of experts with different backgrounds. It is hoped that this book will help control engineers to communicate with other experts in this type of team. To achieve this we have been careful to use standard terminology and notation from the different engineering disciplines in question. Here we deliberately break the tradition evident in many books in the control literature where the emphasis is on having a unified formulation specific to automatic control.

The selection of the material is based on the experience of the authors in teaching and research at the Norwegian University of Science and Technology. In addition to this, material has been selected on the basis of extensive industrial activity through research programs between university and industry, and product development in industry. In this activity there has been close cooperation with experts from other disciplines, and this has given useful experience on how to approach different topics, and on how to interact with other specialists.

The style of modeling used in this book is inspired from the field of robotics where modeling is presented in a precise style based on equations. In addition, quite detailed results and optimized algorithms are included in standard textbook in robotics. As a result of this, the development in our book relies on many equations, but it is our experience that this is well appreciated by most students, as they do no have to waste time on trying to understand long written descriptions on subjects that are easily understood in terms of a series of equations. Moreover, we have experienced that the material presented in this book is suited both for newcomers and for students with prior courses in the topics of the book. In particular we have seen that students with virtually no background in dynamics have been able to master rigid body dynamics after going through

the dynamics chapters of this book. At the same time, students who have taken courses in dynamics also find the material in this book to be useful.

Parts of this book have been taught as a one-semester course at the Norwegian University of Science and Technology. The students are in the third year of their study in electrical engineering with specialization in automatic control, and have taken a basic course in automatic control theory. Standard undergraduate courses in engineering mathematics give a sufficient background in mathematics.

The results presented in this book have been developed and accumulated over a period of 15 years. The first author would like to thank all of his doctoral students over this period for their contributions. Also our colleagues and friends abroad have been important in this work. Thanks are due to Rolf Johansson, Henk Nijmeijer, Rogelio Lozano and Atul Kelkar for discussions on this book. In the writing of the book and in the selection of the material we have benefited from the availability of the lecture notes by Steinar Sælid, Rolf Henriksen and Torleif Iversen that have been used in earlier versions of the course.

We would like to thank doctoral candidate Erlend Kristiansen for his work on simulations, figures and proofreading. We would also like to thank Thor I. Fossen who has been writing a book in parallel, and we have enjoyed all the discussions on writing in general and modeling in particular. Thanks are also due to our colleagues Kristin Y. Pettersen, Tor Arne Johansen and Asgeir Sørensen. We would also like to thank our colleagues at the Department of Engineering Cybernetics for contributing to the stimulating working environment that allowed us to write this book. Also the support from the Norwegian Research Council has been important, as this has made it possible to have a large group of PhD students and Post Docs at the Department. In particular, the Strategic University Program in Marine Control Systems has given us very good working conditions.

Olav Egeland
Jan Tommy Gravdahl
December 2002

Contents

Part I

Modeling

Chapter 1

Model representation

1.1 Introduction

In this chapter we will present model formulations for use in controller analysis and design. The usual type of models in control problems are based on ordinary differential equations with time as the free variable. The two main representations of such models are state-space descriptions, where the model is given as a system of first-order differential equations, and transfer function models using the Laplace transformation. In this setting the signal-flow representation of models is used where each model has a defined set of input variables and a set of output variables. Control theory offers a wide variety of tools and techniques for controller analysis and design based on state-space models and transfer function models.

The signal-flow description has been very successful in control applications. However, in energy-based control analysis and in the development of simulation systems, there is an alternative formulation which is based on an energy-flow description. This formulation is of great use in the development of large simulation systems as it opens up for object-oriented modeling. This is an approach where a model is developed for each physical subsystem, and where the model of the total system is obtained by interconnecting the models of the subsystems using energy-flow variables.

Throughout the book models are developed from physics. This includes physical principles like Newton's laws and balance equations, which are typically based on the conservation of mass, momentum, energy, and electrical charge. In addition, results are derived using the purely mathematical field of kinematics, which is the geometric description of motion. Finally, empirically established constitutive equations are needed to describe material properties like the relation between the force and deformation of a spring, the relation between velocity gradients and viscous tension of a fluid, and the relation between charge and voltage of a capacitive element.

In contrast to this, models may be obtained as black-box model where transfer functions or state space models between inputs and outputs are established from identification experiments . This approach will not be discussed in this book.

This chapter starts with a presentation of state-space models and transfer function models in the signal-flow description, which is the usual formulation in automatic control. Then the energy-flow description is presented. Material on second order mechanical systems and systems described by partial differential equations is also discussed. Background material on control is found in (Kuo 1995), (Chen 1999) and

(Dorf and Bishop 2000), while additional material on linear systems theory is covered by (Rugh 1996) and (Antsaklis and Michel 1997).

1.2 State space methods

1.2.1 State space models

A *state space model*

$$\dot{x}_1 = f_1(x_1, \ldots, x_n, u_1, \ldots u_p, t) \tag{1.1}$$

$$\vdots \tag{1.2}$$

$$\dot{x}_n = f_n(x_1, \ldots, x_n, u_1, \ldots u_p, t) \tag{1.3}$$

which in vector form is written

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{1.4}$$

is a set of first order differential equations describing the dynamics of the *state vector* $\mathbf{x} = (\mathbf{x}_1, \ldots, x_n)^T$ under the action of the *control* or *input vector* is $\mathbf{u} = (\mathbf{u}_1, \ldots, u_p)^T$. The *measurement* or *output vector* $\mathbf{y}$ is is often included in the model formulation, and the state-space model is written

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{1.5}$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}, t) \tag{1.6}$$

An important class of systems for controller design is *linear time-invariant systems* which are written in the form

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \end{aligned} \tag{1.7}$$

A block diagram is shown in Figure 1.1.

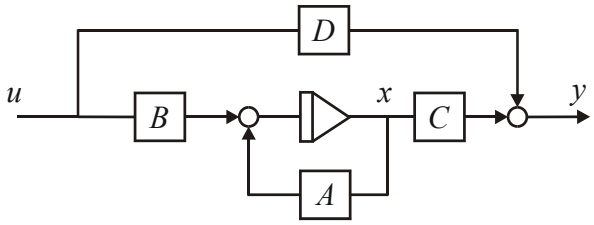

Figure 1.1: Linear time-invariant state space model

Example 1 *Systems that can be written in the form*

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\,\mathbf{u} \tag{1.8}$$

are said to be affine in the control $\mathbf{u}$*, which means that when* $\mathbf{x}$ *is given, then the right side of (1.8) is a constant plus a term that is linear in* $\mathbf{u}$*. This type of system is important in nonlinear control theory where methods are available for this type of model (Isidori 1989), (Nijmeijer and der Schaft 1990).*

1.2.2 Second order models of mechanical systems

Mechanical systems are often described as second order systems in the form

$$\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} + \mathbf{f}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{u} \tag{1.9}$$

where $\mathbf{q}$ is the vector of generalized coordinates and $\mathbf{u}$ is the generalized input force. The matrix $\mathbf{M}(\mathbf{q})$ may be called the mass matrix. Intuitively, this can be regarded as a generalization of Newton's law which states that mass times acceleration is equal to force. This second order model may be written in state space form by defining $\mathbf{x}_1 = \mathbf{q}$, $\mathbf{x}_2 = \dot{\mathbf{q}}$ which gives

$$\begin{pmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{M}^{-1}(\mathbf{x}_1)\left[-\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{u}\right] \end{pmatrix} \tag{1.10}$$

Some mechanical systems have models of a special structure due to the physical properties of the systems. In particular, this is true for vibration problems and for robotics. The model of a robot manipulators is written (Spong and Vidyasagar 1989), (Sciavicco and Siciliano 2000)

$$\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\,\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \tag{1.11}$$

where $\mathbf{M}(\mathbf{q})$ is the symmetric and positive definite mass matrix, $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$ is the Coriolis matrix, $\mathbf{g}(\mathbf{q})$ is the generalized force of gravity, $\mathbf{q} = (q_1, \ldots, q_6)$ is the vector of generalized coordinates, and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_6)$ is the vector of generalized actuator forces. The model is usually left in the second order formulation, as the usual control techniques used for manipulators rely on this formulation.

1.2.3 Linearization of state space models

Many methods and control techniques are available for linear systems. In particular, control methods based on frequency response require a linear model. Therefore, if the modeling of a system results in a nonlinear system

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \\ \mathbf{y} &= \mathbf{h}(\mathbf{x}, \mathbf{u}, t) \end{aligned} \tag{1.12}$$

it may be useful to *linearize* the system. Linearization is done around a solution of the system. A solution of the system is a function $(\mathbf{x}_0(t), \mathbf{u}_0(t))$ that satisfies the system equation

$$\dot{\mathbf{x}}_0 = \mathbf{f}\left[\mathbf{x}_0(t), \mathbf{u}_0(t), t\right] \tag{1.13}$$

We define the perturbations $\Delta\mathbf{x}$, $\Delta\mathbf{u}$ and $\Delta\mathbf{y}$ from the solution by

$$\mathbf{x}(t) = \mathbf{x}_0(t) + \Delta\mathbf{x}(t) \tag{1.14}$$

$$\mathbf{u}(t) = \mathbf{u}_0(t) + \Delta\mathbf{u}(t) \tag{1.15}$$

$$\mathbf{y}(t) = \mathbf{h}\left[\mathbf{x}_0(t), \mathbf{u}_0(t), t\right] + \Delta\mathbf{y}(t) \tag{1.16}$$

Standard Taylor series linearization around the solution $(\mathbf{x}_0(t), \mathbf{u}_0(t))$ gives

$$\dot{\mathbf{x}} = \mathbf{f}\left[\mathbf{x}_0(t), \mathbf{u}_0(t), t\right] + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta\mathbf{x} + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta\mathbf{u} \tag{1.17}$$

$$\mathbf{y} = \mathbf{h}\left[\mathbf{x}_0(t), \mathbf{u}_0(t), t\right] + \left.\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta\mathbf{x} + \left.\frac{\partial \mathbf{h}}{\partial \mathbf{u}}\right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta\mathbf{u} \tag{1.18}$$

where the matrices appearing from the differentiations have elements given by

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \left\{ \frac{\partial f_i}{\partial x_j} \right\}, \quad \frac{\partial \mathbf{f}}{\partial \mathbf{u}} = \left\{ \frac{\partial f_i}{\partial u_j} \right\}, \quad \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \left\{ \frac{\partial h_i}{\partial x_j} \right\}, \quad \frac{\partial \mathbf{h}}{\partial \mathbf{u}} = \left\{ \frac{\partial h_i}{\partial u_j} \right\} \tag{1.19}$$

Insertion of (1.13) into (1.17), and insertion of (1.16) into (1.18) gives the following linearized system:

$$\Delta \dot{\mathbf{x}} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{x} + \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{u} \tag{1.20}$$

$$\Delta \mathbf{y} = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{x} + \left. \frac{\partial \mathbf{h}}{\partial \mathbf{u}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \Delta \mathbf{u} \tag{1.21}$$

It is noted that this model is of the same form as (1.7). The matrices $\mathbf{A}, \mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are seen to be given by

$$\mathbf{A}(t) = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)}, \quad \mathbf{B}(t) = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \tag{1.22}$$

$$\mathbf{C}(t) = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)}, \quad \mathbf{D}(t) = \left. \frac{\partial \mathbf{h}}{\partial \mathbf{u}} \right|_{\mathbf{x}_0(t), \mathbf{u}_0(t)} \tag{1.23}$$

Example 2 *A simplified model for the design of a cruise control system for a car is obtained from Newton's law. Suppose that the forces acting on the car are the air resistance, which is proportional to the square of the velocity, and the motor force, which is assumed to be proportional to the throttle input. Then the model is*

$$m\dot{v} = -\frac{1}{2} C_D \rho A v^2 + K_t u \tag{1.24}$$

where v is the velocity and m is the mass of the car, C_D is the drag coefficient, ρ is the density of air, A is the projected area of the car when seen from the front, K_t is the throttle constant, and u is the throttle input. The control input u_0 corresponding to a constant speed v_0 is found by inserting $m\dot{v}_0 = 0$ in the model, which gives

$$u_0 = \frac{1}{2K_t} C_D \rho A v_0^2 \tag{1.25}$$

We define the perturbations $\Delta v = v - v_0$ and $\Delta u = u - u_0$ and find the linearized model

$$m\Delta\dot{v} = -C_D \rho A v_0 \Delta v + K_t \Delta u \tag{1.26}$$

Example 3 *A standard laboratory demonstration of feedback control is the magnetic levitation experiment where an electromagnet is used to control the vertical position of a steel ball. The equation of motion for the ball is derived in Section 3.7.9 to be*

$$m\ddot{z} = -C\frac{i^2}{z^2} + mg \tag{1.27}$$

where m is the mass, z is the vertical position of the ball in the downwards direction, C is a constant, i is the control input, which is the current of the electromagnet, and g is the

acceleration of gravity. Let z_d be the constant desired position of the ball. The solution (z_d, i_d) is found by inserting $m\ddot{z}_d = 0$ in the model, which gives the constant current

$$i_d = \sqrt{\frac{mg}{C}} z_d \tag{1.28}$$

which will give a lifting force that can hold the ball stationary at position z_d. We define the perturbations $\Delta z = z - z_d$ and $\Delta i = i - i_d$ and get the linearized model

$$m\Delta\ddot{z} = 2C\frac{i_d^2}{z_d^3}\Delta z - 2C\frac{i_d}{z_d^2}\Delta i \tag{1.29}$$

1.2.4 Linearization of second order systems

A second order system

$$\ddot{\mathbf{q}} = \mathbf{f}\left(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u}\right) \tag{1.30}$$

may be linearized by reformulating it as a state-space model (1.4) with $\mathbf{x} = (\mathbf{q}^T, \dot{\mathbf{q}}^T)^T$. However, we may also linearize the system in the second order formulation, which may be advantageous for some systems. Then the system is linearized around a solution $(\mathbf{q}_0(t), \dot{\mathbf{q}}_0(t), \mathbf{u}_0(t))$ which satisfies

$$\ddot{\mathbf{q}}_0 = \mathbf{f}\left(\mathbf{q}_0, \dot{\mathbf{q}}_0, \mathbf{u}_0\right). \tag{1.31}$$

Taylor series expansion of the model around the solution gives

$$\ddot{\mathbf{q}}_0 + \Delta\ddot{\mathbf{q}} = \mathbf{f}\left(\mathbf{q}_0, \dot{\mathbf{q}}_0, \mathbf{u}_0\right) + \frac{\partial \mathbf{f}}{\partial \mathbf{q}}\Delta\mathbf{q} + \frac{\partial \mathbf{f}}{\partial \dot{\mathbf{q}}}\Delta\dot{\mathbf{q}} + \frac{\partial \mathbf{f}}{\partial \mathbf{u}}\Delta\mathbf{u} \tag{1.32}$$

and combination with (1.31) gives the linearized model

$$\Delta\ddot{\mathbf{q}} = \frac{\partial \mathbf{f}}{\partial \mathbf{q}}\Delta\mathbf{q} + \frac{\partial \mathbf{f}}{\partial \dot{\mathbf{q}}}\Delta\dot{\mathbf{q}} + \frac{\partial \mathbf{f}}{\partial \mathbf{u}}\Delta\mathbf{u} \tag{1.33}$$

Example 4 *Consider a pendulum with a point mass m on a massless beam of length L as shown in Figure 1.2. The angle of rotation θ is set to zero when the pendulum is hanging downwards. The equation of motion for the pendulum is*

$$mL^2\ddot{\theta} + mLg\sin\theta = 0 \tag{1.34}$$

which can be written

$$\ddot{\theta} = -\frac{g}{L}\sin\theta \tag{1.35}$$

Linearization around the solution $(\theta, \dot{\theta}) = (0, 0)$ gives the linear model

$$\ddot{\theta} = -\frac{g}{L}\theta \tag{1.36}$$

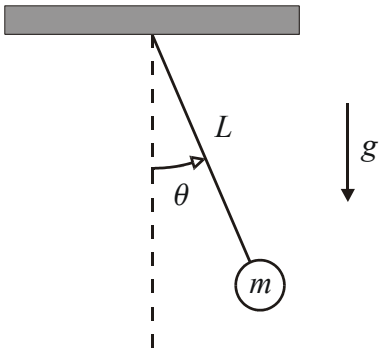

Figure 1.2: Pendulum

1.2.5 Stability with zero input

The concept of *stability* is of fundamental importance in control theory, and it is highly relevant in connection with modelling as we can highlight the stability properties of a system by selecting an appropriate model formulation. We will therefore present results on the stability of a state-space model that are important in connection with modeling. We will focus on the stability of a system with zero input around an *equilibrium state* $\mathbf{x}_e$. A state $\mathbf{x}_e$ is an equilibrium state if the system is at rest in this equilibrium state. With this we mean that if the system state starts in $\mathbf{x} = \mathbf{x}_e$, then the state vector will remain in $\mathbf{x}_e$. The equilibrium is said to be *stable* if it has the property that the state will stay close to the equilibrium whenever the state starts near the equilibrium. If an equilibrium of a system is not stable, then it is said to be *unstable*. If an equilibrium is stable and the state converges to the equilibrium, then the equilibrium is said to be *asymptotically stable*.

Example 5 *Consider the state space model*

$$\frac{d}{dt}\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \dot{\theta} \\ -\frac{g}{L}\sin\theta \end{pmatrix} \tag{1.37}$$

of a pendulum. The states of the pendulum are selected to be θ and $\dot{\theta}$ so that the state vector is

$$\mathbf{x} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \tag{1.38}$$

We see that $\dot{\mathbf{x}} = \mathbf{0}$ whenever $\dot{\theta} = 0$ and $\sin\theta = 0$, which is the case for $\theta = 0$ and $\theta = \pi$. This means that the system has equilibrium states at

$$\mathbf{x}_{e1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_{e2} = \begin{pmatrix} \pi \\ 0 \end{pmatrix} \tag{1.39}$$

Here $\mathbf{x} = \mathbf{x}_{e1}$ is the equilibrium where the pendulum is hanging downwards, and $\mathbf{x} = \mathbf{x}_{e2}$ is the equilibrium where the pendulum is raised upwards with the mass on the top. We know from experience that if the pendulum starts from a state close to the downwards configuration $\mathbf{x}_{e1}$, and with a speed close to zero, then the pendulum will stay close to the downwards configuration. Therefore, the system is stable around the equilibrium

$\mathbf{x}_{e1}$. *Our experience with the equilibrium* $\mathbf{x}_{e2}$ *is that for any small deviation from the equilibrium state the pendulum will fall down and move far away from the equilibrium state. Therefore, the system is unstable around the equilibrium* $\mathbf{x}_{e2}$.

1.2.6 Stability of linear systems

Consider the linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{1.40}$$
$$y = \mathbf{C}\mathbf{x} + \mathbf{D}u \tag{1.41}$$

where u is input, y is output, and $\mathbf{x} \in R^n$ is the state vector. The solution of the state equation are known from basic textbooks in automatic control to be

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}(t_0) + \int_{t_0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}u(\tau)\,d\tau \tag{1.42}$$

The eigenvalues of the $n \times n$ matrix $\mathbf{A}$ are denoted λ_i, $i = 1, \ldots, n$. It is assumed that $\mathbf{A}$ has $m \leq n$ simple eigenvalues λ_i, $i = 1, \ldots, m$, and that the remaining $n-m$ eigenvalues $\lambda_{m+1} = \ldots = \lambda_n$ are coincident. Then, if the input is zero, that is $u = 0$, the solution of the state equation is

$$\mathbf{x}(t) = \left(\sum_{i=1}^{m} \mathbf{K}_i e^{\lambda_i t} + \sum_{i=0}^{n-m-1} \mathbf{K}_{m+i} t^i e^{\lambda_n t} \right) \mathbf{x}(0) \tag{1.43}$$

where $\mathbf{K}_i$, $i = 1, \ldots, n$ are constant matrices depending on $\mathbf{A}$ and $\mathbf{B}$. We see that when the input is zero, then the system is

- Stable whenever all simple eigenvalues have real parts that are not positive, and all multiple eigenvalues have real parts that are negative, that is, if

$$\begin{array}{ll} \mathrm{Re}[\lambda_i] \leq 0, & \lambda_i \text{ simple eigenvalue} \\ \mathrm{Re}[\lambda_i] < 0, & \lambda_i \text{ multiple eigenvalue} \end{array} \tag{1.44}$$

- Asymptotically stable if all eigenvalues are negative, that is, if

$$\mathrm{Re}[\lambda_i], < 0 \quad i = 1, \ldots, n \tag{1.45}$$

1.2.7 Stability analysis using a linearized model

The stability of a nonlinear system around an equilibrium can be studied by analyzing the linearization of the system around the equilibrium. Then (Slotine 1991), (Khalil 1996)

- If the linearized system is asymptotically stable, then the nonlinear system is also asymptotically stable.
- If the linearized system is stable, but with at least one pole at the imaginary axis, then the nonlinear system may be stable or unstable.
- If the linearized system is unstable, then the nonlinear system is unstable.

Example 6 *We will demonstrate this for a pendulum. The nonlinear system is*

$$\ddot{\theta} + 2\zeta\omega_0\dot{\theta} + \omega_0^2 \sin\theta = 0 \tag{1.46}$$

where $\omega_0^2 = g/L$ *and* $2\zeta\omega_0\dot{\theta}$ *is a viscous damping term where* $0 \leq \zeta$*. First, linearization around* $(\theta, \dot{\theta}) = (0, 0)$ *gives*

$$\ddot{\theta} + 2\zeta\omega_0\dot{\theta} + \omega_0^2\theta = 0 \tag{1.47}$$

If $\zeta > 0$ *then the linearized system is asymptotically stable. This implies that the nonlinear system is asymptotically stable around* $(\theta, \dot{\theta}) = (0, 0)$ *for* $\zeta > 0$*. If* $\zeta = 0$*, then the eigenvalues of the systems are at the imaginary axis, and we cannot conclude on the stability of the nonlinear system by analyzing the linear system. Next, consider the equilibrium point* $(\theta, \dot{\theta}) = (\pi, 0)$*. The linearized system is*

$$\ddot{\theta} + 2\zeta\omega_0\dot{\theta} - \omega_0^2\theta = 0 \tag{1.48}$$

which has one pole in the right half plane. The linearized system is therefore unstable, and we may conclude that the nonlinear system is unstable around the equilibrium $(\theta, \dot{\theta}) = (\pi, 0)$*.*

1.3 Transfer function models

1.3.1 Introduction

Linear time-invariant systems may be represented by transfer functions based on the use of the Laplace transform. This makes it possible to use important analysis and design methods in the Laplace description, and it serves as a good starting point for using frequency response techniques. The Laplace transformation is easier to use than the Fourier transformation, and it is a more general and powerful tool in controller design and analysis. Moreover, if the Fourier transformation exists, it is obtained as a special case of the Laplace transformation by using $s = j\omega$ for the complex variable s.

1.3.2 The transfer function of a state-space model

In this section we will derive the transfer function corresponding to a linear time-invariant state-space model, and present some useful results. A linear time-invariant system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \tag{1.49}$$

$$y(t) = \mathbf{C}\mathbf{x}(t) + Du(t) \tag{1.50}$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ can be described by a transfer function using the Laplace transformation. We use the notation

$$\mathbf{x}(s) = \mathcal{L}\{\mathbf{x}(t)\}, \quad u(s) = \mathcal{L}\{u(t)\} \quad \text{and} \quad y(s) = \mathcal{L}\{y(t)\} \tag{1.51}$$

The Laplace transform of the time derivative of the state $\mathbf{x}$ is given by

$$\mathcal{L}\{\dot{\mathbf{x}}(t)\} = s\mathcal{L}\{\mathbf{x}(t)\} - \mathbf{x}(t = 0) \tag{1.52}$$

In the development of transfer function models the initial conditions $\mathbf{x}(t = 0)$ are always set to zero. This can be done as the system is linear and superposition applies. Therefore we set $\dot{\mathbf{x}}(t = 0) = 0$ and get

$$\mathcal{L}\{\dot{\mathbf{x}}(t)\} = s\mathbf{x}(s) \tag{1.53}$$

Then the Laplace transformed state-space model is found to be

$$s\mathbf{x}(s) = \mathbf{A}\mathbf{x}(s) + \mathbf{B}u(s) \tag{1.54}$$
$$y(s) = \mathbf{C}\mathbf{x}(s) + Du(s) \tag{1.55}$$

We eliminate $\mathbf{x}(s)$ using the first equation and insert the expression into the second equation. This gives

$$y(s) = \left[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D\right] u(s) \tag{1.56}$$

We define the transfer function $H(s)$ from $u(s)$ to $y(s)$ as

$$H(s) = \left[\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D\right] \tag{1.57}$$

and write

$$\frac{y}{u}(s) = H(s). \tag{1.58}$$

A block diagram is shown in Figure 1.3.

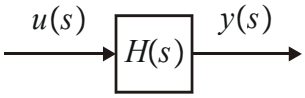

Figure 1.3: Transfer function representaion of system

1.3.3 Rational transfer functions

A transfer function is said to be *rational* if it can be written in the form

$$H(s) = K\frac{P(s)}{Q(s)} \tag{1.59}$$

where the scalar K is called the gain, $P(s)$ is a polynomial in the complex variable s of degree m, and $Q(s)$ is a polynomial in s of degree n. A rational transfer function can be factored in the form

$$H(s) = K\frac{(s+z_1)(s+z_2)\dots(s+z_m)}{(s+p_1)(s+p_2)\dots(s+p_n)} \tag{1.60}$$

The transfer function is said to have m *zeros* at $s = -z_i$ and n *poles* at $s = -p_i$. The poles and the zeros may be real, or they can appear as complex conjugated pairs. We see that the transfer function is defined and continuous for all s except for the poles, which are the singularities of a rational transfer function. The transfer function is said to be *proper* if there are at least as many poles as zeros, that is, if $n \geq m$, and it is said to be *strictly proper* if there are more poles than zeros, that is, if $n > m$. If $m > n$, then the transfer function is said to have $m - n$ poles at infinity. In $n > m$, then the transfer function is said to have $n - m$ zeros at infinity.

It is noted that the transfer function of an n-dimensional state space model is a proper rational transfer function with n poles under the assumption that all states are controllable and observable.

Example 7 *The transfer function $H_1(s) = s$ is not proper, and has one pole at infinity, while the transfer function $H_2(s) = 1/s$ is a strictly proper transfer function with one zero at infinity.*

1.3.4 Impulse response and step response

The Dirac delta function $\delta(t)$, which is referred to as a unit impulse function, has the Laplace transform $\mathcal{L}\{\delta(t)\} = 1$. Thus, the response of the system when the input is a unit impulse is

$$y(s) = H(s) \cdot 1 = H(s) \tag{1.61}$$

which corresponds to the time function

$$y(t) = h(t) := \mathcal{L}^{-1}\{H(s)\} \tag{1.62}$$

Therefore, $h(t)$ is referred to as the *impulse response* of the system.

We define the *unit step function*

$$u_s(t) = \begin{cases} 0, & t < 0 \\ 1, & 0 \leq t \end{cases} \tag{1.63}$$

which has the Laplace transform

$$u_s(s) = \mathcal{L}\{u_s(t)\} = \frac{1}{s} \tag{1.64}$$

The step response of the system, which is the response $y(t)$ resulting from an initial value $y(t=0) = 0$ and the input $u(t) = u_s(t)$, is

$$y(s) = H(s)\frac{1}{s} \tag{1.65}$$

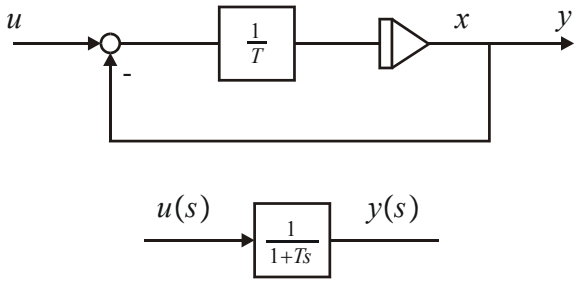

Figure 1.4: A time constant

Example 8 *The dynamic system*

$$\dot{x} = \frac{1}{T}(-x + u) \tag{1.66}$$

$$y = x \tag{1.67}$$

is referred to as a time constant. A block diagram is shown in Figure 1.4. The Laplace transformation gives the transfer function

$$\frac{y}{u}(s) = H(s) = \frac{1}{1+Ts} \tag{1.68}$$

The impulse response of the system is

$$h(t) = \mathcal{L}^{-1}\{H(s)\} = e^{-\frac{t}{T}} \tag{1.69}$$

while the step response of the system is

$$y(t) = \mathcal{L}^{-1}\left\{\frac{H(s)}{s}\right\} = 1 - e^{-\frac{t}{T}} \tag{1.70}$$

Example 9 *By setting all initial values* $\frac{d^i}{dt^i}x(t) = 0$, $i = 1, \ldots, n$ *we find that*

$$\mathcal{L}\left\{\frac{d^n}{dt^n}x(t)\right\} = s^n X(s) \tag{1.71}$$

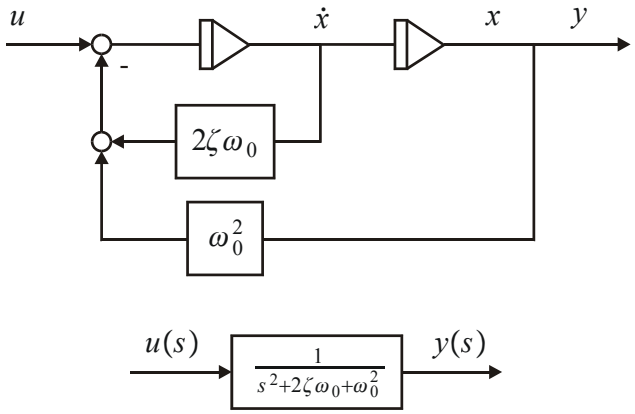

Figure 1.5: Second order oscillatory system

Example 10 *The model*

$$\ddot{x}(t) = -2\zeta\omega_0\dot{x}(t) - \omega_0^2 x(t) + u(t) \tag{1.72}$$
$$y(t) = x(t) \tag{1.73}$$

given as a block diagram in Figure 1.5 is Laplace transformed to

$$s^2 x(s) = -2\zeta\omega_0 s X(s) - \omega_0^2 x(s) + u(s) \tag{1.74}$$

which is solved for $x(s)$ *to give*

$$x(s) = \frac{1}{s^2 + 2\zeta\omega_0 s + \omega_0^2} u(s) \tag{1.75}$$

The transfer function is

$$\frac{y}{u}(s) = H(s) = \frac{1}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{1.76}$$

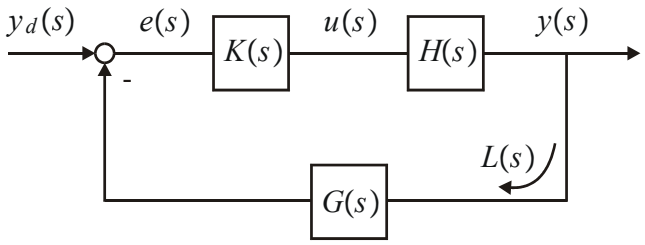

Figure 1.6: Plant $H(s)$ with a series compensation controller $K(s)$ in feedback compensation controller $G(s)$

1.3.5 Loop transfer function

We consider a plant $H(s)$ with a series compensation controller $K(s)$ and feedback compensation controller $G(s)$ as shown in Figure 1.6. The plant is given by

$$y(s) = H(s)\,u(s) \tag{1.77}$$

and the controller is given by

$$u(s) = K(s)\,e(s), \quad e(s) = y_d(s) - G(s)\,y(s) \tag{1.78}$$

where y_d is the input signal to the closed-loop system. Define the *loop transfer function* by

$$L(s) = K(s)\,H(s)\,G(s). \tag{1.79}$$

From

$$e(s) = y_d(s) - G(s)\,y(s) = y_d(s) - G(s)\,H(s)\,K(s)\,e(s) \tag{1.80}$$

it is seen that the transfer function from the input signal $y_d(s)$ to the error signal $e(s)$ is given by

$$S(s) := \frac{e}{y_d}(s) = \frac{1}{1+L(s)} \tag{1.81}$$

where $S(s)$ is called the *sensitivity function* of the closed loop system.

The *closed-loop transfer function* $T(s)$ is defined as the transfer function from the closed-loop input $y_d(s)$ to the output $y(s)$. From the expression

$$y(s) = K(s)\,H(s)\,e(s) = K(s)\,H(s)\,[y_d(s) - G(s)\,y(s)] \tag{1.82}$$

it is possible to solve for $y(s)$ as a function of $y_d(s)$, and the closed-loop transfer function $T(s)$ is found to be

$$T(s) := \frac{y}{y_d}(s) = \frac{K(s)\,H(s)}{1+L(s)} \tag{1.83}$$

The closed-loop transfer function can be written

$$T(s) = \frac{1}{G(s)}\frac{L(s)}{1+L(s)} \approx \begin{cases} \frac{1}{G(s)} & |L(s)| \gg 1 \\ K(s)\,H(s) & |L(s)| \ll 1 \end{cases}$$

This means that if the loop-transfer function is large, that is, if $|L(s)| \gg 1$, then $T(s) = 1/G(s)$.

Example 11 *If unity feedback is used, that is, if* $G(s) = 1$, *then*

$$T(s) = \frac{L(s)}{1 + L(s)} = 1 - S(s)$$

and $T(s)$ *is called the* complementary sensitivity function.

1.3.6 Example: Actuator with dynamic compensation

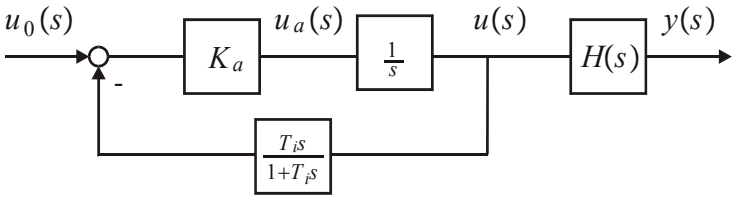

Figure 1.7: System with dynamic feedback control of the actuator.

In many control systems there is a servomotor that acts as an actuator for the main plant. In this case it can be useful to control the servomotor with an inner actuator loop with dynamic feedback to achieve a suitable transfer function in the outer loop. Suppose that the main plant is described by the transfer function model

$$y(s) = H(s)\, u(s) \tag{1.84}$$

while the control input is obtained using a velocity controlled servomotor. The model for the actuator is assumed to be given by

$$u(s) = \frac{1}{s} u_a(s) \tag{1.85}$$

where $u_a(s)$ is the velocity command to the actuator. The actuator is here modeled by an integration which is the transfer function of a motor with a velocity loop where u_a is the desired velocity input. The feedback for the actuator loop is given by

$$u_a(s) = K_a [u_0(s) - G_a(s)\, u(s)] \tag{1.86}$$

where the dynamic feedback is the high-pass filter

$$G_a(s) = \frac{T_i s}{1 + T_i s} \tag{1.87}$$

as shown in Figure 1.7. Then the loop transfer function of the actuator loop is

$$L_a(s) = \frac{K_a T_i}{1 + T_i s} \tag{1.88}$$

and we find that the closed loop transfer function of the actuator loop is given by

$$\frac{u}{u_0}(s) = K_p \frac{1 + T_i s}{T_i s} \frac{1}{1 + T_1 s} \tag{1.89}$$

where

$$K_p = \frac{K_a T_i}{1 + K_a T_i}, \quad T_1 = \frac{T_i}{1 + K_a T_i} \tag{1.90}$$

This is illustrated in Figure 1.8. If K_p and T_i are selected so that break frequency $1/T_1$ is much higher than the crossover frequency of the outer loop, then the actuator loop will introduce integral action in the outer loop according to

$$\frac{y}{u_0}(s) \approx K_p \frac{1 + T_i s}{T_i s} H(s) \tag{1.91}$$

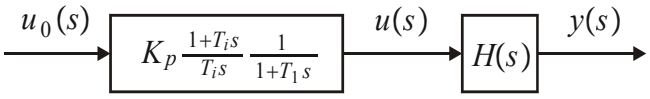

Figure 1.8: System with the closed loop dynamics of the actuator with controller.

1.3.7 Stability of transfer functions

Consider the system

$$y(s) = H(s)u(s) \tag{1.92}$$

where $H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D$ is the transfer function of the system. This system is bounded-input-bonded-output stable, which is termed BIBO stable, if and only if all the poles λ_i of $H(s)$ have real parts that are less than zero, that is, $\operatorname{Re}\lambda_i < 0$, $i = 1, \ldots, n$.

This is shown as follows: The impulse response corresponding to the transfer function $H(s)$ is denoted $h(t)$. Assume that the input is bounded according to

$$|u(t)| \leq U \quad \text{for all } t \tag{1.93}$$

where $U > 0$ is a constant. The output is given by

$$y(t) = \int_0^\infty h(\tau)\, u(t - \tau)\, d\tau \tag{1.94}$$

Taking the absolute values on both sides, we find that

$$\begin{aligned} |y(t)| &= \left| \int_0^\infty h(\tau)\, u(t - \tau)\, d\tau \right| \\ &\leq \int_0^\infty |h(\tau)|\, |u(t - \tau)|\, d\tau \\ &\leq U \int_0^\infty |h(\tau)|\, d\tau \end{aligned} \tag{1.95}$$

Suppose that all the poles are to the left of $-\alpha$. Then there is a constant $k \geq 0$ so that

$$|h(t)| \leq k e^{-\alpha t} \tag{1.96}$$

and, using the fact that for $\alpha > 0$ we have $\int_0^\infty e^{-\alpha t} dt = \alpha^{-1}$, it follows that

$$\int_0^\infty |h(\tau)|\, d\tau \leq \frac{k}{\alpha}, \quad \text{if} \quad \alpha > 0 \tag{1.97}$$

We find that

$$|y(t)| \leq \frac{k_2}{\alpha}U \quad \text{when} \quad \alpha > 0 \tag{1.98}$$

which shows that when $\mathrm{Re}[\lambda_i] < 0$, then $y(t)$ is bounded whenever $u(t)$ is bounded.

Suppose that all poles have real parts less than zero, except one pole in $s = 0$. Then the control input $u(t) = 1$ will for large t give $y(t) \propto t \to \infty$. Next, suppose that all poles have real parts less than zero, except a complex conjugated pole pair in $s = \pm j\omega_0$. Then the control input $u(t) = \cos\omega_0 t$ will give a response which for large t satisfies $y(t) \propto t\cos\omega_0 t \to \infty$ where $\propto$ denotes proportional to. If there are poles with real part larger than zero, than $y(t)$ will be unbounded for a unit step input. This shows that an unbounded output may occur for bounded input when there are poles on the imaginary axis or to the right of the imaginary axis.

1.3.8 Stability of closed loop systems

The stability of a closed loop system can be analyzed by studying the sensitivity function $S(s)$. The closed loop system will be stable if the poles of $S(s)$ have real parts that are negative. This can be checked using one of the standard sufficient conditions on the loop transfer function $L(s)$, which are available from automatic control theory. Note, however, that the conditions on $L(s)$ are typically derived under certain assumptions on the properties of $L(s)$. The fundamental requirement for stability is that the poles of $S(s)$ do not have positive real part, and that multiple poles must have real parts that are less than zero.

Example 12 *Large tankers may be unstable, and the transfer function $H(s)$ from the rudder angle δ to the course angle ψ will then include a pole in the right half plane. An example of this is the following model of a tanker (Blanke 1981), (Fossen 1994)*

$$H(s) = \frac{\psi}{\delta}(s) = K\frac{1 + T_a s}{s(1 + T_1 s)(T_2 s - 1)} \tag{1.99}$$

where $K = 0.022$, $T_a = 38$ s, $T_1 = 16$ s and $T_2 = 192$ s. The integration represented by the factor s in the denominator is due to the integration from angular velocity around the vertical axis to the course angle ψ. The transfer function has a pole at $s = 1/T_2$. An autopilot with a PD controller

$$\delta(s) = K_p\frac{1 + T_1 s}{1 + 0.1T_1 s}(\psi_0 - \psi) \tag{1.100}$$

gives the characteristic equation

$$s(1 + 0.1T_1 s)(T_2 s - 1) + KK_p(1 + T_a s) = 0 \tag{1.101}$$

for the closed loop system. The closed loop poles are therefore at $s = -0.5609$ and $s = -0.0295 \pm j0.0303$. This is found using the MATLAB command

```
roots(conv([1.6 1 0],[192 -1]) + 20*0.022*[0 0 38 1])
```

1.3.9 Partial differential equations

Systems described by partial differential equations will typically lead to *irrational transfer functions*. Irrational transfer functions can be approximated by a rational transfer function with infinitely high order, and because of this such systems may be referred to

as *infinite dimensional systems*. An irrational transfer function is said to be *analytic* in a region if it is defined and continuous in that region. The points where an irrational transfer function ceases to be analytic are called the *singularities* of the transfer function. We recall that for a rational transfer function the singularities are called poles.

We will demonstrate the appearance of irrational transfer functions for systems described by partial differential equations by studying the partial differential equation

$$c\frac{\partial v\left(x,t\right)}{\partial x}=-\frac{\partial v\left(x,t\right)}{\partial t},\ v\left(0,t\right)=v_{1}\left(t\right) \tag{1.102}$$

This is the first order wave equation which describes the propagation of a wave-front with velocity c. The variable $v(x,t)$ has the Laplace transform $v\left(x,s\right)=\mathcal{L}\left\{v\left(x,t\right)\right\}$, and the time derivative has the transform

$$\mathcal{L}\left\{\frac{\partial v\left(x,t\right)}{\partial t}\right\}=s\mathcal{L}\left\{v\left(x,t\right)\right\}=sv\left(x,s\right) \tag{1.103}$$

From this it follows that the partial differential equation has the Laplace transform

$$c\frac{\partial v\left(x,s\right)}{\partial x}=-sv\left(x,s\right),\ v\left(0,s\right)=v_{1}\left(s\right) \tag{1.104}$$

This is an ordinary differential equation of the first order in s, which has the solution

$$v\left(x,s\right)=v\left(0,s\right)\exp(-\frac{x}{c}s) \tag{1.105}$$

The transfer function from $v_1(s)$ to $v_2\left(s\right):=v\left(L,s\right)$ at $x=L$ is then found to be the irrational transfer function

$$\frac{v_{2}}{v_{1}}\left(s\right)=e^{-Ts} \tag{1.106}$$

where $T=L/c$ is the propagation time. We see that the solution at $x=L$ is equal to the solution at $x=0$ with a time delay T.

Example 13 *The time delay in (1.106) can be approximated by a rational Padé approximation $P_{k}^{k}\left(-Ts\right)$ of order k where (Golub and van Loan 1989, p. 557)*

$$P_{k}^{k}\left(s\right) = \frac{Q_{kk}\left(s\right)}{Q_{kk}\left(-s\right)} \tag{1.107}$$

$$Q_{kk}\left(s\right) = 1+\sum_{i=1}^{k}\frac{k!\left(2k-i\right)!}{\left(k-i\right)!\left(2k\right)!}\frac{s^{i}}{i!} \tag{1.108}$$

A third order Padé approximation is found to be given by

$$e^{-Ts}\approx P_{3}^{3}\left(-Ts\right)=\frac{1-\frac{Ts}{2}+\frac{\left(Ts\right)^{2}}{10}-\frac{\left(Ts\right)^{3}}{120}}{1+\frac{Ts}{2}+\frac{\left(Ts\right)^{2}}{10}+\frac{\left(Ts\right)^{3}}{120}} \tag{1.109}$$

By letting k go to infinity we can represent the time delay by a rational transfer function of infinite dimension.

Example 14 *Transmission line dynamics are described by the second order wave equation. A hydraulic transmission line where the outlet is open has the irrational transfer function*

$$\tanh s=\frac{\sinh s}{\cosh s} \tag{1.110}$$

from the input flow to the input pressure. The transfer functions has zeros when the numerator is zero, which is the case when

$$\sinh s = \frac{1}{2}\left(e^{s} - e^{-s}\right) = 0 \quad \Rightarrow \quad e^{-2s} = 1 = e^{j2k\pi} \tag{1.111}$$

This occurs for

$$s = jk\pi \tag{1.112}$$

where $k = 0, \pm 1, \pm 2, \ldots$. *In the same way we find that the singularities appear for*

$$\cosh s = 0 \quad \Rightarrow \quad s = \pm j\left(k + \frac{1}{2}\right)\pi \tag{1.113}$$

It can be shown that the numerator and the denominator can be represented by infinite dimensional polynomials in the complex variable s, *and this gives the following infinite dimensional representation of the transfer function*

$$\tanh s = \frac{s\left(1 + \left(\frac{s}{\pi}\right)^2\right)\left(1 + \left(\frac{s}{2\pi}\right)^2\right)\left(1 + \left(\frac{s}{3\pi}\right)^2\right)\ldots}{\left(1 + \left(\frac{2s}{3\pi}\right)^2\right)\left(1 + \left(\frac{2s}{5\pi}\right)^2\right)\left(1 + \left(\frac{2s}{7\pi}\right)^2\right)\ldots} \tag{1.114}$$

We see that there are infinitely many zeros and singularities along the imaginary axis. Moreover, we see that the zeros and singularities alternate along the imaginary axes. This implies that the phase of $\tanh j\omega$ *is between -90° and +90°.*

1.4 Network description

1.4.1 Introduction

The automatic control literature relies to large extent on the use of models that are based on a signal-flow formulation. This means that different blocks of the model are connected with signals that considered to flow in the direction of the signal arrow. We might say that signal-flow description has *unilateral interconnections*. The reliance on the signal-flow description is obviously due to the many control techniques based on a signal-flow description of the physical plant in the form of state-space models and transfer functions. Because of this, it is clear that modeling techniques for use in controller design and analysis should provide methods for developing signal-flow models. However, there are good reasons for deviating from a strict reliance on signal flow in the development of mathematical models of physical systems. We will mention some arguments for this, and then discuss what the consequences are.

Many physical systems that are important in control applications are conveniently represented in an *energy-flow* description. In this case different blocks of the model are connected so that energy flows in both directions, and we say that the formulation relies on *bilateral interconnections*. The signals flowing between the blocks will then typically be voltage and current in electrical systems, force and velocity in translational mechanical systems, torque and angular velocity in rotational mechanical systems, pressure and volumetric flow in isothermal flow problems, and enthalpy and mass flow in thermal flow problems. Note that in this case it is not clearly defined in which direction a signal propagates. The main advantage of an energy-flow formulation is that it well suited for energy-based controller design using Lyapunov techniques and passivity. Moreover, it

Chapter 2

Model analysis tools

2.1 Frequency response methods

2.1.1 The frequency response of a system

The frequency response of a system can be studied by investigating the properties of the transfer function on the imaginary axis, that is, for $s = j\omega$. The starting point for frequency response analysis is the transfer function description

$$y(s) = H(s)\, u(s) \tag{2.1}$$

Suppose that $H(s)$ is strictly proper and rational, and that all the poles of $H(s)$ have real parts less than zero. Then we find the frequency response function $H(j\omega)$ from the transfer function by inserting $s = j\omega$, which is shown in the following:

The impulse response function corresponding to the transfer function $H(s)$ is $h(t) = \mathcal{L}^{-1}\{H(s)\}$. Physical system are *causal*, which means that they do not give any response to an impulse before the impulse is applied. Therefore, for a causal system the impulse response function $h(t)$ is zero for $t < 0$. Suppose that $H(s)$ is strictly proper and rational, and that all the poles of $H(s)$ have real parts less than zero. Then the impulse response $h(t)$ will decay exponentially, which implies that $\int_0^\infty |h(t)|\, dt$ exists. The frequency response

$$H(j\omega) := \mathcal{F}\{h(t)\} = \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t} dt = \int_0^\infty h(t)\, e^{-j\omega t} dt \tag{2.2}$$

will exist as

$$\left| \int_0^\infty h(t)\, e^{-j\omega t} dt \right| \leq \int_0^\infty |h(t)| \left| e^{-j\omega t} \right| dt \leq \int_0^\infty |h(t)|\, dt \tag{2.3}$$

Moreover, we see that the Fourier transform is given by

$$H(j\omega) = H(s)|_{s=j\omega} \tag{2.4}$$

where $H(s)$ is the Laplace transform of $h(t)$ defined by

$$H(s) := \mathcal{L}\{h(t)\} = \int_0^\infty h(t)\, e^{-st} dt \tag{2.5}$$

It turns out to be a great advantage to work with the Laplace transform, which contains the Fourier transform as a special case.

Example 22 *The frequency response of a time constant* $H(s) = (1+Ts)^{-1}$ *is*

$$H(j\omega) = \frac{1}{1+j\omega T} \tag{2.6}$$

with magnitude and phase given by

$$|H(j\omega)| = \frac{1}{\sqrt{1+(\omega T)^2}} \quad \text{and} \quad \angle H(j\omega) = -\arctan \omega T. \tag{2.7}$$

2.1.2 Second order oscillatory system

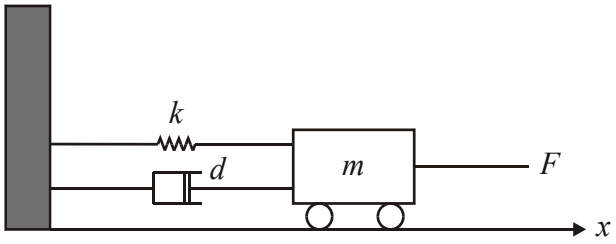

Figure 2.1: Mass-spring-damper system

A mass-spring-damper system (Figure 2.1) has the equation of motion

$$m\ddot{x} + d\dot{x} + kx = F \tag{2.8}$$

where x is the position of the mass m, d is the viscous friction coefficient and k is the spring constant. The input is the force F. The model can be normalized in to the form

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = \frac{1}{m}F \tag{2.9}$$

where the *undamped natural frequency* is

$$\omega_0 = \sqrt{\frac{k}{m}} \tag{2.10}$$

and the relative damping is

$$\zeta = \frac{1}{2\omega_0}\frac{d}{m} = \frac{d}{2\sqrt{km}} \tag{2.11}$$

The transfer function is found by inserting

$$\mathcal{L}\{\ddot{x}(t)\} = s^2\mathcal{L}\{x(t)\}, \quad \mathcal{L}\{\dot{x}(t)\} = s\mathcal{L}\{x(t)\} \tag{2.12}$$

which leads to

$$\begin{aligned} H(s) &= \frac{x}{F}(s) = \frac{1}{m}\frac{1}{s^2+2\zeta\omega_0 s+\omega_0^2} \\ &= \frac{1}{k}\frac{1}{1+2\zeta\frac{s}{\omega_0}+\left(\frac{s}{\omega_0}\right)^2} \end{aligned} \tag{2.13}$$

We assume that $\zeta < 1$ which implies that the poles of the transfer function are complex conjugated and given by

$$\lambda = \left(-\zeta \pm j\sqrt{1-\zeta^2} \right) \omega_0 \tag{2.14}$$

The frequency response is

$$H(j\omega) = \frac{1}{k} \frac{1}{1 - \left(\frac{\omega}{\omega_0}\right)^2 + j2\zeta\frac{\omega}{\omega_0}} \tag{2.15}$$

In particular we find that

$$H(j\omega_0) = \frac{1}{k}\frac{1}{j2\zeta} = -j\frac{1}{2\zeta k} \tag{2.16}$$

This shows that the phase of the frequency response at $\omega = \omega_0$ is $\angle H(j\omega_0) = -90^\circ$, and moreover, that the magnitude is

$$|H(j\omega_0)| = \frac{1}{2\zeta k} \tag{2.17}$$

which is inversely proportional to the relative damping ζ.

2.1.3 Performance of a closed loop system

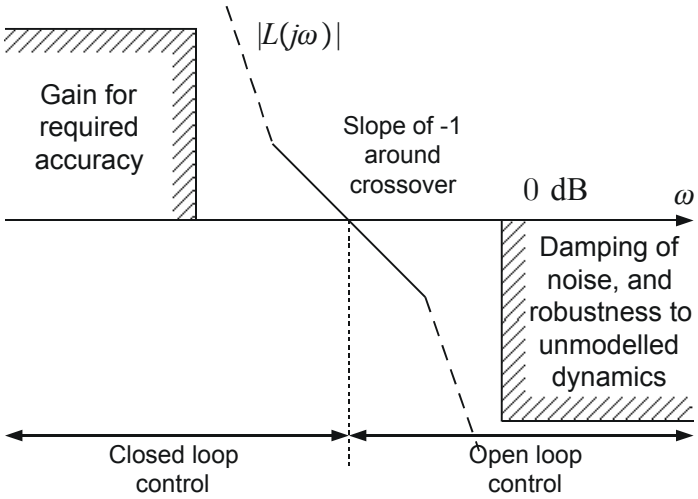

Figure 2.2: Performance requirements on the loop transfer function in a Bode diagram.

Frequency response techniques are well suited to specify the performance of a control system. This involves specifications on the loop transfer function $L(j\omega)$ in a Bode diagram. We recall that the magnitude of the sensitivity function $S(j\omega)$ will satisfy the approximations

$$|S(j\omega)| = \left| \frac{1}{1+L(j\omega)} \right| = \begin{cases} |L(j\omega)|^{-1}, & |L(j\omega)| \gg 1 \\ 1, & 1 \gg |L(j\omega)| \end{cases} \tag{2.18}$$

and that for unity feedback the magnitude of the closed-loop transfer function $T(j\omega)$ can be approximated by

$$|T(j\omega)| = \left|\frac{L(j\omega)}{1+L(j\omega)}\right| = \begin{cases} 1 & |L(j\omega)| \gg 1 \\ |L(j\omega)| & 1 \gg |L(j\omega)| \end{cases} \tag{2.19}$$

Typically, we would like the sensitivity $|S(j\omega)|$ to be small for low frequencies to reduce the effect of disturbances on the system in the low-frequency region. In addition, we would like $|T(j\omega)|$ to be small for high frequencies to reduce the influence of measurement noise and the influence of unmodeled dynamics. This implies that $|L(j\omega)|$ should be large for low frequencies, and small for high frequencies. In addition, there has to be a significant interval around the crossover frequency ω_c, defined by $|L(j\omega_c)| = 1$, where the phase should be around $\angle L(j\omega) \approx -90^\circ$ to ensure a sufficient phase margin. These requirements on the loop transfer function $L(j\omega)$ are indicated in Figure 2.2.

2.1.4 Stability margins

If the loop transfer function $L(s)$ is rational and has no poles with real part larger than zero, then the Bode-Nyquist criterion states that the system is stable if

$$|L(j\omega_{180})| < 1 \quad \text{and} \quad \angle L(j\omega_c) > -180^\circ. \tag{2.20}$$

This can be expressed as conditions on the *gain margin* ΔK and the *phase margin* ϕ as

$$\Delta K := \frac{1}{|L(j\omega_{180})|} > 1 \quad (= 0 \text{ dB}) \tag{2.21}$$

$$\phi := 180^\circ + \angle L(j\omega_c) > 0^\circ \tag{2.22}$$

where ω_c is the crossover frequency defined by $|L(j\omega_c)| = 1$, and ω_{180} is defined by $\angle L(j\omega_{180}) = -180^\circ$. It is possible to specify the performance of a control system around the crossover frequency in terms of the stability margins. This is often done by specifying a phase margin $\phi = 45^\circ$ and a gain margin $\Delta K = 6$ dB.

Example 23 *We note that for any system*

$$\begin{aligned} \Delta K &= 6 \text{ dB} \Rightarrow L(j\omega_{180}) = -\frac{1}{2} \\ &\Rightarrow S(j\omega_{180}) = 2 \text{ and } T(j\omega_{180}) = -1 \end{aligned} \tag{2.23}$$

We also note that for any system we have

$$\begin{aligned} \phi &= 45^\circ \Rightarrow L(j\omega_c) = -\frac{1}{2}\sqrt{2}(1+j) \\ &\Rightarrow |S(j\omega_c)| = |T(j\omega_c)| = 1.3 \end{aligned} \tag{2.24}$$

This may be acceptable if the desired value y_d and the disturbances are of low frequency. However, in high performance motion control like in robotics, the desired value y_d will have a significant frequency content close to the crossover frequency, and bearing in mind that

$$|y(j\omega)| = |T(j\omega)|\,|y_d(j\omega)| \tag{2.25}$$

we see that we will have an amplification of the desired value by a factor of 1.3 close to the crossover frequency ω_c. In robotics this could cause serious problems.

The lesson to be learned from this is that performance specifications on the gain margins are not directly related to the closed loop performance. Thus, for high performance systems it may be useful to study the functions $S(j\omega)$ and $T(j\omega)$ directly.

2.2 Elimination of fast dynamics

2.2.1 Example: The electrical time constant in a DC motor

Consider the following model of a DC motor:

$$T_m \frac{d\omega_m}{dt} = \frac{R_a}{K} i_a \tag{2.26}$$

$$T_a \frac{di_a}{dt} = -i_a - \frac{K}{R_a}\omega_m + \frac{1}{R_a} u_a \tag{2.27}$$

where ω_m is the motor speed, i_a is the armature current, u_a is the armature voltage, T_a is the electrical time constant, and T_m is the mechanical time constant defined by

$$T_a = \frac{L_a}{R_a}, \quad T_m = \frac{JR_a}{K^2} \tag{2.28}$$

The transfer function $H(s)$ from u_a to ω_m is found from the Laplace-transformed model

$$T_m s\omega_m(s) = \frac{R_a}{K} i_a(s) \tag{2.29}$$

$$(1 + T_a s)\, i_a(s) = -\frac{K}{R_a}\omega_m(s) + \frac{1}{R_a} u_a(s) \tag{2.30}$$

to be

$$H(s) = \frac{\omega_m}{u_a}(s) = \frac{1}{K} \frac{1}{1 + T_m s + T_a T_m s} \tag{2.31}$$

Suppose that $T_a \ll T_m$ so that $T_m \approx T_m + T_a$. Then the transfer function can be written

$$H(s) = \frac{1}{K} \frac{1}{(1 + T_m s)(1 + T_a s)} \tag{2.32}$$

Suppose that T_a is small, so that the break frequency $1/T_a$ is much higher than the frequency range where the model will be used. The the transfer function can be approximated with

$$H(s) = \frac{1}{K} \frac{1}{(1 + T_m s)} \tag{2.33}$$

which is obtained using the approximation

$$1 + T_a s \approx 1 \tag{2.34}$$

We will now discuss how the state-space model (2.26, 2.27) should be modified to reflect this approximation. This is best seen from (2.30) where the approximation (2.34) amounts to writing

$$i_a(s) = -\frac{K}{R_a}\omega_m(s) + \frac{1}{R_a} u_a(s) \tag{2.35}$$

This give the approximated state-space model

$$T_m \frac{d\omega_m}{dt} = \frac{R_a}{K} i_a \tag{2.36}$$

$$0 = -i_a - \frac{K}{R_a}\omega_m + \frac{1}{R_a} u_a \tag{2.37}$$

We can use the second equation to eliminate i using

$$i_a = -\frac{K}{R_a}\omega_m + \frac{1}{R_a}u_a \tag{2.38}$$

and the first equation becomes

$$T_m\frac{d\omega_m}{dt} = -\omega_m + \frac{1}{K}u_a \tag{2.39}$$

which is consistent with the simplified transfer function given by (2.33).

2.2.2 Nonlinear system

In the nonlinear case it is not possible to use frequency arguments to eliminate high frequency dynamics. In that case the corresponding model formulation is

$$\dot{x} = f(x, z, t, \epsilon) \tag{2.40}$$
$$\epsilon\dot{z} = g(x, z, t, \epsilon) \tag{2.41}$$

If ϵ is small so that $\epsilon\dot{z} \ll g(x, z, t, \epsilon)$, then it may be possible to approximate the system by inserting $\epsilon = 0$. This gives

$$\dot{x} = f(x, z, t, 0) \tag{2.42}$$
$$0 = g(x, z, t, 0) \tag{2.43}$$

which is a differential-algebraic system.

The differential-algebraic system may be written as a ordinary differential equation if

$$z = z(x, t) \tag{2.44}$$

is a solution to $0 = g(x, z, t, 0)$. In this case the system can be represented by the model

$$\dot{x} = f(x, z(x), t, 0) \tag{2.45}$$

This topic is discussed in great detail in (Khalil 1996)

2.3 Energy-based methods

2.3.1 Introduction

So far controller design based on state-space methods and frequency response has been discussed. These methods form the basis of any fundamental course on automatic control, and there is a wide range of methods, algorithms and software packages. An important formulation that complements state-space and frequency response methods is based on the use of balance laws, and in particular on the use of energy functions. To give the reader an indication on why this can be useful, we briefly consider the following examples: If we are studying robot control around a constant desired position, then the kinetic energy of the robot will increase if the robot is unstable, and the kinetic energy will be reduced if the robot is asymptotically stable. If we are applying active vibration control on a mechanical structure, then the vibration energy increases if the controlled system is unstable, while the energy decreases if the system is asymptotically stable.

On background of this it seems reasonable that the stability properties of a system may be related to the time derivative of some energy function of the system. In this section we will present results that are motivated from energy considerations for mechanical, electrical, and hydrostatic systems, and systems with and thermal flow. These model formulations are very useful in controller design and in analysis of control systems.

2.3.2 The energy function

We define an energy function $V(\mathbf{x}, t) \geq 0$ for the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t). \tag{2.46}$$

The function $V(\mathbf{x}, t)$ may be the total energy of the system, or it may be some other function, usually related to energy. When the system evolves the time derivative of the energy function $V(\mathbf{x}, t)$ is

$$\dot{V} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}, \mathbf{u}, t). \tag{2.47}$$

which follows from the standard rules of time differentiation of a function of two variables. We say that $\dot{V}$ is the time derivative along the solutions of the system. Information about the time derivative of the energy of a system may give valuable insight into properties of the dynamics of the system. In particular, if $\dot{V} \leq 0$, then the energy is monotonically decreasing, which may be important in connection with stability considerations. The analysis of energy functions and their time derivatives along the solutions of the system forms the basis for Lyapunov's stability theory (Slotine 1991), (Khalil 1996). This is an important tool in nonlinear control theory.

Example 24 *Consider the system*

$$\dot{x}_1 = x_2 \tag{2.48}$$
$$\dot{x}_2 = -\omega_0^2 x_1 - 2\zeta\omega_0 x_2 \tag{2.49}$$

and the energy function

$$V = \frac{1}{2}\omega_0^2 x_1^2 + \frac{1}{2} x_2^2 \tag{2.50}$$

The time derivative of V along the solutions of the system is

$$\dot{V} = \omega_0^2 x_1 x_2 + x_2 \left(-\omega_0^2 x_1 - 2\zeta\omega_0 x_2\right) \tag{2.51}$$

which gives

$$\dot{V} = -2\zeta\omega_0 x_2^2 \leq 0 \tag{2.52}$$

Note that the energy of the system decreases proportionally with the relative damping ζ whenever $x_2 \neq 0$.

2.3.3 Second-order systems

If a system is given as a second order system

$$\ddot{x} = f(x, \dot{x}, t) \tag{2.53}$$

then the time derivative of an energy function $V(x, \dot{x}, t)$ along the solutions of the system is

$$\dot{V} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\dot{x} + \frac{\partial V}{\partial \dot{x}} f(x, \dot{x}, t) \tag{2.54}$$

For the system in the Example 24 we could arrive at a second-order description by writing $x_1 = x$ and $x_2 = \dot{x}$. The dynamics could then be presented as the second order system

$$\ddot{x} = -\omega_0^2 x - 2\zeta\omega_0 \dot{x} \tag{2.55}$$

and we find the time derivative of $V = \frac{1}{2}\omega_0^2 x^2 + \frac{1}{2}\dot{x}^2$ along the solutions of the system to be

$$\dot{V} = \omega_0^2 x\dot{x} + \dot{x}\ddot{x} = \omega_0^2 x\dot{x} + \dot{x}\left(-\omega_0^2 x - 2\zeta\omega_0\dot{x}\right) = -2\zeta\omega_0\dot{x}^2 \tag{2.56}$$

This result is the same as the result in (2.52).

2.3.4 Example: Mass-spring-damper

Energy function

A mass m with position x is connected to a fixed point by a spring with spring constant k and a damper with damping constant d as shown in Figure 2.1. The equation of motion is

$$m\ddot{x} + d\dot{x} + kx = 0 \tag{2.57}$$

The potential energy of this system is $U = \frac{1}{2}kx^2$, while the kinetic energy is $T = \frac{1}{2}m\dot{x}^2$. The total energy is

$$V = T + U = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 \tag{2.58}$$

The time derivative of the energy function is

$$\dot{V} = m\dot{x}\ddot{x} + kx\dot{x} \tag{2.59}$$

The time derivative for solutions of the system is found by inserting the equation of motion (2.57). This gives

$$\begin{aligned} \dot{V} &= \dot{x}\left(-d\dot{x} - kx\right) + kx\dot{x} \\ &= -d\dot{x}^2 \end{aligned} \tag{2.60}$$

Two observations are important at this point. First, $\dot{V} < 0$, which means that the energy is not increasing, and second, the energy decreases because of power dissipated in the damper.

From the expression of the energy we see that

$$x(t) \leq \sqrt{\frac{2}{k}V}, \quad \dot{x} \leq \sqrt{\frac{2}{m}V} \tag{2.61}$$

This means that if the energy V decreases to zero, then also the position x and the velocity $\dot{x}$ will decrease to zero. Moreover, because V decreases, $V(t)$ will be less than the initial value V_0. Therefore

$$x(t) \leq \sqrt{\frac{2}{k}V_0}, \quad \dot{x}(t) \leq \sqrt{\frac{2}{m}V_0} \tag{2.62}$$

This means that if the initial energy is small, then also the position and velocity will remain small.

Friction

Now, consider the mass-spring-damper system with a friction force, so that

$$m\ddot{x} + d\dot{x} + kx = -F_f \tag{2.63}$$

Then the time derivative of the energy function is

$$\dot{V} = -F_f\dot{x} - d\dot{x}^2 \tag{2.64}$$

which has two terms, the power $F_f\dot{x}$ dissipated by the system by the friction and the power $d\dot{x}^2$ dissipated in the damper. In its very physical nature, friction work transfers kinetic energy to heat energy. This means that the friction work will decrease the total energy of the system, and it follows that

$$F_f\dot{x} \geq 0 \tag{2.65}$$

and, accordingly,

$$\dot{V} \leq -d\dot{x}^2 \tag{2.66}$$

External force

Suppose that the mass-spring-damper system is actuated with an input force F. Then the equation of motion is

$$m\ddot{x} + d\dot{x} + kx = F \tag{2.67}$$

The time derivative of the energy function is found to be

$$\dot{V} = F\dot{x} - d\dot{x}^2 \tag{2.68}$$

Here the term $F\dot{x}$ is the power that is supplied to the system due to the force F. We see that if $F\dot{x} < 0$, then the energy V will be decreasing.

Example 25 *If F is supplied from a controller, we see that negative velocity feedback*

$$F = -K_d\dot{x} \tag{2.69}$$

will give

$$\dot{V} = -\left(K_d + d\right)\dot{x}^2 \tag{2.70}$$

It is seen that the feedback gain K_d appears as a damping coefficient.

2.3.5 Lyapunov methods

In Lyapunov design for the plant

$$\dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \mathbf{u}\right) \tag{2.71}$$

the main idea is to select a suitable energy function $V\left(\mathbf{x}\right)$, called a Lyapunov function candidate, which is positive definite in the state vector $\mathbf{x}$ in the sense that $V\left(\mathbf{x}\right) = 0$ for $\mathbf{x} = \mathbf{0}$, and $V\left(\mathbf{x}\right) > 0$ for $\mathbf{x} \neq \mathbf{0}$. A typical Lyapunov function candidate is

$$V\left(\mathbf{x}\right) = \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} \tag{2.72}$$

where $\mathbf{P}$ is a positive definite and symmetric matrix. Then

$$\frac{\lambda_{\min}(\mathbf{P})}{2}\mathbf{x}^T\mathbf{x} \leq V(\mathbf{x}) \leq \frac{\lambda_{\max}(\mathbf{P})}{2}\mathbf{x}^T\mathbf{x} \tag{2.73}$$

where $\lambda_{\min}(\mathbf{P}) > 0$ is the smallest eigenvalue of $\mathbf{P}$, and $\lambda_{\max}(\mathbf{P})$ is the largest eigenvalue of $\mathbf{P}$. A control $\mathbf{u}$ is sought to ensure that the function $V(\mathbf{x})$ decreases to zero, which implies that the state vector converges to zero. There is no general method for selecting the Lyapunov function candidate $V(\mathbf{x})$, but for many applications it is possible to select $V(\mathbf{x})$ from some type of energy function.

Example 26 *We investigate the problem of controlling the position of a mass m with position x. The mass is actuated by a force u, and the desired position is $x_d = 0$. The kinetic energy of the mass is $T = \frac{1}{2}m\dot{x}^2$. We may reason as follows: Suppose that a spring with spring constant k_p was fixed to the mass, and, in addition, that a viscous damper with damping constant k_d was fixed to the mass. Then the system would obviously be stable. The potential energy of this spring would be $\frac{1}{2}k_px^2$. Now, if a spring and a damper will stabilize the mass, why not let the force input u set up the same force as the spring and the damper? We accordingly select the control to be the PD controller*

$$u = -k_px - k_d\dot{x} \tag{2.74}$$

and get the closed loop dynamics

$$m\ddot{x} + k_d\dot{x} + k_px = 0. \tag{2.75}$$

This means that the controller defines a virtual spring and a virtual damper. We define the Lyapunov function candidate to be the sum of the kinetic energy of the mass and the potential energy of the virtual spring:

$$V = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}k_px^2 \tag{2.76}$$

The time derivative of V along the solutions of the system (2.75) is

$$\begin{aligned} \dot{V} &= \dot{x}m\ddot{x} + k_px\dot{x} \\ &= -\dot{x}(k_d\dot{x} + k_px) + k_px\dot{x} \\ &= -k_d\dot{x}^2 \end{aligned} \tag{2.77}$$

We see that whenever $\dot{x} \neq 0$, then V will decrease, and it can be shown that V will tend to zero. Then, because m and k_p are positive constants, this implies that x and $\dot{x}$ will tend to zero.

2.3.6 Contraction

We have introduced energy functions to study the stability of nonlinear system around an equilibrium point by calculating the time derivative of the energy function for solutions of the system. A slightly different view is taken in *contraction analysis* (Hartman 1982, p. 537), (Lohmiller and Slotine 1998) where the convergence of different solutions to each other is studied. We will look at two different solutions $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ for the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$$

where the initial conditions are $\mathbf{x}_1(t_0) = \mathbf{x}_{10}$ and $\mathbf{x}_2(t_0) = \mathbf{x}_{20}$. We consider the energy function

$$V = \frac{1}{2}(\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2)$$

which can be seen as a measure of how far the two solutions are from each other. The time derivative of V along the solutions of the system is

$$\dot{V} = (\mathbf{x}_1 - \mathbf{x}_2)^T [\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)]$$

Define the Jacobian matrix

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$$

and consider the line $\alpha\mathbf{x}_1(t) + (1-\alpha)\mathbf{x}_2(t)$, $0 \leq \alpha \leq 1$ between $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$. On this line we have

$$\frac{d}{d\alpha}\mathbf{f}[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2] = \mathbf{J}[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2](\mathbf{x}_1 - \mathbf{x}_2)$$

and we may write

$$\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2) = \int_0^1 \mathbf{J}[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2]\, d\alpha(\mathbf{x}_1 - \mathbf{x}_2)$$

Because of this, the time derivative of V can be expressed as

$$\dot{V} = -(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{Q}(\mathbf{x}_1 - \mathbf{x}_2) \tag{2.78}$$

where

$$\mathbf{Q} = -\frac{1}{2}\int_0^1 \left\{\mathbf{J}[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2] + \mathbf{J}^T[\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2]\right\} d\alpha \tag{2.79}$$

We see that whenever

$$\mathrm{Re}\left[\lambda_i\left(\mathbf{J} + \mathbf{J}^T\right)\right] < 0 \tag{2.80}$$

then $\mathbf{Q}$ is positive definite, and it follows that $\dot{V} \leq 0$. This means that the two solutions $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ will converge to each other as time goes to infinity. As the two solutions were selected freely, this implies that any two solutions will converge to each other as time goes to infinity.

2.3.7 Energy flow in a turbocharged diesel engine

Diesel engines are usually equipped with turbochargers (Heywood 1988), (Kiencke and Nielsen 2000) as shown in Figure 2.3. A turbocharger has a turbine that is driven by the exhaust, and, on the same shaft, a compressor that increases the pressure of the inlet air to the motor. The purpose of this arrangement is to increase the pressure and thereby increase the density of the air into the cylinder. The benefit of this is that by increasing the mass of fresh air into the cylinder, it is possible to increase the amount of injected diesel fuel while still having sufficient oxygen to achieve satisfactory combustion of the fuel. This increases the energy that can be processed in a fixed cylinder volume.

A diesel engine with a turbocharger is a complicated and nonlinear system, still, the energy flow is easy to model and important for understanding the dynamics of the system. The required modeling tools for this is presented in Chapter 12. Energy is

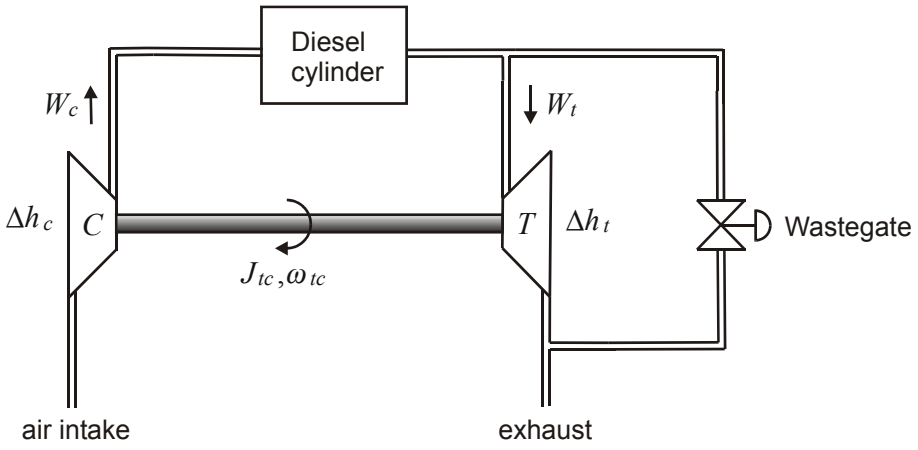

Figure 2.3: Diesel engine with turbocharger

exchanged partly as thermal energy with power $P = wh$ where w is mass flow and h is specific enthalpy, and partly as rotation power $P = \tau\omega$ where τ is torque and ω is angular velocity. The speed ω_{tc} of the turbocharger shaft depends on how much energy that is absorbed by the turbine compared to how much energy that is used to compress the air in the compressor. In other words, the time derivative of the kinetic energy of the turbocharger shaft is equal to the thermal power which is converted to kinetic energy in the turbine, minus the thermal power that is required in the compressor. This is written

$$\frac{d}{dt}\left[\frac{1}{2}J_{tc}\omega_{tc}^2\right] = w_t\Delta h_t - w_c\Delta h_c \tag{2.81}$$

where J_{tc} is the moment of inertia of the combined shaft of the turbine and the compressor, w_t is the mass flow through the turbine, Δh_t is the reduction in specific enthalpy over the turbine, w_c is the mass flow through the compressor, and Δh_c is the increase in specific enthalpy over the compressor.

If the energy carried by the exhaust becomes to high, then the turbocompressor may over-speed. This problem can be avoided by opening a wastegate valve for dumping power from the system. The effect of this is to reduce w_t. Alternatively, inlet guidevanes on the turbine can be used to control the amount of energy that is absorbed by the turbine. The effect of this is to reduce the change of specific enthalpy Δh_t in the turbine. During acceleration it may be desirable to speed up the turbocharger to achieve sufficiently high pressure of the injected air. This can be done by closing the wastegate, or by adjusting the inlet guide vanes for a higher Δh_t. The use of inlet guide vanes gives a faster response than the use of a wastegate, and has become the preferred solution in car engines.

2.4 Passivity

2.4.1 Introduction

The concept of passivity is very useful in control systems analysis and design. Passivity theory provides a set of analysis tools that can be used for a wide range of physical systems with commonly used controllers like the PID controller, and, in addition, for nonlinear

controllers based on adaptive techniques and backstepping. The main observation in passivity theory is: If a system can be described as a parallel or feedback interconnection of passive subsystems, then the total system will be passive, and it will not generate energy.

One practical application of the theory of passivity is within stability analysis based on energy-flow considerations for interconnections of physical subsystems. In this setting, stability can be related to a decreasing total energy of the system. Now, suppose that each physical subsystem is passive in the sense that it can store and dissipate energy, but it cannot produce energy. Then it can be concluded that the total energy of the system will decrease, which under certain assumptions implies that the system is stable. The theory of passivity also provides a generalization of the pure energy-based analysis of interconnected physical subsystems to the analysis of interconnections of general dynamic subsystems, where the storage and dissipation of other functions than energy are studied.

Methods based on passivity can be used for linear time-invariant systems where useful properties of certain transfer functions can be established from simple energy considerations. This leads to very efficient tools for stability analysis. Moreover, passivity is very useful for nonlinear systems as an extension of Lyapunov analysis in the sense that Lyapunov results for an interconnection of subsystems can be inferred from the passivity properties of the individual subsystems.

2.4.2 Definition

The following definition of passivity will be used:

Definition 1 *Consider a system with input u and output y. Suppose that that there is a constant $E_0 \geq 0$ so that for all control time histories u and all $T \geq 0$ the integral of $y(t)\,u(t)$ satisfies*

$$\int_0^T y(t)\,u(t)\,dt \geq -E_0 \tag{2.82}$$

Then the system is said to be passive.

Some remarks can be made to this definition.

1. The definition is based on an input-output description where the input is u and the output is y. There is no reference to the state of the system.
2. Note that it is the system with a specified input u and a specified output y that is passive. Passivity cannot be defined without defining input and output.
3. The definition is valid for both linear and nonlinear systems, and for time-varying and time-invariant systems.
4. If the system is passive with input u and output y, then it will also be passive with input y and output u.

2.4.3 Examples

Time constant

A time constant given by

$$\dot{y} = -ay + u \tag{2.83}$$

is passive if $a \geq 0$. This is seen by inserting $u = \dot{y} + ay$ into the integral of yu, which gives

$$\begin{aligned}\int_0^T yudt &= \int_0^T y(\dot{y} + ay)dt \\ &= \int_{y(0)}^{y(T)} ydy + a\int_0^T y^2 dt \\ &= \frac{1}{2}y^2(T) - \frac{1}{2}y^2(0) + a\int_0^T y^2 dt \end{aligned} \tag{2.84}$$

The first and last term on the right side are positive. It follows that

$$\int_0^T yudt \geq -\frac{1}{2}y^2(0) \tag{2.85}$$

and the system has been shown to be passive with $E_0 := (1/2)y^2(0)$.

Mass, spring and damper

A mass m with position x is connected with a spring and a damper to a fixed point. The equation of motion is

$$m\ddot{x} + B\dot{x} + Kx = F \tag{2.86}$$

where F is the control force acting on the mass. Then the system with input F and output $\dot{x}$ is passive. This is seen from

$$\begin{aligned}\int_0^T F\dot{x}dt &= \int_0^T \left(m\ddot{x} + B\dot{x} + Kx\right)\dot{x}dt \\ &= \int_0^T m\dot{v}vdt + \int_0^T Bv^2 dt + \int_0^T Kx\dot{x}dt \\ &= \frac{1}{2}m\left[v^2(T) - v^2(0)\right] + B\int_0^T v^2 dt + \frac{1}{2}K\left[x^2(T) - x^2(0)\right] \qquad (2.87) \\ &\geq -E_0 + B\int_0^T v^2 dt \qquad (2.88)\end{aligned}$$

where

$$E_0 = \frac{1}{2}mv^2(0) + \frac{1}{2}Kx^2(0) \tag{2.89}$$

is the initial energy in the form of kinetic and potential energy. It follows that

$$\int_0^T F\dot{x}dt \geq -E_0 \tag{2.90}$$

which shows that the system is passive when F is input and $\dot{x}$ is output. Note that the product $F\dot{x}$ between the input and the output is the power supplied to the mass because of the control input F. Moreover, note that the constant E_0 in the passivity inequality (2.90) is the initial energy in the system. It seems intuitively right to describe this type of system as passive as the systems has only passive components in the form of a mass, a spring and a damper. In particular, there is no active element in the system that can produce energy.

Electrical circuit

Consider an electrical circuit with input voltage u, which is the control input, and current i, which is the system output. The circuit is a serial interconnection of a resistor R, a capacitor C and an inductance L. The voltage law gives

$$u = Ri + \frac{1}{C}q + L\frac{di}{dt} \tag{2.91}$$

where q is the capacitor charge which satisfies $\dot{q} = i$. Then passivity from u to i is shown by the calculation

$$\begin{aligned} \int_0^T uidt &= R\int_0^T i^2dt + \frac{1}{C}\int_0^t \dot{q}qdt + L\int_0^t i\frac{di}{dt}dt \\ &= R\int_0^T i^2dt + \frac{1}{2C}\left[q^2(T) - q^2(0)\right] + \frac{L}{2}\left[i^2(T) - i^2(0)\right] \\ &\geq -E_0 + R\int_0^T i^2dt \end{aligned} \tag{2.92}$$

where

$$E_0 = \frac{1}{2C}q^2(0) + \frac{L}{2}i^2(0) \tag{2.93}$$

is the initially stored energy in the circuit. Note that the product ui between the input and the output is the power supplied to the circuit from the control input u. It is also interesting to note that if the current had been take to be the control variable, and the voltage had been the measurement, then the system would still be passive.

2.4.4 Energy considerations

We may think of passivity as a property related to balance equations, and in particular to energy conservation, which we will use to illustrate the meaning of the concept of passivity. Consider a system with input u and output y. Suppose that the product $u(t)\,y(t)$ has the physical dimension of power, and $\int_0^T u(t)\,y(t)\,dt$ is the energy that is supplied to the system due to the control action u. A critical observation is:

- If $\int_0^T u(t)\,y(t)\,dt \geq 0$ for all control histories u and for all $T \geq 0$, then energy is absorbed by the system, and the system cannot supply any energy to the outside. In this case the system is passive according to Definition 1.

- If there exists some $E_0 > 0$ so that the integral $\int_0^T u(t)\,y(t)\,dt \geq -E_0$ for all control histories u and for all $T \geq 0$, then the system may supply a limited quantity of energy to the outside. The energy will typically be energy due to the initial conditions of energy storage elements, which in mechanical systems may be potential energy in springs and kinetic energy of masses, while in electrical systems it will be energy stored in capacitors and inductances. According to Definition 1 the system is passive, which makes sense as the system can only store energy received from the outside, but it cannot produce energy.

- If it is possible to find a control history u so that the integral $\int_0^T u(t)\,y(t)\,dt$ may tend to $-\infty$ for some $T \geq 0$, then this means that the system may supply an unlimited amount of energy to the outside. This is only possible if there is an inexhaustible source of energy inside the system. This agrees with the fact that the system is not passive according to Definition 1.

2.4.5 Positive real transfer functions

It turns out that a system is passive if and only if the transfer function from input to output is *positive real*. This result, which is very useful, will be developed in the following. First the concept of positive real transfer functions will be defined, and a special conditions for rational transfer functions is presented. Then the connection to passivity will be demonstrated. We start by defining positive real transfer functions.

Definition 2 *The rational or irrational function $H(s)$ is positive real if*
1. $H(s)$ is analytic for all $\mathrm{Re}\,[s] > 0$.
2. $H(s)$ is real for all positive and real s.
3. $\mathrm{Re}\,[H(s)] \geq 0$ for all $\mathrm{Re}\,[s] > 0$.

It is emphasized that at this stage in the presentation there is no obvious physical interpretation of this definition. Note that the definition is based on the properties for the transfer function $H(s)$ for $\mathrm{Re}\,[s] > 0$, which is to the right of the imaginary axis.

2.4.6 Positive real rational transfer functions

In the case of rational transfer functions it is convenient to investigate the properties of the transfer function on the imaginary axis by working with $H(j\omega)$. This makes it easier to check if a transfer function is positive real, and it leads to a more intuitive understanding of the positive real property. Now, suppose that the transfer function $H(s)$ is rational. In this case Statement 1 of Definition 2 implies that there are no poles to the right of the imaginary axis. Concerning Statement 3, the result can be formulated on the imaginary axis by observing that the transfer function will be continuous at all s except at the poles. This means that as long as $j\omega$ is not a pole of $H(s)$ the transfer function, then

$$H(j\omega) = \lim_{\substack{\sigma \to 0 \\ \sigma > 0}} H(\sigma + j\omega) \tag{2.94}$$

This implies that $\mathrm{Re}[H(j\omega)] \geq 0$ as long as $j\omega$ is not a pole of $H(s)$. These arguments provide a partial explanation of the following important result:

The rational function $H(s)$ is positive real if and only if
1. All the poles of $H(s)$ have real parts less than or equal to zero.
2. $\mathrm{Re}\,H\,(j\omega) \geq 0$ for all ω so that $j\omega$ is not a pole of $H(s)$.
3. If $j\omega_0$ is pole in $H(s)$, then it is a simple pole, and

$$\mathrm{Res}_{s=j\omega_0}[H(s)] = \lim_{s \to j\omega_0} (s - j\omega_0) H(s) \tag{2.95}$$

is real and positive. If $H(s)$ has a pole at infinity, then it is a simple pole, and

$$R_\infty = \lim_{\omega \to \infty} \frac{H(j\omega)}{j\omega} \tag{2.96}$$

exists, and is real and positive.

The derivation of this result is found in (Anderson and Vongpanitlerd 1973) and (Lozano, Brogliato, Egeland and Maschke 2000).

Example 27 *A time constant has transfer function*

$$H(s) = \frac{1}{1+Ts} \tag{2.97}$$

The frequency response is

$$H(j\omega) = \frac{1}{1+j\omega T} = \frac{1-j\omega T}{1+(\omega T)^2} \tag{2.98}$$

and we see that

$$\mathrm{Re}\, H(j\omega) = \frac{1}{1+(\omega T)^2} > 0 \tag{2.99}$$

In addition, the only pole has a negative real part, and it follows that the transfer function of a time constant is positive real.

Example 28 *Consider the transfer function*

$$H(s) = \frac{s+c}{(s+a)(s+b)} \tag{2.100}$$

where a, b and c are constants greater than zero. Both poles of the transfer function have negative real parts. Then

$$\begin{aligned} H(j\omega) &= \frac{j\omega+c}{(j\omega+a)(j\omega+b)} = \frac{(c+j\omega)(a-j\omega)(b-j\omega)}{(a^2+\omega^2)(b^2+\omega^2)} \\ &= \frac{abc+\omega^2(a+b-c)+j[\omega(ab-ac-bc)-\omega^3]}{(a^2+\omega^2)(b^2+\omega^2)} \end{aligned} \tag{2.101}$$

We find that if $c \leq a+b$, then $\mathrm{Re}[h_2(j\omega)] > 0$ for all ω, and the transfer function is positive real. If $c > a+b$, then $h_2(s)$ is not positive real as $Re[h_2(j\omega)] < 0$ for $\omega > \sqrt{abc/(c-a-b)}$.

Example 29 *The transfer function $H(s) = Ls$ where $L > 0$ has the frequency response $H(j\omega) = j\omega L$, so that $\mathrm{Re}[H(j\omega)] = 0$. The transfer function has only one pole, which is at infinity. As*

$$R_\infty = \lim_{\omega\to\infty} \frac{j\omega L}{j\omega} = L \tag{2.102}$$

is real and positive, it follows that the transfer function is positive real.

Example 30 *Consider the transfer function*

$$H(s) = \frac{s}{s^2+\omega_0^2}, \quad \omega_0 > 0 \tag{2.103}$$

All the poles are simple poles on the imaginary axis in $s = \pm j\omega_0$. The frequency response is

$$H(j\omega) = \frac{j\omega}{\omega_0^2-\omega^2} \tag{2.104}$$

so that $\mathrm{Re}[H(j\omega)] = 0$. The residuals at the poles on the imaginary axis are

$$\mathrm{Res}_{s=\pm j\omega_0} H(s) = \mathrm{Res}_{s=\pm j\omega_0} \frac{s}{(s+j\omega_0)(s-j\omega_0)} = \frac{1}{2} \tag{2.105}$$

The residuals are real and positive. The transfer function is therefore positive real.

Example 31 *Consider the transfer function*

$$H(s) = \frac{s^2 + a^2}{s(s^2 + \omega_0^2)}, \quad a > 0,\ \omega_0 > 0 \tag{2.106}$$

All the poles are simple poles on the imaginary axis in $s = 0$ *and* $s = \pm j\omega_0$*. The frequency response is*

$$H(j\omega) = -j\frac{a^2 - \omega^2}{\omega(\omega_0^2 - \omega^2)} \tag{2.107}$$

so that $\mathrm{Re}[H(j\omega)] = 0$*. The residuals at the poles on the imaginary axis are*

$$\mathrm{Res}_{s=0} H(s) = \frac{a^2}{\omega_0^2}, \quad \mathrm{Res}_{s=\pm j\omega_0} H(s) = \frac{\omega_0^2 - a^2}{2\omega_0^2} \tag{2.108}$$

The residuals are real and positive and the transfer function is positive real if and only if $a < \omega_0$*.*

Example 32 *Consider a proper and rational transfer function*

$$H(s) = \frac{(s + z_1)(s + z_2)\dots}{s(s + p_1)(s + p_2)\dots} \tag{2.109}$$

where $\mathrm{Re}[p_i] > 0$ *and* $\mathrm{Re}[z_i] > 0$*. Then,* $H(s)$ *is positive real if and only if* $\mathrm{Re}[H(j\omega)] \geq 0$ *for all* $\omega \neq 0$*. This follows from*

$$\mathrm{Res}_{s=0} H(s) = \frac{z_1 z_2 \dots}{p_1 p_2 \dots} > 0 \tag{2.110}$$

and from the observation that the $H(s)$ *has one pole in the origin while the remaining poles have negative real parts.*

2.4.7 Positive realness of irrational transfer functions

Introduction

Irrational transfer functions are obtained for systems described by partial differential equations. Consider a linear system $y(s) = H(s)u(s)$ with a irrational transfer function $H(s)$. As for rational transfer functions the system with input u and output y is passive if and only if the transfer function $H(s)$ is positive real (Anderson and Vongpanitlerd 1973). For irrational transfer functions we have to study the properties of the transfer function in the right half plane.

Example: Transmission line

To actuate a valve on the seafloor a hydraulic transmission line can be used. This is a pipe of length L filled with oil. The output side of the transmission line is connected to the valve at the seafloor, while the pressure is controlled on the input side of the pipe. Then the control variable is the volumetric flow q_1 at the input side, and the measurement is the pressure p_1 on the input side. The system can be described with the transfer function

$$\frac{p_1}{q_1}(s) = H_1(s) = \tanh s = \frac{\sinh s}{\cosh s} = \frac{e^s - e^{-s}}{e^s + e^{-s}} \tag{2.111}$$

First we check if the transfer function is analytic in the right half plane. We find that

$$\begin{aligned} e^s + e^{-s} &= 0 \Rightarrow e^{2s} = -1 \\ &\Rightarrow \quad s = j\frac{\pi + 2k\pi}{2}, \quad k = 0, \pm 1, \pm 2, \end{aligned} \tag{2.112}$$

This means that $h(s)$ is analytic for all $\mathrm{Re}\,[s] > 0$. It is trivial that $H_1(s)$ is real for all positive and real s. We then have to establish that $\mathrm{Re}\,[H_1(s)] \geq 0$ for all $\mathrm{Re}\,[s] > 0$ to show that $H_1(s)$ is positive real. To do this we introduce $\sigma = \mathrm{Re}\,[s]$ and $\omega = \mathrm{Im}\,[s]$, so that $s = \sigma + j\omega$, and, using

$$\sinh(\sigma + j\omega) = \sinh\sigma\cos\omega + j\cosh\sigma\sin\omega \tag{2.113}$$

$$\cosh(\sigma + j\omega) = \cosh\sigma\cos\omega + j\sinh\sigma\sin\omega. \tag{2.114}$$

we find that

$$\mathrm{Re}\,[\tanh s] = \frac{\sinh\sigma\cosh\sigma}{|\cosh s|^2} > 0 \quad \text{for all } \sigma > 0. \tag{2.115}$$

We have then showed that $H_1(s) = \tanh s$ is a positive real transfer function.

The transfer function from the pressure p_1 on the input side to the pressure p_2 on the output side is

$$\frac{p_2}{p_1}(s) = H_2(s) = \frac{1}{\cosh s} \tag{2.116}$$

This transfer function is not positive real as the real part of the transfer function is

$$\mathrm{Re}\left[\frac{1}{\cosh s}\right] = \frac{\cosh\sigma\cos\omega}{|\cosh s|^2} \tag{2.117}$$

It follows that for all ω so that $\cos\omega < 0$ the real part will be negative for all $\sigma > 0$.

2.4.8 Passivity and positive real transfer functions

We have the following result:

A linear time-invariant system with input u and output y described with the transfer function model $y(s) = H(s)u(s)$ is passive if and only if the transfer function $H(s)$ is positive real.

To demonstrate that passivity and positive realness are related, we consider the linear system

$$y(s) = H(s)u(s) \tag{2.118}$$

with rational and strictly proper transfer function

$$H(s) = K\frac{(s + z_1)(s + z_2)\ldots(s + z_m)}{(s + p_1)(s + p_2)\ldots(s + p_n)} \tag{2.119}$$

We will now show that passivity is related to the positive realness of the transfer function $H(s)$.

2.4.9 No poles on the imaginary axis

First it is assumed that $\mathrm{Re}[p_i] > 0$, which means that the system is stable, and that all poles are to the left of the imaginary axis. Suppose that the input is

$$u(t) = U \sin \omega_0 t \tag{2.120}$$

Then the output is

$$y(t) = U \left| H\left(j\omega_0\right)\right| \sin\left[\omega_0 t + \angle H\left(j\omega_0\right)\right] + y_t(t) \tag{2.121}$$

where $y_t(t)$ is the transient part of the output. Then the product yu is found to be

$$y\left(t\right)u(t) = \frac{U^2}{2}\,\mathrm{Re}\,H\left(j\omega_0\right) - \frac{U^2}{2}\left|H\left(j\omega_0\right)\right| \cos\left[2\omega_0 t + \angle H\left(j\omega_0\right)\right] + y_t(t) U \sin\omega_0 t \tag{2.122}$$

Integration gives

$$\begin{aligned}\int_0^T y\left(t\right)u(t)dt &= \frac{U^2 T}{2}\,\mathrm{Re}\,H\left(j\omega_0\right) + \frac{U^2}{4\omega_0}\left|H\left(j\omega_0\right)\right| \sin\left[2\omega_0 t + \angle H\left(j\omega_0\right)\right] \\ &\quad + \int_0^T y_t\left(t\right)u(t)dt \end{aligned} \tag{2.123}$$

The second term on the right side will be a sinusoidal signal that is bounded by its amplitude. In the third term on the right side the transient signal $y_t(t)$ will tend exponentially to zero. This leads to

$$\left| + \frac{U^2}{4\omega_0}\left|H\left(j\omega_0\right)\right| \sin\left[2\omega_0 t + \angle H\left(j\omega_0\right)\right] + \int_0^T y_t\left(t\right)u(t)dt \right| \leq E_0 \tag{2.124}$$

for some constant $E_0 \geq 0$. This implies that

$$\left| \int_0^T y\left(t\right)u(t)dt - \frac{U^2 T}{2}\,\mathrm{Re}\,H\left(j\omega_0\right) \right| \leq E_0 \tag{2.125}$$

for all $T \geq 0$. From this result it is seen that the system is passive if and only if $\mathrm{Re}\,H\left(j\omega\right) \geq 0$ for all ω. The if part is obvious. Concerning the only if part, it is seen that if $\mathrm{Re}\,H\left(j\omega_0\right) < 0$ for some ω_0, then there is no lower bound on $\int_0^T y\left(t\right)u(t)dt$ as $\left|U^2 T\,\mathrm{Re}\,H\left(j\omega_0\right)/2\right|$ can be made arbitrarily large by selecting T sufficiently large.

2.4.10 Single poles at the imaginary axis

Assume that the system has all poles to the left of the imaginary axis except two simple poles at $s = \pm ja$ on the imaginary axis. A partial fraction expansion gives

$$\begin{aligned} H(s) &= \frac{\mathrm{Res}_{s=\pm ja} H(s)}{s + ja} + \frac{\mathrm{Res}_{s=\pm ja} H(s)}{s - ja} + G(s) && (2.126)\\ &= 2\frac{s\mathrm{Res}_{s=\pm ja} H(s)}{s^2 + a^2} + G(s) && (2.127)\end{aligned}$$

where $G(s)$ is due to the poles to the left of the imaginary axis. Then, if $\omega_0 \neq a$, the results (2.121) and (2.125) are still valid. If $\omega_0 = a$, then

$$y(s) = H(s)u(s) = 2\frac{s\omega_0 \mathrm{Res}_{s=\pm j\omega_0} H(s)}{\left(s^2 + \omega_0^2\right)^2} + \ldots \tag{2.128}$$

which corresponds to the time function

$$y(t) = t\sin(\omega_0 t)\,\mathrm{Res}_{s=\pm j\omega_0} H(s) + \ldots \tag{2.129}$$

and it follows that

$$\int_0^T y(t)\,u(t)dt = \int_0^T tU\sin^2(\omega_0 t)\,\mathrm{Res}_{s=\pm j\omega_0} H(s)dt + \ldots \tag{2.130}$$

Finally, assume that the system has a simple pole at the origin $s = 0$. Then

$$H(s) = \frac{\mathrm{Res}_{s=0} H(s)}{s} + G_0(s) \tag{2.131}$$

where $G(s)$ is due to the poles to the left of the imaginary axis. If $\omega_0 \neq 0$, the results (2.121) and (2.125) are still valid. If $u(t) = U$, then

$$y(s) = H(s)u(s) = \frac{U\mathrm{Res}_{s=\pm j\omega_0} H(s)}{s^2} + \ldots \tag{2.132}$$

and the time function is

$$y(t) = t\mathrm{Res}_{s=0} H(s) + \ldots \tag{2.133}$$

This gives

$$\int_0^T y(t)\,u(t)dt = \int_0^T t^2\mathrm{Res}_{s=0} H(s)dt + \ldots \tag{2.134}$$

It may then be concluded that if the system has a simple pole in $s = j\omega_0$ at the imaginary axis, then the system is passive if and only if the residual $\mathrm{Res}_{s=j\omega_0}[H(s)]$ is real and positive.

2.4.11 Bounded real transfer functions

Definition 3 *The rational or irrational function $B(s)$ is bonded real if*
1. $B(s)$ is analytic for all $\mathrm{Re}\,[s] > 0$.
2. $|B(s)| \leq 1$ for all positive and real s

The transfer function

$$B(s) = \frac{H(s) - 1}{H(s) + 1} \tag{2.135}$$

is bounded real if and only if $H(s)$ is positive real. This is shown as follows:

Because $H(s)$ is analytic and $\mathrm{Re}\,[H(s)] \geq 0$ in $\mathrm{Re}\,[s] > 0$ it follows that $B(s)$ is analytic in $\mathrm{Re}\,[s] > 0$. It is assumed that $B(s) \neq \pm 1$ in $\mathrm{Re}\,[s] > 0$. Solving for $H(s)$ we get

$$H(s) = \frac{1 + B(s)}{1 - B(s)} \tag{2.136}$$

Then, as $H(s)$ is analytic in $\mathrm{Re}\,[s] > 0$, it follows that $B(s) \neq 1$ in $Re[s] > 0$. Consider the following calculation:

$$\mathrm{Re}\,H(s) = \frac{1}{2}\left[H(s) + H^*(s)\right] = \frac{1}{2}\frac{1 + B(s)}{1 - B(s)} + \frac{1}{2}\frac{1 + B^*(s)}{1 - B^*(s)} \tag{2.137}$$

$$= \frac{1 - B(s)B^*(s)}{\left[1 - B(s)\right]\left[1 - B^*(s)\right]} \tag{2.138}$$

From this computation it is seen that $\mathrm{Re}\,[H(s)] \geq 0$ for all $\mathrm{Re}\,[s] > 0$ if and only if $|B(s)| \leq 1$ in $\mathrm{Re}\,[s] > 0$.

Example 33 *Suppose that the transfer function $H(s)$ is positive real. Then the inverse $H^{-1}(s)$ is positive real. Statement 2 of Definition 2 is trivial in this case. Statements 1 and 3 are shown as follows: It is the possible to conclude from the maximum modulus theorem that $B(s) \neq -1$ in $\mathrm{Re}\,[s] > 0$, and we may express the inverse of $H(s)$ as*

$$G(s) = H^{-1}(s) = \frac{1 - B(s)}{1 + B(s)} \tag{2.139}$$

which is analytic in $\mathrm{Re}\,[s] > 0$ because $B(s)$ is analytic and $B(s) \neq -1$ in this region. This proves Statement 1 for $G(s)$. Finally, statement 3 for $G(s)$ is verified from

$$\begin{aligned} \mathrm{Re}\, G(s) &= \frac{1}{2}\frac{1 - B(s)}{1 + B(s)} + \frac{1}{2}\frac{1 - B^*(s)}{1 + B^*(s)} \\ &= \frac{1 - B(s)B^*(s)}{[1 + B(s)]\,[1 + B^*(s)]} \end{aligned} \tag{2.140}$$

2.4.12 Passivity of PID controllers

A PID controller

$$H_r(s) = K\frac{1 + T_i s}{T_i s}\frac{1 + T_d s}{1 + \alpha T_d s} \tag{2.141}$$

where $K > 0$ and $0 \leq \alpha < 1$ has phase

$$\angle H_r(j\omega) = -\frac{\pi}{2} + \arctan T_i\omega + \arctan T_d\omega - \arctan \alpha T_d\omega. \tag{2.142}$$

From this equation it is seen that the phase must satisfy

$$-\frac{\pi}{2} \leq \angle H_r(j\omega) \leq \frac{\pi}{2}. \tag{2.143}$$

It follows that $\mathrm{Re}\,[H(j\omega)] \geq 0$ for all $\omega \neq 0$. The transfer function has no poles to the right of the imaginary axis, and one single pole at the imaginary axis at $s = 0$. The residual of this pole is found to be

$$\mathrm{Res}_{s=0} H_r(s) = \lim_{s \to 0}\,[sH_r(s)] = \frac{K}{T_i} \tag{2.144}$$

which is real and positive, and it follows that a PID controller is positive real. This implies the following result

A PID controller $u(s) = H_r(s)e(s)$ is a passive system with input e and output u, and the transfer function $H_r(s)$ is positive real.

2.4.13 Closed loop stability of positive real systems

Stability properties can easily be established from passivity arguments for a feedback interconnection of passive systems. Suppose that the system with input u and output y is passive, and that it is given by

$$y(s) = H(s)u(s) \tag{2.145}$$

The transfer function $H(s)$ is then positive real due to the passivity of the system. The input u is generated by the system

$$u(s) = G(s)\left[y_d(s) - y(s)\right] \tag{2.146}$$

where the transfer function $G(s)$ is supposed to be positive real. Then the loop transfer function is $L(s) = G(s)H(s)$. Note the positive realness implies that two transfer functions $G(s)$ and $H(s)$ will have no poles to the right of the imaginary axis. Moreover, the magnitude of the phase of the of $G(j\omega)$ and $H(j\omega)$ will be less than or equal to 90°. This implies that $L(j\omega)$ has phase that is greater than -180°. This means that the system is at least marginally stable.

2.4.14 Storage function formulation

Passivity can be described using storage functions which are closely related to energy functions and Lyapunov functions. In the passivity setting systems can be interconnected in parallel and feedback interconnections, and the resulting system can be analyzed using passivity theory or Lyapunov theory.

We consider the state space model

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \mathbf{u}) && (2.147)\\ \mathbf{y} &= \mathbf{h}(\mathbf{x}) && (2.148) \end{aligned}$$

Suppose that there is a *storage function* $V(\mathbf{x}) \geq 0$ and a dissipation function $g(\mathbf{x}) \geq 0$ so that the time derivative of V for solutions of the system satisfies

$$\dot{V} = \frac{\partial V}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}, \mathbf{u}) = \mathbf{u}^T\mathbf{y} - g(\mathbf{x}) \tag{2.149}$$

for all control inputs $\mathbf{u}$. Then the system with input $\mathbf{u}$ and output $\mathbf{y}$ is said to be passive.

The result follows from the calculation

$$\begin{aligned} \int_0^T \mathbf{y}^T(t)\,\mathbf{u}(t)\,dt &= V(T) - V(0) + \int_0^T g\left[\mathbf{x}(t)\right] dt \\ &\geq -V(0) \end{aligned} \tag{2.150}$$

Example 34 *We consider again the mass-spring-damper system with input force F. The model is*

$$m\ddot{x} + d\dot{x} + kx = F \tag{2.151}$$

The total energy

$$V = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 \tag{2.152}$$

is a candidate for being a storage function. The time derivative of the energy function is found to be

$$\dot{V} = F\dot{x} - d\dot{x}^2 \tag{2.153}$$

where $F\dot{x}$ is the power that is supplied to the system due to the force F. We see that if the input is $u = F$ and the output is selected to be $y = \dot{x}$ then $\dot{V} = yu - d\dot{x}^2$, which means that the system with input F and output $\dot{x}$ is passive.

Remark 1 *Actually, it is sufficient that*

$$\int_0^T g\left[\mathbf{x}(t)\right] dt \geq -E_g \quad \text{for all } T \geq 0 \tag{2.154}$$

for some constant $E_g \geq 0$.

2.4.15 Interconnections of passive systems

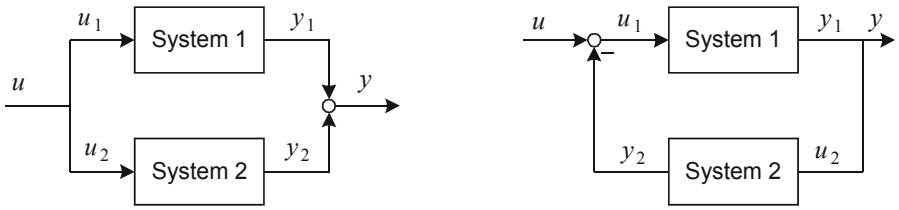

Figure 2.4: Parallel interconnection and feedback interconnection of two passive systems.

We now consider two systems

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1, \mathbf{u}_1), \quad \mathbf{y}_1 = \mathbf{h}_1(\mathbf{x}_1) \tag{2.155}$$

$$\dot{\mathbf{x}}_2 = \mathbf{f}_2(\mathbf{x}_2, \mathbf{u}_2), \quad \mathbf{y}_2 = \mathbf{h}_2(\mathbf{x}_2) \tag{2.156}$$

which are passive in the sense that there are functions $V_1(\mathbf{x}_1) \geq 0$ and $V_2(\mathbf{x}_2) \geq 0$ so that

$$\dot{V}_1 = \frac{\partial V_1}{\partial \mathbf{x}_1}\mathbf{f}_1(\mathbf{x}_1, \mathbf{u}_1) \leq \mathbf{u}_1^T\mathbf{y}_1 - g_1(\mathbf{x}_1) \tag{2.157}$$

$$\dot{V}_2 = \frac{\partial V_2}{\partial \mathbf{x}_2}\mathbf{f}_2(\mathbf{x}_2, \mathbf{u}_2) \leq \mathbf{u}_2^T\mathbf{y}_2 - g_2(\mathbf{x}_2) \tag{2.158}$$

where $g_1(\mathbf{x}_1) \geq -E_{g1}$ and $g_2(\mathbf{x}_2) \geq -E_{g2}$. We will now show that the parallel interconnection and the feedback interconnection shown in Figure 2.4 are passive.

A parallel interconnection

$$\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}, \quad \mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 \tag{2.159}$$

implies that the function V defined by

$$V := V_1 + V_2 \geq 0 \tag{2.160}$$

satisfies

$$\begin{aligned} \dot{V} &= \dot{V}_1 + \dot{V}_2 \\ &= \mathbf{u}_1^T\mathbf{y}_1 - g_1(\mathbf{x}_1) + \mathbf{u}_2^T\mathbf{y}_2 - g_2(\mathbf{x}_2) \\ &= \mathbf{u}^T\mathbf{y} - g_1(\mathbf{x}_1) - g_2(\mathbf{x}_2) \end{aligned} \tag{2.161}$$

which shows that the parallel interconnection with input $\mathbf{u}$ and output $\mathbf{y}$ is passive.

A feedback interconnection

$$\mathbf{y}_1 = \mathbf{u}_2 = \mathbf{y}, \quad \mathbf{u}_1 = \mathbf{u} - \mathbf{y}_2 \tag{2.162}$$

implies that

$$\begin{aligned}\dot{V} &= \dot{V}_1 + \dot{V}_2 \\ &= \mathbf{u}_1^T \mathbf{y}_1 - g_1(\mathbf{x}_1) + \mathbf{u}_2^T \mathbf{y}_2 - g_2(\mathbf{x}_2) \\ &= \mathbf{u}^T \mathbf{y} - g_1(\mathbf{x}_1) - g_2(\mathbf{x}_2) \end{aligned} \tag{2.163}$$

so that also the feedback interconnection with input $\mathbf{u}$ and output $\mathbf{y}$ is passive.

2.4.16 Storage function for PID controller

A PID controller

$$u(s) = H_{pid}(s)e(s) \tag{2.164}$$

where $e(s)$ is the input to the controller, $u(s)$ is the output from the controller, and

$$\begin{aligned} H_{pid}(s) &= K\frac{1+T_i s}{T_i s}(1+T_d s) \\ &= K\left(1 + \frac{T_d}{T_i} + T_d s + \frac{1}{T_i s}\right) \end{aligned} \tag{2.165}$$

can be written in state-space form as

$$\dot{z} = \frac{e}{T_i} \tag{2.166}$$

$$u = K\left[\left(1 + \frac{T_d}{T_i}\right)e + T_d\dot{e} + z\right] \tag{2.167}$$

Consider the nonnegative function

$$V_{pid} = \frac{1}{2}KT_i z^2 + \frac{1}{2}KT_d e^2 \tag{2.168}$$

The time derivative of V along the solutions of the PID controller dynamics is

$$\begin{aligned}\dot{V}_{pid} &= zKT_i\dot{z} + eKT_d\dot{e} \\ &= eK(z + T_d\dot{e}) \\ &= eu - K\left(1 + \frac{T_d}{T_i}\right)e^2 \end{aligned} \tag{2.169}$$

It follows that the PID controller is passive.

2.4.17 Passive plant with PID controller

We consider a passive plant with a PID controller as shown in Figure 2.5. We assume that the passive plant has input u and output y which satisfies

$$\dot{V}_p = yu - g_p \tag{2.170}$$

where $V_p \geq 0$ and $g_p \geq 0$. Consider the nonnegative function

$$V_{cl} = V_p + V_{pid} \tag{2.171}$$

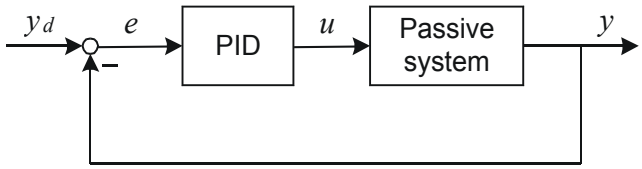

Figure 2.5: Passive plant with PID controller.

The time derivative of V_{cl} along the solutions of the closed-loop system is

$$\begin{aligned} \dot{V}_{cl} &= yu - g_p + eu - K\left(1 + \frac{T_d}{T_i}\right)e^2 \\ &= y_d u - g_p - K\left(1 + \frac{T_d}{T_i}\right)e^2 \end{aligned} \tag{2.172}$$

where it is used that $e = y_d - y$. We see that the closed-loop system with input y_d and output y is passive. In particular we have that if $y_d = 0$, then

$$\dot{V}_{cl} = -g_p - K\left(1 + \frac{T_d}{T_i}\right)e^2 \leq 0 \tag{2.173}$$

2.4.18 Example: Control of mass-spring-damper system

For the mass-spring-damper system with input force F we found that

$$\dot{V} = F\dot{x} - d\dot{x}^2 \tag{2.174}$$

Suppose that the input F is generated by the P controller $F = -K\dot{x}$. Then, with the same storage function we have

$$\dot{V} = -K\dot{x}^2 - d\dot{x}^2 = -\left(K_p - d\right)\dot{x}^2 \leq 0 \tag{2.175}$$

Suppose that the PID controller

$$F(s) = H_{pid}(s)e(s) \tag{2.176}$$

is used where $e(t) = \dot{x}_d(t) - \dot{x}(t)$. Then

$$\dot{V}_{cl} = \dot{V} + \dot{V}_{pid} \tag{2.177}$$

$$= \dot{x}_d F - d\dot{x}^2 - K\left(1 + \frac{T_d}{T_i}\right)e^2 \tag{2.178}$$

In particular, we see that if $\dot{x}_d = 0$, then $\dot{V}_{cl} \leq 0$. Note that the controller is a PID controller from velocity, which corresponds to a PD^2 controller from position.

2.4.19 Example: Active vibration damping

Spacecraft and large space installations are design with lightweight structures, and the lack of atmosphere gives little mechanical damping of vibrations. Even the effect of temperature changes when a satellite passes from the shadow of the earth into the sunlight

may be sufficient to cause unacceptable vibrations in the structure. Because of this the use of feedback for active vibration damping is important (Kelkar and Joshi 1996). Vibration models are usually in the form

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{B}\mathbf{f} \tag{2.179}$$

where $\mathbf{q}$ is the vector of generalized coordinates, which are the elastic deformations, $\mathbf{M}$ is a symmetric mass matrix, $\mathbf{K}$ is a symmetric stiffness matrix, $\mathbf{D}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}}$ is a damping term and $\mathbf{f}$ is the input control force. Detailed vibration models of high accuracy can be derived from the method of finite elements (Bathe 1996). To give the reader an idea richness of the structural properties of this model we briefly mention some issues that will be addressed later in the book: The total energy of the vibration system is $V = K + U$, where

$$K = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} \geq 0 \tag{2.180}$$

is the kinetic energy, and

$$U = \frac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q} \geq 0 \tag{2.181}$$

is the potential energy. The damping term is an energy dissipation term related to friction, and will always satisfy

$$\dot{\mathbf{q}}^T\mathbf{D}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}} \geq 0. \tag{2.182}$$

The time derivative of the total energy along the solutions of the system is

$$\frac{d}{dt}V = \dot{\mathbf{q}}^T\mathbf{M}\ddot{\mathbf{q}} + \dot{\mathbf{q}}^T\mathbf{K}\mathbf{q} = \dot{\mathbf{q}}^{\mathbf{T}}\mathbf{B}\mathbf{f} - \dot{\mathbf{q}}^T\mathbf{D}\dot{\mathbf{q}}. \tag{2.183}$$

If the input force is set to the P controller $\mathbf{f} = -k\mathbf{B}^T\dot{\mathbf{q}}$, which is a velocity feedback, then the time derivative of the energy is

$$\frac{d}{dt}V = -\dot{\mathbf{q}}^T\left(k\mathbf{B}\mathbf{B}^{\mathbf{T}} + \mathbf{D}\right)\dot{\mathbf{q}} \leq 0 \tag{2.184}$$

It is interesting to note that with this velocity feedback the energy will decrease whenever $\dot{\mathbf{q}} \neq \mathbf{0}$. It should be clear from this discussion that the vibration model reflects important physical properties related to energy that may be important in controller design. These properties may be obscured if the model is reformulated in state space. Therefore, when energy-based methods are used, the model is usually kept in the second-order form.

2.4.20 Passive electrical one-port

A *passive electrical one-port* is a circuit with one port and passive components in the form of resistors, capacitors and inductors. Resistors are elements that dissipate energy, while capacitors and inductors are elements that store energy. There are no elements that generate energy. The total energy stored in the circuit is denoted $V(t)$. The stored energy cannot be negative, so we can assume that $V \geq 0$. The port voltage is denoted $u(t)$ and the current into the port is denoted $i(t)$. The power flowing into the one-port is $P(t) = u(t)i(t)$, while the power dissipated in the resistors is $P_r \geq 0$.

The time derivative of the energy V stored in the circuit will be the power ui supplied at the port minus the power loss P_r in the circuit. This is written

$$\dot{V} = ui - P_r \tag{2.185}$$

Integration of this equation gives

$$\int_0^T i\left(t\right) u\left(t\right) dt = V(T) - V(0) + \int_0^T P_r(t)dt \tag{2.186}$$

Here $V(T) \geq 0$ and $P_r(t) \geq 0$, and we find that

$$\int_0^T i\left(t\right) u\left(t\right) dt \geq -V(0) \tag{2.187}$$

This means that for a passive electrical one-port the energy that can be extracted from the circuit over the terminals is less or equal to the energy $V(0)$ that is initially stored in the circuit. Note that (2.187) has the form of a passivity inequality, and that if u is input and i is output, then the system is passive. This implies that the driving point impedance $Z(s) = u(s)/i(s)$ is positive real. On background of this we may conclude that

The driving point impedance of a passive electrical one-port is positive real.

Example 35 *From the passivity inequality (2.187) it is seen that if the current is taken as input and the voltage is considered to be the output, then the system will still be passive, which means that the driving point admittance $Y(s) = i(s)/u(s)$ is passive.*

Example 36 *To illustrate this we consider a passive electrical one-port which is a parallel interconnection of a resistor and a capacitor. The current is given by*

$$i = \frac{1}{R}u + C\dot{u} \tag{2.188}$$

The total energy stored in the circuit is the energy $V = (1/2)Cu^2$ stored by the capacitor. The time derivative of the energy is

$$\dot{V} = Cv\dot{v} = iu - \frac{1}{R}u^2 \tag{2.189}$$

where the loss term is due to the energy dissipation in the resistor. Figure 2.6.

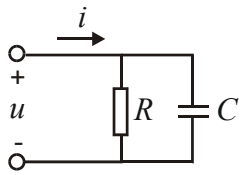

Figure 2.6: Passive electrical one-port.

2.4.21 Electrical analog of PID controller

A PID controller from current to voltage of a one-port is given by

$$u\left(s\right) = -K\left(1 + \frac{T_d}{T_i} + T_d s + \frac{1}{T_i s}\right)\left[i_d\left(s\right) - i\left(s\right)\right] \tag{2.190}$$

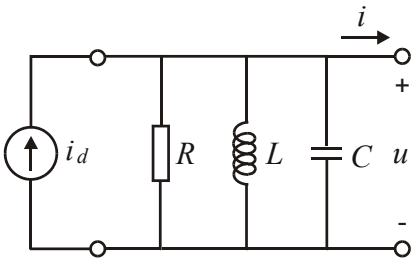

Figure 2.7: Electrical analog of PID controller for an electrical one-port where the current i is controlled with the voltage u.

where i_d is the desired current. This is an electrical one-port where a current source with current i_d is placed in in parallel to a series connection with a resistor $R = K(1+T_d/T_i)$, a capacitor $C = T_i/K$, and an inductor $L = T_d K$ (Figure 2.7). A PI controller is obtained by setting T_d to zero, and in that case there is no inductor in the one-port. In the case that $i_d = 0$, the PID controller is a passive one-port with only passive elements.

2.4.22 Passive electrical two-port

An electrical two-port has two ports, one input port with voltage u_1 and current i_1, and one output port with voltage u_2 and current i_2. The power flowing into the two-port is

$$P(t) = u_1(t)\, i_1(t) + u_2(t)\, i_2(t) \tag{2.191}$$

As for the passive one-port, the energy that can be extracted from a passive n-port cannot be larger than the energy $V_n(0)$ that is initially stored in the capacitors and the inductors. This gives

$$\int_0^T \left[u_1(t)\, i_1(t) + u_2(t)\, i_2(t)\right] dt \geq -V_2(0) \tag{2.192}$$

2.4.23 Termination of electrical two-port

An electrical two-port is said to be terminated if the output port is connected to a one-port so that

$$u_2 = u, \quad i_2 = -i \tag{2.193}$$

Then the two-port with the one-port termination becomes a one-port with port variables u_1 and i_1.

Suppose that both the two-port and the one-port termination are passive. Then the following energy equations are valid

$$\int_0^T \left[u_1(t)\, i_1(t) + u_2(t)\, i_2(t)\right] dt \geq -V_2(0) \tag{2.194}$$

$$\int_0^T i(t)\, u(t)\, dt \geq -V_1(0) \tag{2.195}$$

If we add these equations and insert the connection equations (2.193) we get

$$\int_0^T i_1(t)\, u_1(t)\, dt \geq -\left[V_1(0) + V_2(0)\right] \tag{2.196}$$

The physical interpretation is that the energy that can be extracted from the combined circuits is equal to the sum of the initially stored energy in the two circuits. This result shows clearly shows that if a passive two-port is terminated with a passive-one-port, then the resulting one-port with port variables u_1 and i_1 is passive.

2.4.24 Passive electrical n-ports

An electrical n-port has n ports with voltage u_k and current i_k. The power flowing into the n-port is

$$P(t) = \sum_{k=1}^{n} u_k(t)\, i_k(t) = \mathbf{i}^T(t)\, \mathbf{u}(t) \tag{2.197}$$

where $u = (u_1, \ldots, u_n)^T$ and $\mathbf{i} = (i_1, \ldots, i_n)^T$. As for the passive one-port, the energy that can be extracted from a passive n-port cannot be larger than the energy E_n that is initially stored in the capacitors and the inductors

$$\int_0^T \mathbf{i}^T(t)\, \mathbf{u}(t)\, dt \geq -V_n(0) \tag{2.198}$$

A general n-port with effort vector $\mathbf{e}$ and flow vector $\mathbf{f}$ is passive if the energy that can be extracted from the n-port is limited by the energy that is initially stored in the n-port, that is, if

$$\int_0^T \mathbf{f}^T(t)\, \mathbf{e}(t)\, dt \geq -E_0 \tag{2.199}$$

As in the electrical case this can be expressed in terms of conditions on the impedance $Z(s)$ for a one-port.

2.4.25 Example: Telemanipulation

In a telemanipulation system a manipulator is remotely controlled by a human operator. Early telemanipulation systems were master-slave systems where the operator moved a handle fixed to a master manipulator that was connected to an identical slave manipulator through mechanical linkages. This was used to protect the operator from hazardous environments due to radioactivity or danger of contamination from biological samples. The operator would then typically watch the slave manipulator through a window, and as there was a direct mechanical coupling between the master and the slave manipulators the operator would feel contact forces that resulted when the slave came into contact with a sample or hit against a table. This feature is called force reflection. At a later stage such systems were equipped with computer control. This was done to make it possible to perform telemanipulation in hostile environments for operations in space, and for underwater operations at great depths. A more recent activity is telesurgery.

In telemanipulation systems with computer control the motion of the master is measured by sensors, and the position and velocity commands are transmitted to the slave through a computer, and the slave is driven by DC motors. The opens up for advanced control functions, but it turns out that the system becomes unstable if force reflection is used with time delays of 40 ms or more. This problem was analyzed in an energy flow setting in (Anderson and Moore 1989) and (Niemeyer and Slotine 1991), and it was proposed to transmit wave variables to obtain a closed loop system where the transmission between the master and the slave could be described in terms of passive two-ports. The main idea of the solution is presented in the following.

A human operator that moves a telemanipulation system using a master-slave configuration will expect that the master-slave system in itself will not generate energy that is transferred to the handle that the operator is using to move the master. If the system were to generate energy, then the operator might get the impression that the telemanipulation system had a mind of its own, and the operator might have to struggle against movements that are generated by the telemanipulation system. The operator might even get injured by the master. This means that the telemanipulation system as felt by the operator may store energy and dissipate energy, but it may not generate energy. This means that the handle connected to the master of the telemanipulation system should appear to the operator as a port to a passive mechanical system where the velocities of the handle are the flow variables, and the forces on the handle are the effort variables. If the master and slave are connected with mechanical linkages, then the system will be a passive mechanical system. However, when computer control is added, then the control algorithms must be selected with care so that the system still appears as a passive system to the operator.

What the operator will expect is that the telemanipulation system appears as a passive two-port that transfers the velocity commands from the operator to the slave, and that returns the force from the slave to the operator. A mathematical formulation of this in one dimension is that the master is a two-port

$$m_m \dot{v}_m = F_h - F_m \tag{2.200}$$

with effort F_h and flow v_m on the input port, and effort F_m and flow v_m on the output port. The slave is a passive two-port

$$m_s \dot{v}_s = F_s - F_e \tag{2.201}$$

with effort F_s and flow v_s is the input port and effort F_e and v_s at the output port. The key to a satisfactory system is to have a passive two-port with effort F_m and flow v_m on the input port, and effort F_s and flow v_s is the output port to connect the output port of the master to the input port of the slave. The total energy E of the system will then be

$$E(T) = E(0) + \int_0^T F_h(t) v_m(t) dt + \int_0^T F_e(t) v_s(t) dt \tag{2.202}$$

It is reasonable to assume that the total energy is positive, that is, $E \geq 0$, which implies that

$$\int_0^T F_h(t) v_m(t) dt \geq -E(0) - \int_0^T F_e(t) v_s(t) dt \tag{2.203}$$

The physical interpretation of this is that the energy that is returned to the operator through the handle on the master is less that the energy initially stored in the system plus the energy supplied to the slave from the environment.

First, suppose that the master is directly connected to the slave through a rigid interconnection so that

$$F_s = F_m \quad \text{and} \quad v_s = v_m \tag{2.204}$$

Then the interconnection between the master and the slave is certainly a passive two-port, and the desired passivity of the system is ensured. Moreover, we see that

$$(m_m + m_s)\dot{v}_m = F_h - F_e \tag{2.205}$$

which means that the operator experiences the environmental force F_e on the slave, and moves the combined inertia of master and slave.

Next, consider the case where the slave is driven by DC motors, and that the commanded velocity v_m from the master is measured by a sensor and transmitted electronically without any time delay. Then a possible solution is to control the slave with a PD controller with desired velocity v_d and desired position x_d given by

$$v_d(t) = v_m(t) \quad \text{and} \quad x_d(t) = x_m(t) \tag{2.206}$$

The slave with PD controller is then given by the passive two-port

$$m_s \dot{v}_s = F_s - F_e \tag{2.207}$$
$$F_s = K_s(x_m - x_s) + D_s(v_m - v_s) \tag{2.208}$$

with effort F_s and flow v_m at the input port and effort F_e and flow v_s at the output port. Force reflection to the master is achieved by setting up the force $F_m(t) = F_s(t)$ in the master, which gives the following two-port for the master:

$$m_m \dot{v}_m = F_h - F_s \tag{2.209}$$

The resulting telemanipulation system is passive with a mechanical analog as shown in Figure 2.8 where the transmission between the master and the slave is a spring with stiffness K_s in parallel with a damper with coefficient D_s.

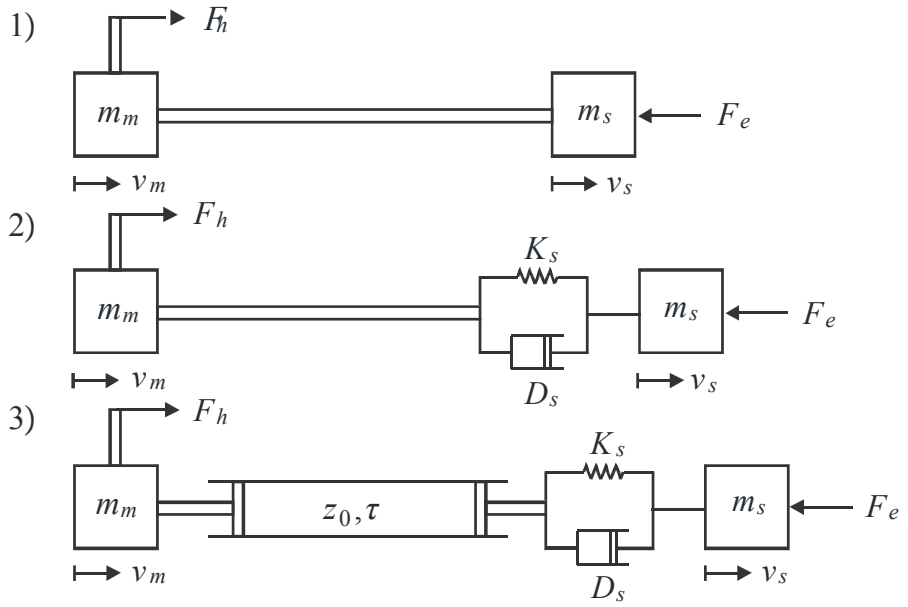

Figure 2.8: Mechanical analogs of telemanipulation systems with force reflection. The master is represented by a mass m_m that is moved with a force F_h from the human operator. The slave is represented by a mass m_s which is exposed to a force F_e from the environment. In 1) there is a direct mechanical coupling between the master and the salve. In 2) the slave is controlled by a DC motor with PD control, and there is no time delay in the signal transmission. In 3) the slave is controlled by a DC motor with PD control, and the signals are transmitted with time delay using wave variables.

Finally, suppose that there is a time delay τ in the electronic transmission between master and slave. Early attempts involved using the same solution as presented above, but with transmission given by

$$v_d(t) = v_m(t-\tau) \quad \text{and} \quad F_m(t) = F_s(t-\tau) \tag{2.210}$$

The resulting system is not passive, as there is no passive mechanical two-port that delays the velocity in the forward direction and that delays the force in the opposite direction. Stability problems where experienced for such solutions already at time delays of 40 ms. A different solution must therefore be sought. An elegant solution to this problem is to use a lossless transmission line to interconnect master and slave. The key to this solution is that a lossless transmission line transmits the wave variables with a time delay that is the transport time τ of the transmission line. We therefore introduce the wave variables

$$a_m = F_m + z_0 v_m \quad \text{and} \quad b_m = F_m - z_0 v_m \tag{2.211}$$

for the output port of the master, and the wave variables

$$a_s = F_s - z_0 v_d \quad \text{and} \quad b_s = F_s + z_0 v_d \tag{2.212}$$

for the input port of the slave where z_0 is the characteristic impedance of the transmission line. The wave variables are transmitted according to

$$b_s(t) = a_m(t-\tau) \quad \text{and} \quad b_m(t) = a_s(t-\tau) \tag{2.213}$$

In terms of forces and velocities this gives the following description of the passive two-port

$$v_d(t) = v_m(t-\tau) + \frac{1}{z_0}[F_m(t-\tau) - F_s(t)] \tag{2.214}$$

$$F_m(t) = F_s(t-\tau) + z_0[v_m(t) - v_0(t-\tau)] \tag{2.215}$$

with effort F_m and v_m at the input and effort F_s and flow v_s at the output port. The slave may then be controlled with the PD controller

$$F_s = K_s(x_d - x_s) + D_s(v_d - v_s) \tag{2.216}$$

as in the case of no time delay. The mechanical analog is as when there is no time delay, but with a transmission corresponding to a lossless hydraulic transmission line with a compressible fluid shown in Figure 2.8.

2.4.26 Passivity and gain

Consider a system with input u and output y. Define the variable

$$r = u + \lambda y \tag{2.217}$$

Then

$$\int_0^T r^2 dt = \int_0^T u^2 dt + 2\lambda \int_0^T uy dt + \lambda^2 \int_0^T y^2 dt \tag{2.218}$$

From this equation it is seen that

$$\int_0^T uy dt \geq -E_0 \tag{2.219}$$

is equivalent for all $\lambda > 0$ to

$$\int_0^T r^2 dt + 2\lambda E_0 \geq \int_0^T u^2 dt \quad and \quad \int_0^T r^2 dt + 2\lambda E_0 \geq \lambda^2 \int_0^T y^2 dt \tag{2.220}$$

This shows that passivity of the system with input u and output y is equivalent to a small gain condition in the L_2 norm (Khalil 1996) from r to u, and from r to y. A related result is used in semi-group theory (Pazy 1983, p. 14).

Example 37 *This result was used in attitude control in (Egeland and Godhavn 1994) where*

$$\mathbf{r} = \boldsymbol{\omega} + \lambda \boldsymbol{\epsilon} \tag{2.221}$$

was used. A controller was designed so that $\mathbf{r} \in L_2$*, and then (2.220) was used to show that the passivity of the system with input* $\boldsymbol{\omega}$ *and output* $\boldsymbol{\epsilon}$ *implied that the mapping from* $\mathbf{r}$ *to* $\boldsymbol{\omega}$*, and the mapping from* $\mathbf{r}$ *to* $\boldsymbol{\epsilon}$ *were* L_2 *stable.*

Example 38 *In the adaptive tracking controller in (Slotine and Li 1988), stability in the variable*

$$\mathbf{r} = \dot{\mathbf{q}} + \lambda \mathbf{q} \tag{2.222}$$

was used to establish convergence in $\mathbf{q}$ *and* $\dot{\mathbf{q}}$*. In (Kelly, Carelli and Ortega 1989) the same controller was analyzed, and it was shown that* $\mathbf{r} \in L_2$*, and linear theory was used to show that this implied convergence in* $\mathbf{q}$ *and* $\dot{\mathbf{q}}$*.*

2.5 Uncertainty in modeling

2.5.1 General state space models

In a general state space model

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u, t) \tag{2.223}$$
$$y = h(\mathbf{x}, t) \tag{2.224}$$

or, in the case of linear models,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \tag{2.225}$$
$$y = \mathbf{C}\mathbf{x} + Du \tag{2.226}$$

there may be uncertainties related to

- Parameter values
- Model structure
- System order

There are techniques for describing such uncertainties (Skogestad and Postlethwaite 1996), which may be based on additive uncertainties. Such additive terms may be related to model properties.

Chapter 3

Electromechanical systems

3.1 Introduction

This chapter deals with mathematical models of electrical motors, and models of electromechanical sensors and actuators. These electromechanical systems are based on energy conversion between electrical energy and mechanical energy due to capacitive and inductive effects. This type of electromechanical systems are important, as they are vital components in most control systems. Special attention is given to the DC motor with constant field, which is a basic building block in many control systems. This motor is described by a simple model, and it is possible to control the motor torque directly. Because of its importance and simplicity the chapter starts with the model of a DC motor, and presents typical load configurations for the DC motor. Then selected topics from the general theory of electromechanical energy conversion is presented with emphasis on energy functions. This provides us with the necessary background to derive more advanced models of electrical motors. This includes the model of a DC motor with externally controlled field, the model of a general AC motor, and models for induction motors.

3.2 Electrical motors

3.2.1 Introduction

An electrical motor with rotary motion has a stationary part called the stator. The rotary part of the motor is called the rotor. The rotor is fixed to the motor shaft which drives the load. The motion of the rotor is due to the motor torque which is set up by electromagnetic Lorentz forces acting on the rotor. There are many different ways of setting up an appropriate Lorentz force, and electrical motors are characterized depending on how this is done. Electrical motors are divided into DC motors and AC motors. DC motors are well suited for control applications, as the torque of the motor can be accurately controlled. The recent development in power electronics, however, has made it possible to control the torque also for AC motors, and, consequently, AC motors are now used for accurate control. A basic reference on electrical motors is (Fitzgerald, Kingsley and Umans 1983), while a more advanced textbook including control methods is (Leonhard 1996).

3.2.2 Basic equations

A rotary motor has a motor shaft that rotates with angular velocity ω_m, and it has some device for setting up a motor torque T so that the motor shaft has the following equation of motion:

$$J_m\dot{\omega}_m = T - T_L \tag{3.1}$$

Here T_L is the load torque acting on the shaft. The mechanical power delivered from the motor to the shaft is

$$P_m = T\omega_m \tag{3.2}$$

while the mechanical power delivered to the load is

$$P_L = T_L\omega_m \tag{3.3}$$

The motor shaft dynamics can be described as a two-port with effort T and flow ω_m at the input port, and effort T_L and flow ω_m at the output port. Different types of motors are characterized according to how the motor torque T is generated. In electrical motors the torque is due to electromagnetic forces, in a hydraulic motor of the hydrostatic type it is due to the pressure force from a pressurized fluid, while in a turbine the torque is set up by the forces that result from the change of momentum in the flowing fluid.

The speed of a motor is commonly given in revolutions per minute (rev/min). The relation to the SI unit rad/s is

$$1\frac{\text{rev}}{\text{min}} = \frac{2\pi}{60}\frac{\text{rad}}{\text{s}} = 0.105\frac{\text{rad}}{\text{s}} \tag{3.4}$$

3.2.3 Gear model

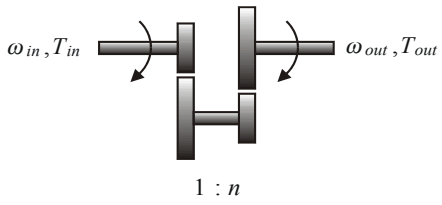

Figure 3.1: Reduction gear

An electrical motor will typically have a speed range from zero up to about 3000 rev/min. Specially designed electrical motors may run up to 12000 rev/min. In comparison to this, car engines run from 800–6000 rev/min. For many applications the required speed range of the load is significantly less than the speed range of the motor, and a reduction gear must be used. This gives a lower speed of the load, and, more importantly, it gives a higher torque.

A reduction gear with gear ratio n (Figure 3.1) is described by

$$\omega_{out} = n\omega_{in} \tag{3.5}$$

where ω_{in} is the angular velocity of the shaft on the input side of the gear, and ω_{out} is the angular velocity of the shaft on the output side of the gear. For a reduction gear $n < 1$, and a gear is said to have a gear ratio of, say, 10 if $n = 1/10$.

The relation between the input torque T_{in} and the output torque T_{out} is found by comparing power in and power out for the gear. Suppose that the gear is lossless. Then power in is equal to power out, that is,

$$T_{in}\omega_{in} = T_{out}\omega_{out} \tag{3.6}$$

Inserting the expression for ω_{out} we find that

$$T_{out} = \frac{1}{n}T_{in} \tag{3.7}$$

This means that a reduction gear reduces the speed by a factor n, while it amplifies the torque by a factor $1/n$.

A gear with gear ratio n may be described as a two-port

$$\omega_{out} = n\omega_{in} \tag{3.8}$$

$$T_{out} = \frac{1}{n}T_{in} \tag{3.9}$$

with variables T_{in} and ω_{in} at the input port, and variables T_{out} and ω_{out} at the output port.

3.2.4 Motor and gear

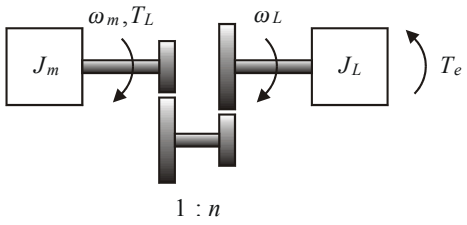

Figure 3.2: Motor and gear.

Consider a motor with equation of motion

$$J_m\dot{\omega}_m = T - T_L \tag{3.10}$$

that drives a load over a reduction gear with gear ratio n (Figure 3.2). Then the load has a shaft speed $\omega_L = n\omega_m$, and is driven by the output torque of the gear, which is T_L/n. The inertia of the load is J_L, and it is assumed that an external torque T_e acts on the load. Then the equation of motion for the load is

$$J_L\dot{\omega}_L = \frac{1}{n}T_L - T_e \tag{3.11}$$

If the load equation (3.11) is multiplied by n and added to the equation of motion of the motor (3.10), then the result is the equation of motion for the system referred to the motor side. Alternatively, the motor equation (3.10) can be divided by n and added to the load equation (3.11). This will give the equation of motion of the system referred to the load side. To sum up:

The equation of motion for motor, gear and load referred to the motor side is

$$(J_m + n^2 J_L)\dot{\omega}_m = T - nT_e \tag{3.12}$$

The equation of motion for motor, gear and load referred to the load side is

$$\left(\frac{1}{n^2}J_m + J_L\right)\dot{\omega}_L = \frac{1}{n}T - T_e \tag{3.13}$$

3.2.5 Transformation of rotation to translation

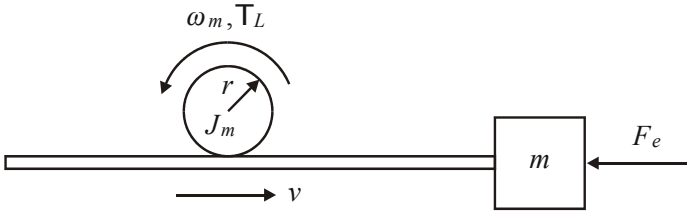

Figure 3.3: Transmission from rotation to translation.

Rotational motion of a shaft can be transformed to translational motion and vice versa by mounting a wheel that rolls on a surface as shown in Figure 3.3. This type of transmission is commonly seen in rack-and-pinion drives, friction gears, pulleys, and between car wheels and the road. Suppose that the wheel has radius r, shaft speed ω_m, and torque T_L. Then the translational velocity will be $v = r\omega_m$. Denote the force acting from the wheel on the translating part by F. Then the input power will be $\omega_m T_L$ and the output power will be vF. The gear does not store energy, and it follows that $F = T_L/r$. This shows that:

A rotation to translation transmission can be described by the two-port

$$v = r\omega_m \tag{3.14}$$

$$F = \frac{1}{r}T_L \tag{3.15}$$

with variables T_L and ω_m at the input port, and variables F and v at the output port.

Consider a motor which drives a mass m in translational motion over a wheel with radius r. The load is assumed to have equation of motion

$$m\dot{v} = F - F_e \tag{3.16}$$

where F_e is an external force acting on the load. A motor described by

$$J_m\dot{\omega}_m = T - T_L \tag{3.17}$$

is used. By combining these two equations the following result is found:

The equation of motion for motor and load referred to the motor side is

$$(J_m + mr^2)\dot{\omega}_m = T - rF_e \tag{3.18}$$

The equation of motion for motor and load referred to the load side is

$$(\frac{1}{r^2}J_m + m)\dot{v} = \frac{1}{r}T - F_e \tag{3.19}$$

3.2.6 Torque characteristics

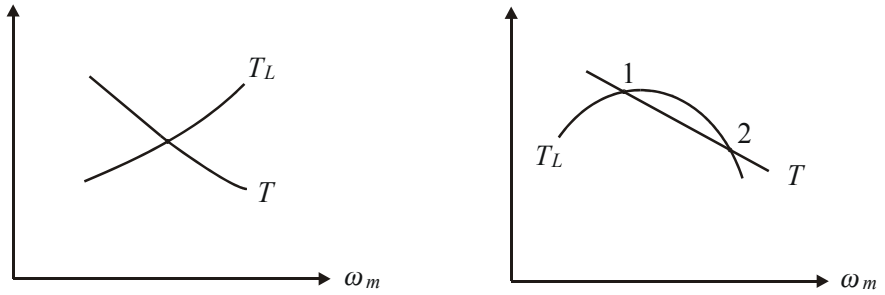

Figure 3.4: To the left is shown a stable system where the load torque T_L is increasing for increasing motor velocity ω_m. To the right is shown a system with two equilibrium points. Equilibrium 1 is stable, while equilibrium 2 is unstable as the load torque T_L decreases faster than the motor torque T when the motor velocity ω_m increases.

In many applications the load torque T_L will depend on the motor speed. An example of this is shown in the left diagram of Figure 3.4, where the load torque increases with increasing speed. This will be the case for systems where the friction increases with the velocity, like the air resistance of a car or a bicycle. Moreover, the motor torque will typically be a decreasing function of the motor shaft speed ω_m due to increasing energy loss in the motor. It turns out that if both the motor torque and the load torque are functions of the motor speed so that $T = T(\omega_m)$ and $T_L = T_L(\omega_m)$, then the stability of the motor and load can be investigated in a torque-speed diagram. This is done by linearization of the motor model (3.1), which gives

$$J_m\Delta\dot{\omega}_m = k\Delta\omega_m \tag{3.20}$$

where

$$k = \left(\frac{\partial T}{\partial \omega_m} - \frac{\partial T_L}{\partial \omega_m}\right)\Big|_{\omega_{mo}} \tag{3.21}$$

is a linearization constant. From linear stability theory we see that the system is stable if and only if $k \leq 0$. This can be investigated graphically in a torque-speed diagram as shown in Figure 3.4.

Example 41 *Suppose that a motor is connected to a load which is simply a friction*

torque T_L. The friction is given by

$$T_L(\omega_m) = \left\{ T_c + (T_s - T_c) \exp\left[-\left(\frac{\omega_m}{\omega_s} \right)^2 \right] \right\} \operatorname{sgn}(\omega_m) + B\omega_m \tag{3.22}$$

where T_c is the Coulomb friction and T_s is the static friction and

$$\operatorname{sgn}(\omega_m) = \begin{cases} -1 & \omega_m < 0 \\ 1 & 0 < \omega_m \end{cases} \tag{3.23}$$

The constant ω_s is the characteristic velocity of the Stribeck effect, and B is the coefficient of the viscous friction. For further details on this friction characteristic, see the Chapter 5. The motor torque is directly controlled, so that T is a constant. The equation of motion is then

$$J_m \dot{\omega}_m = T - \left\{ T_c + (T_s - T_c) \exp\left[-\left(\frac{\omega_m}{\omega_s} \right)^2 \right] \right\} sgn(\omega_m) - B\omega_m \tag{3.24}$$

For simplicity it is assumed that $\omega_m \geq 0$ so that $sgn(\omega_m) = 1$. Linearization gives

$$J_m \Delta\dot{\omega}_m = \left(2\frac{\omega_m}{\omega_s}(T_s - T_c) \exp\left[-\left(\frac{\omega_m}{\omega_s} \right)^2 \right] - B \right) \Delta\omega_m \tag{3.25}$$

This shows that the system is unstable for constant motor torque T at the speed ω_m if

$$B < 2\frac{\omega_m}{\omega_s}(T_s - T_c) \exp\left[-\left(\frac{\omega_m}{\omega_s} \right)^2 \right] \tag{3.26}$$

3.2.7 The four quadrants of the motor

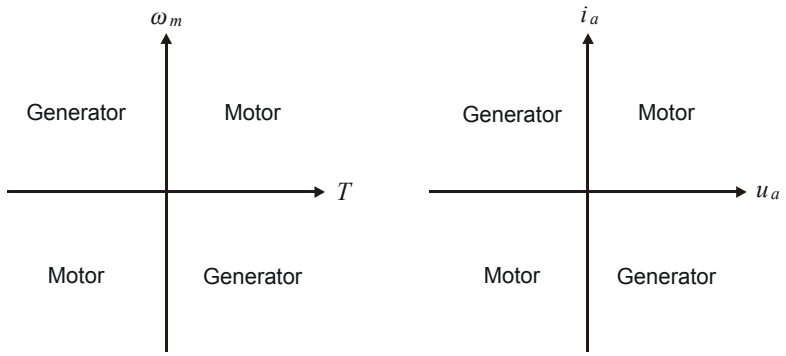

Figure 3.5: The four quadrants of the motor (to the left), and the four quadrants of the power amplifier (to the right).

A motor delivers the mechanical power $T\omega_m$ through the motor shaft, where T is the motor torque and ω_m is the motor speed. If T and ω_m has the same signs under stationary operation, then the motor delivers power. In this case the motor transforms electrical power to mechanical power, and is said to work as a motor. If T and ω_m has opposite signs under stationary operation, then the motor receives mechanical power and transforms it into electrical power, and is said to work as a generator. This is illustrated in Figure 3.5.

3.3 The DC motor with constant field

3.3.1 Introduction

The DC motor with constant field has a simple dynamic model, and has been controlled accurately with simple electronics from the early period of automatic control. Because of this it has been a very important component in servomechanisms, which are control systems involving fast and accurate control of motion. In modern servomechanisms, the DC motor will always be used as a current controlled DC motor, where a high gain current loop is integrated with the motor. This makes it possible to control the motor torque directly, and this is one of the reasons for the success of the motor. More recently, advanced power electronics has made it possible to control other types of electrical motors with the same fast response as the DC motor. This will typically involve some method to control the motor torque, which leads to the same dynamic model as for the current controlled DC motor. Therefore, the models and the analysis results presented for the current controlled DC motor in this section is also valid for other types of electrical motors where the motor torque can be controlled.

3.3.2 Model

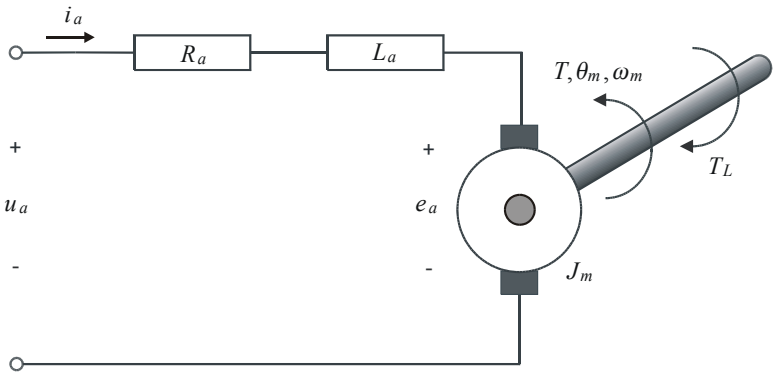

Figure 3.6: Armature circuit of DC motor with constant field.

A DC motor with constant field is described by an armature circuit with current i_a and input voltage u_a. The armature circuit is a serial connection of the armature resistance R_a, the armature inductance L_a, and the electromechanical energy conversion unit with induced voltage e_a. This voltage e_a is induced by the motor speed ω_m in combination with a constant electromagnetic field that is set up either by a field circuit with a constant field current i_e, or by a permanent magnet which replaces the field circuit. An important characteristic of the DC motor with constant field is that the motor torque is proportional to the armature current, and is given by

$$T = K_T i_a \tag{3.27}$$

where K_T is the torque constant.

The DC motor with constant field can be described as a serial interconnection of three passive two-ports. The first two-port is the armature circuit where the input port has variables u_a and i_a, and the output port has variables e_a and i_a. The second two-port is

the electromechanical energy conversion unit with an electrical port with port variables e_a and i_a, and a mechanical port with port variables T and ω_m. Finally, the third two-port is the motor shaft with input port with variables T and ω_m and output port with variables T_L and ω_m. There is no energy storage in the electromechanical energy conversion unit, which implies that the power $e_a i_a$ of the electrical port equals the power $T\omega_m$ of the mechanical port. This gives

$$e_a i_a = T\omega_m = K_T i_a \omega_m \Rightarrow e_a = K_E \omega_m \tag{3.28}$$

where $K_E = K_T$ is the field constant. The dynamic model can then be found from the voltage law of the armature circuit and the equation of motion for the motor shaft:

A DC motor with constant field has the dynamic model

$$\begin{aligned} L_a \frac{d}{dt} i_a &= -R_a i_a - K_E \omega_m + u_a && (3.29)\\ J_m \dot{\omega}_m &= K_T i_a - T_L && (3.30)\\ \dot{\theta}_m &= \omega_m && (3.31) \end{aligned}$$

The block diagram is shown in Figure 3.7.

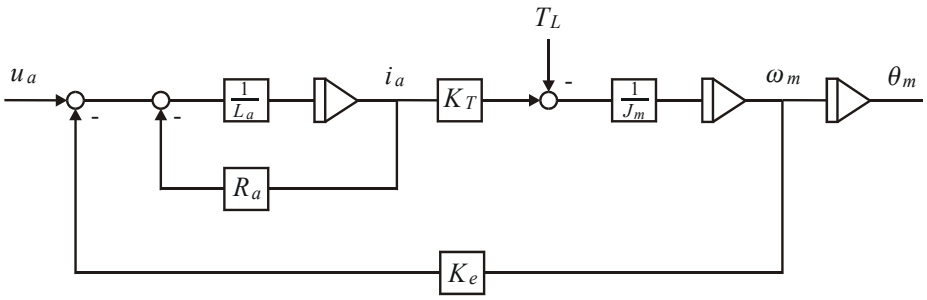

Figure 3.7: Voltage controlled DC motor.

3.3.3 Energy function

The total energy V of the motor is the sum of inductive energy stored in the armature inductance L_a and the kinetic energy of the motor shaft. This gives

$$V = \frac{1}{2} L_a i_a^2 + \frac{1}{2} J_m \omega_m^2 \geq 0 \tag{3.32}$$

The time derivative of the energy function V along the solutions of the system is

$$\begin{aligned} \dot{V} &= i_a L_a \frac{di_a}{dt} + \omega_m J_m \dot{\omega}_m \\ &= i_a \left(-R_a i_a - K_E \omega_m + u_a\right) + \omega_m \left(K_T i_a - T_L\right) \\ &= i_a u_a - \omega_m T_L - R_a i_a^2 \\ &\leq i_a u_a - \omega_m T_L \end{aligned} \tag{3.33}$$

Now, suppose that the load model with input ω_m and output T_L is passive, and that there is a storage function $V_L \geq 0$ so that

$$\dot{V}_L \leq \omega_m T_L \tag{3.34}$$

Then the total energy $V_t := V + V_L$ is greater that or equal to zero, and the time derivative of V_t along the solution of the system is

$$\dot{V}_t \leq i_a u_a \tag{3.35}$$

We have then established the following result:

The DC motor model (3.29–3.31) with input u_a and output i_a is passive if the load with input ω_m and output T_L is passive.

Example 42 *Suppose that the load is simply a damper with viscous friction so that $T_L = B_L\omega_m$. Then the system with input ω_m and output T_L is passive with storage function $V_L = 0$ as this gives*

$$\dot{V}_L = 0 \leq B_L\omega_m^2 = \omega_m T_L \tag{3.36}$$

It follows that the DC motor with input u_a and output i_a is passive with this load.

Example 43 *Suppose that the load is an inertia J_L with shaft angle θ_L connected to the motor shaft by a spring with torque*

$$T_L = K(\theta_m - \theta_L) + D(\dot{\theta}_m - \dot{\theta}_L) \tag{3.37}$$

The equation of motion for the inertia is

$$J_L\ddot{\theta}_L = T_L \tag{3.38}$$

Then the system with input ω_m and output T_L is passive with storage function equal to the total energy

$$V_L = \frac{1}{2}J_L\dot{\theta}_L^2 + \frac{1}{2}K(\theta_m - \theta_L)^2 \geq 0 \tag{3.39}$$

because the time derivative of V_L along the solutions of the system is

$$\begin{aligned}
\dot{V}_L &= J_L\ddot{\theta}_L\dot{\theta}_L + K(\theta_m - \theta_L)(\dot{\theta}_m - \dot{\theta}_L) \\
&= T_L\dot{\theta}_L - K(\theta_m - \theta_L)(\dot{\theta}_m - \dot{\theta}_L) \\
&= T_L\dot{\theta}_m + T_L(\dot{\theta}_L - \dot{\theta}_m) - K(\theta_m - \theta_L)(\dot{\theta}_m - \dot{\theta}_L) \\
&= T_L\dot{\theta}_m - D(\dot{\theta}_m - \dot{\theta}_L)^2
\end{aligned} \tag{3.40}$$

3.3.4 Laplace transformed model

Laplace transformation of the DC motor model (3.29–3.31) gives

$$si_a(s) = \frac{1}{L_a}\left[-R_a i_a(s) - K_E\omega_m(s) + u_a(s)\right] \tag{3.41}$$

$$s\omega_m(s) = \frac{K_T}{J_m}i_a(s) - \frac{1}{J_m}T_L(s) \tag{3.42}$$

$$s\theta_m(s) = \omega_m(s) \tag{3.43}$$

The equation of motion gives

$$s^2\theta_m(s) = \frac{K_T}{J_m} i_a(s) - \frac{1}{J_m} T_L(s) \tag{3.44}$$

while the armature equation gives

$$(L_a s + R_a)\, i_a(s) = -K_E s\theta_m(s) + u_a(s) \tag{3.45}$$

Insertion of (3.45) in (3.44) gives

$$s^2\theta_m(s) = \frac{K_T}{J_m} \frac{1}{L_a s + R_a} \left(-K_E s\theta_m(s) + u_a(s)\right) - \frac{1}{J_m} T_L(s) \tag{3.46}$$

and finally

$$\theta_m(s) = \frac{\frac{1}{K_E} u_a(s) - \frac{R_a}{K_E K_T}\left(1 + \frac{L_a}{R_a} s\right) T_L(s)}{s\left(\frac{J_m L_a}{K_E K_T} s^2 + \frac{J_m R_a}{K_E K_T} s + 1\right)} \tag{3.47}$$

This can be written

$$\theta_m(s) = \frac{\frac{1}{K_E} u_a(s) - \frac{R_a}{K_T K_E}(1 + T_a s) T_L(s)}{s(T_a T_m s^2 + T_m s + 1)} \tag{3.48}$$

where

$$T_a = \frac{L_a}{R_a} \tag{3.49}$$

is the electrical time constant of the motor, and

$$T_m = \frac{J_m R_a}{K_E K_T} \tag{3.50}$$

is the mechanical time constant. Usually, one may assume that the electrical time constant is much less than the mechanical time constant so that the model can be written

$$\theta_m(s) = \frac{\frac{1}{K_E} u_a(s)}{s(1 + T_m s)(1 + T_a s)} - \frac{\frac{R_a}{K_E K_T} T_L(s)}{s(1 + T_m s)} \tag{3.51}$$

This leads to the following result:

The transfer function from the input u_a to the angle θ_m is

$$H_p(s) = \frac{\theta_m}{u_a}(s) = \frac{\frac{1}{K_E}}{s(1 + T_m s)(1 + T_a s)} \tag{3.52}$$

3.4 DC motor control

3.4.1 Introduction

A DC motor used in a servomechanism will more or less always have a current control loop integrated in the motor, and it will normally have a speed loop outside of the current loop. These feedback loops are often seen as an integrated part of the DC motor, and

it is therefore useful to present models of the DC motor with current control and with speed control. The advantage of these feedback loops is that the current loop will have very high bandwidth, and it will therefore suppress nonlinearities in the power amplifier. The velocity loop can also be given a high bandwidth, and will tend to eliminate the effect of friction on the motor. The outer position control loop will normally have to be slower than the first mechanical resonance, and this limits the gain in the position loop. Usually a PI controller will be used in the current loop, and a PI controller with limited integral action will used in the velocity loop. In the presentation here we use P controllers to simplify the expressions. The main results will still be valid.

3.4.2 Current controlled DC motor

The transfer function from the input $u_a(s)$ to the current $i_a(s)$ of the armature circuit can be found from

$$\frac{i_a}{u_a}(s) = \frac{i_a}{\theta_m}(s)\frac{\theta_m}{u_a}(s) \tag{3.53}$$

From the Laplace transformed model we see that

$$s^2\theta_m = \frac{K_T}{J_m}i_a \quad \Rightarrow \quad \frac{i_a}{\theta_m}(s) = \frac{J_m s^2}{K_T} \tag{3.54}$$

and we find that

$$H_a(s) := \frac{i_a}{u_a}(s) = \frac{\frac{J_m}{K_E K_T}s}{1 + T_m s + T_m T_a s^2} \tag{3.55}$$

where it is used that $K_E = K_T$. The following current controller is used:

$$u_a = K_i\left(i_d - i_a\right) \tag{3.56}$$

This gives the closed loop dynamics

$$\frac{i_a}{i_d}(s) = \frac{K_i H_a(s)}{1 + K_i H_a(s)} \tag{3.57}$$

In practice it is possible to select a very high gain K_i. This is a consequence of the passivity of the system when u_a is input and i_a is output, which implies that the transfer function $H_a(s)$ is positive real with phase satisfying $|\angle H_a(j\omega)| \leq 90^\circ$. Therefore we may let K_i approach infinity in the expression, which gives the approximation

$$i_a(s) = i_d(s) \tag{3.58}$$

Insertion in (3.44) gives the following result:

The model of a current controlled DC motor is given by the double integrator model

$$\theta_m(s) = \frac{1}{J_m s^2}\left[K_T i_d(s) - T_L(s)\right] \tag{3.59}$$

where the input is the desired current i_d.

The block diagram is shown in Figure 3.8.

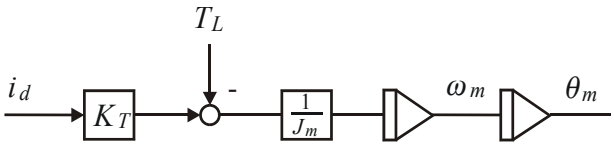

Figure 3.8: Current controlled DC motor.

Example 44 *Consider a PI controller*

$$u_a = K_i T_i \frac{1 + T_i s}{T_i s}(i_0 - i_a) \tag{3.60}$$

for armature current control where $T_i = T_m$. Assume that $T_a \ll T_m$ so that the denominator of $L(s)$ can be factored as $(1 + T_a s)(1 + T_m s)$. Then the loop transfer function is

$$L(s) = K_i \frac{J_m}{K_E K_T} \frac{1}{1 + T_a s} \tag{3.61}$$

which shows that the controller is effective also at low frequencies. Also in this case the model (3.59) results for realistic gains K_i. This is the controller that is used in practice.

3.4.3 Velocity controlled DC motor

The speed ω_m of the motor satisfies $s\theta_m = \omega_m$, and it follows that

$$\frac{\omega_m}{i_d}(s) = \frac{\omega_m}{\theta_m}(s)\frac{\theta_m}{i_d}(s) = \frac{K_T}{J_m s} \tag{3.62}$$

This transfer function is the product of a gain K_T/J_m and an integrator $1/s$. The velocity controller

$$i_d = K_\omega(\omega_d - \omega_m) \tag{3.63}$$

gives gives the closed-loop dynamics

$$\frac{\omega_m}{\omega_d}(s) = \frac{\frac{K_a}{s}}{1 + \frac{K_a}{s}} = \frac{1}{1 + \frac{s}{K_a}} \tag{3.64}$$

where

$$K_a = \frac{K_T K_\omega}{J_m} \tag{3.65}$$

is the acceleration constant of the system. We see that the velocity loop is stable as it has only one pole, which is at $s = -K_a$.

3.4.4 Position controlled DC motor

A position feedback loop is closed around the velocity loop as shown in Figure 3.9. The transfer function from the velocity ω_m to the angle θ_m is an integrator, which leads to

$$\frac{\theta_m}{\omega_d}(s) = \frac{1}{s\left(1 + \frac{s}{K_a}\right)} \tag{3.66}$$

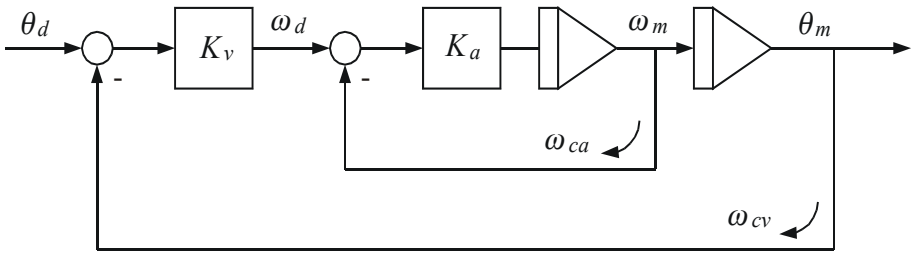

Figure 3.9: Current controlled DC motor with velocity loop and position loop. The crossover frequency of the velocity loop is $\omega_{ca} = K_a$, and the crossover frequency of the position loop is $\omega_{cv} = K_v$ provided that $K_a \gg K_v$.

The position controller

$$\omega_d = K_v (\theta_d - \theta_m) \tag{3.67}$$

where K_v is the velocity constant gives the closed-loop dynamics

$$\frac{\theta_m}{\theta_d}(s) = \frac{1}{1 + \frac{1}{K_v}s + \frac{1}{K_v K_a}s^2} \tag{3.68}$$

Usually, K_a can be selected to be several hundred rad/s, while K_v is usually limited by the first resonance in the system, which will typically occur in the range 10–100 rad/s. Therefore, we may assume that $K_a \gg K_v$, and we get

$$\frac{\theta_m}{\theta_d}(s) = \frac{1}{\left(1 + \frac{s}{K_v}\right)\left(1 + \frac{s}{K_a}\right)} \tag{3.69}$$

3.5 Motor and load with elastic transmission

3.5.1 Introduction

A situation that is often seen in control applications is that a motor is used to move some inertial load using an elastic interconnection. The elasticity may be due to a flexibility in shaft or in a gearbox, or it may be that the motor and load are interconnected by wires or with a crane that is not completely rigid. This type of system will be modelled and analyzed in this section. It turns out that the transfer functions of the system have very interesting properties that have great significance in the selection of controller structure for such systems. It will be shown that some of these properties can be explained from passivity arguments where the energy formulation can be used efficiently. The results are useful both from a practical perspective, and, in addition, the results provide valuable insight into passivity-based controller design.

3.5.2 Equations of motion

We consider a motor driving a load through an elastic transmission as shown in Figure 3.10. The equation of motion for the motor and load are given by

$$J_m \ddot{\theta}_m = T_m - T_L \tag{3.70}$$

$$J_L \ddot{\theta}_L = T_L \tag{3.71}$$

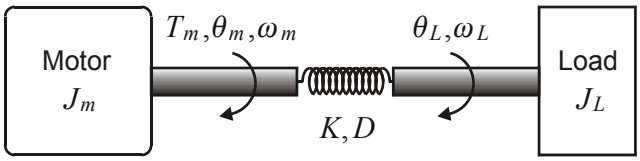

Figure 3.10: Motor with elastic transmission

where θ_m is the motor shaft angle, θ_L is the load shaft angle, T_m is the motor torque, and T_L is the load torque from the transmission on the motor shaft. The elastic transmission and the load inertia is modelled as a torsional spring with spring constant K in parallel with a torsional damper with damping coefficient D. The resulting torque is

$$T_L = -K\theta_e - D\dot{\theta}_e \tag{3.72}$$

where

$$\theta_e = \theta_L - \theta_m \tag{3.73}$$

is the elastic deflection of the the transmission. In the derivation of the transfer functions it is helpful to introduce the variable

$$\theta_r = \theta_m + \frac{J_L}{J_m}\theta_L. \tag{3.74}$$

and derive the model in terms of the variables θ_e and θ_r.

The equations of motion for θ_e and θ_r are found by combining the equations (3.70–3.72). This gives

$$\ddot{\theta}_e + \frac{D}{J_e}\dot{\theta}_e + \frac{K}{J_e}\theta_e = -\frac{1}{J_m}T_m \tag{3.75}$$

$$\ddot{\theta}_r = \frac{T_m}{J_m} \tag{3.76}$$

where

$$J_e = \frac{J_m J_L}{J}, \quad J = J_m + J_L \tag{3.77}$$

3.5.3 Transfer functions

We see from (3.75) and (3.76) that the dynamic model of the elastic deflection θ_e is a second order oscillatory system, while the rigid motion θ_r results from a double integrator from the motor torque T_m. The transfer functions from the input T_m to θ_e and θ_r are found to given by

$$\frac{\theta_e}{T_m}(s) = -\frac{1}{J_m}\frac{\left(\frac{1}{\omega_1}\right)^2}{1 + 2\zeta_1\frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \tag{3.78}$$

$$\frac{\theta_r}{T_m}(s) = \frac{1}{J_m s^2} \tag{3.79}$$

where

$$\omega_1 = \sqrt{\frac{K}{J_e}} \quad \text{and} \quad \zeta_1 = \frac{D}{2}\frac{1}{\sqrt{J_e K}} \tag{3.80}$$

The transfer functions for the original variables are found by solving (3.73) and (3.74), which gives

$$\theta_m = \frac{J_m}{J}\left(\theta_r - \frac{J_L}{J_m}\theta_e\right) \tag{3.81}$$

$$\theta_L = \frac{J_m}{J}\left(\theta_r + \theta_e\right) \tag{3.82}$$

This gives

$$\begin{aligned}\frac{\theta_m}{T_m}(s) &= \frac{J_m}{J}\left[\frac{\theta_r}{T_m}(s) - \frac{J_L}{J_m}\frac{\theta_e}{T_m}(s)\right] \\ &= \frac{1}{J}\left[\frac{1}{s^2} + \frac{\frac{J_L}{J_m}\left(\frac{1}{\omega_1}\right)^2}{1 + 2\zeta_1\frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2}\right]\end{aligned} \tag{3.83}$$

and

$$\begin{aligned}\frac{\theta_L}{T_m}(s) &= \frac{J_m}{J}\left[\frac{\theta_r}{T_m}(s) + \frac{\theta_e}{T_m}(s)\right] \\ &= \frac{J_m}{J}\left[\frac{1}{J_m s^2} - \frac{1}{J_m}\frac{\left(\frac{1}{\omega_1}\right)^2}{1 + 2\zeta_1\frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2}\right]\end{aligned} \tag{3.84}$$

After some work the following result is found:

The motor and elastic load with elastic transmission is described by the two transfer functions

$$H_{\theta m}(s) := \frac{\theta_m}{T_m}(s) = \frac{1}{Js^2}\frac{1 + 2\zeta_a\frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1\frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \tag{3.85}$$

$$H_{\theta L}(s) := \frac{\theta_L}{T_m}(s) = \frac{1}{Js^2}\frac{1 + 2\zeta_1\frac{s}{\omega_1}}{1 + 2\zeta_1\frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \tag{3.86}$$

where the parameters are given by

$$\zeta_1 = \frac{D}{2}\frac{1}{\sqrt{J_e K}}, \quad \omega_1 = \sqrt{\frac{K}{J_e}} \tag{3.87}$$

$$\zeta_a = \sqrt{\frac{J_m}{J}}\zeta_1, \quad \omega_a = \sqrt{\frac{J_m}{J}}\omega_1 < \omega_1 \tag{3.88}$$

The transfer functions are often formulated in terms of the shaft speeds $\omega_m(s) =$

$s\theta_m(s)$ and $\omega_L(s) = s\theta_L(s)$. Then the transfer functions are

$$H_{\omega m}(j\omega) \quad : \quad = \frac{\omega_m}{T_m}(s) = \frac{1}{Js}\frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \tag{3.89}$$

$$H_{\omega L}(j\omega) \quad : \quad = \frac{\omega_L}{T_m}(s) = \frac{1}{Js}\frac{1 + 2\zeta_1 \frac{s}{\omega_1}}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2} \tag{3.90}$$

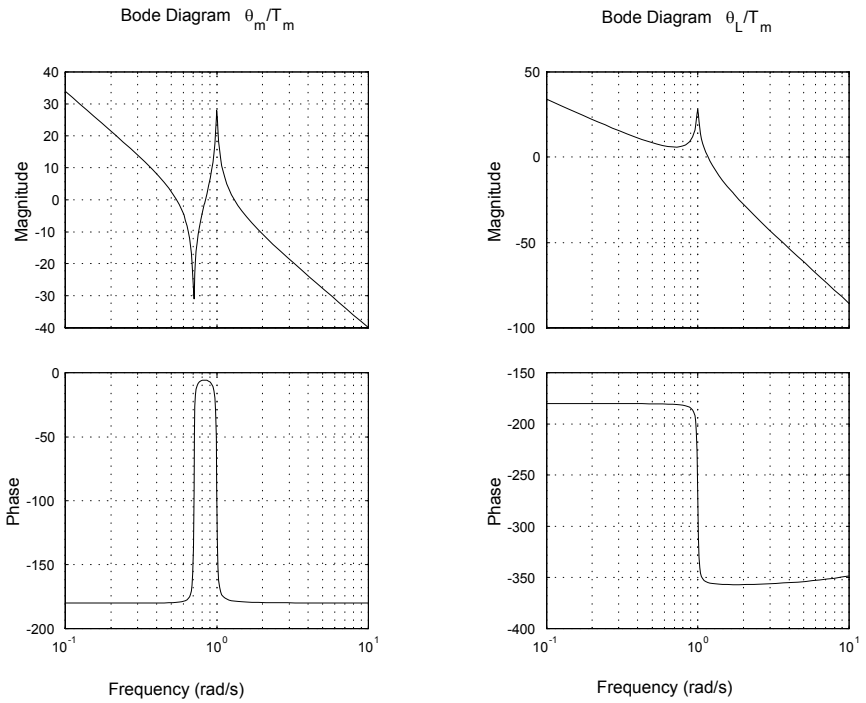

Figure 3.11: Frequency response from the motor torque T_m to the motor angle θ_m (to the left), and frequency response from the motor torque T_m to the load angle θ_L (to the right).

An important observation is that $\omega_a < \omega_1$, which means that the break frequency of the zeros in is smaller than the break frequency of the poles in $\theta_m/T_m(j\omega)$. The frequency responses are shown in Figure 3.11 for $K_1 = 0.5$, $J_m = J_1 = 1$ and $D_1 = 0.01$. Note that the frequency response $H_{\theta m}(j\omega)$ of the motor angle does not have any negative phase contribution from the elasticity, whereas the frequency response $H_{\theta L}(j\omega)$ of the load angle drops 180° because of the resonance. Obviously, this has serious consequences for controller design, and for the achievable bandwidth when feedback is taken from either $\dot{\theta}_m$ or $\dot{\theta}_L$. In practice it means that when feedback is taken from $\dot{\theta}_L$ the crossover frequency must be less than ω_1. In contrast to this, the crossover may be selected above ω_1 when feedback is taken from $\dot{\theta}_m$. Experience shows that feedback loops from $\dot{\theta}_m$ are very robust, and can be given a very high crossover frequency. Feedback from $\dot{\theta}_L$ gives an upper limit on the crossover frequency at ω_1.

3.5.4 Zeros of the transfer function

The zeros of the transfer function $H_{\theta m}(s)$ are the roots of

$$1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2 = 0 \tag{3.91}$$

Under the assumption that $\zeta_a \ll 1$ the transfer function $H_{\theta m}(s)$ will have zeros close to $\pm j\omega_a$. This means that a nonzero torque input $T_m(j\omega_a)$ with frequency ω_a will give a small $\theta_m(j\omega_a)$ as $\pm j\omega_a$ are close to the zeros of $H_{\theta m}(s)$.

From (3.85) and (3.86) it is seen that the transfer function from the motor angle to the load angle is given by

$$\frac{\theta_L}{\theta_m}(s) = \frac{\theta_L}{T_m}(s)\,\frac{T_m}{\theta_m}(s) = \frac{1 + 2\zeta_1 \frac{s}{\omega_1}}{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2} \tag{3.92}$$

This means that poles of $\theta_L(s)/\theta_m(s)$ are equal to the zeros of $H_{\theta m}(s)$ which are close to $\pm j\omega_a$. This means that the system $\theta_L(s)/\theta_m(s)$ will have resonances near $\pm j\omega_a$, so that a large amplitude in $\theta_L(j\omega_a)$ can occur with a small $\theta_m(j\omega_a)$. This agrees with the fact that the zeros of $H_{\theta m}(s)$ which are close to $\pm j\omega_a$.

3.5.5 Energy analysis

The sum of kinetic and potential energy for the motor, transmission and load is

$$V = \frac{1}{2}J_m\omega_m^2 + \frac{1}{2}J_L\omega_L^2 + \frac{1}{2}K\theta_e^2 \tag{3.93}$$

The time derivative of the energy as the system evolves will be the power $\dot{\theta}_m T_m$ supplied by the input T_m minus the power $D\dot{\theta}_e^2$ dissipated in the rotational damper. This is written

$$\dot{V} = \dot{\theta}_m T_m - D\dot{\theta}_e^2 \tag{3.94}$$

This implies that the system with input T_m and output $\dot{\theta}_m$ is passive. This again implies that the transfer function $H_{\omega m}(s)$ from the input T_m to $\omega_m = \dot{\theta}_m$ is positive real, which means that

$$\mathrm{Re}[H_{\omega m}(j\omega)] \geq 0 \qquad \text{for all } \omega. \tag{3.95}$$

Thus, from simple energy arguments we can establish that

$$|\angle H_{\omega m}(j\omega)| \leq 90^\circ \tag{3.96}$$

which implies that

$$-180^\circ \leq \angle H_{\theta m}(j\omega) \leq 0^\circ. \tag{3.97}$$

This result in agreement with the plot in Figure 3.11

3.5.6 Motor with several resonances in the load

We may connect an additional degree of freedom in the load by modifying the load into a two-port with dynamics

$$J_L\dot{\omega}_L = T_L - T_1 \tag{3.98}$$

and by adding the mechanical two-port

$$\begin{aligned} J_2\dot{\omega}_2 &= T_1 - T_2 && (3.99)\\ \frac{d}{dt}(\theta_1 - \theta_2) &= (\omega_1 - \omega_2) && (3.100)\\ T_1 &= D_1(\omega_1 - \omega_2) + K_1(\theta_1 - \theta_2). && (3.101) \end{aligned}$$

where ω_2 is the shaft speed, and J_2 is the inertia. The transmission is modelled as a torsional spring with spring constant K_1 in parallel with a torsional damper with damping coefficient D_1. The input port has effort T_1 and flow ω_1, while the output port has effort T_2 and flow ω_2. We may add on any number of additional degrees of freedom as two-ports

$$\begin{aligned} J_i\dot{\omega}_i &= T_{i-1} - T_i && (3.102)\\ \frac{d}{dt}(\theta_{i-1} - \theta_i) &= (\omega_{i-1} - \omega_i) && (3.103)\\ T_{i-1} &= D_{i-1}(\omega_{i-1} - \omega_i) + K_{i-1}(\theta_{i-1} - \theta_i) && (3.104) \end{aligned}$$

with port variables T_{i-1} and ω_{i-1} at the input and T_i and ω_i at the output. In a computational setting the inputs are ω_{i-1} and T_i, while the outputs are ω_i and T_{i-1}.

The sum of kinetic and potential energy for a motor with n degrees of freedom in the load is

$$V = \frac{1}{2}J_m\omega_m^2 + \sum_{i=1}^{n}\frac{1}{2}J_i\omega_i^2 + \frac{1}{2}K\theta_e^2 + \sum_{i=1}^{n-1}\frac{1}{2}K_i(\theta_i - \theta_{i+1})^2 \qquad (3.105)$$

The time derivative of the energy for the solutions of the system will be the power $\omega_m T_m$ supplied by the input T_m minus the power dissipated in the rotational dampers. This is written

$$\dot{V} = \omega_m T_m - D\dot{\theta}_e^2 - \sum_{i=1}^{n-1} D_i(\omega_i - \omega_{i+1})^2 \qquad (3.106)$$

This implies that the system with input T_m and output ω_m will still be passive with n degrees of freedom in the load.

3.5.7 Two motors driving an elastic load

Consider an inertia J_L with rotation angle θ_L that is driven by two motors. Motor 1 has shaft angle θ_1, inertia J_1 and motor torque T_{m1}, while motor 2 has shaft angle θ_2, inertia J_2 and motor torque T_{m2}. The motors are connected to the load using gears with gear ratio n.

The model of the system is derived by first establishing the equations of motion for the two motors and for the load, and then connecting the motors and the load by deriving expressions for the connecting torques. The equations of motion for the motors are

$$\begin{aligned} J_1\ddot{\theta}_1 &= T_{m1} - T_{g1} && (3.107)\\ J_2\ddot{\theta}_2 &= T_{m2} - T_{g2} && (3.108) \end{aligned}$$

where T_{g1} is the torque from gear 1 on motor 1, and T_{g2} is the torque from gear 2 on motor 2. The equation of motion for the load is

$$J_L\ddot{\theta}_L = \frac{1}{n}(T_{g1} + T_{g2}) - T_e \qquad (3.109)$$

where T_e is an external disturbance torque.

The elastic deformation of the gears referenced to the motor side are given by

$$\phi_1 = \theta_1 - \frac{1}{n}\theta_L, \quad \phi_2 = \theta_2 - \frac{1}{n}\theta_L$$

The gears can then be modeled as springs and dampers with torques

$$T_{g1} = K_1\phi_1 + D_1\dot{\phi}_1, \quad T_{g2} = K_2\phi_2 + D_1\dot{\phi}_2 \tag{3.110}$$

3.5.8 Energy analysis of two motors and load

The system of two motors and a load can be regarded as an interconnection of three two-ports, where the load is a two-port connected to motor 1 through a port with input $\dot{\theta}_1$ and output T_{g1}, and to motor 2 through a port with input $\dot{\theta}_2$ and output T_{g2}. The total energy of the system is

$$V = \frac{1}{2}\left(J_1\dot{\theta}_1^2 + J_2\dot{\theta}_2^2 + J_L\dot{\theta}_L^2\right) + \frac{1}{2}\left(K_1\phi_1^2 + K_2\phi_2^2\right) \geq 0 \tag{3.111}$$

The time derivative of the energy will be the power suppled by the motor torques minus the power dissipated in the dampers. This gives

$$\dot{V} = T_{m1}\dot{\theta}_1 + T_{m2}\dot{\theta}_2 - D_1\dot{\phi}_1^2 - D_2\dot{\phi}_2^2 \tag{3.112}$$

This shows that if a passive controller from $\dot{\theta}_2$ to T_{m2} is used for motor 2, then the system with input T_{m1} and output $\dot{\theta}_1$ will be passive.

Example 45 *Suppose that a PD controller*

$$T_{m2} = -K_{p2}\theta_2 - K_{d2}\dot{\theta}_2 \tag{3.113}$$

is used for motor 2. This controller is passive when $\dot{\theta}_2$ is considered to be the input and T_{m2} is the output. In agreement with this, the controller has a mechanical analog which is a spring with stiffness K_{p2} and a damper with coefficient K_{d2}. The system can then be analyzed using the energy function of the system including the mechanical analog. The energy function for this system is

$$V_a = \frac{1}{2}\left(J_1\dot{\theta}_1^2 + J_2\dot{\theta}_2^2 + J_L\dot{\theta}_L^2\right) + \frac{1}{2}\left(K_1\phi_1^2 + K_2\phi_2^2 + K_{p2}\theta_2^2\right) \geq 0 \tag{3.114}$$

which will have time derivative along the solutions of the system given by

$$\dot{V}_a = T_{m1}\dot{\theta}_1 - D_1\dot{\phi}_1^2 - D_2\dot{\phi}_2^2 - K_{d2}\dot{\theta}_2^2 \tag{3.115}$$

This shows that the system with input T_{m1} and output $\dot{\theta}_1$ is passive when the PD controller (3.113) is used.

3.6 Motor and load with deadzone in the gear

3.6.1 Introduction

In this section we will study the problem of a motor that drives a load through a gear with a deadzone. In the deadzone there is no physical contact between the input axis and

the output axis of the gear, and as a consequence of this there is no torque transmitted through the gear in the deadzone. The modeling of a gear with deadzone requires some care. In the following it will be seen that if it is assumed that there is elasticity in the gear, then the modeling is simplified. In case of a rigid gear with deadzone it is necessary to switch between two models of that have a different number of states.

3.6.2 Elastic gear with deadzone

The equations of motion for the motor and load are given by

$$J_m\ddot{\theta}_m = T_m - T_{gm} \tag{3.116}$$
$$J_L\ddot{\theta}_L = T_{gL} \tag{3.117}$$

where T_{gm} is the gear torque on the motor side, and T_{gL} is the gear torque on the load side. The gear ratio is n. The deflection between the motor and the load is given by

$$\phi = \theta_m - \frac{1}{n}\theta_L \tag{3.118}$$

The gear has a deadzone δ. This means that the gear torque is zero when $|\phi| < \delta$. Suppose that the gear is elastic, and that is can be described by a spring with stiffness K outside of the deadzone. Then the gear torques T_{gm} and T_{gL} are given as functions of the gear deflection ϕ according to

$$T_{gm}(\phi) = \begin{cases} K\left(\phi+\delta\right), & \phi \le -\delta \\ 0, & -\delta \le \phi \le \delta \\ K\left(\phi-\delta\right), & \delta \le \phi \end{cases}, \quad T_{gL}(\phi) = \frac{1}{n}T_{gm}(\phi) \tag{3.119}$$

The gear is then a mechanical two-port in impedance form where port 1 has input $\dot{\theta}_m$ and output T_{gm}, and port 2 with input $\dot{\theta}_L$ and output T_{gL}. Port 1 of the gear will then be compatible with the motor port, which has output $\dot{\theta}_m$ and input T_{gm}, and in the same way port 2 of the gear can be connected with the port of the load. The interconnection of the motor, gear and load is then straightforward, and a simulation model is given by the equations (3.116–3.119).

3.6.3 Rigid gear with deadzone

If it is assumed that the gear is rigid, then the system will have two independent degrees of freedom $\dot{\theta}_L$ and $\dot{\theta}_m$ in the deadzone, and only one degree of freedom $\dot{\theta}_L = \dot{\theta}_m$ outside of the deadzone. This means that the system changes the number of degrees of freedom from two to one when it leaves the dead-zone. In this case the gear torques are functions of the deflection when the system is inside the deadzone as

$$\left.\begin{array}{l} T_{gm}(\phi) = 0 \\ T_{gL}(\phi) = 0 \end{array}\right\}, \quad |\phi| < \delta \tag{3.120}$$

This can be regarded as an impedance model with inputs $\dot{\theta}_m$ and $\dot{\theta}_L$ and outputs T_{gm} and T_{gL}. Outside the deadzone the gear is defined by the usual gear equations

$$\left.\begin{array}{l} \dot{\theta}_L = n\dot{\theta}_m \\ T_{gm} = nT_{gL} \end{array}\right\}, \quad |\phi| = \delta \tag{3.121}$$

In terms of inputs and outputs this can be regarded as a hybrid model with inputs $\dot{\theta}_m$ and T_{gL} and outputs $\dot{\theta}_L$ and T_{gm}, or it can be seen as a cascade model with inputs $\dot{\theta}_m$ and T_{gm} and outputs $\dot{\theta}_L$ and T_{gL}. Note, however, that the gear model for $|\phi| = \delta$ cannot be put in impedance form.

The impedance model (3.120) that is valid for $|\phi| < \delta$ is a two-port with inputs and outputs that are compatible with the two-port formulation of the motor and load where shaft speed is output and torque is output. In contrast to this, the hybrid model (3.121) does not have inputs and output that can be connected to the motor and load. In fact, the system of motor, gear and load has only one degree of freedom in this case, and the models of the motor and load must be combined into one model.

The way to handle this in a simulation system is to switch between three models, where one model is valid inside the deadzone, and there is one model on each side of the deadzone. At the negative side of the deadzone where $\phi = -\delta$ the load angle is $\theta_L = n(\theta_m + \delta)$, and $J_L\ddot{\theta}_L = T_{gL}$ must be negative if the system is to stay at $\phi = -\delta$. This implies that $\ddot{\theta}_m$ must be negative, which is the case if $T_m < 0$, while the systems enters the deadzone if $T_m > 0$ At the positive side of the deadzone where $\phi = \delta$ then $T_m > 0$ will give positive acceleration, and the system will stay at $\phi = \delta$. If $T_m < 0$, then the system enters the deadzone. The three models are therefore given by

$$\left.\begin{array}{r} \left(J_m + n^2 J_L\right)\ddot{\theta}_m = T_m \\ \theta_L = n(\theta_m + \delta) \end{array}\right\} \quad \text{when } \phi = -\delta \text{ and } T_m < 0 \tag{3.122}$$

$$\left.\begin{array}{r} \left(J_m + n^2 J_L\right)\ddot{\theta}_m = T_m \\ \theta_L = n(\theta_m - \delta) \end{array}\right\} \quad \text{when } \phi = \delta \text{ and } T_m > 0 \tag{3.123}$$

$$\left.\begin{array}{r} J_m\ddot{\theta}_m = T_m \\ J_L\ddot{\theta}_L = 0 \end{array}\right\} \quad \text{otherwise} \tag{3.124}$$

In simulations it is necessary to have an event-detection method to determine when the system enters and leaves the deadzone.

3.6.4 Two motors with deadzone and load

Large space antennas need to be rotated with high accuracy at a very low speed. This will normally require a reduction gear between the motor and the antenna, where gear will typically have a deadzone. Because of this, the motor and antenna may oscillate because of the deadzone, and this may prevent the system from achieving the specified accuracy. A typical configuration for such systems is to use two motors that are connected with gears to the antenna. With this type of solution the chattering may be eliminated by pretensioning the motors in opposite directions so that both gears are loaded for moderate control torques. The equations of motion for the motors are

$$J_1\ddot{\theta}_1 = T_{m1} - T_{g1} \tag{3.125}$$

$$J_2\ddot{\theta}_2 = T_{m2} - T_{g2} \tag{3.126}$$

where T_{g1} is the torque from gear 1 on motor 1, and T_{g2} is the torque from gear 2 on motor 2. The equation of motion for the load is

$$J_L\ddot{\theta}_L = \frac{1}{n}(T_{g1} + T_{g2}) - T_e \tag{3.127}$$

where T_e is an external disturbance torque.

The gears are supposed to have a spring constant K and a deadzone δ. The deviation angle of the gear referenced to the motor side are given by

$$\phi_1 = \theta_1 - \frac{1}{n}\theta_L, \quad \phi_2 = \theta_2 - \frac{1}{n}\theta_L$$

The gears can then be modeled as a spring with a deadzone, which gives the gear torques

$$T_{gi} = \begin{cases} K\left(\phi_i + \delta\right), & \phi_i \leq -\delta \\ 0, & -\delta \leq \phi_i \leq \delta \\ K\left(\phi_i - \delta\right), & \delta \leq \phi_i \end{cases}, \quad i = 1, 2 \tag{3.128}$$

The motors can then be connected to the load with the gear equations. The system is shown in Figure 3.12.

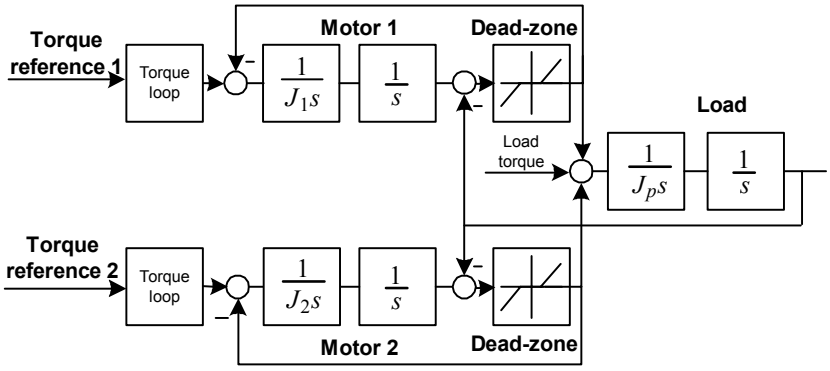

Figure 3.12: Block diagram of two motors driving a load through elastic gears with deadzone.

Due to the deadzone and the lack of damping in the gear the load may oscillate. This can be eliminated by pretensioning the gears by controlling the motors with an offset torque in opposite directions. Alternatively, damping of ϕ_1 and ϕ_2 can be achieved by including rate feedback from $\dot{\phi}_1$ and $\dot{\phi}_2$ (Leonhard 1996), (Gavronski, Beech-Brandt, Ahlstrom and Manieri 2000).

3.7 Electromechanical energy conversion

3.7.1 Introduction

Electrical motors and various electrical sensors and actuators are based on energy conversion between electrical energy and mechanical energy. This energy conversion typically takes place due to inductive and capacitive effects. The presentation that follows starts with a presentation of energy functions for inductive and capacitive circuit elements, and proceeds by extending these results to electromechanical systems.

Chapter 4

Hydraulic motors

4.1 Introduction

Hydraulic motors are widely used because of the low weight and small size of hydraulic motors compared to electrical motors with the same power. Typically a hydraulic motor can have 10 times as high power as an electrical motor of the same dimensions. Hydraulic systems can be divided into *hydrostatic* and *hydrodynamic* systems. Hydrostatic motors are motors that are driven by pressure work of a flowing fluid. In contrast to this, hydrodynamic motors are driven by the exchange of momentum of a fluid that flows past the turbine blades. We will use the term hydraulic systems for hydrostatic systems in the following. In this chapter dynamic models for hydraulic systems will be presented and analyzed. The main reference for the material is (Merritt 1967). We mention the following conversion rules between commonly used physical units:

- 1 bar = 10^5 Pa
- 1 atm = $1.01325 \cdot 10^5$ Pa
- 1 psi = 1 pound/inch^2 = 6897 Pa = 0.068 bar
- 1 Pa = 1 N/m^2

4.2 Valves

4.2.1 Introduction

Valves are important components of hydraulic systems, and are used to control flow. In this section background material and models for typical valves will be developed.

4.2.2 Flow through a restriction

The flow through a restriction or orifice in a valve is generally turbulent and is given by

$$q = C_d A \sqrt{\frac{2}{\rho} \Delta p} \tag{4.1}$$

where A is the cross section of the orifice, and Δp is the pressure drop over the orifice, and ρ is the density of the fluid. The discharge coefficient C_d is a constant. Under

the assumption of zero loss of energy, and that the flow area is not smaller than A, the discharge coefficient is found to be $C_d = 1$ from the continuity equation and Bernoulli's equation in Section 11.2.8. In practice, there will be some loss of energy, and the cross section of the flow will be somewhat smaller than the cross section A. This will reduce the discharge coefficient C_d to be in the range 0.60 – 0.65 for orifices with sharp edges, and in the range 0.8 – 0.9 when the edges are rounded.

The Reynolds number for flow through a restriction is given by

$$\mathrm{Re} = \frac{D}{A\nu} q \tag{4.2}$$

where D is the diameter of the restriction, A is the cross sectional area of the flow, and ν is the kinematic viscosity. For hydraulic oil the kinematic viscosity is approximately $\nu \approx 30 \times 10^{-6}$ m$^2/s$. The flow may be assumed to be turbulent and given by (4.1) for Reynolds numbers larger than 1000.

For a narrow restriction with low volumetric flow q the Reynolds number becomes small. If the Reynolds number becomes less than 10 the flow may be assumed to be laminar and given by

$$q = C_l \Delta p \tag{4.3}$$

where C_l is a constant and Δp is the pressure drop. This is the case for leakage flows through narrow openings, and for typical restrictions in pressure feedback channels.

When Re > 1000 the flow through a restriction will be turbulent and proportional to the square root of the pressure difference according to (4.1). When Re < 10 the flow will be laminar and proportional to the pressure drop as in (4.3).

Example 56 *The leakage flow coefficient of laminar flow through a circular tube (Hagen-Poiseuille flow) is (Merritt 1967):*

$$C_l = \frac{r^2}{8\mu L} A \tag{4.4}$$

where $\mu = \nu\rho$ is the absolute viscosity

4.2.3 Regularization of turbulent orifice flow

The turbulent flow characteristic in (4.1) is often used to describe the flow through an orifice for all Reynolds numbers. This is not physically justified, and, moreover, it creates problems in simulations as the derivative of the characteristic (4.1) is infinite at the origin where the flow approaches zero. As discussed in Section 4.2.2, the Reynolds number becomes small when the flow tends to zero, and this means that the flow is actually laminar around zero flow. On background of this it is recommended that the valve characteristic is modified so that the flow is modeled as laminar around zero flow and turbulent for high Reynolds numbers. In the following it is shown how this can be done.

First it is noted that the laminar flow characteristic (4.3) can be written in the same form as the turbulent flow characteristic (4.2) by defining a threshold constant Re_{tr} for the Reynolds number by

$$\mathrm{Re}_{tr} = 2\frac{C_d^2 DA}{C_l \mu} \tag{4.5}$$

and by expressing the laminar flow characteristic (4.3) according to

$$q = C_d \sqrt{\frac{\mathrm{Re}}{\mathrm{Re}_{tr}}} A \sqrt{\frac{2}{\rho}\Delta p} \tag{4.6}$$

This result is verified by squaring (4.6) and inserting the Reynolds number from (4.2), which gives

$$q = \frac{C_d^2}{\mathrm{Re}_{tr}} \frac{2DA}{\mu} \Delta p \tag{4.7}$$

Then (4.3) is recovered when Re_{tr} is defined by (4.5).

This means that we should seek a flow characteristic that satisfies

$$q = \begin{cases} C_d \sqrt{\frac{\mathrm{Re}}{\mathrm{Re}_{tr}}} A \sqrt{\frac{2}{\rho}\Delta p} & \mathrm{Re} \ll \mathrm{Re}_{tr} \\ C_d A \sqrt{\frac{2}{\rho}\Delta p} & \mathrm{Re}_{tr} \ll \mathrm{Re} \end{cases} \tag{4.8}$$

To obtain a solution which is defined for all Δp a smooth transition between the laminar and turbulent regimes was introduced in (Ellman and Piché 1999) using

$$q = \begin{cases} \frac{3\nu\,\mathrm{Re}_{tr}}{4} \frac{A}{D} \frac{\Delta p}{p_{tr}} \left(3 - \frac{\Delta p}{p_{tr}}\right) & \Delta p \le p_{tr} \\ C_d A \sqrt{\frac{2}{\rho}\Delta p} & p_{tr} \le \Delta p \end{cases} \tag{4.9}$$

Here the threshold pressure

$$p_{tr} = \frac{9\,\mathrm{Re}_{tr}^2\,\rho\nu^2}{8C_d^2} \frac{1}{D^2} \tag{4.10}$$

corresponds to a given threshold Re_{tr} for the Reynolds number. Assuming a circular orifice with diameter D, we have

$$A = \frac{\pi D^2}{4} \;\Rightarrow\; D^2 = \frac{4A}{\pi} \;\Rightarrow\; \frac{A}{D} = \frac{\sqrt{\pi}}{2}\sqrt{A} \tag{4.11}$$

Moreover, we define the constant

$$F_{tr} = p_{tr} A = p_{tr} \frac{\pi D^2}{4} = \frac{9\,\mathrm{Re}_{tr}^2\,\rho\nu^2}{8C_d^2} \frac{\pi}{4} \tag{4.12}$$

Then the following result has been established:

The flow through a restriction can be described by the regularized flow characteristic

$$q(A, \Delta p) = \begin{cases} \frac{3\nu\,\mathrm{Re}_{tr}}{4} \frac{\sqrt{\pi}}{2}\sqrt{A} \frac{A\Delta p}{F_{tr}} \left(3 - \frac{A\Delta p}{F_{tr}}\right) & A\Delta p \le F_{tr} \\ C_d A \sqrt{\frac{2}{\rho}\Delta p} & F_{tr} \le A\Delta p \end{cases} \tag{4.13}$$

where

$$F_{tr} = \frac{9\,\mathrm{Re}_{tr}^2\,\rho\nu^2}{8C_d^2} \frac{\pi}{4} \tag{4.14}$$

The regularized characteristic (4.13) describes the flow as laminar according to (4.3) for low Reynolds numbers, and turbulent as given by (4.1) for high Reynolds numbers. There is a smooth transition between the laminar and the turbulent flow regimes.

Numerical values for hydraulic oil are according to (Ellman and Piché 1999): $\mathrm{Re}_{tr} = 1000$, $\rho = 900$ kg/m^3, $\nu = 30 \times 10^{-6}$ m^2/s =30 cSt, $C_d = 0.6$. The regularized characteristic (4.13) is well suited for numerical simulation as it it physically justified, and it eliminates the problems that are experienced with the turbulent characteristic (4.1).

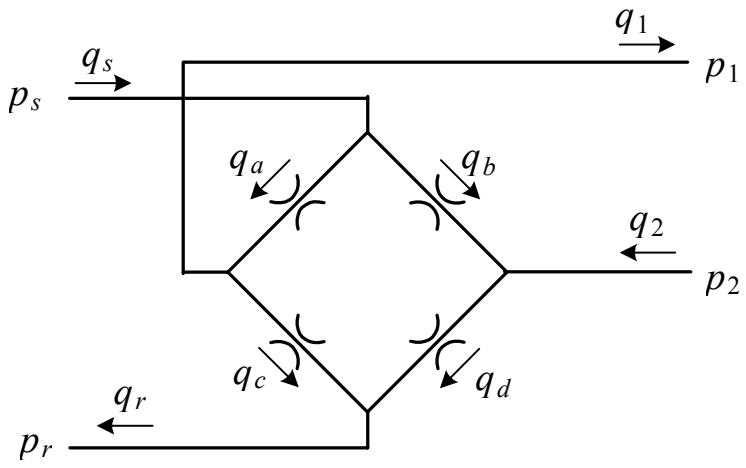

Figure 4.1: Four-way valve

4.2.4 Four-way valve

Typical flow control valves used in hydraulic systems have four orifices, and the flow is controlled by varying the flow areas of the orifices. This type of valve is termed a four-way valve. In this section we will derive the flow equations for the four-way valve shown schematically in Figure 4.1. The valve is connected to the rest of the hydraulic system through four ports, where each port has pressure as the effort variable and volumetric flow as the flow variable. The supply port is connected to the pressure supply with pressure p_s and flow q_s, the return port is connected to the return tank with pressure $p_r = 0$ and flow q_r, port 1 with pressure p_1 and flow q_1 is connected to input side of the load, and port 2 with pressure p_2 and flow q_2 is connected to the output side of the load. The volumetric flows through the orifices a, b, c and d are given by the orifice equations

$$\begin{aligned} q_a &= C_d A_a(x_v)\sqrt{\tfrac{2}{\rho}(p_s - p_1)} \\ q_b &= C_d A_b(x_v)\sqrt{\tfrac{2}{\rho}(p_s - p_2)} \\ q_c &= C_d A_c(x_v)\sqrt{\tfrac{2}{\rho}(p_1 - p_r)} \\ q_d &= C_d A_d(x_v)\sqrt{\tfrac{2}{\rho}(p_2 - p_r)} \end{aligned} \tag{4.15}$$

where the opening areas $A_a(x_v)$, $A_b(x_v)$, $A_c(x_v)$ and $A_d(x_v)$ of the orifices are assumed to be functions of the spool position x_v, and the turbulent flow characteristic (4.1) has been used to keep the equations simple. The more elaborate flow model (4.13) should be used in simulations. The port flows are related to the orifice flows through the equations

$$q_s = q_a + q_b, \quad q_r = q_c + q_d \tag{4.16}$$

$$q_1 = q_a - q_c, \quad q_2 = q_d - q_b \tag{4.17}$$

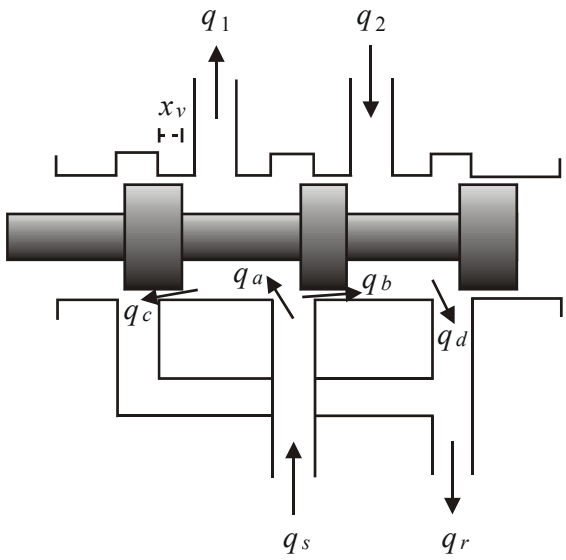

Figure 4.2: A matched and symmetric four-way valve.

4.2.5 Matched and symmetrical four-way valve

In a spool-controlled four-way valve the port openings are controlled by controlling the spool position x_v (Figure 4.2). A matched and symmetrical valve is designed so that

$$A_a(x_v) = A_d(x_v) = A_b(-x_v) = A_c(-x_v) \tag{4.18}$$

If a matched and symmetric valve is equipped with a critical spool and rectangular orifices, then the port areas are given by

$$A_a(x_v) = A_d(x_v) = \begin{cases} 0, & x_v \leq 0 \\ bx_v, & x_v \geq 0 \end{cases} \tag{4.19}$$

$$A_b(x_v) = A_c(x_v) = \begin{cases} -bx_v, & x_v \leq 0 \\ 0, & x_v \geq 0 \end{cases} \tag{4.20}$$

A matched and symmetric valve with open center spool and rectangular orifices has port openings given by

$$A_a(x_v) = A_d(x_v) = b(U + x_v), \quad |x_v| \leq U \tag{4.21}$$

$$A_b(x_v) = A_c(x_v) = b(U - x_v), \quad |x_v| \leq U \tag{4.22}$$

4.2.6 Symmetric motor and valve with critical spool

The characteristic of a four-way valve is given by the orifice equations (4.15). These equations can be combined into one characteristic if the valve is assumed to be matched and symmetric, and if it is assumed that the load is symmetric in the sense that

$$q_1 = q_2 \tag{4.23}$$

The symmetric load assumption (4.23) implies that the load does not accumulate fluid, which means that compressibility effects are not accounted for. This assumption is

therefore not consistent with the assumptions that will used in the derivation of models of hydraulic motors in the following. However, the assumption of a matched and symmetric valve and a symmetric load leads to a very useful transfer function model for valve controlled hydraulic motors, and in spite of the inconsistent assumptions introduced in the modeling, the resulting transfer function model turns out to represent the dynamics of the system with sufficient accuracy. In this connection it is it is interesting to note that major textbooks on hydraulic control systems like (Merritt 1967) and (Watton 1989) rely to a great extent on the use of the symmetric load assumption (4.23) in the analysis of control systems for valve controlled motors and cylinders.

The symmetric load assumption (4.23) together with the orifice equations (4.15) and the matching conditions (4.19, 4.20) imply the equations

$$q_a = q_d, \quad q_b = q_c \tag{4.24}$$

which in turn imply that

$$p_s + p_r = p_1 + p_2 \tag{4.25}$$

In the symmetric load case it is convenient to define the load pressure

$$p_L = p_1 - p_2 \tag{4.26}$$

and the load flow

$$q_L = \frac{1}{2}(q_1 + q_2) \tag{4.27}$$

We then find that the pressures p_1 and p_2 can be expressed as

$$p_1 = \frac{p_s + p_L}{2}, \quad p_2 = \frac{p_s - p_L}{2} \tag{4.28}$$

and the load flow can be found to be

$$q_L = C_d A_a(x_v)\sqrt{\frac{1}{\rho}(p_s - p_L)} - C_d A_b(x_v)\sqrt{\frac{1}{\rho}(p_s + p_L)} \tag{4.29}$$

If a valve with a critical spool and rectangular ports is connected to a symmetric load, then the port areas are given by

$$A_a(x_v) = \begin{cases} 0 & x_v \leq 0 \\ bx_v & x_v \geq 0 \end{cases}, \quad A_b(x_v) = \begin{cases} -bx_v & x_v \leq 0 \\ 0 & x_v \geq 0 \end{cases} \tag{4.30}$$

This leads to the following result:

The load flow of a matched and symmetric valve with a symmetric load can be expressed by valve characteristic

$$q_L = C_d b x_v \sqrt{\frac{1}{\rho}(p_s - \operatorname{sgn}(x_v) p_L)} \tag{4.31}$$

The valve characteristic (4.31) is usually written in the nondimensional form

$$\frac{q_L}{C_d b x_{v\max}\sqrt{p_s/\rho}} = \frac{x_v}{x_{v\max}}\sqrt{1 - \operatorname{sgn}(x_v)\frac{p_L}{p_s}} \tag{4.32}$$

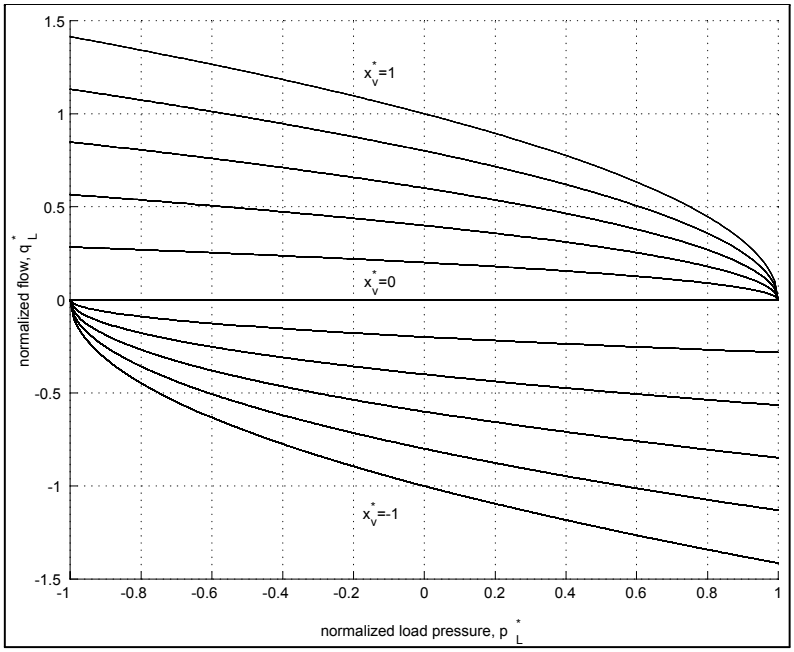

Figure 4.3: Valve characteristic

which is commonly plotted using pressure-flow curves as in (Figure 4.3), where the normalized values are

$$q_L^* = \frac{q_L}{C_d b x_{v\max}\sqrt{\frac{p_s}{\rho}}}, \quad x_v^* = \frac{x_v}{x_{v\max}}, \quad p_L^* = \frac{p_L}{p_s} \tag{4.33}$$

Valve-controlled hydraulic motors are usually designed so the load pressure p_L is limited according to $|p_L^*| < \frac{2}{3}$, and in this range the pressure-flow curves are close to linear. The valve characteristic can be linearized to give

$$q_L = K_q x_v - K_c p_L \tag{4.34}$$

where

$$K_q = \frac{\partial q_L}{\partial x_v} = C_d b\sqrt{\frac{1}{\rho}(p_s - \mathrm{sgn}(x_v)p_L)} \tag{4.35}$$

and

$$K_c = -\frac{\partial q_L}{\partial p_L} = \frac{C_d b x_v\sqrt{(1/\rho)(p_s - \mathrm{sgn}(x_v)p_L)}}{2(p_s - \mathrm{sgn}(x_v)p_L)} \tag{4.36}$$

At zero flow, zero load pressure and zero spool position, that is, at $q_L = 0$, $p_L = 0$ and $x_v = 0$ the constants of linearization are

$$K_{q0} = C_d b\sqrt{\frac{p_s}{\rho}} \tag{4.37}$$

$$K_{c0} = 0 \tag{4.38}$$

If the system is designed so that the load pressure satisfies the condition $|p_L| < \frac{2}{3}p_s$, then

$$|p_L| < \frac{2}{3}p_s \quad \Rightarrow \quad 0.577K_{q0} \leq K_q \leq 1.29K_{q0} \tag{4.39}$$

The calculated value for K_c is not consistent with what is found in practice. A more realistic value for the constant K_c is obtained by setting the spool in its zero position ($x_v = 0$) and measuring the leakage flow q_l as a function of the load pressure p_L. The flow-pressure coefficient K_{c0} is then found from $K_{c0} = q_l/p_L$.

The valve characteristic (4.31) is only valid when a matched and symmetrical valve with critical spool is connected to a symmetric load as defined by (4.23). If the load is not symmetric, then the valve must be modelled with the orifice equations (4.15).

Example 57 *A regularization of the characteristic (4.31) for simulation is found from (4.13) by defining*

$$\widetilde{p} := p_s - \mathrm{sgn}(x_v)p_L \tag{4.40}$$

and inserting $\Delta p = \tilde{p}/2$ into the expression of (4.13). This gives

$$q_L = \begin{cases} \frac{3\nu\,\mathrm{Re}_{tr}}{4}\frac{\sqrt{\pi}}{2}\sqrt{A}\frac{A\widetilde{p}}{2F_{tr}}\left(3 - \frac{A\widetilde{p}}{2F_{tr}}\right) & A\widetilde{p} \leq 2F_{tr} \\ C_d A_v \sqrt{\frac{1}{\rho}\widetilde{p}} & 2F_{tr} \leq A\widetilde{p} \end{cases} \tag{4.41}$$

where

$$F_{tr} = \frac{9\,\mathrm{Re}_{tr}^2\,\rho\nu^2}{8C_d^2}\frac{\pi}{4}, \quad A = bx_v, \quad \widetilde{p} = p_s - \mathrm{sgn}(x_v)p_L \tag{4.42}$$

4.2.7 Symmetric motor and valve with open spool

A matched and symmetric valve with open spool with rectangular ports and symmetric load gives the load flow

$$\frac{q_L}{C_d b U\sqrt{p_s/\rho}} = \left(1 + \frac{x_v}{U}\right)\sqrt{1 - \frac{p_L}{p}} - \left(1 - \frac{x_v}{U}\right)\sqrt{1 + \frac{p_L}{p}}, \quad |x_v| \leq U \tag{4.43}$$

4.2.8 Flow control using pressure compensated valves

Flow control valves can be designed with an additional pressure compensation spool that is designed to keep the pressure across the main spool constant. This is done with hydraulic feedback interconnections in the valve as shown in Figure 4.4. Let the valve have an input port with pressure p_1 and flow q_1, and an output port with pressure p_2 and flow q_2. The motion of the pressure compensation spool is controlled by a spring with force

$$F_c = -K_c x_c + F_{c0} \tag{4.44}$$

and by two compensation chambers with pressures p_3 and p_4 acting on the spool cross section A_c. Chamber 3 is connected to the pressure p_c through a restriction with laminar flow constant C_3, and chamber 4 is connected to the output pressure p_2 with a restriction with laminar flow constant C_4, and is connected to the output pressure p_2 on the on the

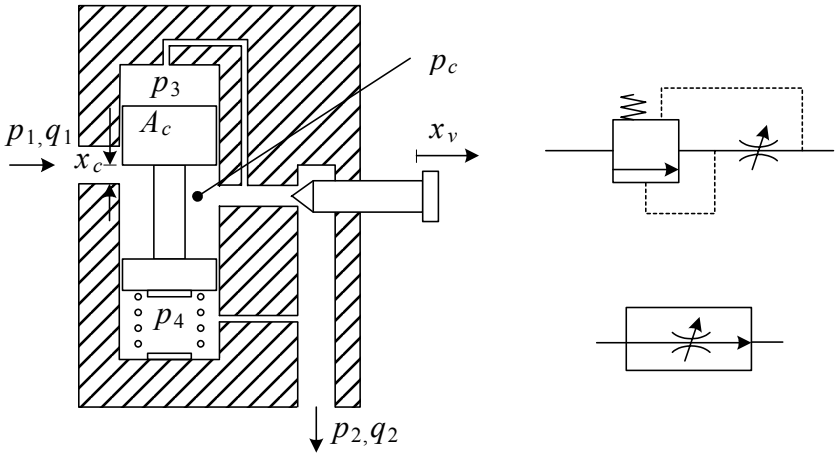

Figure 4.4: The figure shows the mechanical design and symbols for a pressure compensated valve for volume flow control. The compensation spool with position x_c which is positive in the upwards direction will be positioned so that the pressure difference $p_2 - p_c$ over the main spool is approximately constant for varying input variables q_1, p_1 and output variables q_2, p_2. The position x_v of the main spool may be controlled with a electric actuator or a pilot valve.

spring side, and to the internal pressure p_c on the opposite side. The input and output flows are given by the orifice equations

$$q_1 = C_d b_c x_c \sqrt{\frac{2}{\rho}(p_1 - p_c)} \tag{4.45}$$

$$q_2 = C_d b x_v \sqrt{\frac{2}{\rho}(p_c - p_2)} \tag{4.46}$$

where x_c is the position of the spool in the pressure compensation valve, and x_v is the position of the main valve.

To analyze the dynamics of the pressure correction valve we use the equation of motion for the pressure compensation spool

$$m_c s^2 x_c(s) + K_c x_c(s) = A_c[p_4(s) - p_3(s)] + F_{c0} \tag{4.47}$$

and the pressure dynamics of the compensation chambers

$$\frac{V_3}{\beta}\dot{p}_3 = +A_c\dot{x}_c + C_3(p_c - p_3) \tag{4.48}$$

$$\frac{V_4}{\beta}\dot{p}_4 = -A_c\dot{x}_c + C_4(p_2 - p_4) \tag{4.49}$$

The Laplace transformed pressure equations are

$$C_3\left(1 + \frac{V_3}{C_3\beta}s\right)p_3 = +A_c s x_c + C_3 p_c \tag{4.50}$$

$$C_4\left(1 + \frac{V_4}{C_4\beta}s\right)p_4 = -A_c s x_c + C_4 p_2 \tag{4.51}$$

Under the assumption that the time constants $V_3/(C_3\beta)$ and $V_4/(C_4\beta)$ are sufficiently small, we may use the approximations

$$C_3 p_3 = +A_c s x_c + C_3 p_c \Rightarrow p_3 = p_c + \frac{A_c}{C_3} s x_c \tag{4.52}$$

$$C_4 p_4 = -A_c s x_c + C_4 p_2 \Rightarrow p_4 = p_2 - \frac{A_c}{C_4} s x_c \tag{4.53}$$

Insertion in the equation of motion gives

$$m_c s^2 x_c(s) + A_c^2 \frac{C_3 + C_4}{C_3 C_4} s x_c(s) + K_c x_c(s) = A_c[p_2(s) - p_c(s)] + F_{c0} \tag{4.54}$$

which is Laplace transformed and rearranged to

$$\begin{aligned} p_c(s) - p_2(s) &= +\frac{F_{c0}}{A_c} - \frac{K_c}{A_c}\left(\frac{m_c}{K_c} s^2 + \frac{A_c^2}{K_c}\frac{C_3 + C_4}{C_3 C_4} s + 1\right) x_c(s) \\ &= +\frac{F_{c0}}{A_c} - \frac{K_c}{A_c}\left(\frac{s^2}{\omega_c^2} + 2\zeta_c \frac{s}{\omega_c} + 1\right) x_c(s) \end{aligned} \tag{4.55}$$

where $\omega_c^2 = K_c/m_c$. It can be seen that for frequencies well below ω_c, the compensation spool dynamics will satisfy

$$p_c - p_2 = \frac{F_{c0}}{A_c} - \frac{K_c x_c}{A_c} \tag{4.56}$$

It follows that if K_c/A_c is sufficiently small, then the pressure difference $p_c - p_2$ over the main spool will be approximately constant, and the flow (4.46) through the valve can be approximated by

$$q_1 = q_2 = C_d \sqrt{\frac{2}{\rho}\frac{F_{c0}}{A_c}}\, b x_v \tag{4.57}$$

This means that the use of an additional pressure compensated stage in the valve, the flow through the valve becomes proportional to the orifice area bx_v of the main spool.

4.2.9 Balance valve

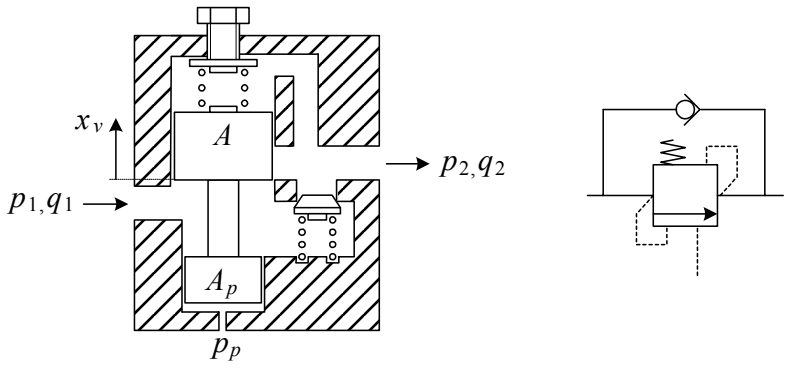

Figure 4.5: Balance valve: Mechanical design and symbol

Balance valves are used in heavy lifting operations to ensure that a hanging load will not fall down if the supply pressure is lost. In a balance valve a preset spring will push the spool towards the closed position. The inlet pressure will set up a force that will push the spool to the open position, while a high output pressure will tend to close the spool. In addition, a pilot pressure is used to assist in the opening of the valve. Consider a balance valve with inlet pressure p_1, outlet pressure p_2 and pilot pressure p_p. The balance valve has a spool with cross section A at the spring end and cross section A_p at the pilot end. Define the area

$$A_r = A - A_p \tag{4.58}$$

and the pilot ratio

$$R = \frac{A_p}{A_r}$$

The spring force on the spool is $F = F_0 + K_e x_v$ where F_0 is the pretensioning of the spring, K_e is the spring stiffness, and x_v is the spool position. We define $x_v = 0$ in the closed position, while the valve is open for $x_v > 0$. A preset pressure $p_0 = F_0/A_r$ is defined for convenience of notation. The equation of motion for the spool is

$$m_v \ddot{x}_v = p_p A_p + (p_1 - p_0) A_r - K x_v - p_2 A \tag{4.59}$$

where m_v is the mass of the spool. For a properly selected balance valve, the spool dynamics will be stable, and in the frequency range of the rest of the system it can be represented by the static characteristic

$$x_v = \frac{A_r}{K} \left[p_1 - p_0 + R p_p - p_2 (R + 1) \right], \quad 0 \leq x_v \leq x_{v,\max}$$

It is seen that the valve will open when the input pressure p_1 and the pilot pressure p_p are sufficiently high in comparison to the preset pressure p_0 and the outlet pressure p_2. The influence of the pilot pressure increases when the area ration R increases. If $p_1 > p_2$ then there will be flow in the positive direction if the spool opens. If the pressures reverse so that $p_2 > p_1$, then the flow is lead through the relief which can be considered to have a flow area A_c when it is open. The resulting flow is given by

$$q_1 = \begin{cases} C_d x_v b \sqrt{\frac{2}{\rho} (p_1 - p_2)} & p_1 > p_2 \\ -C_d A_c \sqrt{\frac{2}{\rho} (p_2 - p_1)} & p_1 < p_2 \end{cases} \tag{4.60}$$

Again the regularized orifice flow model (4.13) should be used in simulations.

4.3 Motor models

4.3.1 Mass balance

The compressibility effect of the working fluid is significant for hydraulic motors. This means that the density ρ is a function of the pressure p. A customary assumption is:

The relation between the differential $d\rho$ in density and the differential dp in pressure is given by

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \tag{4.61}$$

where β is the *bulk modulus*.

We see that the bulk modulus β has the physical dimension of pressure. Usually a numerical value of $\beta = 7 \times 10^8$ Pa =7000 bar (which corresponds to 10^5 psi) is assumed for the bulk modulus, although the value can change by a factor of 10.

The mass balance for a volume V is given by

$$\frac{d}{dt}(\rho V) = w_{in} - w_{out} \tag{4.62}$$

Here $w_{in} = \rho q_{in}$ is the mass flow and q_{in} is the volumetric flow into the volume, while $w_{out} = \rho q_{out}$ is the mass flow and q_{out} is the volumetric flow out of the volume. The density is assumed to be a function of time only. This leads to

$$\dot{\rho} V + \rho \dot{V} = \rho(q_{in} - q_{out}) \tag{4.63}$$

and insertion of the expression (4.61) leads to the following result:

The mass balance of a hydraulic volume V is

$$\frac{V}{\beta}\dot{p} + \dot{V} = q_{in} - q_{out} \tag{4.64}$$

Example 58 *The differential pressure work on a volume V of constant mass m is*

$$pdV = pd\left(\frac{m}{\rho}\right) = -pV\frac{d\rho}{\rho} = -\frac{pV}{\beta}dp \tag{4.65}$$

This means that the stored energy in a volume V due to a pressure p is

$$W_p = \int_0^p \frac{V}{\beta} p' dp' = \frac{1}{2}\frac{V}{\beta}p^2 \tag{4.66}$$

4.3.2 Rotational motors

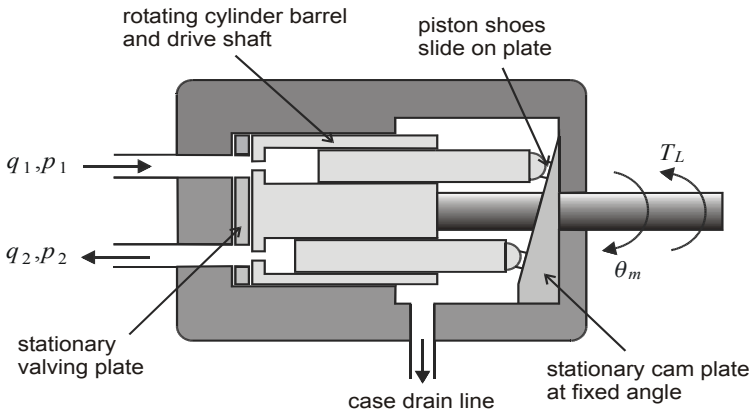

Figure 4.6: Hydraulic motor

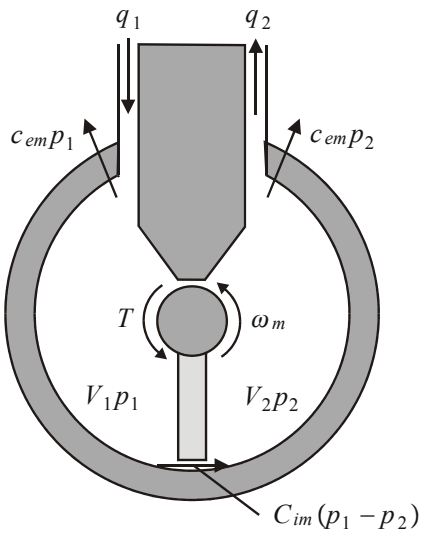

Figure 4.7: Rotational hydraulic motor of the single vane type with limited travel.

Rotational hydraulic motors are available in many designs, and are made with limited travel and continuous travel (Merritt 1967). A limited travel motor (Figure 4.7) will have a maximum rotational angle that will be slightly less than 180° or slightly less than 360°, while a motor with continuous travel (Figure 4.6) there is no limit on the rotational angle. A hydraulic motor may also run as a pump. The dynamic model is the same for a motor and pump operation.

In this section the dynamic model for a motor with limited travel will be derived. A schematic diagram of the motor is shown in Figure 4.7. The resulting model is equal to the model for motors of continuous travel. A motor with limited travel has one inlet chamber and one outlet chamber. The inlet chamber has volume V_1 and pressure p_1, and the flow into the chamber is q_1. The outlet chamber has volume V_2 and pressure p_2, and the flow out of the chamber is q_2. A motor torque is set up by the pressure difference between the two chambers, and the motor torque drives the motor shaft. A dynamic model for a rotational hydraulic motor can be derived from the mass balances of chambers 1 and 2, and the equation of motion for the motor shaft. The mass balance for the inlet and outlet chambers are

$$\dot{V}_1 + \frac{V_1}{\beta}\dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em}p_1 + q_1 \tag{4.67}$$

$$\dot{V}_2 + \frac{V_2}{\beta}\dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em}p_2 - q_2 \tag{4.68}$$

where C_{im} is the coefficient for the internal leakage and C_{em} is the coefficient for leakage out of the motor. β is the bulk modulus, and V_1 and V_2 are the volumes of the two chambers. The rate of change of the chamber volumes are proportional to the angular velocity ω_m of the motor:

$$\dot{V}_1 = -\dot{V}_2 = D_m\omega_m \tag{4.69}$$

The constant D_m is called the displacement. The shaft angle is denoted θ_m.

The motor torque T is proportional to the pressure difference, and by equating the

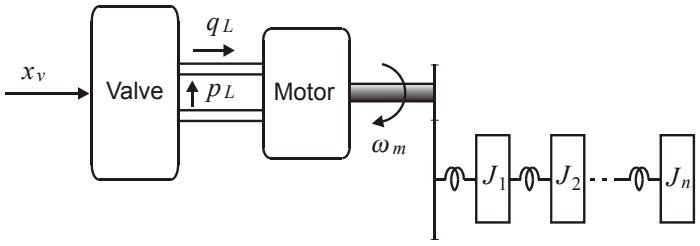

Figure 4.8: Valve controlled motor with elastic modes in the load.

power of the motor torque with the power of the working fluid we find that

$$T\omega_m = p_1\dot{V}_1 + p_2\dot{V}_2 = (p_1 - p_2)\, D_m\omega_m \tag{4.70}$$

It follows that the motor torque is

$$T = D_m\,(p_1 - p_2) \tag{4.71}$$

The equation of motion for the motor shaft is therefore

$$J_t\dot{\omega}_m = -B_m\omega_m + D_m\,(p_1 - p_2) - T_L \tag{4.72}$$

where J_t is the moment of inertia of the motor, B_m is the viscous friction coefficient, and T_L is the load torque. To sum up:

The model of a rotational hydraulic motor is given by

$$\frac{V_1}{\beta}\dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em}p_1 - D_m\omega_m + q_1 \tag{4.73}$$

$$\frac{V_2}{\beta}\dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em}p_2 + D_m\omega_m - q_2 \tag{4.74}$$

$$J_t\dot{\omega}_m = -B_m\omega_m + D_m\,(p_1 - p_2) - T_L \tag{4.75}$$

The rotational hydraulic motor can be described with a two-port for each chamber, and a three-port for the shaft dynamics. Chamber 1 has one port with effort p_1 and flow q_1, and one port with effort $T_1 = D_m p_1$ and flow ω_m. Chamber 2 has one port with effort p_2 and flow q_2, and one port with effort $T_2 = D_m p_2$ and flow ω_m. The shaft model has one port with effort T_1 and flow ω_m, one port with effort $-T_2$ and flow ω_m, and one port with effort T_L and flow ω_m. In terms of computation the systems can be interconnected if the variables q_1 and ω_m are inputs to chamber 1, q_2 and ω_m are inputs to chamber 2, and $T = T_1 - T_2$ and T_L are inputs to the shaft dynamics.

4.3.3 Elastic modes in the load

In many applications there will be elastic resonances in the load. If there is one resonance, then this can be modelled with an elastic transmission and an inertia. This can be modelled as a mechanical two-port

$$J_1\dot{\omega}_1 = T_L - T_1 \tag{4.76}$$

$$\dot{\theta}_1 = \omega_1 \tag{4.77}$$

$$T_L = D_1\,(\omega_m - \omega_1) + K_1\,(\theta_m - \theta_1) \tag{4.78}$$

where the input port has been connected to the motor shaft. The inputs to the two-port are ω_m and T_1, while T_L and ω_1 are outputs. We may add on any number of additional degrees of freedom as two-ports

$$J_i\dot{\omega}_i = T_{i-1} - T_i \tag{4.79}$$
$$T_i = D_i(\omega_{i-1} - \omega_i) + K_i(\theta_{i-1} - \theta_i) \tag{4.80}$$

with port variables T_{i-1} and ω_{i-1} at the input and T_i and ω_i at the output.

4.3.4 Hydraulic cylinder

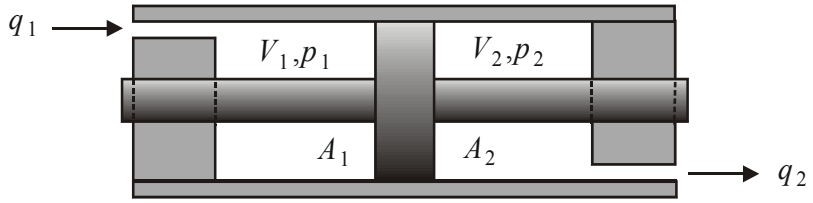

Figure 4.9: Symmetric hydraulic cylinder

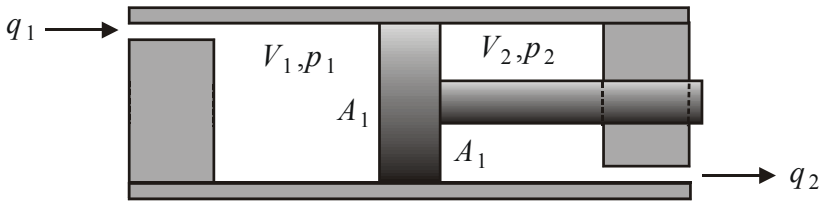

Figure 4.10: Single-rod hydraulic piston

The model of a hydraulic cylinder, which is a linear hydraulic motor, is found in same way as for a rotational motor. The cylinder will have an inlet chamber with volume $V_1 = V_{10} + A_1 x_p$ and pressure p_1, and an outlet chamber with volume with $V_2 = V_{20} - A_2 x_p$ and pressure p_2. Here V_{10} and V_{20} the chamber volumes when the piston position x_p is zero. Suppose that the piston has cross sectional area A_p, and that the piston is connected to a rod with cross section A_r.

1. If the rod goes through both chambers as in Figure 4.9, then the cylinder is said to be symmetric, and the areas A_1 and A_2 are equal and given by

$$A_1 = A_2 = A_p - A_r \tag{4.81}$$

2. If the rod goes through chamber 2 but not chamber 1 as in Figure 4.10, then the cylinder is said to have a single-rod piston and the areas are given by

$$A_1 = A_p, \quad A_2 = A_p - A_r \tag{4.82}$$

The motor force acting on the piston will be $F = A_1 p_1 - A_2 p_2$. The mass balance for the inlet and outlet chambers and the equation of motion for the piston will then give the model.

The dynamic model for a hydraulic cylinder is

$$\frac{V_{10} + A_1 x_p}{\beta} \dot{p}_1 = -C_{im}(p_1 - p_2) - C_{em} p_1 - A_1 \dot{x}_p + q_1 \tag{4.83}$$

$$\frac{V_{20} - A_2 x_p}{\beta} \dot{p}_2 = -C_{im}(p_2 - p_1) - C_{em} p_2 + A_2 \dot{x}_p - q_2 \tag{4.84}$$

$$m_t \ddot{x}_p = -B_p \dot{x}_p + A_1 p_1 - A_2 p_2 - F_L \tag{4.85}$$

Here q_1 is the flow into chamber 1, q_2 is the flow out of chamber 2, C_{im} is the coefficient for the internal leakage and C_{em} is the coefficient for leakage out of the motor, m_t is the mass of the piston and load, B_p is the viscous friction coefficient, F_L is the load force.

4.4 Models for transfer function analysis

4.4.1 Matched and symmetric valve and symmetric motor

Valve controlled hydraulic motors are used for servomechanisms where high accuracy and high bandwidth are the primary objectives. The power efficiency is moderate or low for such systems, so that for systems where power efficiency is important it is usual to have pump controlled hydraulic motors, which will be addressed in a later section. If the load is assumed to be symmetric in the sense that it satisfies the symmetric load condition (4.23), and the valve is matched and symmetric and satisfies (4.31), then it is possible to combine the two mass balances of the motor into one single mass balance, where the load flow q_L is input and the load pressure p_L is output. This is very useful in transfer function analysis of the valve controlled motor.

We consider the motor in Figure 4.6. It is assumed that when the shaft angle is zero, then the volumes are both equal to V_0. The volumes may then be written

$$V_1 = V_0 + D_m \theta_m, \quad V_2 = V_0 - D_m \theta_m \tag{4.86}$$

Subtraction of the mass balance (4.74) for chamber 2 from the mass balance (4.73) for chamber 1 gives

$$2D_m \omega_m + \frac{V_0}{\beta}(\dot{p}_1 - \dot{p}_2) + \frac{D_m \theta_m}{\beta}(\dot{p}_1 + \dot{p}_2) = q_1 + q_2 - 2C_{im}(p_1 - p_2) - C_{em}(p_1 - p_2) \tag{4.87}$$

In this expression we have the pressures and flows of the individual chambers. It is recalled that according to (4.25) the sum of the chamber pressures p_1 and p_2 are equal to the constant supply pressure p_s, and it follows that $\dot{p}_1 + \dot{p}_2 = 0$. It is then possible to reformulate (4.87) using the load pressure p_L defined in (4.26) and the load flow q_L defined in (4.27). This gives

$$\frac{V_t}{4\beta} \dot{p}_L = -C_{tm} p_L - D_m \omega_m + q_L \tag{4.88}$$

where $V_t = V_1 + V_2 = 2V_0$ is the total volume and $C_{tm} = C_{im} + \frac{1}{2} C_{em}$ is the leakage coefficient. Combining this with the equation of motion (4.72) we get the following result:

The model of a symmetric hydraulic motor with a matched and symmetric valve is given by

$$\frac{V_t}{4\beta}\dot{p}_L = -C_{tm}p_L - D_m\omega_m + q_L \tag{4.89}$$

$$J_t\dot{\omega}_m = -B_m\omega_m + D_m p_L - T_L \tag{4.90}$$

Example 59 *An energy function of the motor is*

$$V = \frac{1}{2}J_t\omega_m^2 + \frac{1}{2}\frac{V}{4\beta}p_L^2 \tag{4.91}$$

The time derivative is

$$\begin{aligned}\dot{V} &= \omega_m J_t\dot{\omega}_m + p_L\frac{V_t}{4\beta}\dot{p}_L \\ &= -B_m\omega_m^2 - \omega_m T_L + \omega_m D_m p_L - C_{tm}p_L^2 - p_L D_m\omega_m + p_L q_L \\ &= p_L q_L - \omega_m T_L - B_m\omega_m^2 - C_{tm}p_L^2 \end{aligned} \tag{4.92}$$

We see that is the load dynamics is passive, then the system with input q_L and output p_L is passive. However, this does not have much relevance for the controller design for this system.

4.4.2 Valve controlled motor: Transfer function

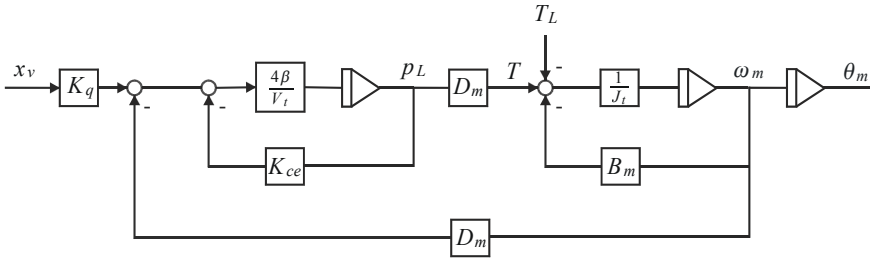

Figure 4.11: Valve controlled hydraulic motor.

A linearized dynamic model for a valve controlled motor is found by inserting the linearized valve characteristic (4.34) into the model (4.89, 4.90). The result is

$$\frac{V_t}{4\beta}\dot{p}_L = -K_{ce}p_L - D_m\omega_m + K_q x_v \tag{4.93}$$

$$J_t\dot{\omega}_m = -B_m\omega_m + D_m p_L - T_L \tag{4.94}$$

$$\dot{\theta}_m = \omega_m \tag{4.95}$$

where $K_{ce} = K_c + C_{tm}$ is the leakage coefficient for motor and valve, B_m is the viscous friction coefficient, while θ_m is the angle of rotation of the motor shaft. A block diagram is shown in Figure 4.11. Note the similarity to Figure 3.7.

The Laplace transformed model is found by Laplace transformation of the model (4.93, 4.94). This gives

$$K_{ce}\left(1+\frac{V_t}{4\beta K_{ce}}s\right)p_L = \left(-D_m s\theta_m + K_q x_v\right) \tag{4.96}$$

$$\left(J_t s^2 + B_m s\right)\theta_m = D_m p_L - T_L \tag{4.97}$$

Insertion of the mass balance (4.96) into the equation of motion (4.97) gives

$$\begin{aligned} K_{ce}\left(1+\frac{V_t}{4\beta K_{ce}}s\right)\left(J_t s^2 + B_m s\right)\theta_m &= -D_m^2 s\theta_m + D_m K_q x_v \\ &\quad -K_{ce}\left(1+\frac{V_t}{4\beta K_{ce}}s\right)T_L \end{aligned} \tag{4.98}$$

which can be rearranged as

$$\theta_m(s) = \frac{\frac{K_q}{D_m}x_v(s) - \frac{K_{ce}}{D_m^2}\left(1+\frac{V_t}{4\beta K_{ce}}s\right)T_L(s)}{s\left[\frac{V_t J_t}{4\beta D_m^2}s^2 + \left(\frac{K_{ce}J_t}{D_m^2}+\frac{B_m V_t}{4\beta D_m^2}\right)s + \left(1+\frac{B_m K_{ce}}{D_m^2}\right)\right]} \tag{4.99}$$

Under the assumption that $B_m = 0$ the standard formulation of this expression is obtained:

The Laplace transformed model of a symmetric hydraulic motor with matched and symmetric valve and $B_m = 0$ is given by

$$\theta_m(s) = \frac{\frac{K_q}{D_m}x_v(s) - \frac{K_{ce}}{D_m^2}\left(1+\frac{s}{\omega_t}\right)T_L(s)}{s\left(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}\right)} \tag{4.100}$$

where ω_h the hydraulic undamped natural frequency, ζ_h is the the relative damping, and ω_t is the break frequency of the pressure dynamics defined by

$$\omega_h^2 = \frac{4\beta D_m^2}{V_t J_t}, \quad \zeta_h = \frac{K_{ce}}{D_m}\sqrt{\frac{\beta J_t}{V_t}}, \quad \omega_t = \frac{4\beta K_{ce}}{V_t} \tag{4.101}$$

We note that

$$2\zeta_h\omega_h = \frac{4\beta K_{ce}}{V_t} = \omega_t \tag{4.102}$$

The transfer function from the spool position x_v to the shaft angle θ_m is given by

$$H_m(s) = \frac{\theta_m}{x_v}(s) = \frac{\frac{K_q}{D_m}}{s\left(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}\right)} \tag{4.103}$$

The magnitude of the frequency response $H_m(j\omega)$ is shown in Figure 4.12 with the parameters $K_q/D_m = 40$, $\omega_h = 400$ rad/s and $\zeta_h = 0.1$.

The transfer function $H_m(s)$ has a pole in $s = 0$, which corresponds to the integrator from angular velocity to the valve angle. This means that for low frequencies where

$$(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}) \approx 1 \tag{4.104}$$

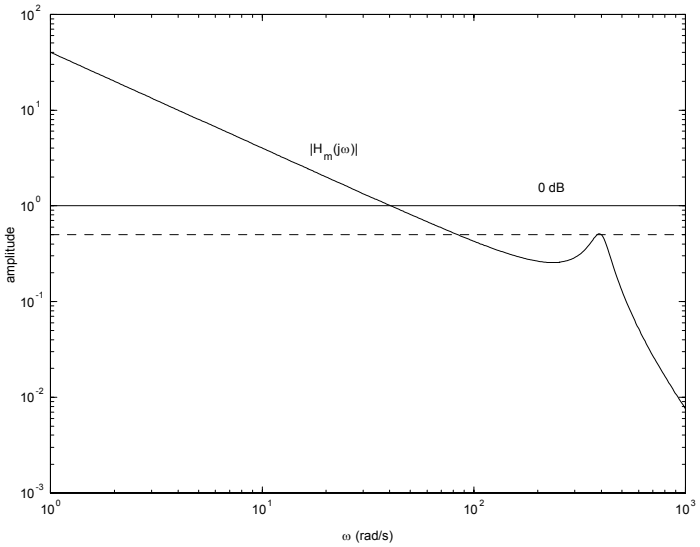

Figure 4.12: Magnitude of the frequency response from the valve spool position x_v to the motor angle θ. Numerical values are $K_q/D_m = 40$, $\omega_h = 400$ rad/s and $\zeta_h = 0.1$. The dashed line is drawn at -6 dB.

the motor velocity ω_m will be proportional to the spool position x_v. The gain K_q/D_m of the transfer function is the flow gain K_q divided by the displacement D_m. The displacement is given by the geometry of the motor, and will be available with high accuracy. Thus, variations in the gain will only depend on the flow gain K_q. The flow gain will vary with the factor $\sqrt{p_s - p_L}/\sqrt{p_s}$, and under the usual design rule $|p_L| < \frac{2}{3}p_s$ the flow gain will be between 57.7 % and 129 % of the nominal value.

The hydraulic undamped natural frequency ω_h is an important parameter in the design of electrohydraulic servomechanisms. The undamped natural frequency is given by β, D_m and J_t. The parameters J_t and D_m can be found with high accuracy, while the bulk modulus β may vary. However, the numerical value $\beta = 7.0 \cdot 10^8$ Pa ($= 10^5$ psi) (Merritt 1967) will in many cases be reasonably accurate when the working fluid is hydraulic oil. It turns out that the leakage coefficient K_{ce} will be dominated by the valve.

The transfer function to the load torque T_L to the shaft angle θ_m is

$$\frac{\theta_m}{T_L}(s) = \frac{-\frac{K_{ce}}{D_m^2}\left(1 + \frac{s}{\omega_t}\right)}{s\left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \tag{4.105}$$

4.4.3 Hydraulic motor with P controller

With a proportional controller

$$x_v = K_p(\theta_d - \theta_m) \tag{4.106}$$

the loop transfer function for a hydraulic motor is

$$L(s) = K_p H_m(s) = \frac{K_v}{s\left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \tag{4.107}$$

where

$$K_v = \frac{K_p K_q}{D_m} \tag{4.108}$$

is the velocity constant of the closed loop system. The loop transfer function $L(s)$ has a pole in $s = 0$ and two complex conjugated poles as the numerical value of the relative damping is typically in the range $0.1 < \zeta_h < 0.5$.

An important parameter in the control design is the gain of the loop transfer function at ω_{180}, which is the frequency where the frequency response $L(j\omega_h)$ has a phase of 180°. A closer inspection of the loop transfer function $L(s)$ reveals that $\omega_{180} = \omega_h$, and that

$$|L(j\omega_h)| = \frac{K_v}{2\zeta_h \omega_h}, \quad \angle L(j\omega_h) = -180^\circ \tag{4.109}$$

Using the expression for ζ_h in (4.101) we find that

$$|L(j\omega_{180})| = \frac{K_v}{2\zeta_h \omega_h} \tag{4.110}$$

Thus, a gain margin of $\Delta K = 6$ dB, which occurs for $|L(j\omega_{180})| = 1/2$, is achieved with

$$K_v = \zeta_h \omega_h \Rightarrow K_p = \frac{D_m}{K_q} \zeta_h \omega_h \tag{4.111}$$

For the numerical values in Figure 4.12 a gain margin of 6 dB will be obtained with $K_v = \zeta_h \omega_h = 40$, which corresponds to a gain of $K_p = K_v D_m / K_q = 1$, and it follows that in Figure 4.12 we have $L(j\omega) = H_m(j\omega)$ if $K_p = 1$. The dashed line in the figure, which is drawn at -6 dB, will therefore indicate $|L(j\omega_{180})| = K_v/(2\zeta_h\omega_h) = 0.5$.

Then, from Nyquist stability theory it may be concluded that:

A rotation motor with matched and symmetric valve that is controlled with a proportional controller $x_v = K_p(\theta_d - \theta_m)$ will be stable if the velocity constant satisfies

$$K_v = \frac{K_p K_q}{D_m} \leq 2\zeta_h \omega_h \Rightarrow K_p \leq 2\frac{D_m}{K_q} \zeta_h \omega_h \tag{4.112}$$

A gain margin of 6 dB is achieved with

$$K_v = \zeta_h \omega_h \Rightarrow K_p = \frac{D_m}{K_q} \zeta_h \omega_h \tag{4.113}$$

Example 60 *Suppose that the leakage coefficient K_{ce} is determined by the valve, which is the typical situation as the leakage in motors are usually negligible. Then if two different motors are used with the same valve and the same fluid, the constants K_q, K_{ce} and β will be unchanged. It follows that the stability limit will be proportional to V_t^{-1}.*

Example 61 *If a nonzero B_m is used, the undamped natural frequency and the relative damping are given by*

$$\omega_h^2 = \frac{4\beta D_m^2}{V_t J_t}\left(1 + \frac{B_m K_{ce}}{D_m^2}\right)$$

and

$$\zeta_h = \left(\frac{K_{ce}}{D_m}\sqrt{\frac{\beta J_t}{V_t}} + \frac{B_m}{4D_m}\sqrt{\frac{V_t}{\beta J_t}}\right)\left(1 + \frac{B_m K_{ce}}{D_m^2}\right)^{-\frac{1}{2}}$$

We note that in this case

$$2\zeta_h\omega_h = \left(\frac{K_{ce}J_t}{D_m^2} + \frac{B_m V_t}{4\beta D_m^2}\right)\frac{4\beta D_m^2}{V_t J_t} = \frac{4\beta K_{ce}}{V_t} + \frac{B_m}{J_t} \tag{4.114}$$

4.4.4 Symmetric cylinder with matched and symmetric valve

A symmetric cylinder, which is a cylinder with a symmetric piston, has a dynamic model that is similar to a rotary motor. Therefore, the model of a symmetric cylinder with matched and symmetric valve is found by combining the two mass balances and using the equation of motion for the mass. This gives

$$\frac{V_t}{4\beta}\dot{p}_L = -C_{tp}p_L - A_p\dot{x}_p + q_L \tag{4.115}$$

$$m_t\ddot{x}_p = -B_p\dot{x}_p + A_p p_L - F_L \tag{4.116}$$

The Laplace transform of a symmetric cylinder with matched and symmetric valve is

$$x_p(s) = \frac{\frac{K_q}{A_p}x_v(s) - \frac{K_{ce}}{A_p^2}\left(1 + \frac{s}{\omega_t}\right)F_L(s)}{s\left(1 + 2\zeta_h\frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \tag{4.117}$$

where

$$\omega_h^2 = \frac{4\beta A_p^2}{V_t m_t}, \quad \zeta_h = \frac{K_{ce}}{A_p}\sqrt{\frac{\beta m_t}{V_t}}, \quad \omega_t = \frac{4\beta K_{ce}}{V_t} \tag{4.118}$$

if it is assumed that $B_p = 0$.

Example 62 *With a proportional controller $x_v = K_p(x_d - x_p)$, the stability limit for the gain K_p is found in the same way as for the rotation motor with matched and symmetric valve. Moreover, to have a gain margin of 6 dB the gain should be selected as*

$$K_p = \frac{A_p}{K_q}\zeta_h\omega_h \tag{4.119}$$

Example 63 *A hydraulic cylinder is to be selected so that it can generate a force F_0 for a given supply pressure p_s, and so that the position x_p of the piston can be changed between zero and $\bar{x}_p$. The cross sectional area of the cylinder must then be $A_p = F_0/p_s$, and the volume is found from $V_t = A_p\bar{x}_p = F_0\bar{x}_p/p_s$. Note that the required volume can be found if the force F_0, the stroke $\bar{x}_p$ and the supply pressure p_s is given. Suppose that a similar installation with the same valve has volume V_s and bandwidth ω_s as defined by the crossover frequency. Then the bandwidth of the system with volume V_t will be*

$$\omega_c = \omega_s\frac{V_s}{V_t} \tag{4.120}$$

4.4.5 Pump controlled hydraulic drive with P controller

The model of a pump controlled motor is derived in Section 4.7.2. At this point we simply state that the Laplace transformed model of a pump controlled motor with constant motor displacement D_m and constant pump speed ω_p is

$$\theta_m(s) = \frac{\frac{k_p\omega_p}{D_m}\phi_p(s) - \frac{C_t}{D_m^2}\left(1+\frac{s}{\omega_0}\right)T_L(s)}{s\left(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}\right)} \tag{4.121}$$

where

$$\omega_h^2 = \frac{\beta D_m^2}{V_0 J_t}, \quad \zeta_h = \frac{C_t}{2D_m}\sqrt{\frac{\beta J_t}{V_0}}, \quad \omega_0 = \frac{\beta C_t}{V_0} \tag{4.122}$$

It is seen that the dynamic model of a pump controlled hydraulic motor has the same structure as a valve controlled motor. The main differences are:

- The volume V_0 includes the high pressure pipe and the high pressure chamber of the motor and pump. Only the high pressure side is considered to be driving the motor, and because of this the volume term in ω_h^2 is V_0 for the pump controlled system instead of the $4V_t$ term which appears for the valve controlled motor.
- The gain $k_p\omega_p/D_m$ does not vary and can be found with high accuracy.
- The relative damping of the system may be very small compared to a valve controlled motor where the main leakage is in the valve. Additional leakage may be introduced in the system to make it less oscillatory. This will give loss of power, but it may be necessary to achieve satisfactory performance.

The usual controller for this system is a proportional feedback from the motor shaft angle θ_m:

$$u_p = K_p\left(\theta_d - \theta_m\right) \tag{4.123}$$

The loop transfer function $L(s)$ is seen to be

$$L(s) = \frac{K_v}{s\left(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}\right)} \tag{4.124}$$

where $K_v = K_p k_p \omega_p / D_m$ is the velocity constant. The system is seen to be stable if and only if

$$K_v \leq 2\zeta_h\omega_h \tag{4.125}$$

where $2\zeta_h\omega_h = \omega_0$. Typically, a gain margin equal to 2 will be used, in which case the velocity constant is set to

$$K_v = \zeta_h\omega_h \;\Rightarrow\; K_p = \frac{D_m}{k_p\omega_p}\zeta_h\omega_h \tag{4.126}$$

4.4.6 Transfer functions for elastic modes

Suppose that the load is driven by the motor through an elastic transmission as shown in Figure 4.8. We restrict our analysis to one mechanical resonance in the load, which is the case when an inertia is connected to the motor shaft through a spring and a damper.

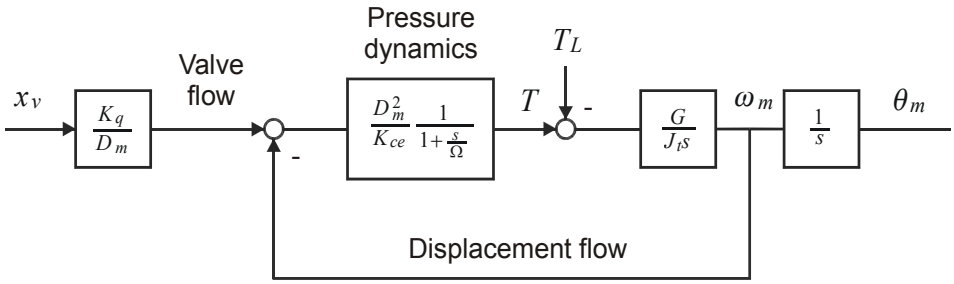

Figure 4.13: Block diagram for valve controlled motor with elastic modes in the load.

Then the transfer function from the motor torque T to the motor angle θ_m is given by (3.85) to be

$$Js^2\theta_m(s) = G(s)\,T(s) \tag{4.127}$$

where

$$G(s) = \frac{1 + 2\zeta_a \frac{s}{\omega_a} + \left(\frac{s}{\omega_a}\right)^2}{1 + 2\zeta_1 \frac{s}{\omega_1} + \left(\frac{s}{\omega_1}\right)^2}, \quad \omega_a < \omega_1 \tag{4.128}$$

The pressure dynamics will still be given by (4.96), while the equation of motion is found from (4.127) and $T = D_m p_L$. This gives

$$K_{ce}\left(1 + \frac{V_t}{4\beta K_{ce}} s\right) p_L = (-D_m s\theta_m + K_q x_v) \tag{4.129}$$

$$J_t s^2 \theta_m = G(s)\, D_m p_L \tag{4.130}$$

Insertion of the first equation into the second gives

$$K_{ce}\left(1 + \frac{V_t}{4\beta K_{ce}} s\right) J_t s^2 \theta_m = G(s)\left(-D_m^2 s\theta_m + D_m K_q x_v\right) \tag{4.131}$$

This is more or less the same equation as (4.98) except for the appearance of $G(s)$, and the transfer function is found to be

$$H_e(s) = \frac{\theta_m}{x_v}(s) = \frac{G(s)\,\frac{K_q}{D_m}}{s\left(G(s) + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \tag{4.132}$$

which reduces to the transfer function $H_m(s)$ given by (4.103) for the rigid case if $G(s) = 1$. The block diagram is given in Figure 4.13.

1. In the frequency ranges $\omega \ll \omega_a$ and $\omega \gg \omega_1$ we will have $G(j\omega) \approx 1$ and therefore $H_s(j\omega) \approx H_m(j\omega)$. This means that in the frequency range below ω_a and above ω_1 the frequency response is the same for the rigid and the elastic case.

2. If $\omega_1 \ll \omega_h$, then

$$H_e(s) \approx \frac{\frac{K_q}{D_m}}{s\left(1 + 2\zeta_h \frac{s}{\omega_h} + \frac{s^2}{\omega_h^2}\right)} \tag{4.133}$$

3. If $\omega_h \ll \omega_a$, then

$$H_e(s) \approx \frac{G(s)\frac{K_q}{D_m}}{s\left(1+2\zeta_h\frac{s}{\omega_h}+\frac{s^2}{\omega_h^2}\right)} \tag{4.134}$$

4.4.7 Mechanical analog

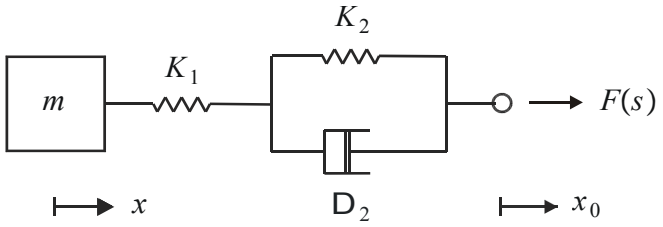

Figure 4.14: Mechanical analog of hydraulic motor with proportional position controller.

We consider a mechanical analog of a valve controlled motor with P controller. The analog is modelled as in Figure 4.14 with a spring S_1 of stiffness K_1 in series with a parallel interconnection of a spring S_2 of stiffness K_2 and a damper with coefficient D_2. The spring S_1 is connected to a mass m in position x, while the spring S_2 is connected to a moving attachment of position x_0. The force from the spring S_1 on the mass is

$$F_1(s) = K_p\frac{\left(1+\frac{s}{\omega_1}\right)}{\left(1+\frac{s}{\omega_2}\right)}\left[x_0(s)-x(s)\right] \tag{4.135}$$

which clearly shows that this corresponds to a PD controller with limited derivative action, and where the constants are given by.

$$K_p = \frac{K_1K_2}{K_1+K_2}, \quad \omega_1 = \frac{K_2}{D_2}, \quad \omega_2 = \frac{K_1+K_2}{D_2} \tag{4.136}$$

Suppose that a mass m with position x and friction coefficient B is actuated by the force F_1 from the mechanical interconnection, and that the mass is subject to the load force F_L. Then the equation of motion will be

$$(ms^2+Bs)x(s) = K_p\frac{\left(1+\frac{s}{\omega_1}\right)}{\left(1+\frac{s}{\omega_2}\right)}\left[x_0(s)-x(s)\right] - F_L \tag{4.137}$$

Consider a hydraulic motor with equation of motion given by (4.98)

$$\left(J_ts^2+B_ms\right)\theta_m = \frac{D_m^2}{K_{ce}}\frac{\frac{K_q}{D_m}x_v - s\theta_m}{1+\frac{s}{\omega_t}} - T_L$$

where

$$\omega_t = \frac{4\beta K_{ce}}{V_t} \tag{4.138}$$

Then, with proportional feedback $x_v = K_p(\theta_0 - \theta_m)$ this becomes

$$\left(J_t s^2 + B_m s\right)\theta_m = -K_v \frac{D_m^2}{K_{ce}} \frac{1+\frac{s}{K_v}}{1+\frac{s}{\omega_t}}\theta_m + K_v \frac{D_m^2}{K_{ce}} \frac{1}{1+\frac{s}{\omega_t}}\theta_0 - T_L \tag{4.139}$$

where $K_v = K_p K_q / D_m$ is the velocity constant. If we introduce the variable θ_d defined by

$$\theta_0(s) = \frac{1}{1+\frac{s}{K_v}}\theta_d(s) \tag{4.140}$$

then equation (4.139) can be written

$$\left(J_t s^2 + B_m s\right)\theta_m = K_v \frac{D_m^2}{K_{ce}} \frac{1+\frac{s}{K_v}}{1+\frac{s}{\omega_t}} \left[\theta_d(s) - \theta_m(s)\right] - T_L$$

and we find that the dynamics are the same as for the mechanical analog if the constants satisfy

$$K_v \frac{D_m^2}{K_{ce}} = \frac{K_1 K_2}{K_1 + K_2}, \quad K_v = \frac{K_2}{D_2}, \quad \omega_t = \frac{K_1 + K_2}{D_2} \tag{4.141}$$

We may solve for the parameters of the mechanical analog, which are found to be

$$K_1 = \frac{4\beta D_m^2}{V_t}, \quad K_2 = \frac{K_v}{\omega_t - K_v}\frac{4\beta D_m^2}{V_t}, \quad D_2 = \frac{1}{\omega_t - K_v}\frac{4\beta D_m^2}{V_t} \tag{4.142}$$

Note that the mechanical analog is passive if and only if $K_v \leq \omega_t$, which is also the condition for stability of the closed loop system. It is interesting to see that in the unstable case when $K_v > \omega_t$, then the mechanical analog is no longer passive as K_2 and D_2 become negative for $K_v > \omega_t$. This means that the closed loop system has a passive mechanical analog if and only if the closed loop system is stable. This result also applies to pump-controlled motors, which have the same transfer transfer functions as valve controlled motors with minor adjustments in the parameters.

4.5 Hydraulic transmission lines

4.5.1 Introduction

The mass balance of a volume V was found in (4.64) to be given by

$$\frac{V}{\beta}\dot{p} + \dot{V} = q_{in} - q_{out} \tag{4.143}$$

In the derivation of this equation it was assumed that the pressure would be the same over the volume, which means that the pressure $p = p(t)$ is a function of time only. Pressure changes will propagate with the speed of sound, which is about $c = 1000$ m/s for hydraulic oil. If the volume is reasonably small so that the pressure only propagates less than one meter, then pressure differences in the volume will disappear after 1 ms. It is then normally justified to assume that the pressure is the same over the volume.

However, there are systems where the spatial variations of the pressure must be taken into account by describing the pressure as a function of position and time. If the volume V is a pipe of length L, then the time for a pressure change to propagate through the pipe will be $T = L/c$. Long pipes are used in large hydraulic installations where pipes

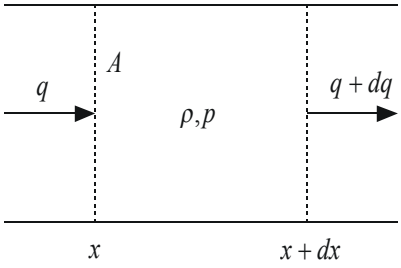

Figure 4.15: Volume element for hydraulic transmission line.

of length up to 10 m are not uncommon. Moreover, in offshore oil and gas production pipes of several hundred meters may be used. A propagation time of $T = 10$ ms will result if $L = 10$ m. This introduces a time delay that may be significant if bandwidths up to 100 rad/s (16 Hz) are required. The propagation time will increase to $T = 0.5$ s for a pipe with length $L = 500$ m. In addition to problems associated with time delays, a severe problem with long hydraulic pipes is that pressure pulses may be reflected at the end of the pipe. This may cause strong pressure fluctuations in the system that will limit bandwidth, and that will increase the risk of mechanical damage to the system.

On background of this there is a need to describe the pressure and flow dynamics of long hydraulic pipes. It will be shown that the dynamics of such systems are described by partial differential equations in the form of the wave equation, and that this is a special case of the theory of transmission lines. The relevant models, analysis tools and simulation algorithms will be presented in the following. Basic references are (Goodson and Leonard 1972), (Stecki and Davis 1986a) and (Stecki and Davis 1986b).

4.5.2 PDE Model

A hydraulic transmission line is a pipe filled with a compressible liquid which may be water or mineral oil. The pipe is of length L and has a cross section of area A, and the length coordinate along the pipe is denoted x. The pressure of the liquid is $p(x,t)$, the volumetric flow is $q(x,t)$, the density is $\rho(x,t)$, and the bulk modulus is β.

The dynamic model is developed in detail in Section 11.2.7. The model is found from the mass balance and momentum balance of a differential control volume Adx where A is the cross sectional area of the pipe and x is the length coordinate along the pipe. The velocity along the pipe is denoted v, and the volumetric flow is $q = A\bar{v}$, where $\bar{v}$ is the velocity v averaged over the cross section A. The friction force on the volume element is Fdx where $F = F(q)$ is assumed to be a function of the volumetric flow q. Then, assuming that the velocity $\bar{v}$ is small, and that the density can be considered to be a constant ρ_0, the following model is found from the mass balance and the momentum balance:

The model for a hydraulic transmission line can be written as the partial differential equations

$$\frac{\partial p(x,t)}{\partial t} = -cZ_0 \frac{\partial q(x,t)}{\partial x} \tag{4.144}$$

$$\frac{\partial q(x,t)}{\partial t} = -\frac{c}{Z_0}\frac{\partial p(x,t)}{\partial x} - \frac{F[q(x,t)]}{\rho_0} \tag{4.145}$$

where the sonic velocity c and the line impedance Z_0 are defined by

$$c = \sqrt{\frac{\beta}{\rho_0}}, \quad Z_0 = \frac{\rho_0 c}{A} = \frac{\sqrt{\rho_0 \beta}}{A} \tag{4.146}$$

4.5.3 Laplace transformed model

The PDE model (4.144, 4.145) can be Laplace transformed to give

$$\frac{\partial q(x,s)}{\partial x} = -\frac{s}{cZ_0} p(x,s) \tag{4.147}$$

$$\frac{\partial p(x,s)}{\partial x} = -\frac{Z_0 s}{c} q(x,s) - \frac{Z_0 F[q(x,s)]}{c\rho_0} \tag{4.148}$$

The friction force $F[q(x,s)]$ will depend on the volumetric flow $q(x,s)$, and different models will result depending on the friction model that is used. It is commonly assumed that the friction $F[q(x,s)]$ is a linear function of $q(x,s)$. This makes it possible to define the propagation operator $\Gamma(s)$ according to

$$\frac{Z_0 \Gamma(s)^2}{LTs} q(x,s) = \frac{Z_0 s}{c} q(x,s) + \frac{Z_0 F[q(x,s)]}{c\rho_0} \tag{4.149}$$

where $T = L/c$ is the propagation time.

The transmission line model can be written

$$\frac{\partial q(x,s)}{\partial x} = -\frac{Ts}{LZ_0} p(x,s) \tag{4.150}$$

$$\frac{\partial p(x,s)}{\partial x} = -\frac{Z_0 \Gamma(s)^2}{LTs} q(x,s) \tag{4.151}$$

where $\Gamma(s)$ is the wave propagation operator, Z_0 is the line impedance, and T is the propagation time.

To complete the model the friction model $F = F[q(x,s)]$ must be specified so that the wave propagation operator $\Gamma(s)$ can be found from (4.149). This will be done in the following with three different friction models.

Example 64 *The equations of the transmission line model (4.150) and (4.151) can be combined so the Laplace transformed model can be written as a wave equation in pressure or flow as given by the two equations*

$$L^2 \frac{\partial^2 p(x,s)}{\partial x^2} - \Gamma^2 p(x,s) = 0 \tag{4.152}$$

$$L^2 \frac{\partial^2 q(x,s)}{\partial x^2} - \Gamma^2 q(x,s) = 0 \tag{4.153}$$

Example 65 *The series impedance $X(s)$ and the parallel admittance $Y(s)$ are given by*

$$X(s) = \frac{Z_0 \Gamma(s)^2}{LTs}, \quad Y(s) = \frac{Ts}{L(s) Z_0} \tag{4.154}$$

The characteristic impedance $Z_c(s)$ is then found to be

$$Z_c(s) = \sqrt{\frac{X(s)}{Y(s)}} = Z_0 \frac{\Gamma(s)}{Ts} \tag{4.155}$$

4.5.4 Lossless model

First it is assumed that there is no friction in the pipe, which means that $F = 0$. The transmission line model becomes

$$\frac{\partial q(x,s)}{\partial x} = -\frac{s}{cZ_0} p(x,s) \tag{4.156}$$

$$\frac{\partial p(x,s)}{\partial x} = -\frac{Z_0 s}{c} q(x,s) \tag{4.157}$$

Comparison with the general case (4.151) and (4.155) shows that:

In the lossless case the propagation operator $\Gamma(s)$ and the characteristic impedance $Z_c(s)$ are given by

$$\Gamma(s) = Ts, \quad Z_c(s) = Z_0 \tag{4.158}$$

4.5.5 Linear friction

Loss terms in the form of friction in the pipe can be modelled using the Hagen-Poiseuille equation (White 1999) by assuming laminar flow. Then the friction force is

$$F = \rho_0 B q \tag{4.159}$$

where the friction coefficient B is

$$B = \frac{8\nu_0}{r_0^2} \tag{4.160}$$

where r_0 is the radius of the pipe, and ν_0 is the kinematic viscosity. The model (4.147, 4.148) becomes

$$\frac{\partial q(x,s)}{\partial x} = -\frac{s}{cZ_0} p(x,s) \tag{4.161}$$

$$\frac{\partial p(x,s)}{\partial x} = -\frac{Z_0}{c} (s + B)\, q(x,s) \tag{4.162}$$

Then the propagation operator can be found according to (4.150) to be given by

$$\Gamma^2 = T^2 s(s + B) \tag{4.163}$$

From this result and (4.155) it is seen that:

With linear friction the propagation operator and the characteristic impedance are

$$\Gamma = Ts\sqrt{\frac{s+B}{s}}, \quad Z_c = Z_0\sqrt{\frac{s+B}{s}} \tag{4.164}$$

In this the case the wave equation from the lossless case is modified to

$$\frac{\partial^2 p}{\partial t^2} + B\frac{\partial p}{\partial t} - c^2\frac{\partial^2 p(x,s)}{\partial x^2} = 0 \tag{4.165}$$

or, using the Laplace transform,

$$L^2\frac{\partial^2 p(x,s)}{\partial x^2} = T^2 s\,(s+B)\,p(x,s) \tag{4.166}$$

4.5.6 Nonlinear friction

In the case of nonlinear friction the PDE model is (Goodson and Leonard 1972)

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0}\frac{\partial p}{\partial x} + \frac{\mu_0}{\rho_0}\left(\frac{\partial^2 q}{\partial r^2} + \frac{1}{r}\frac{\partial q}{\partial r}\right) \tag{4.167}$$

$$\frac{\partial \rho}{\partial t} = -\frac{\rho_0}{A}\frac{\partial q}{\partial x} - \rho_0\left(\frac{\partial v}{\partial r} + \frac{v}{r}\right) \tag{4.168}$$

$$\frac{\partial T}{\partial t} = \alpha_0\left(\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r}\right) \tag{4.169}$$

In this case viscosity and heat transfer effects has been added to the model. Without further explanation we state that the propagation operator is

$$\Gamma^2 = (Ts)^2\,\frac{1}{N\left(r\sqrt{\frac{s}{\nu}}\right)} \tag{4.170}$$

where the function N is given by the two alternative expressions

$$N(z) = 1 - \frac{2J_1(jz)}{jzJ_0(jz)} = \frac{I_2(z)}{I_0(z)} \tag{4.171}$$

Here J_0 and J_1 are Bessel functions of the first kind of order 0 and 1, respectively, and I_0 and I_2 are modified Bessel functions of the first kind of order 0 and 1, respectively. Note that the propagation operator is irrational. Details are found in (Goodson and Leonard 1972).

4.5.7 Wave variables

The transmission line model given by (4.150) and (4.151) can be written

$$\frac{\partial}{\partial x}\begin{pmatrix} q(x,s) \\ p(x,s) \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & -\frac{Ts}{L(s)Z_0} \\ -\frac{Z_0\Gamma(s)^2}{LTs} & 0 \end{pmatrix}}_{\mathbf{A}}\begin{pmatrix} q(x,s) \\ p(x,s) \end{pmatrix} \tag{4.172}$$

By taking the eigenvectors of the matrix $\mathbf{A}$ it is found that the system can be made diagonal a change of variables.

The transmission line can be modeled with the wave variables

$$a(x,s) = p(x,s) + Z_c(s)q(x,s) \tag{4.173}$$

$$b(x,s) = p(x,s) - Z_c(s)q(x,s) \tag{4.174}$$

Here

$$Z_c = Z_0 \frac{\Gamma}{Ts} \tag{4.175}$$

is the characteristic impedance of the transmission line. From 4.172) the model for the wave variables is found to be

$$\frac{\partial a(x,s)}{\partial x} = -\frac{\Gamma}{L} a(x,s) \tag{4.176}$$

$$\frac{\partial b(x,s)}{\partial x} = \frac{\Gamma}{L} b(x,s) \tag{4.177}$$

The solutions for the wave variables are given by

$$a(x,s) = \exp\left(-\Gamma \frac{x}{L}\right) a(0,s) \tag{4.178}$$

$$b(x,s) = \exp\left(-\Gamma \frac{L-x}{L}\right) b(L,s) \tag{4.179}$$

In the lossless case

$$\Gamma = Ts; \quad Z_c = Z_0 = \frac{\rho_0 c}{A} \tag{4.180}$$

The solutions are then

$$a(x,s) = \exp\left(-\frac{x}{L} Ts\right) a(0,s) \tag{4.181}$$

$$b(x,s) = \exp\left(-\frac{L-x}{L} Ts\right) b(L,s) \tag{4.182}$$

It is seen that a describes a wave moving in the positive x direction, and b describes a wave moving in the negative x direction.

The inverse relations are

$$p(x,s) = \frac{1}{2}\left[a(x,s) + b(x,s)\right] \tag{4.183}$$

$$q((x,s) = \frac{1}{2Z_c(s)}\left[a(x,s) - b(x,s)\right] \tag{4.184}$$

Suppose that the transmission line is terminated with an impedance $Z_L(s)$ so that

$$p(L,s) = Z_L(s) q(L,s) \tag{4.185}$$

The boundary conditions for the wave variables are given by

$$a(0,s) = a_1(s) \tag{4.186}$$

$$b(L,s) = G_L(s) a(L,s) \tag{4.187}$$

where

$$G_L(s) = \frac{Z_L(s) - Z_c}{Z_L(s) + Z_c} \tag{4.188}$$

The transfer function from $a(0,s)$ to $b(0,s)$ is then found to be

$$\frac{b(0,s)}{a(0,s)} = \exp\left(-2\Gamma\right) G_L(s) \tag{4.189}$$

In the lossless case this gives the transfer function

$$\frac{b(0,s)}{a(0,s)} = \exp(-2Ts)\, G_L(s) \tag{4.190}$$

which is a time delay of $2T$ multiplied multiplied with $G_L(s)$.

4.5.8 Example: Lossless pipe

In the lossless case the transfer function from the inlet pressure $p(0,s)$ to the outlet pressure $p(L,s)$ is given by (1.210), but an alternative expression can be found from (4.183) to be

$$\frac{p(L,s)}{p(0,s)} = \frac{a(L,s)+b(L,s)}{a(0,s)+b(0,s)} = \frac{[1+G_L(s)]\exp(-Ts)}{[1+\exp(-2Ts)\, G_L(s)]} \tag{4.191}$$

while the transfer function from $q(0,s)$ to $p(0,s)$ is given by

$$\begin{aligned} \frac{p(0,s)}{q(0,s)} &= Z_c \frac{a(0,s)+b(0,s)}{a(0,s)-b(0,s)} = Z_c \frac{1+\frac{b(0,s)}{a(0,s)}}{1-\frac{b(0,s)}{a(0,s)}} = Z_c \frac{\exp(Ts)+\exp(-Ts)\, G_L(s)}{\exp(Ts)-\exp(-Ts)\, G_L(s)} \\ &= Z_c \frac{Z_L \cosh Ts + Z_c \sinh Ts}{Z_c \cosh Ts + Z_L \sinh Ts} \end{aligned} \tag{4.192}$$

Consider a lossless pipe which is open at the outlet $x = L$. Then $p(L,s) = 0$ and therefore $Z_L(s) = 0$ and $G_L(s) = -1$. It follows that

$$\frac{b(0,s)}{a(0,s)} = -\exp(-2Ts) \tag{4.193}$$

while

$$\frac{p(L,s)}{p(0,s)} = 0 \tag{4.194}$$

and

$$\frac{p(0,s)}{q(0,s)} = Z_c \tanh Ts \tag{4.195}$$

Next, consider a pipe which is closed at the outlet at $x = L$. Then $q(L,s) = 0$, $Z_L(s) = \infty$ and $G_L(s) = 1$ The transfer functions become

$$\frac{b(0,s)}{a(0,s)} = \exp(-2Ts) \tag{4.196}$$

and

$$\frac{p(L,s)}{p(0,s)} = \frac{1}{\cosh(Ts)} \tag{4.197}$$

and

$$\frac{p(0,s)}{q(0,s)} = Z_c \frac{1}{\tanh Ts} \tag{4.198}$$

Finally, consider impedance matching which is achieved with a restriction giving $p(L,s) = Z_c q(L,s)$, that is with $Z_L = Z_c$. Then $G_L(s) = 0$, and

$$\frac{b(0,s)}{a(0,s)} = 0 \tag{4.199}$$

$$\frac{p(L,s)}{p(0,s)} = \exp\left(-Ts\right) \tag{4.200}$$

and

$$\frac{p(0,s)}{q(0,s)} = Z_c \tag{4.201}$$

4.5.9 Linear network models of transmission lines

From an input-output perspective a transmission lines can be modeled as a passive system with one inlet port at $x = 0$ with pressure p_1 and flow q_1 into the line, and one outlet port at $x = L$ with pressure p_2 and flow q_2 out of the volume. The dynamics can then be described by transfer functions. If the ports are connected to valves at both sides, then the valves will normally be describe with pressures as inputs and flows as outputs. This means that the inputs to the transmission line model will be flows, and the transfer function should be given in impedance form

$$\begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} = \mathbf{Z}(s) \begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} \tag{4.202}$$

If the transmission line is connected to volumes at both sides, where the volumes may be chambers in a pump, a hydraulic motor or a cylinder, then flows will be inputs and pressures will be outputs of the mass balance models of the volumes. This means that pressures will be input variables at the transmission line ports, and an admittance form

$$\begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} = \mathbf{Y}(s) \begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} \tag{4.203}$$

is is the appropriate model formulation. If the transmission line is connected to a valve at port 1 and a volume at port 2, then the flow q_1 will be the input at port 1 and the pressure p_2 will be the input variable at port 2. The the model should be formulated as a hybrid model

$$\begin{pmatrix} p_1(s) \\ q_2(s) \end{pmatrix} = \mathbf{H}(s) \begin{pmatrix} q_1(s) \\ p_2(s) \end{pmatrix} \tag{4.204}$$

The impedance model, the admittance model and the hybrid models are well suited for analysis and simulation models.

To describe a cascade of components it may seem to be a good idea to use the cascade form

$$\begin{pmatrix} p_2(s) \\ q_2(s) \end{pmatrix} = \mathbf{B}(s) \begin{pmatrix} p_1(s) \\ q_1(s) \end{pmatrix} \tag{4.205}$$

However, the cascade form leads to an ill-conditioned formulation. This is due to the fact that the solution of the wave equation can be described as the sum of two waves that travel in opposite directions. The cascade form is only suited to describe solutions that propagate from port 1 to port 2.

4.5.10 Rational approximations of transfer function models

The dynamic model of a transmission line is given by partial differential equations. Because of this, the transfer function matrices $\mathbf{Z}(s)$, $\mathbf{Y}(s)$ and $\mathbf{H}(s)$ will have irrational entries. It is necessary to find rational approximations to derive simulation models based on the transfer function description. Methods for finding rational approximations of the transfer functions are developed in the following sections. The material is taken from (Piché and Ellman 1996) and (Mäkinen et al. 2000). Alternative solutions are presented in (Yang and Tobler 1991).

4.5.11 Rational series expansion of impedance model

The dynamic model for a hydraulic transmission line is given by partial differential equations, and the transfer functions are therefore irrational. For use in simulation models there is a need for an approximation in the form of ordinary differential equations corresponding to rational transfer functions. There are several ways of doing this. In this section we will present a rational series expansion of the irrational transfer functions. The results are taken from (Piché and Ellman 1996), and the point of departure is the impedance form of the transfer functions where volumetric flows are inputs, and pressures are outputs. The impedance model is given by (1.201) as

$$\begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} = Z_c(s) \begin{pmatrix} \frac{\cosh\Gamma}{\sinh\Gamma} & \frac{1}{\sinh\Gamma} \\ \frac{1}{\sinh\Gamma} & \frac{\cosh\Gamma}{\sinh\Gamma} \end{pmatrix} \begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} \tag{4.206}$$

The development is simplified by a change of variables into symmetric variables p_s and q_s, and antisymmetric variables p_a and q_a according to

$$q_s = \frac{1}{2}(q_1 - q_2), \quad p_s = \frac{1}{2}(p_1 + p_2) \tag{4.207}$$

$$q_a = \frac{1}{2}(q_1 + q_2), \quad p_a = \frac{1}{2}(p_1 - p_2) \tag{4.208}$$

The transfer function model of a hydraulic transmission line can be written in the impedance form

$$p_s(s) = Z_s(s)q_s(s) \tag{4.209}$$

$$p_a(s) = Z_a(s)q_a(s) \tag{4.210}$$

where the transfer functions or impedance functions are found from (4.206–4.208) to be given by

$$Z_s(s) = \frac{Z_0\Gamma(s)}{Ts}\left(\frac{\cosh\Gamma(s)+1}{\sinh\Gamma(s)}\right) \tag{4.211}$$

$$Z_a(s) = \frac{Z_0\Gamma(s)}{Ts}\left(\frac{\cosh\Gamma(s)-1}{\sinh\Gamma(s)}\right) \tag{4.212}$$

The transfer functions $Z_s(s)$ and $Z_a(s)$ both have singularities for

$$\sinh\Gamma = 0 \Leftrightarrow e^{2\Gamma} = 1 \Leftrightarrow \Gamma = j\omega_{sk} \tag{4.213}$$

where the natural frequencies are

$$\omega_{sk} = k\pi, \quad k = 0, \pm1, \pm2, \ldots \tag{4.214}$$

Note that there are infinitely many singularities with an even spacing of π along the imaginary axis.

The following partial fraction expansions of the impedance functions can be used to arrive at the rational series with infinitely many terms:

$$Z_s(s) = \frac{2Z_0}{Ts} + \sum_{k=2,4,\ldots}^{\infty} \frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+\omega_{sk}^2\right)}, \quad Z_a(s) = \sum_{k=1,3,\ldots}^{\infty} \frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+\omega_{sk}^2\right)} \tag{4.215}$$

where $\omega_{sk} = k\pi$

It is possible to develop a rational transfer function model by using a truncated model where terms up to $k = N$ are included.

Example 66 *The partial fraction expansion is done by expanding* $(\cosh\Gamma + 1)/\sinh\Gamma$ *and* $(\cosh\Gamma - 1)/\sinh\Gamma$ *using the formula*

$$\frac{f(s)}{g(s)} = \frac{A_1}{s-a_1} + \frac{A_2}{s-a_2} + \ldots \Rightarrow A_i = \frac{f(a_i)}{g'(a_i)} \tag{4.216}$$

We can then find the coefficients of the partial fraction expansions from

$$\frac{\cosh\Gamma + 1}{\cosh\Gamma} = \frac{e^{2\Gamma} + 2e^{\Gamma} + 1}{e^{2\Gamma} + 1} = \begin{cases} 0 & \Gamma = j\omega_{sk}, k = \text{odd} \\ 2 & \Gamma = j\omega_{sk}, k = \text{even} \end{cases} \tag{4.217}$$

The coefficient for the second order terms are found for even k from

$$\frac{2}{\Gamma + j\omega_{sk}} + \frac{2}{\Gamma - j\omega_{sk}} = \frac{4\Gamma}{\Gamma^2 + \omega_{sk}^2} \tag{4.218}$$

In the same way we calculate

$$\frac{\cosh\Gamma - 1}{\cosh\Gamma} = \frac{e^{2\Gamma} - 2e^{\Gamma} + 1}{e^{2\Gamma} + 1} = \begin{cases} 2 & \Gamma = j\omega_{sk}, k = \text{odd} \\ 0 & \Gamma = j\omega_{sk}, k = \text{even} \end{cases} \tag{4.219}$$

and find that for odd k we have

$$\frac{2}{\Gamma + j\omega_{sk}} + \frac{2}{\Gamma - j\omega_{sk}} = \frac{4\Gamma}{\Gamma^2 + \omega_{sk}^2} \tag{4.220}$$

4.5.12 Rational series expansion of admittance model

The admittance for of the transmission line model is given by (1.202) as

$$\begin{pmatrix} q_1(s) \\ -q_2(s) \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} \frac{\cosh\Gamma}{\sinh\Gamma} & -\frac{1}{\sinh\Gamma} \\ -\frac{1}{\sinh\Gamma} & \frac{\cosh\Gamma}{\sinh\Gamma} \end{pmatrix} \begin{pmatrix} p_1(s) \\ p_2(s) \end{pmatrix} \tag{4.221}$$

Again the model is simplified by using symmetric and asymmetric variables defined in (4.207) and (4.208). Then the transfer functions become

$$q_s(s) = Y_s(s)q_s(s) \tag{4.222}$$
$$q_a(s) = Y_a(s)q_a(s) \tag{4.223}$$

where the admittances are

$$Y_s(s) = \frac{Ts}{Z_0\Gamma}\frac{\cosh\Gamma + 1}{\sinh\Gamma} \tag{4.224}$$
$$Y_a(s) = \frac{Ts}{Z_0\Gamma}\left(\frac{\cosh\Gamma - 1}{\sinh\Gamma}\right) \tag{4.225}$$

Using partial fraction expansion in the same way as for the impedance model we find the following rational representation of the infinite-dimensional admittances:

$$Y_s(s) = \sum_{k=1,3,\ldots}^{\infty} \frac{4Ts}{Z_0\left(\Gamma^2 + \omega_{sk}^2\right)}, \quad Y_a(s) = \frac{2Ts}{Z_0\Gamma^2} + \sum_{k=2,4,\ldots}^{\infty} \frac{4Ts}{Z_0\left(\Gamma^2 + \omega_{sk}^2\right)} \tag{4.226}$$

4.5.13 Galerkin derivation of impedance model

An alternative and more general approach to find a rational model of the transmission line dynamics is based the use of Galerkin's method (Mäkinen et al. 2000). Shape functions $\phi_k(x)$ are then used to express the pressure as

$$\bar{p}(s,x) = \sum_{k=0}^{N} P_k(s)\phi_k(x) \tag{4.227}$$

This is used in combination with the transmission line model (4.157) which is

$$\Gamma^2 p(s) - L^2 \frac{d^2 p(s)}{dx^2} = 0 \tag{4.228}$$

The boundary conditions are supposed to be

$$\frac{\partial p(0,s)}{\partial x} = -\frac{Z_0\Gamma(s)^2}{LTs} q(0,s), \quad \frac{\partial p(L,s)}{\partial x} = -\frac{Z_0\Gamma(s)^2}{LTs} q(L,s) \tag{4.229}$$

where $q(0,s)$ and $q(L,s)$ are inputs to the model.

The pressure shape functions in the lossless case with zero flow at the end-points are

$$\phi_k(x) = \cos\left(\frac{k\pi}{L}x\right) \tag{4.230}$$

which is a well-established result for the wave equation. These shape functions are orthogonal in the sense that

$$\int_0^L \phi_k(x)\phi_j(x)dx = \int_0^L \cos\left(\frac{k\pi}{L}x\right)\cos\left(\frac{j\pi}{L}x\right)dx = \frac{L}{2}\delta_{kj} \tag{4.231}$$

Moreover, the derivatives of the shape functions are orthogonal and satisfies

$$\int_0^L \phi_k'(x)\phi_j'(x)dx = \int_0^L \sin\left(\frac{k\pi}{L}x\right)\sin\left(\frac{j\pi}{L}x\right)dx = \frac{(k\pi)^2}{2}\delta_{kj} \tag{4.232}$$

These shape functions will be used as assumed modes in a Ritz approximation as in (Mäkinen et al. 2000) to derive a rational model with Galerkin's method. This is done by multiplying the shape function $\phi_k(x)$ with the model (4.228) using $p = \bar{p}$, and then integrating over the length of the transmission line. This gives

$$I := \int_0^L \phi_k(x)\left(\Gamma\bar{p}(s,x) - L^2\frac{d^2\bar{p}(s,x)}{dx^2}\right)dx = 0 \tag{4.233}$$

As usual in the Galerkin approach the expression for the integral is developed using partial integration:

$$\begin{aligned} I &= \int_0^L \phi_k\left(\Gamma^2\phi_k(x)\bar{p}(s,x) + L^2\frac{d\phi_k(x)}{dx}\frac{d\bar{p}(s,x)}{dx}\right)dx \\ &\quad + \phi_k(x)L^2\frac{\partial\bar{p}(s,x)}{\partial x}\Big|_0^L \\ &= \int_0^L \left(\Gamma^2\phi_k(x)\bar{p}(s,x) + L^2\frac{d\phi_k(x)}{dx}\frac{d\bar{p}(s,x)}{dx}\right)dx \\ &\quad + \frac{z_0\Gamma^2 L}{Ts}\left[\phi_k(L)q(L) - \phi_k(0)q(0)\right] \end{aligned} \tag{4.234}$$

Using the orthogonality of $\phi_k(x)$ and $\phi'_k(x)$ as stated in (4.231) and (4.232) we find that

$$P_k(s)\frac{L}{2}\left(\Gamma^2+(k\pi)^2\right)-\frac{z_0\Gamma^2 L}{Ts}\left[(-1)^k q(L)+q(0)\right]=0 \tag{4.235}$$

which means that the pressure coefficients are given by

$$P_k(s)=\frac{2z_0\Gamma^2}{Ts\left(\Gamma^2+(k\pi)^2\right)}\left[q(0)+(-1)^k q(L)\right] \tag{4.236}$$

Then the impedance functions $Z_s(s)$ and $Z_a(s)$ in the model

$$p_s(s) = Z_s(s)q_s(s) \tag{4.237}$$
$$p_a(s) = Z_a(s)q_a(s) \tag{4.238}$$

for the symmetric variables are found to be

$$Z_s(s)=\frac{2Z_0}{Ts}+\sum_{k=2,4,\ldots}^{\infty}\frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+(k\pi)^2\right)},\quad Z_a(s)=\sum_{k=1,3,\ldots}^{\infty}\frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+(k\pi)^2\right)} \tag{4.239}$$

This result is the same as the result (4.215) that was obtained by series expansion of the transfer functions.

4.5.14 Galerkin derivation of the admittance model

The Galerkin solution when the pressures are inputs is found in a similar way as in the case where the flows are inputs. In this case the flow is represented by shape functions $\phi_k(x)$ so that

$$\bar{q}(s,x)=\sum_{k=0}^{\infty}Q_k(s)\phi_k(x) \tag{4.240}$$

and the transmission line model is given by (4.157) as

$$\Gamma^2 q(s)-L^2\frac{d^2q(s)}{dx^2}=0 \tag{4.241}$$

with boundary conditions

$$\frac{\partial q(0,s)}{\partial x}=-\frac{s}{cZ_0}p(0,s),\quad \frac{\partial q(L,s)}{\partial x}=-\frac{s}{cZ_0}p(L,s) \tag{4.242}$$

where $p(0,s)$ and $p(L,s)$ are inputs to the model. The shape functions are taken to be the orthogonal eigenfunctions of the lossless wave equation when the pressures are zero at both ends. This gives

$$\phi_k(x)=\cos\left(\frac{k\pi}{L}x\right) \tag{4.243}$$

The Galerkin approach gives

$$I:=\int_0^L\phi_k(x)\left(\Gamma^2\bar{q}(x)-L^2\frac{d^2\bar{q}(x)}{dx^2}\right)dx=0 \tag{4.244}$$

and

$$\begin{aligned} I &= \int_0^L \phi_k \left(\Gamma^2 \phi_k(x) \bar{q}(s) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{q}(x)}{dx} \right) dx \\ &\quad + \phi_k(x) L^2 \frac{\partial q(x)}{\partial x} \Big|_0^L \\ &= \int_0^L \left(\Gamma^2 \phi_k(x) \bar{q}(x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{q}(x)}{dx} \right) dx \\ &\quad + \frac{LTs}{z_0} \left[\phi_k(L) p(L) - \phi_k(0) p(0) \right] \end{aligned} \tag{4.245}$$

and due to the orthogonality of $\phi_k(x)$ and $\phi_k'(x)$, it follows that

$$Q_k(s) \frac{L}{2} \left(\Gamma^2 + (k\pi)^2 \right) = \frac{LTs}{z_0} \left[p(0) + (-1)^k p(L) \right] \tag{4.246}$$

so that

$$Q_k(s) = \frac{2Ts}{z_0 \left(\Gamma^2 + (k\pi)^2 \right)} \left[p(0) + (-1)^k p(L) \right] \tag{4.247}$$

The admittance functions of the symmetric and asymmetric variables

$$q_s(s) = Y_s(s) p_s(s) \tag{4.248}$$

$$q_a(s) = Y_a(s) p_a(s) \tag{4.249}$$

are found to be

$$Y_s(s) = \sum_{k=1,3,\ldots}^{\infty} \frac{4Ts}{Z_0 \left(\Gamma^2 + (k\pi)^2 \right)}, \quad Y_a(s) = \frac{2Ts}{Z_0 \Gamma^2} + \sum_{k=2,4,\ldots}^{\infty} \frac{4Ts}{Z_0 \left(\Gamma^2 + (k\pi)^2 \right)} \tag{4.250}$$

which is the same result as the result (4.226) that was found from the transfer functions.

4.5.15 Galerkin derivation of the hybrid model

Finally the hybrid case is investigated where the pressure is given at one end, and where flow is given at the other end. Then the model is

$$\Gamma^2 p(s) - L^2 \frac{d^2 p(s)}{dx^2} = 0 \tag{4.251}$$

with boundary condition

$$p(0, s) = p_1, \quad \frac{\partial p(L, s)}{\partial x} = -\frac{Z_0 \Gamma(s)^2}{LTs} q(L, s) \tag{4.252}$$

where p_1 and $q(L, s)$ inputs to the model. The pressure is represented by

$$\bar{p}(s, x) = p_1 + \sum_{k=1}^{\infty} P_k(s) \phi_k(x) \tag{4.253}$$

where the shape functions

$$\phi_k(x) = \sin \left[\left(k - \frac{1}{2} \right) \frac{\pi x}{L} \right] \tag{4.254}$$

are the orthogonal eigenfunctions of the lossless wave equation when the pressure at the input is zero and the flow at the output is zero. The Galerkin method gives

$$I := \int_0^L \phi_k(x) \left(\Gamma \bar{p}(s,x) - L^2 \frac{d^2 \bar{p}(s,x)}{dx^2} \right) dx = 0 \tag{4.255}$$

and

$$\begin{aligned} I &= \int_0^L \phi_k \left(\Gamma^2 \phi_k(x) \bar{p}(s,x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s,x)}{dx} \right) dx \\ &\quad + \phi_k(x) L^2 \frac{\partial \bar{p}(s,x)}{\partial x} \Big|_0^L \\ &= \int_0^L \left(\Gamma^2 \phi_k(x) \bar{p}(s,x) + L^2 \frac{d\phi_k(x)}{dx} \frac{d\bar{p}(s,x)}{dx} \right) dx \\ &\quad + \frac{Z_0 \Gamma^2 L}{Ts} \left[(-1)^k q(L) \right] \end{aligned} \tag{4.256}$$

and orthogonality of the shape functions gives

$$P_k(s) \frac{L}{2} \left(\Gamma^2 + \left[\left(k - \frac{1}{2} \right) \pi \right]^2 \right) + \frac{\Gamma^2 L}{\left(k - \frac{1}{2} \right) \pi} p_1 - \frac{Z_0 \Gamma^2 L}{Ts} \left[(-1)^k q(L) \right] = 0 \tag{4.257}$$

and the pressure coefficients are found to be

$$P_k(s) = -\frac{2\Gamma^2}{\Gamma^2 + \left[\left(k - \frac{1}{2} \right) \pi \right]^2} \left(\frac{1}{\left(k - \frac{1}{2} \right) \pi} p_1 + (-1)^k \frac{Z_0}{Ts} q_2 \right) \tag{4.258}$$

The output variables $q_1(s)$ and $p_2(s)$ are then found to be

$$p_2(s) = p_1 + \sum_{k=1}^{\infty} (-1)^{k+1} P_k(s) \tag{4.259}$$

$$q_1(s) = -\frac{Ts}{Z_0 \Gamma^2} \sum_{k=1}^{\infty} \left(k - \frac{1}{2} \right) \pi P_k(s) \tag{4.260}$$

4.5.16 Rational simulation models

To find a simulation model it is necessary to develop a model with finite dimension. This can be done by truncating the infinite series (4.215). Then the model is

$$p_s(s) = Z_s(s) q_s(s) \tag{4.261}$$

$$p_a(s) = Z_a(s) q_a(s) \tag{4.262}$$

where the impedances are given by the truncated versions

$$Z_s = \frac{2Z_0}{Ts} + \sum_{k=2,4,\ldots}^{N} \frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2 + \omega_{sk}^2\right)}, \quad Z_a = \sum_{k=1,3,\ldots}^{N-1} \frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2 + \omega_{sk}^2\right)} \tag{4.263}$$

of the rational transfer functions where N terms are included. It is assumed that N is an even number. To find a state-space formulation for this model it is necessary to investigate the terms of $Z_s(s)$ and $Z_a(s)$ closer. In the lossless case $\Gamma = Ts$ and

$$\frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2 + \omega_{sk}^2\right)} = \frac{4Z_0 Ts}{T^2 s^2 + \omega_{sk}^2} \tag{4.264}$$

which is straightforward to represent as a second-order system. Moreover, with linear friction $\Gamma^2 = T^2 s(s+B)$ and we find that

$$\frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+\omega_{sk}^2\right)} = \frac{4Z_0T(s+B)}{T^2s^2+BT^2s+\omega_{sk}^2} \tag{4.265}$$

which is also a second-order system.

With nonlinear friction it is necessary to introduce a rational approximation of Γ^2. An approximation due to (Woods 1983) is

$$\Gamma^2 = \frac{(Ts)^2}{1-\left(1+2\frac{Ts}{B}\right)^{-1/2}} \tag{4.266}$$

In (Piché and Ellman 1996) this approximation is used to get a rational approximation which is accurate at the natural frequencies. This is done with

$$\frac{4Z_0\Gamma^2}{Ts\left(\Gamma^2+\omega_{sk}^2\right)} = \frac{4T(s+B)}{(Ts)^2+B_kT^2s+\omega_k^2} \tag{4.267}$$

where

$$B_k = \frac{1}{2}\sqrt{\omega_{sk}B}+\frac{B}{8}, \quad \omega_k = \omega_{sk}-\frac{B_k}{2} \tag{4.268}$$

Then the transfer function from $q_s(s)$ to $p_s(s)$ can be expressed as a parallel interconnection of an integrator and $N/2$ second order systems that can be given a state-space realization. In the same way the transfer function from $q_a(s)$ to $p_a(s)$ will be a parallel interconnection of $N/2$ second order systems.

Solutions computed from such a truncated model will give spurious and non-physical oscillations in the face of discontinuities like a step change in an input. This is known as the Gibb's phenomenon, and resembles problems that appear with data windows in digital signal processing. A solution to the problem (Piché and Ellman 1996) is to use a data window to modify the truncated transfer function to

$$Z_s = \frac{2Z_0}{Ts}+\sum_{k=2,4,\ldots}^{N}\frac{4Z_0\Gamma^2\sigma_k}{Ts\left(\Gamma^2+\omega_{sk}^2\right)}, \quad Z_a = \sum_{k=1,3,\ldots}^{N-1}\frac{4Z_0\Gamma^2\sigma_k}{Ts\left(\Gamma^2+\omega_{sk}^2\right)} \tag{4.269}$$

where

$$\sigma_k = \frac{\sin\beta_k}{\beta_k}, \quad \beta_k = \frac{\omega_{sk}}{N+1} \tag{4.270}$$

are the coefficients of a Riemann window. It is also possible to use a Hann window with $\sigma_k = (1-\cos\beta_k)/2$ or a Hamming window with $\sigma_k = 0.54+0.46\cos\beta_k$. Moreover, a steady-state correction is necessary to achieve correct steady-state pressure reduction, which is

$$p_2 = p_1 - \varepsilon Z_0 q \tag{4.271}$$

where it is assumed that $q_1 = q_2 = q$. To obtain this steady-state result with a truncated approximation it is sufficient to insert $b_N B$ for B in Z_a where

$$b_N = \left(8\sum_{k=1,3,\ldots}^{N-1}\frac{\sigma_k}{\omega_k^2}\right)^{-1} \tag{4.272}$$

Γ^2	$Z_s(s)/Z_0$	$Z_a(s)/Z_0$
$(Ts)^2$	$\frac{2}{Ts} + \sum_{k=2,4,\ldots}^{N} \frac{4\sigma_k Ts}{(Ts)^2+(k\pi)^2}$	$\sum_{k=1,3,\ldots}^{N-1} \frac{4\sigma_k Ts}{(Ts)^2+(k\pi)^2}$
$T^2 s\,(s+B)$	$\frac{2}{Ts} + \sum_{k=2,4,\ldots}^{N} \frac{4\sigma_k T(s+B)}{(Ts)^2+BT^2 s+(k\pi)^2}$	$\sum_{k=1,3,\ldots}^{N-1} \frac{4\sigma_k T(s+b_N B)}{(Ts)^2+BT^2 s+(k\pi)^2}$
$(Ts)^2 \frac{1}{N(r\sqrt{\frac{s}{\nu}})}$	$\frac{2}{Ts} + \sum_{k=2,4,\ldots}^{N} \frac{4\sigma_k T(s+B)}{(Ts)^2+B_k T^2 s+\omega_k^2}$	$\sum_{k=1,3,\ldots}^{N-1} \frac{4\sigma_k T(s+b_N B)}{(Ts)^2+B_k T^2 s+\omega_k^2}$

Table 4.1: Rational approximations of infinite dimensional transfer function with three different friction models.

The rational truncated models are summarized in the following Table 4.1. SIMULINK models are available on the web (Mäkinen et al. 2000).

The constants of the models are given by

$$T = \frac{L}{c}, \quad B = \frac{8\nu_0}{r_0^2}, \quad B_k = \frac{1}{2}\sqrt{k\pi B} + \frac{B}{8}, \quad \omega_k^2 = k\pi - \frac{B_k}{2} \tag{4.273}$$

$$\sigma_k = \frac{\sin\beta_k}{\beta_k}, \quad \beta_k = \frac{\omega_{sk}}{N+1}, \quad b_N = \left(8 \sum_{k=1,3,\ldots}^{N-1} \frac{\sigma_k}{\omega_k^2}\right)^{-1} \tag{4.274}$$

Numerical values for a transmission line is $L = 20$ m, $\rho = 870$ kg/m^3, $c = 1400$ m/s, $\nu_0 = 8 \times 10^{-5}$ m^2/s and $r_0 = 6 \times 10^{-3}$ m. This corresponds to a propagation time of $T = L/c = 14$ ms.

4.6 Lumped parameter model of hydraulic line

4.6.1 Introduction

The hydraulic transmission line has been described by distributed parameter models, which are models that are formulated by partial differential equations. Transfer functions that describe distributed parameter models of transmission lines are irrational with terms like $\cosh Ts$, $\sinh Ts$, $\tanh Ts$ and $\exp(-Ts)$. We recall that transfer functions with irrational terms are called infinite dimensional as they can be expressed by a series expansion in the complex variable s with an infinite number of terms. For analysis and control design it may be desirable to obtain finite-dimensional models of transmission lines. This can be done by some numerical discretization scheme or by truncating a series expansion of an irrational model. In this section we will follow a different path. We will reformulate the model by describing the physics of the system with a lumped parameter model by describing the transmission line as a series of control volumes of finite size instead of using infinitesimal control volumes. This type of model is based on the same assumptions that are used in the Helmholtz resonator model, and we will therefore briefly present the Helmholtz resonator, which is the physical system in fluid flow that is analog to a mass-spring-damper system in flexible mechanical systems.

4.6.2 Helmholtz resonator model

A Helmholtz resonator consists of a volume V that is connected to a pipe of length h and cross section A (Figure 12.4). To develop the mathematical model of the system the following assumptions are made:

1. The velocity of the fluid in the volume is sufficiently small to assume that the pressure p is the same over the volume.

2. The compressibility effects in the pipe are negligible, so that the volumetric flow q is the same along the pipe.

This means that the Helmholtz resonator is modeled by a pipe with incompressible fluid flow that is connected to a volume with compressibility effects. The mass balance of the volume is

$$\frac{V}{\beta}\dot{p} = q \tag{4.275}$$

while the momentum balance of the pipe is

$$h\rho_0\dot{q} = -Ap \tag{4.276}$$

where the inlet pressure of the pipe has been set to zero. By differentiating the mass balance (4.275) with respect to time and inserting the momentum equation (4.276) the harmonic oscillator

$$\ddot{p} + \omega_H^2 p = 0 \tag{4.277}$$

is obtained, where

$$\omega_H^2 = \frac{A\beta}{Vh\rho_0} = \frac{Ac_0^2}{Vh} \tag{4.278}$$

is the Helmholtz frequency. Here $c_0^2 = \beta/\rho_0$ is the sonic speed corresponding to the constant density ρ_0.

4.6.3 Model formulation

In this section a chain of Helmholtz resonators will be used to model a hydraulic transmission line. Consider a hydraulic transmission line of length L and cross section A. The model is developed by connecting Helmholtz resonators where the model of Helmholtz resonator i is established by a mass balance

$$\frac{Ah}{\beta}\dot{p}_i = q_{i-1} - q_i \tag{4.279}$$

for a volume $V_h = hA$ with pressure p_i. Here q_{i-1} is the volumetric flow into the volume, q_i is the volumetric flow out of the volume, and $\beta = c_0^2\rho_0$ is the bulk modulus. In addition, the model includes a momentum balance

$$h\rho_0\dot{q}_{i-1} = A(p_{i-1} - p_i) - Fh \tag{4.280}$$

for an incompressible fluid with density ρ_0 in a pipe of length h with volumetric flow q_{i-1}. Here F is the friction force per unit length. This gives the following model for Helmholtz resonator i:

A hydraulic transmission line can be modeled by a chain of N Helmholtz resonators with model

$$\dot{p}_i = \frac{c^2\rho_0}{Ah}(q_{i-1} - q_i) \tag{4.281}$$

$$\dot{q}_{i-1} = \frac{A}{h\rho_0}(p_{i-1} - p_i) - \frac{F}{\rho_0} \tag{4.282}$$

Note that when h tends to zero, then the model will converge to the transmission line model

$$\frac{\partial p}{\partial t} = -\frac{c^2\rho_0}{A}\frac{\partial q}{\partial x} \tag{4.283}$$

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0}\frac{\partial p}{\partial x} - \frac{F}{\rho_0} \tag{4.284}$$

which equivalent to the transmission line model (4.144, 4.145). This means that the lumped parameter model converges to the partial differential equation model when h tends to zero.

The model (4.281, 4.282) describes a two-port with input variables q_i and p_{i-1} and output variables q_{i-1} and p_i. This means that this is a model in hybrid form. We denote the port variables of the transmission line at $x = 0$ as p_{in} and q_{in}, while the port variables at the line end are $x = L$ are p_{out} and q_{out}.Depending on which of the port variables that are selected for inputs and outputs the model will be in admittance form, impedance form, or hybrid form. Equations for these three cases will be presented in the next sections.

4.6.4 Admittance model

If the input variables to the model are the pressures p_{in} and p_{out}, then the transmission line can be modeled with an admittance model. The Helmholtz resonator model (4.144, 4.145) is in hybrid form, and because of this an extra pipe of length h and volumetric flow q_N must be connected to the outlet of a chain of N Helmholtz resonators $i = 1, \ldots, N$. This is shown in Figure 4.16 for $N = 3$. The model has N volumes so that volume i

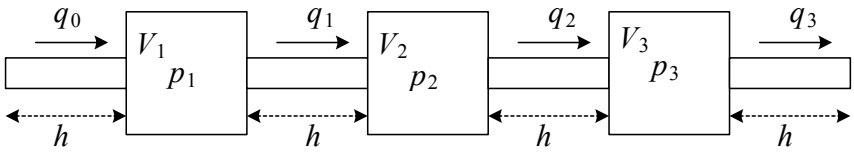

Figure 4.16: A chain of interconnected Helmholtz resonators with ducts of length h representing a transmission line in the admittance form.

is centered at $x_i = ih$ for $i = 1, \ldots, N$, and there are $(N + 1)$ pipes where pipe i is centered at $x_{i+1/2} = (i + 1/2)h$ for $i = 0, \ldots, N$ as shown in Figure 4.17. We note that $V = (N+1)V_h$ and $L = (N+1)h$, so that the number of volumes N will tend to infinity

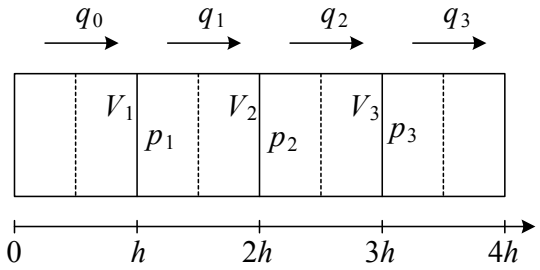

Figure 4.17: Spatial discretization of a transmission line with pressure inputs at both sides to get a description in the form of a chain of Helmholtz resonators.

when h tends to zero. The model is

$$\dot{p}_i = \frac{c^2\rho_0}{Ah}\left(q_{i-1} - q_i\right), \quad i = 1, \ldots, N \tag{4.285}$$

$$\dot{q}_{i-1} = \frac{A}{h\rho_0}(p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 1, \ldots, N+1 \tag{4.286}$$

$$p_0 = p_{in}, \quad p_{N+1} = p_{out} \tag{4.287}$$

4.6.5 Impedance model

Suppose that the input variables at the line ends are the volumetric flows q_{in} and q_{out}. In this case the model has N volumes, but the pipe of the first resonator must be removed to have the right input variable. This means that volume 1 is at the start of the line and volume N is at the end of the line. Volume i is centered at $x_{i-1/2} = (i-1/2)h$ for $i = 1, \ldots, N$, and is connected with pipes of length h and cross section A. There are $N-1$ pipes, where pipe i is centered at $x_i = ih$ for $i = 1, \ldots, N-1$. We note that $V = NV_h$ and $L = Nh$. The model is

$$\dot{p}_i = \frac{c^2\rho_0}{Ah}\left(q_{i-1} - q_i\right), \quad i = 1, \ldots, N \tag{4.288}$$

$$\dot{q}_{i-1} = \frac{A}{h\rho_0}(p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 2, \ldots, N \tag{4.289}$$

$$q_0 = q_{in}, \quad q_N = q_{out} \tag{4.290}$$

4.6.6 Hybrid model

Suppose that the input variables to the transmission line model at the inlet side is the pressures p_{in}, and that the input variable at the outlet side is q_{out}. Then the model is in hybrid form. The Helmholtz resonator model (4.144, 4.145) is also in hybrid form, and because of this the transmission line can be represented by a chain of N Helmholtz resonators $i = 1, \ldots, N$ with model

$$\dot{p}_i = \frac{c^2\rho_0}{Ah}\left(q_{i-1} - q_i\right), \quad i = 1, \ldots, N \tag{4.291}$$

$$\dot{q}_{i-1} = \frac{A}{h\rho_0}(p_{i-1} - p_i) - \frac{F}{\rho_0}, \quad i = 1, \ldots, N \tag{4.292}$$

$$p_0 = p_{in}, \quad q_N = q_{out} \tag{4.293}$$

In this case volume i will be centered at $x_i = ih$ for $i = 1, \ldots, N$, and there are N pipes where pipe i is centered at $x_{i+1/2} = (i + 1/2)h$ for $i = 0, \ldots, N - 1$. We note that $V = (N + 1/2)V_h$ and $L = (N + 1/2)h$.

4.6.7 Natural frequencies

We may think of each Helmholtz resonator as a two-port with port variables q_{i-1} and p_{i-1} for port 1 and port variables q_i and p_i on port 2. Note that $q_i p_i$ has the physical dimension power. The system can be seen as a system with input variables q_i and $p_{i-1}$and output variables equal to the states q_{i-1} and p_i. The hybrid transfer function model is

$$\begin{pmatrix} q_{i-1}(s) \\ p_i(s) \end{pmatrix} = \begin{pmatrix} \frac{A}{h\rho_0\omega_H^2}\frac{s}{1+\frac{s^2}{\omega_H^2}} & -\frac{1}{1+\frac{s^2}{\omega_H^2}} \\ \frac{1}{1+\frac{s^2}{\omega_H^2}} & \frac{h\rho_0}{A}\frac{s}{1+\frac{s^2}{\omega_H^2}} \end{pmatrix} \begin{pmatrix} p_{i-1}(s) \\ -q_i(s) \end{pmatrix} \tag{4.294}$$

where the Helmholtz frequency ω_H is given by

$$\omega_H^2 = \frac{Ac^2}{V_h h} \quad \Rightarrow \quad \omega_H = \frac{c}{h} \tag{4.295}$$

Through the discretization we have introduced Helmholtz resonator i with oscillatory poles at $s = \pm j\omega_H$.

Example 67 *Consider a transmission line with both ends closed. This means that the inlet and outlet flow are given as inputs, so that an impedance model should be used. With two volumes, that is, with $L = 2h$, the dynamics of the model are given by*

$$\dot{p}_1 = -\frac{c^2\rho_0}{Ah}q \tag{4.296}$$

$$\dot{p}_2 = \frac{c^2\rho_0}{Ah}q \tag{4.297}$$

$$\dot{q} = \frac{A}{\rho_0 h}(p_1 - p_2) \tag{4.298}$$

Laplace transformation of the model, and insertion of the pressure equations in the mass flow equation leads to

$$\left(s^2 + 2\frac{c^2}{h^2}\right) q(s) = 0 \tag{4.299}$$

This system has undamped natural frequency

$$\omega_0 = \sqrt{2}\frac{c}{h} = 2\sqrt{2}\frac{c}{L} = 2.82\frac{c}{L} \tag{4.300}$$

while the exact value for the first resonance of the partial differential equation model is found from the period $T_1 = 2L/c$, which gives

$$\omega_1 = \frac{2\pi}{T_1} = \pi\frac{c}{L} = 3.14\frac{c}{L} \tag{4.301}$$

This means that the resonance frequency of this simplified model is about 11% lower than the exact resonance frequency.

Example 68 *For $N = 3$ volumes we have $L = 3h$, and the state space model of the impedance model is*

$$\begin{aligned} A\dot{p}_i &= \omega_H^2 \left(h\rho_0 q_{i-1} - h\rho_0 q_i\right), \quad i = 1,2,3 && (4.302)\\ h\rho_0 \dot{q}_{i-1} &= \left(Ap_{i-1} - Ap_i\right), \quad i = 2,3 && (4.303)\end{aligned}$$

$$\frac{d}{dt}\begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \\ Ap_2 \\ h\rho_0 q_2 \\ Ap_3 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_H^2 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & \omega_H^2 & 0 & -\omega_H^2 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & \omega_H^2 & 0 \end{pmatrix}\begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \\ Ap_2 \\ h\rho_0 q_2 \\ Ap_3 \end{pmatrix} \tag{4.304}$$

The resonance frequencies of the system matrix are

$$\omega_H = \frac{c}{h} = 3\frac{c}{L} \text{ and } \sqrt{3}\omega_H = \frac{3\sqrt{3}c}{L} = 5.2\frac{c}{L} \tag{4.305}$$

while the first and second resonance of the exact model are

$$\omega_1 = \pi\frac{c}{L} = 3.14\frac{c}{L} \text{ and } \omega_2 = 2\pi\frac{c}{L} = 6.28\frac{c}{L} \tag{4.306}$$

Example 69 *The input flow is zero and the outlet flow is zero, then a hybrid model should be used. With $N = 1$, then $L = 3h/2$, and the model is*

$$\frac{d}{dt}\begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_H^2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} Ap_1 \\ h\rho_0 q_1 \end{pmatrix} \tag{4.307}$$

The resonance is at

$$\omega_1 = \frac{3}{2}\frac{c}{L} = 1.5\frac{c}{L} \tag{4.308}$$

while the exact value for the first resonance is

$$\omega = \frac{\pi}{2}\frac{c}{L} = 1.57\frac{c}{L} \tag{4.309}$$

4.7 Object oriented simulation models

4.7.1 Introduction

The final section of the chapter will present examples on how subsystem models can be connected to a model, and how subsystem models may be added or changed without causing extensive work. This will be done using subsystem models that are interconnected with effort and flow variables.

4.7.2 Pump controlled hydraulic motor

Pump controlled hydraulic motors are used in applications where high power is required as the power efficiency may be as high as 90% for such systems. Such systems are used in vehicles and in cranes for heavy lifting operations.

We consider the system shown in Figure 4.18 which depicts an arrangement which is called a *hydrostatic gear* or *hydraulic gear*. The system has a pump with a variable displacement, which is driven by a motor with constant speed. The pump is connected to a motor which may have a fixed or variable displacement. The speed and direction of

Chapter 5

Friction

5.1 Introduction

5.1.1 Background

Friction is the tangential reaction force between two surfaces in contact. The friction force is dependent on a number of factors, such as contact geometry, properties of the surface materials, displacement, relative velocity and lubrications. Friction is a highly complex phenomenon, composed of several physical phenomena in combination. As a result of this, models of friction are to a large extent empirical, which means that the models are constructed in order to reproduce effects observed in experiments. On the other hand, some of the dynamic friction models aim at modelling the physics behind the phenomenon.

A macroscopic smooth surface is far from smooth when viewed at a microscopical scale. The small features of the surface are called *asperities*. When two surfaces are brought into contact, the true contact occur between the asperities in what is called *asperity junctions*. An example of this is shown in Figure 5.1. In engineering materials, the slope of the asperities are typical in the range $5° - 10°$, and the width is typical 10 μm. When two bodies in contact are brought into relative motion by an external force, the asperities will behave like springs, and there will be an elastic deformation of the asperities. This motion is referred to as *pre-sliding displacement* or the *Dahl effect*.

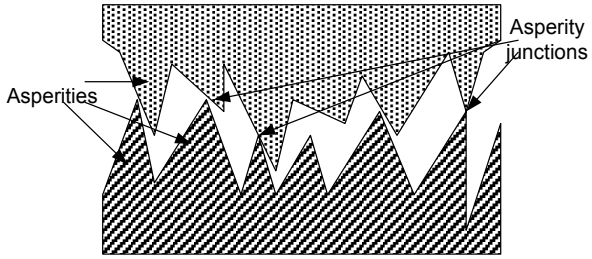

Figure 5.1: The asperities and junctions of two bodies in contact, viewed at a microscopical scale

The tangential force $F_t(x)$ can in this regime be approximated by

$$F_t(x) = -k_t x \tag{5.1}$$

where k_t is the stiffness of the contact and x is the relative displacement.

By increasing the external force the asperities will undergo plastic, irreversible deformation and then rupture. The force needed to break the junctions is referred to as the *break-away force* F_b, and the phenomena itself is called *break-away*. The stiffness k_t can be found from

$$k_t = \frac{F_b}{x_b} \tag{5.2}$$

where x_b is the break-away displacement. Before beak-away the system is said to *stick,* while after break-away the system is said to *slip,* and the term stick-slip friction is used to characterize the phenomenon.

5.2 Static friction models

Static friction models present the friction force as a function of velocity. This function can be characterized by the following for regimes:

I. Static friction Elastic deformation of the asperities, the Dahl effect.

II. Boundary lubrication For very low velocities, no fluid lubrication occur, and the friction is dominated by shear forces in the solid boundary film.

III. Partial fluid lubrication The Stribeck effect

IV. Full fluid lubrication A lubricant film thicker than the size of the asperities is maintained, and no solid contact occurs. The friction is purely viscous.

The resulting map is referred to as the *generalized Stribeck curve*, which is shown in Figure 5.2. Static friction models will represent these regimes to a varying extent. A selection of static friction models that are commonly used is shown in Figure 5.3.

5.2.1 Models for the individual phenomena

Coulomb friction

The classical model of friction where the friction force is proportional to load, opposes the motion, and is independent of contact area is known as Coulomb friction. The friction force in the Coulomb model is given by

$$F_f = F_c \text{sgn}(v), \quad v \neq 0 \tag{5.3}$$

where the Coulomb force F_c is given by

$$F_c = \mu F_N \tag{5.4}$$

Here μ is the friction coefficient and F_N is the load. Equation (5.4) can be derived as follows. It is assumed that there is no contamination, such as lubrication, of the contact surfaces. The friction is then referred to as *dry friction.* In this context friction can

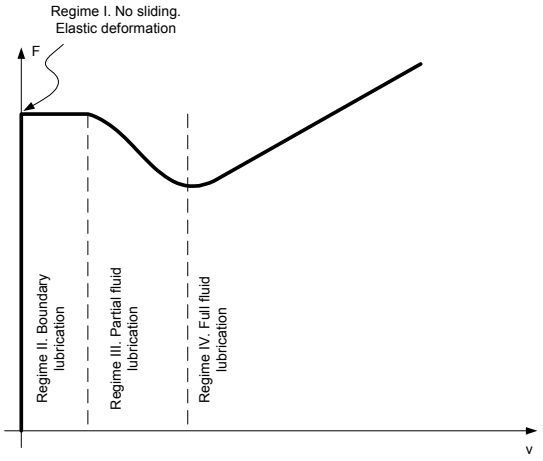

Figure 5.2: The generalized Stribeck curve, showing friction as a function of velocity for low velocities, (Armstrong-Hélouvry et al. 1994).

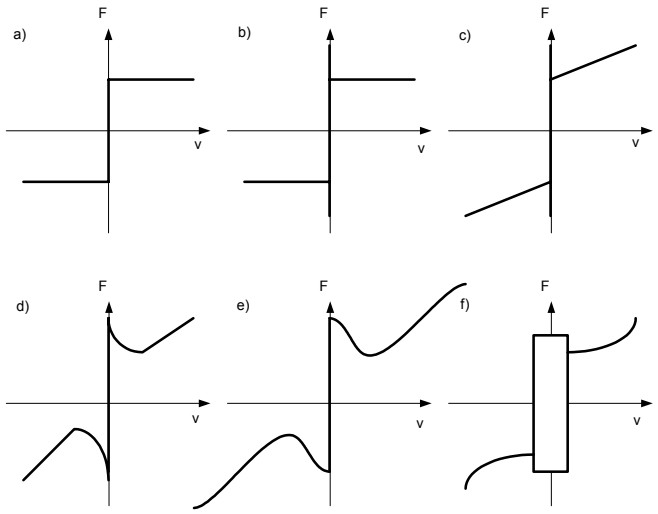

Figure 5.3: Static friction models: a) Colomb friction b) Coulomb+stiction c) Coulomb+stiction+viscous d) Stribeck effect e) Hess and Soom; Armstrong f) Karnopp model

be defined as the shear strength of the asperity junction areas, and the friction force is proportional to the true area of contact A_c

$$F_f = A_c f_s \tag{5.5}$$

where f_s is the shear force per unit area, a constant material property. The true area of contact A_c can be found from

$$A_c = \frac{F_N}{p_y} \tag{5.6}$$

where F_N is the load, and p_y is the yield pressure, a constant material property. Combining (5.6) with (5.5) gives

$$F_f = \frac{F_N}{p_y} f_s = \mu F_N \tag{5.7}$$

where the friction coefficient μ is found as $\mu = f_s/p_y$, and thus (5.4) holds. From the derivation, it is seen that A_c is cancelled out of the expression, and so friction is independent of contact area.

According to Armstrong-Hélouvry et al. (1994), F_c is also dependent on lubricant viscosity and contact geometry. The nature of Coulomb friction was known to Leonardo Da Vinci, and his results were further developed by Coulomb. The friction force given by (5.3) is not necessary symmetric in v, that is, F_f may take different values for different directions of the velocity.

Static friction

Static friction is also known as stiction and models the fact that in some cases the friction force is larger in magnitude for zero velocity than for a non-zero velocity. According to the stiction model the system sticks if the velocity is zero and $|F_f| < F_s$, and it breaks away if $|F_f| = F_s$ where $F_s > F_c$ is the stiction force, which is larger in magnitude than the Coulomb force F_c.

During a pre-sliding displacement, some motion is possible even when a mechanism is stuck in static friction. If the applied force returns to zero, the position returns to its initial value, possibly after a transient of pre-sliding displacement.

Viscous friction

Viscous friction is present in fluid lubricated contacts between solids. The concept of viscous friction was first introduced by Reynolds(1886). A viscous friction model takes into account that due to hydrodynamic effects, the friction force depends on the magnitude of the velocity, and not only its direction. The usual linear model is given by

$$F_{fv} = F_v v \tag{5.8}$$

where the viscous friction is proportional to velocity. The constant of proportionality F_v depends on lubricant viscosity, loading and contact geometry. Generally, viscous friction exhibits a non-linear behavior, and a nonlinear version of (5.8) is

$$F_{fv} = F_v \left|v\right|^{\delta_v} \operatorname{sgn}(v) \tag{5.9}$$

where $0 < \delta_v \leq 1$ is a constant.

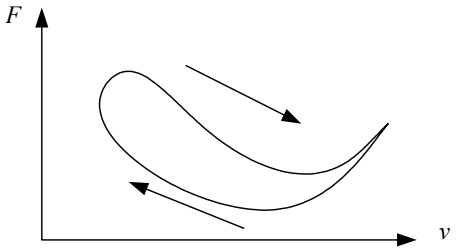

Figure 5.4: Frictional lag

Decreasing viscous friction: The Stribeck effect

Decreasing viscous friction is also known as Stribeck friction or the *Stribeck effect*, and was first described in Stribeck (1902). The Stribeck effect has its background in partial fluid lubrication. In this case some of the load is carried by the fluid film created by the lubricant, and some by elastic and plastic deformation of the asperities. The fluid film thickness increases with velocity, and the resulting tangential force decreases since the shear forces of the film are smaller than the shear forces of the asperities. Several static models have been proposed for the Stribeck effect. Armstrong-Hélouvry (1990) propose to use

$$F_f = \left[F_c + (F_s - F_c)e^{-(v/v_s)^2}\right] \text{sgn}(v), \quad v \neq 0 \tag{5.10}$$

where F_s is the static friction, F_c is the Coulomb friction and v_s is denoted the characteristic velocity of the Stribeck friction. Equation (5.10) will model Coulomb friction, stiction and the Stribeck effect. The model is not defined for zero velocity, but $|F_f| \to |F_c|$ when the velocity tends to zero. Hess and Soom (1990) propose the expression

$$F_f = \left[F_c + \frac{(F_s - F_c)}{1 + (v/v_s)^2}\right] \text{sgn}(v), \quad v \neq 0 \tag{5.11}$$

for modeling the same effect. It is important to notice that models such as (5.10) or (5.11) are not based on the physics of the phenomenon, but is rather a curve fit to experimental data as shown in Figures 5.3 d) and f).

Other friction related phenomena

In the friction experiments of Hess and Soom (1990) it was observed that friction force was lower for decreasing velocities than for increasing. This is a hysteresis effect, and is referred to as *frictional lag*, or frictional memory, see Figure 5.4. Moreover, the break-away force has been found to depend on the rate of change of externally applied force. Larger rates gives smaller break-away force.

5.2.2 Combination of individual models

The static models presented above can be combined to produce models that take several of these phenomena into account. For instance, the most commonly used model in engineering, which is the Coulomb+viscous friction model can be found by adding (5.3) and (5.8) together to form

$$F = F_c \text{sgn}\,(v) + F_v v \tag{5.12}$$

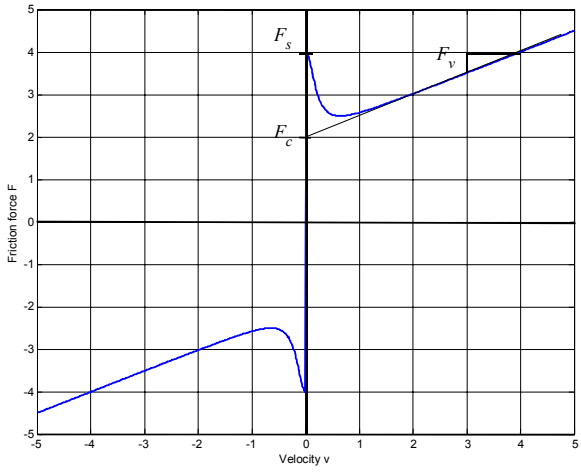

Figure 5.5: An example of a friction map

The Stribeck effect can also be included using (5.11), which gives (Hess and Soom 1990)

$$F = \left(F_c + \frac{(F_s - F_c)}{1 + (v/v_s)^2} \right) \operatorname{sgn}(v) + F_v v \tag{5.13}$$

A plot of the Friction curve given by (5.13) is shown in Figure 5.5

5.2.3 Problems with the static models

There are two main problems connected to the use of static friction models for simulation and control applications:

1. They are dependent on the detection of zero velocity, as the model rely on switching at zero velocity.

2. They do not describe all observed dynamic effects, such as pre-sliding displacement, varying break-away force and frictional lag..

The zero velocity problem can be handled by a static model known as the Karnopp model, Karnopp (1985), where a zero velocity interval, $|v| < \eta$ is used. Outside this interval,that is for $|v| \geq \eta$, friction force is the usual function of velocity, but within the interval, the velocity is considered to be zero and friction is a function of other forces in the system:

$$F_f(v, F) = \begin{cases} F_f(v), & |v| \geq \eta \\ F_f(F), & |v| < \eta \end{cases} \tag{5.14}$$

where F represents the sum of other forces in the system and $\eta > 0$ is the small constant defining the Karnopp zero interval. This is advantageous in simulations, but the zero interval does not agree with real friction, and the model strongly depends on the rest of the system through $F_f(F)$. A plot of the static Karnopp model is shown in Figure 5.3 f).

A friction model known as Armstrong's 7-parameter model Armstrong-Hélouvry et al. (1994) is capable of modelling some of the phenomena not included in the classical static models. The model consists of two equations, one for sticking and one for sliding:

$$F(x) = \sigma_0 x, \tag{5.15}$$

when sticking, and

$$F = \left(F_c + F_s(\gamma, t_d) \frac{1}{1 + (v(t-\tau_l)/v_s)^2} \right) \mathrm{sgn}(v) + F_v v \tag{5.16}$$

where

$$F_s(\gamma, t_d) = F_{s,a} + \left(F_{s,\infty} - F_{s,a} \frac{t_d}{t_d + \gamma} \right) \tag{5.17}$$

when sliding. $F_{s,a}$ is the Stribeck friction at the end of the previous sliding period and t_d is the dwell time, the time in stick. Although the Armstrong 7-parameter model describes more phenomena than the classical models, it still requires switching between different equations. The problems with static friction models in connection with simulation and control systems design has led to the use of *dynamic friction models* for high precision friction modeling.

5.2.4 Problems with signum terms at zero velocity

In static friction models the friction term will typically include a term which more or less looks like the Coulomb friction model $F_f = F_c \mathrm{sgn}(v)$. This model is not defined at zero velocity, however, it is not unusual that the model is extended to be valid at zero velocity using the model

$$F_c \mathrm{sgn}(v) = \begin{cases} -F_c, & v < 0 \\ 0 & v = 0 \\ F_c, & 0 < v \end{cases} \tag{5.18}$$

This is e.g. done in the Simulink block for Coulomb and viscous friction. It is not difficult to see that this model does not reflect the physics of the problem. To makes this clear we consider a mass m with position x, velocity $v = \dot{x}$ and an active force F_a acting on the mass. The friction force is F_f, and Newton's law gives

$$m\dot{v} = F_a - F_f \tag{5.19}$$

Now, suppose that $v = 0$. If the friction force is given by $F_f = F_c \mathrm{sgn}(v)$ as defined by 5.18, then $F_f = 0$ for $v = 0$, and the system will not stick, but rather accelerate according to $m\dot{v} = F_a$. If F_a is positive and nonzero, then the velocity becomes positive, and the friction force will at the next time step be $F_f = -F_c$. If $F_a < F_c$ this will cause a reversal of the acceleration, and at the next step the velocity may have changed sign so that $F_f = F_c$. It is clear that this will give strong oscillations in the system, and that these oscillations have nothing to do with the physics of the system. In a simulation with fixed time-step, there will be strong oscillations in the numerical solution, whereas a simulation with variable time-step will more or less stop as the time step will be made very small to in order to reach the specified accuracy.

The conclusion to this discussion is that models that contain a signum term like the one in (5.18) will not give results that agree with the physics of the problem at zero velocity. Moreover, serious problems are introduced in simulations.

5.2.5 Karnopp's model of Coulomb friction

Karnopp's friction model extends the basic Coulomb friction model for dry friction to be valid also for zero velocity. It is straightforward to extend this model to include stiction and the Stribeck effect. Karnopp's model can be explained by considering a mass m with velocity v that is pushed on a flat surface with a force F_a. The friction force on the mass is F_f so that the equation of motion is given by (5.19) According to the Coulomb friction model the friction force F_f is of magnitude F_c in the opposite direction of the velocity as long as the velocity is nonzero. This may be modelled as

$$m\dot{v} = \begin{cases} F_a + F_c, & v < 0 \\ F_a - F_c, & v > 0 \end{cases} \tag{5.20}$$

In this formulation the friction force is a function of the velocity. Note, however, that this model is undefined for zero velocity, so there is a need for refining the model. This is done in Karnopp's model by observing that the physical behavior of the system at zero velocity is that the velocity remains equal to zero as long as the force F_a is less than F_c in magnitude. This can be written

$$m\dot{v} = 0, \quad v = 0 \text{ and } |F_a| \leq F_c \tag{5.21}$$

Combining this with the equation of motion we see that $F_f = F_a$ when $v = 0$ and $|F_a| \leq F_c$. Thus, for zero velocity, the friction force is a function of the force acting on the mass. Define the saturation function $\text{sat}(x, S)$ so that $\text{sat}(x, S) = x$ when $|x| \leq S$ and $\text{sat}(x, S) = S\text{sgn}(x)$ when $|x| \geq S$.

The Karnopp friction model for Coulomb friction is given by

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } v = 0 \\ F_c\text{sgn}(v) & \text{else} \end{cases} \tag{5.22}$$

Note that the computational input of the Karnopp model at the input port is F_a when $v = 0$ and $|F_a| \leq F_c$, and that the computational input changes to v when the condition does no longer hold.

5.2.6 More on Karnopp's friction model

The main contribution of Karnopp's friction model is the handling of the sticking phenomenon at zero speed. This can also be applied to other friction models with signum terms. The model (5.10) which includes sticking and the Stribeck effect can be modeled with Karnopp's method with the friction force

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } v = 0 \\ \left[F_c + (F_s - F_c)e^{-(v/v_s)^2}\right]\text{sgn}(v) & \text{else} \end{cases} \tag{5.23}$$

where F_a is the applied force and v is the velocity.

In simulations with Karnopp's model there must be a switch between the two regimes in (5.22) or (5.23). This requires some method for detecting that the velocity is zero. Some simulation systems with variable-step integration methods will have event-detection mechanisms that can be used for this purpose. This type of event detection is included

in MATLAB and Simulink. This method is used in friction models in Modelica, where the computational inputs are switched at zero velocity.

Alternatively, a dead-zone around zero velocity can be used where the velocity is treated as is it were zero in the computation of the friction force. Then the friction model (5.22 will be modified to

$$F_f = \begin{cases} \text{sat}(F_a, F_c) & \text{when } |v| \leq \delta \\ F_c \text{sgn}(v) & \text{else} \end{cases} \tag{5.24}$$

where the magnitude δ of the dead-zone will have to be selected depending on the size of the time-step, and on the maximum acceleration that can be expected in the system. The model (5.24) is straightforward to implement in MATLAB and Simulink. If the friction model is used in an observer in a control system , then the time-step will have to be fixed, and a dead-zone must be used. It is clear that the introduction of a dead-zone will be an approximation that will introduce some error. However, the performance of this solution is vastly superior to the naive implementation in (5.18) of the signum term, and Karnopp's model should be the standard way of modeling friction unless dynamic phenomena like pre-sliding and frictional hysteresis are the dominant physical effects.

5.2.7 Passivity of static models

Friction in its very nature is a dissipative phenomenon as the friction force cause dissipation of energy whenever the velocity is nonzero. An exception to this is elastic deformation in the pre-sliding region where energy is stored as in a spring, and the system will still be passive although energy is not dissipated. Because of this a friction model should be passive in the sense that the system with velocity as input and friction force as output should be passive to reflect the physics of the system. In this section we establish the passivity properties of the static friction models.

For any of the static friction models presented above, the friction force $F_f(v)$ is a sector nonlinearity, that is $F_f(v)$ satisfies

$$F_f(v) \in \text{sector}(k_1, k_2) \Longleftrightarrow k_1 v^2 < F_f(v)v < k_2 v^2 \tag{5.25}$$

A plot of the static model (5.13) is shown in Figure 5.5, and as can be seen $F_f(v)$ is located in the first and third quadrants, that is $F_f(v) \in \text{sector}[0, \infty)$. Moreover, by studying Figure 5.5, it can be seen that

$$F_f(v) \in \text{sector}[k_1, \infty), \text{where } 0 < k_1 \leq F_v \tag{5.26}$$

Calculating the power $F_f(v)v$ for a static friction model where $F_f(v)$ satisfies (5.26), and integrating, we get

$$\int_0^T F_f(v)v dt > \int_0^T k_1 v^2 dt, \text{where } 0 < k_1 \leq F_v \tag{5.27}$$

It follows from (5.27), that the system with input v and output F_f is passive. This result can be generalized to any sector nonlinearity. Karnopp's model is identical to the static models except at zero velocity. Therefore the integral $\int_0^T F_f(v)v dt$ will be the same as for the static methods. This implies that Karnopp's model is passive. When the dead-zone is included passivity cannot be established.

5.3 Dynamic friction models

5.3.1 Introduction

There are two problems connected to the use of static friction models for simulation and control applications. First, static models are dependent on the detection of zero velocity, as the models rely on switching between different models at zero velocity. Second, static models do not describe dynamic effects such as pre-sliding displacement, varying break away force and frictional lag. Because of this, dynamic modes of friction have been proposed.

5.3.2 The Dahl model

The model of Dahl (1968) was developed for the purpose of simulating control systems with friction. A dynamic friction model can be developed by differentiating the friction force with respect to time:

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial x}\frac{dx}{dt} \tag{5.28}$$

Then if $\partial F/\partial t = 0$, the expression

$$\frac{dF}{dt} = \frac{dF}{dx}\frac{dx}{dt} \tag{5.29}$$

is found. For small displacements the friction is determined by the pre-sliding elastic deformation of the asperities. This can be modeled as a linear spring:

$$x \ll 1 \Rightarrow |F| = |\sigma x| \ll F_c \tag{5.30}$$

with σ being the spring stiffness. For large displacements, the model should behave like a Coulomb model. Dahl (1976) found that a model that satisfies this is given by

$$\begin{aligned} \frac{dF}{dx} &= \sigma \left| 1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt} \right|^{\alpha} \text{sgn}\left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right) \\ &= \sigma \left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right)^{\alpha} \text{sgn}^{\alpha+1}\left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right) \end{aligned} \tag{5.31}$$

By using the fact that $F < F_c$ it follows that $\text{sgn}\left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right) > 0$, and (5.31) simplifies to

$$\frac{dF}{dx} = \sigma \left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right)^{\alpha} \tag{5.32}$$

The constant α depends on the material of the solid, $\alpha \geq 1$ describe ductile type materials, while $\alpha < 1$ describes brittle type materials. Applications of the model, however, typically employ $\alpha = 1$, so that

$$\frac{dF}{dx} = \sigma \left(1 - \frac{F}{F_c}\text{sgn}\frac{dx}{dt}\right) \tag{5.33}$$

It follows from (5.33) that in steady state

$$F_{ss} = F_c \text{sgn}\frac{dx}{dt} = F_c \text{sgn}(v) \tag{5.34}$$

which can be compared to (5.3). The Dahl model includes the phenomena Coulomb friction and pre-sliding displacement. From (5.33) it is seen that for small displacements the friction force can be approximated by $F \approx \sigma x$, which is the model of a linear spring with σ being the spring stiffness. For large displacements the friction for can be approximated by $F \approx F_c$ as for a static Coulomb model. However, as the model (5.33) is rate independent, it is not capable of describing such phenomena as Stribeck-effect.

Combination of (5.29) and (5.33) then leads to the following result

The Dahl friction model is given by the dynamic model

$$\frac{dF}{dt} = \sigma\left(v - |v|\frac{F}{F_c}\right) \tag{5.35}$$

The Dahl model can be written in the form

$$\frac{dF}{dt} = \sigma\frac{|v|}{F_c}\left(F_f - F\right) \tag{5.36}$$

where $F_f = F_c\text{sgn}(v)$. This resembles a low pass filter where the computed friction force F tracks the Coulomb friction $F_c\text{sgn}(v)$ with a pole at $|v|/F_c$ which corresponds to a time constant $T = F_c/|v|$. The pole of the low pass filter tends to zero when the velocity tends to zero, and this is advantageous in simulation as the apparent gain around zero is small in spite of the step due to the signum function. However, this advantage does not come for free. The low gain of the model at low speeds may create problems in the modeling of the sticking regime. In particular, if the system (5.19) is excited by an oscillatory force F_a where $|F_a| < F_c$, then the system may drift if the friction force F_f is modelled with the Dahl model, although sticking would be expected from the physics of the system.

To conclude, the Dahl model is very simple to implement compared to Karnopp's model as there is no switching in the Dahl model. Moreover, there are no problems with oscillations around zero velocity with Dahl's model. However, the model may give drift in the sticking region, so for models where correct sticking behavior is important it is recommended to use Karnopp's model.

Example 75 *It is interesting to note that the differential equation in (5.33) can be solved explicitly. Assuming F_c to be a constant and considering only forward motion, that is $\dot{x} > 0$, we find by integrating (5.33) that*

$$\int \frac{dF}{\sigma\left(1 - \frac{F}{F_c}\right)} = \int dx$$

This gives

$$-\frac{F_c}{\sigma}\ln\left(\sigma - \frac{\sigma}{F_c}F\right) = x + C' \tag{5.37}$$

where C' is a constant of integration, and where it has been assumed that $F < F_c$. Solving for F we find that

$$F = F_c - \frac{F_c C''}{\sigma}e^{-\frac{\sigma}{F_c}x}$$

where C'' is another constant. Finally, by using the fact that $F(0) = 0$, we find that $C'' = \sigma$, and consequently Dahl's friction model for forward speed can be written in the form

$$F = F_c\left(1 - e^{-\frac{\sigma}{F_c}x}\right) \tag{5.38}$$

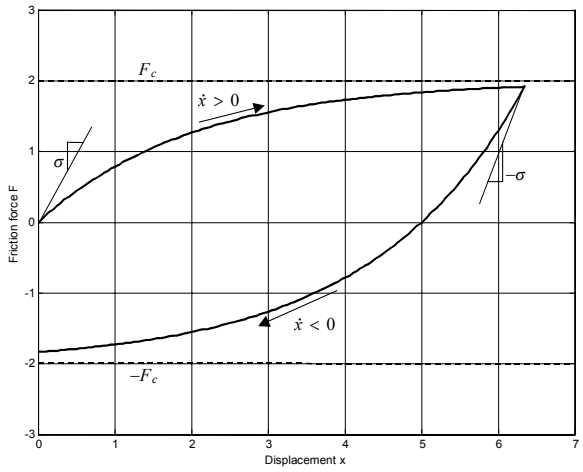

Figure 5.6: The Dahl friction model for positive and negative velocity.

By doing the same calculations for $\dot{x} < 0$, *we find that*

$$F = F_c \text{sgn}(v) \left(1 - e^{-\frac{\sigma}{F_c} x sign(v)}\right) \tag{5.39}$$

which is shown for $F_c = 2$ *and* $\sigma = 0.5$ *in Figure 5.6.*

5.3.3 Passivity of the Dahl model

For dynamic Dahl model the friction force is given by

$$\frac{dF}{dt} = \sigma_0 \left(v - \frac{F|v|}{F_c}\right), F_c > 0$$

we consider the storage function

$$V = \frac{F^2}{2\sigma_0}$$

The time derivative along the solutions of the system is

$$\dot{V} = \frac{1}{\sigma_0} F\dot{F} = Fv - \frac{F^2|v|}{F_c} \tag{5.40}$$

It is seen that for the Dahl model the system with input v and output F is passive.

5.3.4 The Bristle and LuGre model

The Bristle model to friction was introduced by (Haessig, Jr. and Friedland 1991). The basic assumption behind the Bristle friction model is that asperity junctions can be modeled of elastic bristles as shown in Figure 5.7. As the surfaces move relative to each

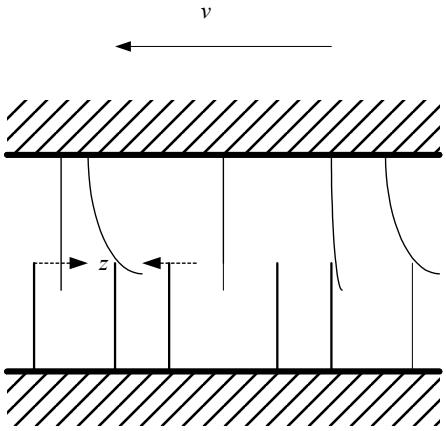

Figure 5.7: The asperity junctions of two bodies in contact are modeled as elastic bristles. For simplicity, only the upper body is moving in this figure, and the bristles of the lower body are assumed rigid.

other the strain in the bond increases and the bristles acts as springs giving rise to the friction force. The force is given by

$$F = \sum_{i=1}^{N} \sigma_0 \left(x_i - b_i \right) \tag{5.41}$$

where N is the number of bristles, σ_0 is the stiffness of the bristles, x_i is the relative position of the bristle and b_i is the location where a bond was formed. In simulations, a bond will snap when $|x_i - b_i| = \delta_s$, and then a new will be formed at a random location relative to the previous location. The complexity of this model increases with N. The stiffness of the bristles σ_0 can be made velocity dependent. An interesting property of this model is that it attempts to capture the random nature of friction. However, it is inefficient in simulations due to its complexity. Also it may give rise to oscillatory motion in stick due to the lack of damping. The LuGre (Lund-Grenoble) model (Canudas de Wit, Olsson, Åström and Lischinsky 1995) is based on the same idea as the bristle model, but the friction force is generated by a dynamic equation reminiscent of the Dahl model describing the average deflection of the bristles, thereby reducing the complexity introduced by the sum in (5.41).

The LuGre dynamic friction model is given by the dynamic system

$$\dot{z} = v - \sigma_0 \frac{|v|}{g(v)} z \tag{5.42}$$

where the function $g(v)$ is selected to be

$$\sigma_0 g(v) = F_c + (F_s - F_c) e^{-(v/v_s)^2} \tag{5.43}$$

to account for stiction and the Stribeck effect as in (5.10). The friction force is given by

$$F = \sigma_0 z + \sigma_1 \dot{z} + \sigma_2 v. \tag{5.44}$$

The following property of the model is noted: It is seen from (5.42) that the model for z can be written

$$\dot{z} = \frac{|v|}{g(v)} \left[g(v)\text{sgn}(v) - \sigma_0 z \right] \tag{5.45}$$

From this formulation it is seen that if the initial value of z satisfies $|\sigma_0 z(0)| \leq g_{\max}$, where $g_{\max}$ is the maximum value of $g(v)$, then the absolute value of the state $z(t)$ is upper bounded according to

$$|\sigma_0 z(t)| \leq g_{\max} \tag{5.46}$$

It is also noted that

$$F_c \leq \sigma_0 g(v) \leq F_s \tag{5.47}$$

It is seen that the stationary solution of (5.42) is

$$z_{ss} = g(v)\text{sgn}(v). \tag{5.48}$$

This shows that the therm $\sigma_2 v$ in (5.44) will tend to represent stiction and the Stribeck effect, while $\sigma_2 v$ term will represent viscous friction. The is characterized by the six parameters σ_0, σ_1, σ_2, v_s, F_s and F_c. By comparing the LuGre model with the Dahl model as given by (5.35) it is clear that the LuGre model is a generalization of the Dahl model, where the Dahl model appears in the case that $\sigma_0 z = F$, $\sigma_1 = \sigma_2 = 0$ and $g(v) = F_c$. The LuGre model has the potential of being more accurate than the Dahl model as it includes the Stribeck effect, and as it may represent frictional lag. As with the Dahl model the LuGre model may drift in the sticking region.

5.3.5 Passivity of the LuGre model

To find out if the model is passive from velocity v to friction force F we follow the procedure of (Barabanov and Ortega 2000) and investigate the integral

$$\int_0^T vF dt = \int_0^T \sigma_0 vz dt + \int_0^T v \left(\sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt \tag{5.49}$$

The first term on the right side corresponds to the Dahl part of the LuGre model. We find that this term is not problematic as

$$\begin{aligned} \int_0^T \sigma_0 vz dt &= \int_0^T \sigma_0 z \left(\dot{z} + \sigma_0 \frac{|v|}{g(v)} z \right) dt \\ &= \frac{\sigma_0}{2} \left[z^2(T) - z^2(0) \right] + \int_0^T z^2 \sigma_0 \frac{|v|}{g(v)} dt \\ &\geq -\frac{\sigma_0}{2} z^2(0) \end{aligned} \tag{5.50}$$

The second term on the right side of (5.49) is somewhat more involved. We find that

$$\int_0^T v \left(\sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt = \int_0^T v \left((\sigma_1 + \sigma_2) v - \sigma_1 \frac{|v|}{g(v)} \sigma_0 z \right) dt \tag{5.51}$$

Using (5.46) we find that

$$\int_0^T v \left(\sigma_1 \frac{dz}{dt} + \sigma_2 v \right) dt \geq \int_0^T v^2 \left((\sigma_1 + \sigma_2) - \sigma_1 \frac{g_{\max}}{g_{\min}} \right) dt \tag{5.52}$$

This implies that the LuGre model is passive from v to F if

$$(\sigma_1 + \sigma_2) - \sigma_1 \frac{g_{\max}}{g_{\min}} \geq 0 \Rightarrow \sigma_1 \leq \sigma_2 \frac{g_{\min}}{g_{\max} - g_{\min}} \tag{5.53}$$

Insertion of the maximum and minimum values of g according to (5.47) leads to the conclusion that the LuGre model is passive if

$$\sigma_1 \leq \sigma_2 \frac{F_c}{F_s - F_c} \tag{5.54}$$

In (Barabanov and Ortega 2000) it was shown that this is a necessary and sufficient condition for passivity from v to F.

5.3.6 The Elasto-Plastic model

The LuGre model does not render true stiction. Therefore a new dynamic friction model was proposed in (Dupont, Hayward, Armstrong and Altpeter 2002) with the aim of having a model that accounts for both true stiction and pre-sliding. The model is a generalization of the LuGre model. The model is written

$$\dot{z} = v \left(1 - \alpha(z, v) \frac{\sigma_0 \text{sgn}(v)}{g(v)} z \right)^i \tag{5.55}$$

$$F = \sigma_0 z + \sigma_1 \frac{dz}{dt} + \sigma_2 v \tag{5.56}$$

and the term $\alpha(z, v)$, which is the new feature of the model when compared to LuGre, is used to render true stiction. The piecewise continuous function $\alpha(z, v)$ is defined as

$$\alpha(z,v) = \begin{cases} \begin{cases} 0 & |z| \leq z_b \\ 0 < \alpha < 1 & z_b < |z| < z_{\max}(v) \\ 1 & |z| \geq z_{\max}(v) \end{cases} & , \text{sgn}(v) = \text{sgn}(z) \\ 0 & , \text{sgn}(v) \neq \text{sgn}(z) \end{cases}, \tag{5.57}$$

where

$$0 < z_b < z_{\max}(v) = \frac{g(v)}{\sigma_0}, \forall v \in \mathbb{R}. \tag{5.58}$$

An example of the term $\alpha(z, v)$ is

$$\alpha(z,v) = \begin{cases} \begin{cases} 0 & |z| \leq z_b \\ \frac{1}{2} \sin \left(\pi \frac{z - \left(\frac{z_{\max} + z_b}{2} \right)}{z_{\max} - z_b} \right) + \frac{1}{2} & z_b < |z| < z_{\max}(v) \\ 1 & |z| \geq z_{\max}(v) \end{cases} & , \text{sgn}(v) = \text{sgn}(z) \\ 0 & , \text{sgn}(v) \neq \text{sgn}(z) \end{cases}$$

In the Elasto-Plastic model, the body displacement

$$x = z + w.$$

is decomposed into its elastic and plastic (inelastic) components z and w. Stiction corresponds to the existence of a breakaway displacement $z_b > 0$ such that for $|z| \leq z_b$ all motion of the friction interface consists entirely of elastic displacement. In this context,

elastic displacement z corresponds to pre-sliding displacement and plastic (inelastic) displacement w corresponds to sliding displacement. The choice of $\alpha(z, v) = 0, |z| \leq z_b$ in (5.57), directly implies that the Elasto-Plastic model has a true stiction phase. The Elasto-Plastic model is relatively complicated to implement. Therefore it is recommended to use Karnopp's model if the objective is to have a model with true stiction, and accurate modeling of pre-sliding is not important, which normally will be the case. However, for problems of very high accuracy where pre-sliding is the dominant frictional phenomenon the Elasto-Plastic model can be used.

5.3.7 Passivity of the Elasto-Plastic model

As the Elasto-Plastic model differs from the LuGre model only in the inclusion of the term $\alpha(z, v)$ in (5.55), the passivity analysis of the LuGre applies. It follows that the system with input v and output F will be passive provided that

$$\sigma_1 \leq \sigma_2 \frac{F_c}{F_s - F_c} \tag{5.59}$$

Part III

Dynamics

Chapter 6

Rigid body kinematics

6.1 Introduction

Rigid body dynamics is important for a wide range of control applications, and is essential in robot control, ship control, the control of aircraft and satellites, and vehicle control in automotive systems. The field of rigid body dynamics is old and is very rich in results. Important results date back to Newton in the 17th century, Euler in the 18th century and Lagrange, Hamilton and Rodrigues in the 19th century. Because of the development in control applications like robotics, aerospace, and the development in numerical simulation, the selection of topics and method to be presented in rigid body dynamics has developed quite a lot the last two decades, and this text attempts to reflect this change. The material is based on general texts like (Kane and Levinson 1985) and (Robertson and Schwertassek 1988), texts on spacecraft dynamics like (Kane, Likins and Levinson 1983) and (Hughes 1986), and robotics books like (Spong and Vidyasagar 1989) and (Sciavicco and Siciliano 2000).

6.2 Vectors

6.2.1 Vector description

Forces, torques, velocities and accelerations are well-known entities that can be described by vectors. A vector $\vec{u}$ can be described by its magnitude $|\vec{u}|$ and its direction. Note that

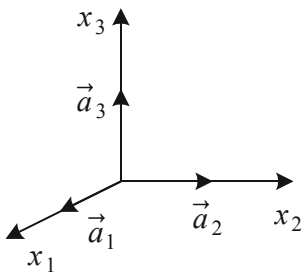

Figure 6.1: The coordinate frame a.

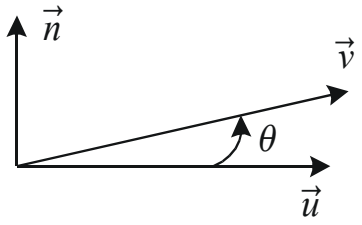

Figure 6.2: Vectors $\vec{u}$, $\vec{v}$ and $\vec{n}$.

this description of a vector does not rely on the definition of any coordinate frame. In this respect the description may be said to be coordinate-free. Alternatively, a Cartesian coordinate frame can be introduced, and the vector can be described in terms of its components in the Cartesian coordinate frame. Let the Cartesian coordinate frame a be defined by three orthogonal unit vectors $\vec{a}_1$, $\vec{a}_2$ and $\vec{a}_3$ that are unit vectors along the x_1, x_2, x_3 axes of a (Figure 6.1). Then the vector $\vec{u}$ can be expressed as a linear combination of the orthogonal unit vectors $\vec{a}_1$, $\vec{a}_2$ and $\vec{a}_3$ by

$$\vec{u} = u_1\vec{a}_1 + u_2\vec{a}_2 + u_3\vec{a}_3 \tag{6.1}$$

where

$$u_i = \vec{u} \cdot \vec{a}_i, \quad i \in \{1, 2, 3\} \tag{6.2}$$

are the unique *components* or *coordinates* of $\vec{u}$ in a. A related description of the vector is the *coordinate vector* form where the coordinates of the vector are written as a column vector

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \tag{6.3}$$

6.2.2 The scalar product

The *scalar product* between two vectors $\vec{u}$ and $\vec{v}$ is given in the coordinate-free description by

$$\vec{u} \cdot \vec{v} = |\vec{u}||\vec{v}| \cos\theta \tag{6.4}$$

where θ is the angle between the two vectors (Figure 6.2). With reference to the frame a we may then represent the vectors $\vec{u}$ and $\vec{v}$ by their coordinate vectors

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \tag{6.5}$$

where $u_i = \vec{u} \cdot \vec{a}_i$ and $v_i = \vec{v} \cdot \vec{a}_i$. The scalar product in terms of coordinate vectors is

$$\begin{aligned} \vec{u} \cdot \vec{v} &= (u_1\vec{a}_1 + u_2\vec{a}_2 + u_3\vec{a}_3) \cdot (v_1\vec{a}_1 + v_2\vec{a}_2 + v_3\vec{a}_3) \\ &= u_1v_1 + u_2v_2 + u_3v_3 \\ &= \mathbf{u}^T\mathbf{v} \end{aligned}$$

where it is used that $\vec{a}_i \cdot \vec{a}_j = \delta_{ij}$ which is equal to unity when $i = j$ and zero otherwise.

The scalar product can be written in the three alternative forms

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^{3} u_i v_i = \mathbf{u}^T \mathbf{v} \tag{6.6}$$

6.2.3 The vector cross product

The *vector cross product* is given in the coordinate-free form by

$$\vec{u} \times \vec{v} = \vec{n} |\vec{u}| |\vec{v}| \sin\theta \tag{6.7}$$

where $0 \leq \theta \leq \pi$ and $\vec{n}$ is a unit vector that is orthogonal to both $\vec{u}$ and $\vec{v}$ and defined so that $(\vec{u}, \vec{v}, \vec{n})$ forms a right-hand system (Figure 6.2).

With reference to a Cartesian frame a the vector cross product can be evaluated from

$$\vec{w} = \vec{u} \times \vec{v} = \begin{vmatrix} \vec{a}_1 & \vec{a}_2 & \vec{a}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \tag{6.8}$$

In component form this may alternatively be expressed by introducing the *permutation symbol*

$$\varepsilon_{ijk} = \begin{cases} 1 & \text{when } i,j,k \text{ is a cyclic permutation} \\ -1 & \text{when } i,j,k \text{ is not a cyclic permutation} \\ 0 & \text{when } i=j,\ i=k \text{ or } j=k \end{cases} \tag{6.9}$$

Here, as the indices $\{i, j, k\}$ is a cyclic permutation if they are equal to $\{1, 2, 3\}$, $\{2, 3, 1\}$ or $\{3, 1, 2\}$, and not a cyclic permutation if they are $\{1, 3, 2\}$, $\{2, 1, 3\}$ or $\{3, 2, 1\}$. It is noted that the definition implies that

$$\varepsilon_{ijk} = -\varepsilon_{jik} = -\varepsilon_{ikj} \tag{6.10}$$

$$\varepsilon_{ijk} = \varepsilon_{jki} = \varepsilon_{kij} \tag{6.11}$$

Then the components of $\vec{w} = \vec{u} \times \vec{v}$ are given by

$$w_i = \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} u_j v_k \tag{6.12}$$

and the vector may be written

$$\vec{w} = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} \vec{a}_i u_j v_k \tag{6.13}$$

In coordinate vector notation we introduce the *skew-symmetric form* of the coordinate vector $\mathbf{u}$ defined by

$$\mathbf{u}^\times := \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix} \tag{6.14}$$

Then the vector cross product can be written in coordinate vector form as

$$\mathbf{w} = \mathbf{u}^\times \mathbf{v} = \begin{pmatrix} u_2 v_3 - u_3 v_2 \\ u_3 v_1 - u_1 v_3 \\ u_1 v_2 - u_2 v_1 \end{pmatrix}$$

We sum up this result with the following three equivalent representations of the vector cross product:

The vector cross product has the following three equivalent representations:

$$\vec{w} = \vec{u} \times \vec{v} \;\Leftrightarrow\; w_i = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \varepsilon_{ijk} u_j v_k \;\Leftrightarrow\; \mathbf{w} = \mathbf{u}^{\times}\mathbf{v} \tag{6.15}$$

Example 76 *Orthogonal unit vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$ satisfy*

$$\vec{a}_i \cdot \vec{a}_j = \delta_{ij} = \begin{cases} 1, & \text{when } i = j \\ 0, & \text{when } i \neq j \end{cases} \tag{6.16}$$

and

$$\vec{a}_1 \times \vec{a}_2 = \vec{a}_3, \quad \vec{a}_1 \times \vec{a}_3 = -\vec{a}_2 \tag{6.17}$$
$$\vec{a}_2 \times \vec{a}_3 = \vec{a}_1, \quad \vec{a}_2 \times \vec{a}_1 = -\vec{a}_3 \tag{6.18}$$
$$\vec{a}_3 \times \vec{a}_1 = \vec{a}_2, \quad \vec{a}_3 \times \vec{a}_2 = -\vec{a}_1 \tag{6.19}$$

Example 77 *The relation between the components of the skew symmetric form $\mathbf{u}^{\times}$ and the vector form of $\mathbf{u}$ can be expressed in terms of the permutation symbol as*

$$\left(\mathbf{u}^{\times}\right)_{ik} = \varepsilon_{ijk} u_j \tag{6.20}$$
$$u_j = \frac{1}{2}\varepsilon_{ijk}\left(\mathbf{u}^{\times}\right)_{ik} \tag{6.21}$$

Example 78 *For three arbitrary vectors $\vec{a}, \vec{b}, \vec{c}$ the vector cross product satisfies*

$$\vec{a} \times (\vec{b} \times \vec{c}) = \vec{b}\vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b}\vec{c} \tag{6.22}$$

This can be shown by calculation of the components on both sides. Let $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ be the coordinate representations of $\vec{a}$, $\vec{b}$ and $\vec{c}$ in some coordinate frame. The coordinate form of (6.22) is

$$\mathbf{a}^{\times}\mathbf{b}^{\times}\mathbf{c} = \mathbf{b}\mathbf{a}^T\mathbf{c} - \mathbf{a}^T\mathbf{b}\mathbf{c} = (\mathbf{b}\mathbf{a}^T - \mathbf{a}^T\mathbf{b}\mathbf{I})\mathbf{c} \tag{6.23}$$

which implies

$$\mathbf{a}^{\times}\mathbf{b}^{\times} = \mathbf{b}\mathbf{a}^T - \mathbf{a}^T\mathbf{b}\mathbf{I} \tag{6.24}$$

In particular we note that

$$\mathbf{a}^{\times}\mathbf{a}^{\times} = \mathbf{a}\mathbf{a}^T - \mathbf{a}^T\mathbf{a}\mathbf{I}. \tag{6.25}$$

Example 79 *From (6.25) it follows that*

$$\mathbf{a}^{\times}\mathbf{a}^{\times}\mathbf{a}^{\times} = \mathbf{a}^{\times}\left(\mathbf{a}\mathbf{a}^T - \mathbf{a}^T\mathbf{a}\mathbf{I}\right) = -\left(\mathbf{a}^T\mathbf{a}\right)\mathbf{a}^{\times} \tag{6.26}$$

where it is used that $\mathbf{a}^{\times}\mathbf{a} = \mathbf{0}$. In particular, if $\mathbf{k}$ is a unit vector, then $\mathbf{k}^T\mathbf{k} = 1$, and

$$\mathbf{k}^{\times}\mathbf{k}^{\times}\mathbf{k}^{\times} = -\mathbf{k}^{\times} \tag{6.27}$$

Example 80 *Let $\vec{a}$, $\vec{b}$ and $\vec{c}$ be three arbitrary vectors. The Jacobi identity is written*

$$\vec{a} \times (\vec{b} \times \vec{c}) + \vec{b} \times (\vec{c} \times \vec{a}) + \vec{c} \times (\vec{a} \times \vec{b}) = \vec{0} \tag{6.28}$$

This identity is established from (6.22) which gives

$$\vec{b}\vec{a} \cdot \vec{c} - \vec{a} \cdot \vec{b}\vec{c} + \vec{c}\vec{b} \cdot \vec{a} - \vec{b} \cdot \vec{c}\vec{a} + \vec{a}\vec{c} \cdot \vec{b} - \vec{c} \cdot \vec{a}\vec{b} = \vec{0} \tag{6.29}$$

The coordinate form of the Jacobi identity is

$$\mathbf{a}^{\times}\mathbf{b}^{\times}\mathbf{c} + \mathbf{b}^{\times}\mathbf{c}^{\times}\mathbf{a} + \mathbf{c}^{\times}\mathbf{a}^{\times}\mathbf{b} = \mathbf{0} \tag{6.30}$$

Example 81 *The Jacobi identity implies that*

$$(\vec{a} \times \vec{b}) \times \vec{c} = \vec{a} \times (\vec{b} \times \vec{c}) - \vec{b} \times (\vec{a} \times \vec{c}) \tag{6.31}$$

In coordinate form this is written

$$(\mathbf{a}^{\times}\mathbf{b})^{\times}\mathbf{c} = \mathbf{a}^{\times}\mathbf{b}^{\times}\mathbf{c} - \mathbf{b}^{\times}\mathbf{a}^{\times}\mathbf{c} \tag{6.32}$$

which implies that

$$(\mathbf{a}^{\times}\mathbf{b})^{\times} = \mathbf{a}^{\times}\mathbf{b}^{\times} - \mathbf{b}^{\times}\mathbf{a}^{\times} \tag{6.33}$$

Example 82 *The following problem is investigated: To what extent can the vector $\vec{v}$ be determined when*

$$\vec{w} = \vec{u} \times \vec{v} \tag{6.34}$$

and $\vec{w}$ and $\vec{u}$ are given? In coordinate form this is written

$$\mathbf{w} = \mathbf{u}^{\times}\mathbf{v} \tag{6.35}$$

The skew symmetric matrix $\mathbf{u}^{\times}$ is singular, which is obvious from the identity $\vec{u} \times \vec{u} = \vec{0}$ which implies that $\mathbf{u}^{\times}\mathbf{u} = \mathbf{0}$. This means that it is not possible to solve for $\mathbf{v}$. However, it is possible to find two equations for $\vec{v}$. First, it is clear that $\vec{w} \cdot \vec{v} = 0$, which means that $\vec{v}$ is in the plane orthogonal to $\vec{w}$. Second, it is found that

$$\vec{w} = \frac{\vec{w}}{|\vec{w}|} \sin\theta |\vec{u}||\vec{v}| \Rightarrow |\vec{v}| = \frac{|\vec{w}|}{|\vec{u}| \sin\theta} \tag{6.36}$$

This shows that if the angle θ between $\vec{u}$ and $\vec{v}$ is selected to be some value, then the length of $\vec{v}$ is given by (6.36).

6.3 Dyadics

6.3.1 Introduction

The idea of using vectors in mathematical modelling of physical systems is well known. Also the use of column vectors to represent vectors is easy to accept. In analogy with this it turns out that certain matrices can be the representation of physical quantities described by *pairs of vectors*. Such matrices play an important role in rigid body dynamics and fluid mechanics, and it is worthwhile to invest some time in developing the required formalism.

6.3.2 Introductory example: The inertia dyadic

The angular momentum of a rigid body about its center of mass, which will be discussed in great detail in Section 7.3.2, can be written in coordinate-free form as a vector $\vec{h}$, or it may be written in terms of its coordinates as a column vector $\mathbf{h}$ or the generic component h_i, where

$$\vec{h} = \sum_{i=1}^{3} h_i \vec{a}_i, \quad \mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} \tag{6.37}$$

Likewise the angular velocity can be represented by a vector $\vec{\omega}$, by a column vector $\boldsymbol{\omega}$, or by the generic component ω_i, where

$$\vec{\omega} = \sum_{i=1}^{3} \omega_i \vec{a}_i, \quad \boldsymbol{\omega} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \tag{6.38}$$

A standard result in rigid body dynamics (Section 7.3.2) is that the angular momentum can be expressed by the angular velocity according to the two alternative formulations

$$\mathbf{h} = \mathbf{M}\boldsymbol{\omega}, \quad h_i = \sum_{j=1}^{3} m_{ij}\omega_j \tag{6.39}$$

where $\mathbf{M} = \{m_{ij}\}$ is the inertia matrix of the rigid body about its center of mass. The first formulation gives the relation between the column vectors $\mathbf{h}$ and $\boldsymbol{\omega}$, and the other formulation presents the relation between the generic components h_i and ω_j. At this stage one might wonder: Is there a corresponding equation for the relation between $\vec{h}$ and $\vec{\omega}$ in coordinate-free form? This turns out to be the case, but to be able to do this we need to introduce the concept of a dyadic, which is the sum of pairs of vectors. We define the *inertia dyadic* by

$$\vec{M} := \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\vec{a}_i\vec{a}_j \tag{6.40}$$

Note that $\vec{a}_i\vec{a}_j$ is a pair of vectors which should not be confused with the scalar product $\vec{a}_i \cdot \vec{a}_j$. Consider the following calculation:

$$\begin{aligned} \vec{M} \cdot \vec{\omega} &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\vec{a}_i\vec{a}_j \cdot \sum_{k=1}^{3} \omega_k \vec{a}_k \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} m_{ij}\vec{a}_i \left(\vec{a}_j \cdot \omega_k \vec{a}_k\right) \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\omega_j \vec{a}_i \end{aligned} \tag{6.41}$$

Here we have used the result $\vec{a}_j \cdot \vec{a}_k = \delta_{jk}$. Comparing with (6.37) and (6.39) we see that this implies that

$$\vec{h} = \vec{M} \cdot \vec{\omega} \tag{6.42}$$

which is the relation between $\vec{h}$ and $\vec{\omega}$ in vector notation. This result is equivalent to the expressions in (6.39). We note that the dyadic $\vec{M}$ represents the same physical quantity as the inertial matrix $\mathbf{M}$ and the components m_{ij}. We conclude that:

The angular momentum vector $\vec{h}$ can be expressed by the angular velocity vector $\vec{\omega}$ with the three equivalent formulations

$$\vec{h} = \vec{M} \cdot \vec{\omega} \quad \Leftrightarrow \quad \mathbf{h} = \mathbf{M}\boldsymbol{\omega} \quad \Leftrightarrow \quad h_i = \sum_{j=1}^{3} m_{ij}\omega_j \tag{6.43}$$

where $\vec{M}$ is the inertia dyadic and $\mathbf{M}$ is the inertia matrix, which is the matrix representation of the inertia dyadic.

6.3.3 Matrix representation of dyadics

We define the *dyadic* $\vec{D}$ to be a linear combination of pairs of vectors $\vec{a}_i\vec{a}_j$ given by

$$\vec{D} = \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}\vec{a}_i\vec{a}_j \tag{6.44}$$

where

$$d_{ij} = \vec{a}_i \cdot \vec{D} \cdot \vec{a}_j \tag{6.45}$$

are the components of the dyadic $\vec{D}$ in frame a. The matrix

$$\mathbf{D} = \{d_{ij}\} \tag{6.46}$$

is said to be the the matrix representation of the dyadic $\vec{D}$ in frame a. Scalar premultiplication with a vector, that is the scalar product of the vector $\vec{u}$ with the dyadic $\vec{D}$ gives a vector according to

$$\begin{aligned} \vec{w} &= \vec{u} \cdot \vec{D} = \sum_{k=1}^{3} u_k\vec{a}_k \cdot \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}\vec{a}_i\vec{a}_j \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}u_i\vec{a}_j \end{aligned} \tag{6.47}$$

Scalar postmultiplication with a vector, which is the scalar product of a dyadic with a vector gives the vector

$$\vec{z} = \vec{D} \cdot \vec{u} = \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}\vec{a}_i\vec{a}_j \cdot \sum_{k=1}^{3} u_k\vec{a}_k \tag{6.48}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}u_j\vec{a}_i \tag{6.49}$$

We define the column vectors $\mathbf{w} = (w_1, w_2, w_3)^T$ and $\mathbf{z} = (z_1, z_2, z_3)^T$ corresponding to the vectors $\vec{w}$ and $\vec{z}$. We may then write the equivalent expressions

$$\vec{w} = \vec{u} \cdot \vec{D} \quad \Leftrightarrow \quad \mathbf{w}^T = \mathbf{u}^T\mathbf{D} \tag{6.50}$$

$$\vec{z} = \vec{D} \cdot \vec{u} \quad \Leftrightarrow \quad \mathbf{z} = \mathbf{D}\mathbf{u} \tag{6.51}$$

The *identity dyadic* is defined by

$$\vec{I} := \sum_{i=1}^{3}\sum_{j=1}^{3} \delta_{ij}\vec{a}_i\vec{a}_j = \vec{a}_1\vec{a}_1 + \vec{a}_2\vec{a}_2 + \vec{a}_3\vec{a}_3 \tag{6.52}$$

where δ_{ij} is equal to unity when $i = j$, and zero otherwise. This implies that for any vector $\vec{u}$

$$\vec{I}\cdot\vec{u} = \vec{u}\cdot\vec{I} = \vec{u} \tag{6.53}$$

and for any dyadic $\vec{D}$ we have

$$\vec{I}\cdot\vec{D} = \vec{D}\cdot\vec{I} = \vec{D} \tag{6.54}$$

The equivalent matrix form of these equations are

$$\mathbf{Iu} = \left(\mathbf{u}^T\mathbf{I}\right)^T = \mathbf{u} \tag{6.55}$$

$$\mathbf{ID} = \mathbf{DI} = \mathbf{D} \tag{6.56}$$

Example 83 *Let $\vec{\omega}$ be a vector and let $\vec{M}$ be a dyadic. Define the scalar*

$$K = \frac{1}{2}\vec{\omega}\cdot\vec{M}\cdot\vec{\omega} \tag{6.57}$$

which is defined independently of any coordinate frame. Let $\vec{\omega}$ be given in the a frame by $\vec{\omega} = \omega_1\vec{a}_1 + \omega_2\vec{a}_2 + \omega_3\vec{a}_3$, and let the dyadic be given by

$$\vec{M} = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}\vec{a}_i\vec{a}_j. \tag{6.58}$$

Then the quadratic form is found to be

$$K = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \omega_i\omega_j m_{ij} \tag{6.59}$$

The corresponding representation in matrix form is given by

$$K = \frac{1}{2}\boldsymbol{\omega}^T\mathbf{M}\boldsymbol{\omega} \tag{6.60}$$

where $\mathbf{M} = \{m_{ij}\}$ is the matrix representation of the dyadic $\vec{M}$.

Example 84 *Consider the dyadic $\vec{K} := \vec{k}\vec{k}$. Let $\vec{w}$ be an arbitrary vector. Then*

$$\vec{w}\cdot\vec{K} = (\vec{w}\cdot\vec{k})\vec{k} \quad \text{and} \quad \vec{K}\cdot\vec{w} = \vec{k}(\vec{k}\cdot\vec{w}) \tag{6.61}$$

The coordinate form is

$$\mathbf{w}^T\mathbf{K} = \mathbf{w}^T\mathbf{k}\mathbf{k}^T \quad \text{and} \quad \mathbf{Kw} = \mathbf{k}\mathbf{k}^T\mathbf{w} \tag{6.62}$$

It follows that the matrix form of $\vec{K} = \vec{k}\vec{k}$ is

$$\mathbf{K} = \mathbf{k}\mathbf{k}^T \tag{6.63}$$

Example 85 *The dyadic form of the vector cross product is*

$$\vec{u} \times \vec{v} = \vec{u}^{\times} \cdot \vec{v} = \vec{u} \cdot \vec{v}^{\times} \tag{6.64}$$

where

$$\vec{u}^{\times} = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} u_j \vec{a}_i \vec{a}_k \tag{6.65}$$

is the dyadic form of the skew symmetric form $\mathbf{u}^{\times}$*, and*

$$\vec{v}^{\times} = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} v_j \vec{a}_i \vec{a}_k \tag{6.66}$$

We may then check that

$$\vec{u}^{\times} \cdot \vec{v} = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} u_j \vec{a}_i \vec{a}_k \cdot \sum_{p=1}^{3} v_p \vec{a}_p = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} \vec{a}_i u_j v_k = \vec{u} \times \vec{v} \tag{6.67}$$

$$\vec{u} \cdot \vec{v}^{\times} = \sum_{p=1}^{3} u_p \vec{a}_p \cdot \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} v_j \vec{a}_i \vec{a}_k = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} \vec{a}_k u_i v_j \tag{6.68}$$

$$= \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{kij} \vec{a}_k u_i v_j = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \varepsilon_{ijk} \vec{a}_i u_j v_k = \vec{u} \times \vec{v} \tag{6.69}$$

The skew symmetric form used in the coordinate vector form is consistent with (6.64) as

$$\mathbf{u}^{\times}\mathbf{v} = -\mathbf{v}^{\times}\mathbf{u} = \left(\mathbf{v}^{\times}\right)^{T} \mathbf{u} = \left(\mathbf{u}^{T}\mathbf{v}^{\times}\right)^{T} \tag{6.70}$$

This shows that the matrix representation of cross product dyadic $\vec{u}^{\times}$ *is the skew symmetric form* $\mathbf{u}^{\times}$*.*

Example 86 *The dyadic form of the triple cross product (6.22) is*

$$\vec{a}^{\times} \cdot \vec{b}^{\times} \cdot \vec{c} = \left[\vec{b}\vec{a} - (\vec{a} \cdot \vec{b}\vec{I})\right] \cdot \vec{c} \tag{6.71}$$

and it follows that

$$\vec{a}^{\times} \cdot \vec{b}^{\times} = \vec{b}\vec{a} - \vec{a} \cdot \vec{b}\vec{I} \tag{6.72}$$

In particular, we note that

$$\vec{a}^{\times} \cdot \vec{a}^{\times} = \vec{a}\vec{a} - \vec{a} \cdot \vec{a}\vec{I} \tag{6.73}$$

Example 87 *The triple scalar product satisfies*

$$\left(\vec{d} \times \vec{a}\right) \cdot \vec{w} = \vec{d} \cdot (\vec{a} \times \vec{w})$$

for any vectors $\vec{d}$*,* $\vec{a}$ *and* $\vec{w}$*, and it follows from (6.71) that*

$$\left(\vec{d} \times \vec{a}\right) \cdot \left(\vec{b} \times \vec{c}\right) = \vec{d} \cdot \left(\vec{a} \times \left(\vec{b} \times \vec{c}\right)\right) = \vec{d} \cdot \left(\vec{a} \cdot \vec{c}\vec{I} - \vec{c}\vec{a}\right) \cdot \vec{b} \tag{6.74}$$

The same result could have been obtained using

$$\left(\vec{d} \times \vec{a}\right) \cdot \left(\vec{b} \times \vec{c}\right) = -\left(\vec{d} \times \vec{a}\right) \cdot \left(\vec{c} \times \vec{b}\right) = -\vec{d} \cdot \vec{a}^{\times} \cdot \vec{c}^{\times} \cdot \vec{b} = -\vec{d} \cdot \left(\vec{a}^{\times} \cdot \vec{c}^{\times}\right) \cdot \vec{b} \tag{6.75}$$

In the development of the kinetic energy for a rigid body the following special case will be used:

$$(\vec{\omega} \times \vec{r}) \cdot (\vec{\omega} \times \vec{r}) = \vec{\omega} \cdot \left(\vec{r} \cdot \vec{r}\vec{I} - \vec{r}\vec{r}\right) \cdot \vec{\omega} = -\vec{\omega} \cdot \left(\vec{r}^{\times} \cdot \vec{r}^{\times}\right) \cdot \vec{\omega} \tag{6.76}$$

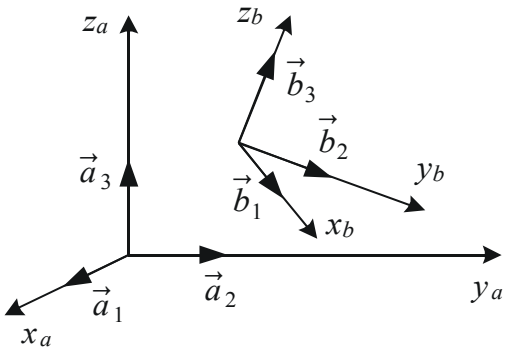

Figure 6.3: Frames a and b.

6.4 The rotation matrix

6.4.1 Coordinate transformations for vectors

It was shown that a vector can be described in terms of its component in a coordinate frame a with orthogonal unit vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$. Dynamic models for use in robotics, car dynamics, aerospace, marine systems, and navigation typically involve several Cartesian frames, so that a vector may have to be described in more than one frame. To investigate this we introduce a second coordinate frame b with orthogonal unit vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$ along the axes. A vector $\vec{v}$ may then be represented with respect to any of the systems a and b. We use the notation

$$\vec{v} = \sum_{i=1}^{3} v_i^a \vec{a}_i \quad \text{and} \quad \vec{v} = \sum_{i=1}^{3} v_i^b \vec{b}_i \tag{6.77}$$

where

$$v_i^a = \vec{v} \cdot \vec{a}_i \tag{6.78}$$

are the coordinates of $\vec{v}$ in a, and

$$v_i^b = \vec{v} \cdot \vec{b}_i \tag{6.79}$$

are the coordinates of $\vec{v}$ in b. To distinguish the column vectors of coordinates in frame a from the column vector of coordinates in frame b we write

$$\mathbf{v}^a = \begin{pmatrix} v_1^a \\ v_2^a \\ v_3^a \end{pmatrix} \quad \text{and} \quad \mathbf{v}^b = \begin{pmatrix} v_1^b \\ v_2^b \\ v_3^b \end{pmatrix} \tag{6.80}$$

where superscript a denotes that the vector is given by the the coordinates in a, and the superscript b denotes that the vector is given by the coordinates in b.

To find the relation between the coordinate vectors $\mathbf{v}^a$ and $\mathbf{v}^b$ in frames a and b the following calculation is used:

$$\begin{aligned} v_i^a &= \vec{v} \cdot \vec{a}_i = (v_1^b \vec{b}_1 + v_2^b \vec{b}_2 + v_3^b \vec{b}_3) \cdot \vec{a}_i \\ &= \sum_{j=1}^{3} v_j^b \left(\vec{a}_i \cdot \vec{b}_j \right) \end{aligned} \tag{6.81}$$

This leads to the following result:

The coordinate transformation from frame b to frame a is given by

$$\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b \tag{6.82}$$

where

$$\mathbf{R}_b^a = \{\vec{a}_i \cdot \vec{b}_j\} \tag{6.83}$$

is called the *rotation matrix* from a to b. The elements $r_{ij} = \vec{a}_i \cdot \vec{b}_j$ of the rotation matrix $\mathbf{R}_b^a$ are called the *direction cosines.*

We see that the rotation matrix from a to b transforms a coordinate vector in b to a coordinate vector in a. Because of this the matrix may also be called the *coordinate transformation matrix* from b to a.

6.4.2 Properties of the rotation matrix

The rotation matrix has a number of useful properties that will be discussed in this section. First it is noted that the rotation matrix from b to a can be found in the same way as the rotation matrix from a to b by simply interchanging a and b in the expressions. This gives

$$\mathbf{R}_a^b = \{\vec{b}_i \cdot \vec{a}_j\} \tag{6.84}$$

For all $\mathbf{v}^b$ we have

$$\mathbf{v}^b = \mathbf{R}_a^b \mathbf{v}^a = \mathbf{R}_a^b \mathbf{R}_b^a \mathbf{v}^b \tag{6.85}$$

This implies that

$$\mathbf{R}_a^b \mathbf{R}_b^a = \mathbf{I}, \tag{6.86}$$

and it follows that

$$\mathbf{R}_a^b = (\mathbf{R}_b^a)^{-1} \tag{6.87}$$

A comparison of the elements in the matrices in (6.83) and (6.84) leads to the conclusion that $\mathbf{R}_a^b = (\mathbf{R}_b^a)^T$. Combining these results we arrive at the first result:

The rotation matrix is orthogonal and satisfies

$$\mathbf{R}_a^b = (\mathbf{R}_b^a)^{-1} = (\mathbf{R}_b^a)^T \tag{6.88}$$

Consider a vector $\vec{p}$ with coordinate vector $\mathbf{p}^a$ in frame a. Define the vector $\vec{q}$ defined by its coordinate vector

$$\mathbf{q}^a = \mathbf{R}_b^a \mathbf{p}^a \tag{6.89}$$

Note that the vector $\vec{q}$ is defined by the vector $\vec{p}$ and the rotation matrix $\mathbf{R}_b^a$. The coordinate vector $\mathbf{q}^b$ in b is according to the usual coordinate transformation rule

$$\mathbf{q}^b = \mathbf{R}_a^b \mathbf{q}^a = \mathbf{R}_a^b \mathbf{R}_b^a \mathbf{p}^a = \mathbf{p}^a \tag{6.90}$$

which means that the coordinates of $\vec{q}$ in b are equal to the coordinates of $\vec{p}$ in a. This is the second result: The rotation matrix from a to b rotates the vector $\vec{p}$ to the vector $\vec{q}$ so that $\mathbf{q}^b = \mathbf{p}^a$.

The rotation matrix $\mathbf{R}_b^a$ from a to b has two interpretations:

1. Let the vector $\vec{v}$ have coordinate vector $\mathbf{v}^b$ in b and coordinate vector $\mathbf{v}^a$ in a. Then the rotation matrix $\mathbf{R}_b^a$ transforms the coordinate vector in b to the coordinate vector in a according to
$$\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b \tag{6.91}$$
In this equation $\mathbf{R}_b^a$ acts as a coordinate transformation matrix.

2. The vector $\vec{p}$ with coordinate vector $\mathbf{p}^a$ in a is rotated to the vector $\vec{q}$ with coordinate vector $\mathbf{q}^b = \mathbf{p}^a$ by
$$\mathbf{q}^a = \mathbf{R}_b^a \mathbf{p}^a \tag{6.92}$$
In this equation $\mathbf{R}_b^a$ acts as a rotation matrix.

As a special case of this the rotation matrix rotates the orthogonal unit vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3$ in a to the orthogonal unit vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$ in b which is seen from

$$\mathbf{a}_1^a = \mathbf{b}_1^b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2^a = \mathbf{b}_2^b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{a}_3^a = \mathbf{b}_3^b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{6.93}$$

Moreover, from $\mathbf{b}_i^a = \mathbf{R}_b^a \mathbf{a}_i^a$ it follows that the columns of the rotation matrix are the coordinate vectors $\mathbf{b}_i^a$ of $\vec{b}_i$ in frame a, that is

$$\mathbf{R}_b^a = \begin{pmatrix} \mathbf{b}_1^a & \mathbf{b}_2^a & \mathbf{b}_3^a \end{pmatrix} \tag{6.94}$$

which is the third result.

The determinant of the rotation matrix $\mathbf{R}_b^a$ is found by direct calculation to be

$$\begin{aligned} \det \mathbf{R}_b^a &= r_{11}(r_{22}r_{33} - r_{32}r_{23}) + r_{21}(r_{32}r_{13} - r_{12}r_{33}) + r_{31}(r_{12}r_{23} - r_{22}r_{13}) \\ &= (\mathbf{b}_1^a)^T \left[(\mathbf{b}_2^a)^\times \mathbf{b}_3^a \right] = (\mathbf{b}_1^a)^T \mathbf{b}_1^a = 1 \end{aligned}$$

where it is used that $(\mathbf{b}_2^a)^\times \mathbf{b}_3^a = \mathbf{b}_1^a$, and that $\mathbf{b}_1^a$ is a unit vector. We have then shown the fourth result: The rotation matrix has a determinant equal to unity, that is

$$\det \mathbf{R}_b^a = 1 \tag{6.95}$$

Finally, the set $SO(3)$ is defined. We have established that the rotation matrix is orthogonal and has a determinant equal to unity. The set of all matrices that are orthogonal and with a determinant equal to unity is denoted by $SO(3)$, that is,

$$SO(3) = \{\mathbf{R} | \mathbf{R} \in R^{3\times 3}, \quad \mathbf{R}^T\mathbf{R} = \mathbf{I} \quad \text{and} \quad \det \mathbf{R} = 1\} \tag{6.96}$$

Here $R^{3\times 3}$ is the set of all 3×3 matrices with real elements. A matrix $\mathbf{R}$ is a rotation matrix if and only if it is an element of the set $SO(3)$.

6.4.3 Composite rotations

The rotation from frame a to a frame c may be described a a *composite rotation* made up by a rotation from a to b, and then a rotation from b to c. The transformation of $\mathbf{v}^c$

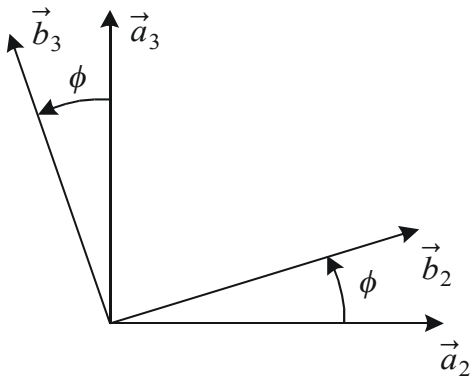

Figure 6.4: A rotation by an angle ϕ around $\vec{a}_1$.

to b and to a is given by

$$\begin{aligned} \mathbf{v}^b &= \mathbf{R}^b_c \mathbf{v}^c \\ \mathbf{v}^a &= \mathbf{R}^a_c \mathbf{v}^c \end{aligned}$$

Combining these two equations we get

$$\mathbf{v}^a = \mathbf{R}^a_b \mathbf{v}^b = \mathbf{R}^a_b \mathbf{R}^b_c \mathbf{v}^c$$

This shows that:

The rotation matrix of a composite rotation is the product of the rotation matrices:

$$\mathbf{R}^a_c = \mathbf{R}^a_b \mathbf{R}^b_c$$

This shows that the rotation matrix for the composite rotation $\mathbf{R}^a_c$ is simply the product of the rotation matrices $\mathbf{R}^a_b$ from a to b and $\mathbf{R}^b_c$ from b to c. It is straightforward to extend this result to the composite rotation of three or more rotations. In the case of three rotations we have

$$\mathbf{R}^a_d = \mathbf{R}^a_b \mathbf{R}^b_c \mathbf{R}^c_d \tag{6.97}$$

6.4.4 Simple rotations

A rotation about a fixed axis is called a *simple rotation*. We will here derive the rotation matrices corresponding to simple rotations about the x, y and z axes. Consider a rotation by an angle ϕ about the x_a axis from a frame a to a frame b. The resulting rotation matrix is denoted $\mathbf{R}_x(\phi)$. In the same way we define $\mathbf{R}_y(\theta)$ to be the rotation by an angle θ about the y axis, and $\mathbf{R}_z(\psi)$ to be the rotation by an angel ψ about the z axis.

For the rotation $\mathbf{R}_x(\phi)$ we see from Figure 6.4 that $\vec{a}_1 = \vec{b}_1$, so that $\vec{a}_1 \cdot \vec{b}_1 = 1$, while

$$\vec{a}_1 \cdot \vec{b}_2 = \vec{a}_1 \cdot \vec{b}_3 = \vec{a}_2 \cdot \vec{b}_1 = \vec{a}_3 \cdot \vec{b}_1 = 0 \tag{6.98}$$

$$\vec{a}_2 \cdot \vec{b}_2 = \cos\phi, \quad \vec{a}_3 \cdot \vec{b}_3 = \cos\phi \tag{6.99}$$

$$\vec{a}_3 \cdot \vec{b}_2 = \sin\phi, \quad \vec{a}_2 \cdot \vec{b}_3 = -\sin\phi \tag{6.100}$$

In the same way we can find the elements of the matrices $\mathbf{R}_y(\theta)$ and $\mathbf{R}_z(\psi)$. This results in

$$\mathbf{R}_x(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \tag{6.101}$$

$$\mathbf{R}_y(\theta) = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \tag{6.102}$$

$$\mathbf{R}_z(\psi) = \begin{pmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6.103}$$

6.4.5 Coordinate transformations for dyadics

A coordinate vector can be transformed from a frame a to a frame b through multiplication with the rotation matrix. An important property of dyadics is that the matrix representation transform from frame a to frame b with a similarity transformation using the rotation matrix. The dyadic $\vec{D}$ can be expressed in frames a and b by

$$\vec{D} = \sum_{i=1}^{3}\sum_{j=1}^{3} d_{ij}^a \vec{a}_i \vec{a}_j = \sum_{p=1}^{3}\sum_{q=1}^{3} d_{pq}^b \vec{b}_p \vec{b}_q \tag{6.104}$$

where d_{ij}^a are the components in frame a and d_{pq}^b are the components in frame b. The matrix representation in the two frames are denoted

$$\mathbf{D}^a = \{d_{ij}^a\}, \quad \mathbf{D}^b = \{d_{ij}^b\} \tag{6.105}$$

Let the vector $\vec{z}$ be given by

$$\vec{z} = \vec{D} \cdot \vec{u} \tag{6.106}$$

Then, in frames a and b this may be written in matrix form as

$$\mathbf{z}^a = \mathbf{D}^a \mathbf{u}^a, \quad \mathbf{z}^b = \mathbf{D}^b \mathbf{u}^b \tag{6.107}$$

We then find that

$$\mathbf{D}^a \mathbf{u}^a = \mathbf{z}^a = \mathbf{R}_b^a \mathbf{z}^b = \mathbf{R}_b^a \mathbf{D}^b \mathbf{u}^b = \mathbf{R}_b^a \mathbf{D}^b \mathbf{R}_a^b \mathbf{u}^a \tag{6.108}$$

and, since $\mathbf{u}^a$ is arbitrary, this implies that

The matrix representation of a dyadic transforms by a similarity transform with the rotation matrix according to

$$\mathbf{D}^a = \mathbf{R}_b^a \mathbf{D}^b \mathbf{R}_a^b \tag{6.109}$$

Example 88 *In rigid body dynamics a frame b with orthogonal unit vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$ is fixed in the rigid body. Then the inertia dyadic can be written*

$$\vec{M} = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}^b \vec{b}_i \vec{b}_j \tag{6.110}$$

and the corresponding matrix representation is

$$\mathbf{M}^b = \left\{ m_{ij}^b \right\} \tag{6.111}$$

An important result in rigid body dynamics is that when frame b is fixed in the rigid body, and therefore moves with the rigid boy, then $\mathbf{M}^b$ *is a constant matrix. In contrast to this, the matrix representation* $\mathbf{M}^a$ *in a stationary coordinate frame a will be given by*

$$\mathbf{M}^a = \mathbf{R}_b^a \mathbf{M}^b \mathbf{R}_a^b \tag{6.112}$$

Example 89 *The relation between the skew symmetric forms of a vector is given by*

$$\left(\mathbf{u}^b\right)^\times \mathbf{v}^b = \mathbf{w}^b = \mathbf{R}_a^b \mathbf{w}^a = \mathbf{R}_a^b \left(\mathbf{u}^a\right)^\times \mathbf{v}^a = \mathbf{R}_a^b \left(\mathbf{u}^a\right)^\times \mathbf{R}_b^a \mathbf{v}^b \tag{6.113}$$

which implies that

$$\left(\mathbf{u}^b\right)^\times = \mathbf{R}_a^b \left(\mathbf{u}^a\right)^\times \mathbf{R}_b^a \tag{6.114}$$

that is, the skew symmetric form of the vector $\mathbf{u}^a$ *transforms to frame b by a similarity transformation. This is a consequence of the fact that* $\left(\mathbf{u}^a\right)^\times$ *is the matrix representation of the dyadic* $\vec{u}^\times$*.*

6.4.6 Homogeneous transformation matrices

By now we are familiar with the notion of a rotation matrix which specifies the orientation of a coordinate frame with respect to some other frame. To extend our set of mathematical tools we introduce the concept of a *homogeneous transformation matrix* which is a matrix that describes the position and orientation of a coordinate frame with respect to a reference frame. To be precise we consider a frame a and a frame b, and let $\mathbf{R}_b^a$ be the rotation matrix from a to b, while $\mathbf{r}_{ab}^a$ is the position in a coordinates of the origin of frame b relative to the origin of frame a.

The position and orientation of frame b relative to frame a is given by the homogeneous transformation matrix

$$\mathbf{T}_b^a = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \in SE(3) \tag{6.115}$$

Here the set $SE(3)$ is the Special Euclidean Group of dimension 3 defined by

$$SE(3) = \left\{ \mathbf{T} \mid \mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{r} \\ \mathbf{0}^T & 1 \end{pmatrix},\ \mathbf{R} \in SO(3),\ \mathbf{r} \in R^3 \right\} \tag{6.116}$$

The inverse of $\mathbf{T}_b^a$ is found by matrix inversion to be

$$\left(\mathbf{T}_b^a\right)^{-1} = \begin{pmatrix} \left(\mathbf{R}_b^a\right)^T & -\left(\mathbf{R}_b^a\right)^T \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_a^b & \mathbf{r}_{ba}^b \\ \mathbf{0}^T & 1 \end{pmatrix} = \mathbf{T}_a^b \tag{6.117}$$

This means that

$$\left(\mathbf{T}_b^a\right)^{-1} = \mathbf{T}_a^b \tag{6.118}$$

Composite homogenous transformation matrices give

$$\begin{aligned}\mathbf{T}_b^a\mathbf{T}_c^b &= \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix}\begin{pmatrix} \mathbf{R}_c^b & \mathbf{r}_{bc}^b \\ \mathbf{0}^T & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{R}_b^a\mathbf{R}_c^b & \mathbf{r}_{ab}^a + \mathbf{R}_b^a\mathbf{r}_{bc}^b \\ \mathbf{0}^T & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{R}_c^a & \mathbf{r}_{ac}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \\ &= \mathbf{T}_c^a \end{aligned} \tag{6.119}$$

We conclude that

$$\mathbf{T}_c^a = \mathbf{T}_b^a\mathbf{T}_c^b \tag{6.120}$$

Example 90 *In the description of robotic manipulators, a coordinate frame is fixed to each link of the arm. The manipulator is made up by rigid bodies called links that are connected by joints. Each joint is assumed to have one degree of freedom that is either a translation or a rotation. In a typical manipulator design with rotary joints we may think of link 1 as the torso that is connected to the upper arm (link 2) by a shoulder joint with a horizontal axis of rotation. The upper arm is in turn connected to the lower arm (link 3) by an elbow joint. The lower arm is connected with the robot hand (joint 6) trough three rotary joints that form the robotic wrist. In general, the base frame* 0 *is fixed to the floor, frame 1 is fixed to the first link, frame 2 to the second link and so on, and for a six link manipulator frame 6 is fixed to the robot hand. The position and orientation of frame* $i+1$ *relative to frame* i *can then be specified in terms of a homogeneous transformation matrix*

$$\mathbf{T}_{i+1}^i = \begin{pmatrix} \mathbf{R}_{i+1}^i & \mathbf{r}_{i,i+1}^i \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{6.121}$$

and the position and orientation of the hand is given by

$$\mathbf{T}_6^0 = \begin{pmatrix} \mathbf{R}_6^0 & \mathbf{r}_{06}^0 \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{6.122}$$

which is computed from

$$\mathbf{T}_6^0 = \mathbf{T}_1^0\mathbf{T}_2^1\ldots\mathbf{T}_6^5. \tag{6.123}$$

In the Denavit-Hartenberg convention the transformation from frame i *to frame* $i+1$ *is given by the Denavit-Hartenberg parameters* $\alpha_i, a_i, d_i, \theta_i$ *according to*

$$\begin{aligned}\mathbf{T}_{i+1}^i &= \begin{pmatrix} \mathbf{R}_z(\theta_i) & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix}\begin{pmatrix} \mathbf{I} & a_i\mathbf{e}_1 + d_i\mathbf{e}_3 \\ \mathbf{0}^T & 1 \end{pmatrix}\begin{pmatrix} \mathbf{R}_x(\alpha_i) & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{R}_z(\theta_i)\mathbf{R}_x(\alpha_i) & a_i\mathbf{R}_z(\theta_i)\mathbf{e}_1 + d_i\mathbf{e}_3 \\ \mathbf{0}^T & 1 \end{pmatrix} \end{aligned} \tag{6.124}$$

where $\mathbf{e}_1 = (1,0,0)^T$ *and* $\mathbf{e}_3 = (0,0,1)^T$*. The joint variable is* d_i *for translational joints and* θ_i *for rotational joints.*

6.5 Euler angles

6.5.1 Introduction

A rotation matrix describes the orientation of a frame b with respect to a frame a. The rotation matrix is a 3×3 matrix with nine elements. The orthogonality of the

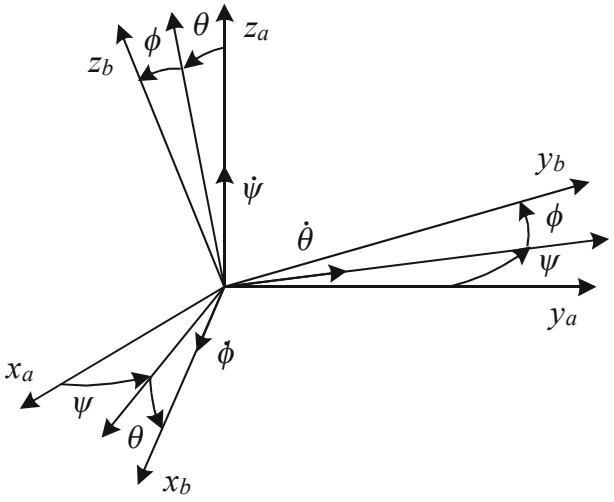

Figure 6.5: Roll-pitch-yaw Euler angles.

matrix gives six constraints on the elements of the matrix, so that there are only three independent parameters that describes the rotation matrix. Therefore, it is of great interest to investigate if it is possible to find three parameters that give a parameterization of the rotation matrix.

A widely used set of parameters for the rotation matrix is the Euler angles. In this description the rotation matrix is given as a composite rotation of selected combinations of rotations about the x, y and z axes. There are many possible permutations of x, y and z rotations, and a description of this is given in (Kane et al. 1983). Here we will present the two sets of Euler angles that are the most often seen, namely the roll-pitch-yaw angles, and the classical Euler angles.

6.5.2 Roll-pitch-yaw

The Euler angles of the roll-pitch yaw type are commonly used to describe the motion of rigid bodies that move freely, like aeroplanes, spacecraft, ships and underwater vehicles. The rotation from a to b is described as a rotation ψ about the z_a axis, then a rotation θ about the current (rotated) y axis, and finally a rotation ϕ about the current (rotated) x axis as shown in Figure 6.5. The resulting rotation matrix is

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_x(\phi) \tag{6.125}$$

This formulation is very useful as it makes it possible to describe the rotation of e.g. an airplane as a sequence of a roll rotation about the longitudinal axis of the plane, then a pitch rotation about a lateral axis of the plane, and finally a yaw rotation about the vertical axis of the plane. Obviously, it is easier to interpret this sequence of simple rotations angle than a rotation matrix.

To derive and remember the expression (6.125) it is convenient to use the rotation matrix interpretation of the simple rotations $\mathbf{R}_z(\psi)$, $\mathbf{R}_y(\theta)$ and $\mathbf{R}_x(\phi)$. In this interpretation the rotation from a to b is in the same sequence as the matrices are written in (6.125), namely, first $\mathbf{R}_z(\psi)$, then $\mathbf{R}_y(\theta)$, and finally $\mathbf{R}_x(\phi)$. Note that in a coordinate

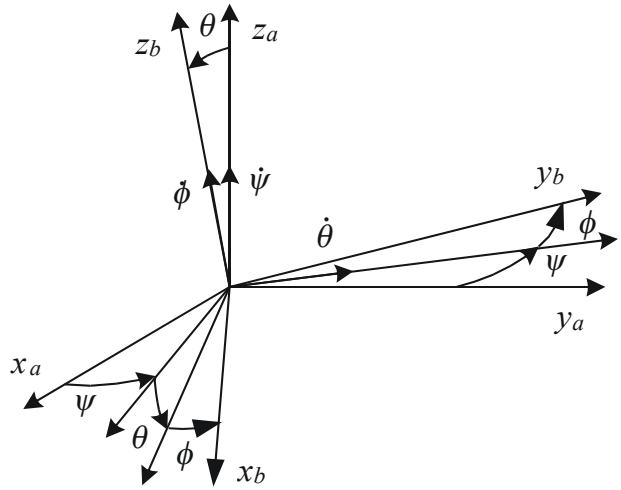

Figure 6.6: Classical Euler angles.

transformation interpretation the matrix $\mathbf{R}_b^a$ transforms a vector $\mathbf{v}^b$ to a vector $\mathbf{v}^a$ according to $\mathbf{v}^a = \mathbf{R}_b^a \mathbf{v}^b = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_x(\phi)\mathbf{v}^b$. Then the vector $\mathbf{v}^b$ is first transformed by $\mathbf{R}_x(\phi)$, then by $\mathbf{R}_y(\theta)$ and finally by $\mathbf{R}_z(\psi)$.

6.5.3 Classical Euler angles

The classical Euler angles are used to describe the rotation of rigid bodies that are connected to a fixed base by three joints. Typically this involves robotic wrist joints, platforms stabilized by gyroscopes in inertial navigation, and pointing devices. In this description the rotation consists of a rotation ψ about the z_a axis, then a rotation θ about the current (rotated) y axis, and finally a rotation ϕ about the current (rotated) z axis as shown in Figure 6.6. The resulting rotation matrix is

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_z(\phi) \tag{6.126}$$

6.6 Angle-axis description of rotation

6.6.1 Introduction

In the previous section it was shown that the rotation matrix can be represented by Euler angles, which are very useful in some applications. In particular this is the case for a robotic wrist joint where the hand is connected to the arm through three rotational joints. Also in ship dynamics it is convenient to describe the rotation of the ship in terms of the Euler angles roll, pitch and yaw. Likewise airplane dynamics rely on a characterization based on the roll angle, the pitch angle and the sideslip angle, which are the Euler angles from the wind frame to the airplane frame. The motivation for this is that the forces acting on the plane are functions of these Euler angles. However, in many other applications involving rotation there is no clear physical motivation for introducing Euler angles. The use of Euler angles in the equations of motion may then introduce complicated expressions with inherent singularities. There are alternative descriptions

of rotation that avoid these problems, and that are well suited for simulation as well as for controller design and analysis. On background of this it may be argued that Euler angles have been over-emphasized in the dynamics literature. In the following we will study the angle-axis parametrization of the rotation, which is a very useful tool in the development of kinematic models and equations of motion for use in control systems.

6.6.2 Angle-axis parameters

A rotation matrix $\mathbf{R}_b^a$ is orthogonal with determinant equal to unity. It can be shown (Angeles 1988), (McCarthy 2000) that this implies that one of the eigenvalues to the matrix is equal to one, and that the corresponding unit eigenvector $\mathbf{k}$ satisfies

$$\mathbf{R}_b^a \mathbf{k} = \mathbf{k} \tag{6.127}$$

This purely algebraic result can be given a geometric interpretation which is the basis for the *angle-axis parameterization* of the rotation matrix $\mathbf{R}_b^a$. The geometric interpretation that will be used is that the eigenvector $\mathbf{k}$ is the coordinate vector of a unit vector $\vec{k}$, where $\vec{k}$ is defined by its coordinate vector

$$\mathbf{k}^a = \mathbf{k} \tag{6.128}$$

in frame a. The transformation rule

$$\mathbf{k}^a = \mathbf{R}_b^a \mathbf{k}^b \tag{6.129}$$

then implies that

$$\mathbf{k}^a = \mathbf{k}^b = \mathbf{k} \tag{6.130}$$

which means that $\vec{k}$ has the same coordinates in frames a and b. It is therefore possible to describe the rotation from a to b as a simple rotation by an angle θ about the vector $\vec{k}$ which is fixed in both a and b. On background of this $(\theta, \vec{k})$ is called the *angle-axis parameterization* of the rotation matrix $\mathbf{R}_b^a$. Note that this gives four parameters and one constraint equation, namely the angle θ plus the three coordinates of the unit vector $\vec{k}$, and the constraint equation $\vec{k} \cdot \vec{k} = 1$.

6.6.3 Derivation of rotation dyadic

We will here derive the expression for the rotation matrix given by the angle θ and the vector $\vec{k}$. The derivation is taken from (Kane et al. 1983). It was shown in Section 6.4.2 that the rotation matrix $\mathbf{R}_b^a$ rotates a vector $\vec{p}$ in a to a vector $\vec{q}$ in b so that the coordinates of $\vec{p}$ in a are equal to the coordinates of $\vec{q}$ in b. This result is used to find an expression for the rotation matrix $\mathbf{R}_b^a$ in terms of θ and $\vec{k}$. We will do this by deriving an expression where $\vec{q}$ is given by $\vec{p}$, θ and $\vec{k}$. To simplify the derivation two additional frames c and d are used where $\mathbf{R}_c^a = \mathbf{R}_d^b$.

Consider two frames c and d that initially coincide. Let $\vec{c}_1, \vec{c}_2, \vec{c}_3$ be the orthogonal unit vectors in c, and let $\vec{d}_1, \vec{d}_2, \vec{d}_3$ be the orthogonal unit vectors in d. The frames are selected so that $\vec{c}_3 = \vec{d}_3 = \vec{k}$. Frame d is obtained by rotating frame c by an angle θ about $\vec{k}$ as shown in Figure 6.7. Let the vector $\vec{p}$ be a fixed vector in frame c, and let the vector $\vec{q}$ be a fixed vector in frame d so that $\vec{p}$ and $\vec{q}$ coincide before the rotation. Then it is possible to express $\vec{q}$ after the rotation by $\vec{p}$, θ and $\vec{k}$.

The vectors $\vec{p}$ and $\vec{q}$ can be written

$$\vec{p} = x\vec{c}_1 + y\vec{c}_2 + z\vec{k} \tag{6.131}$$

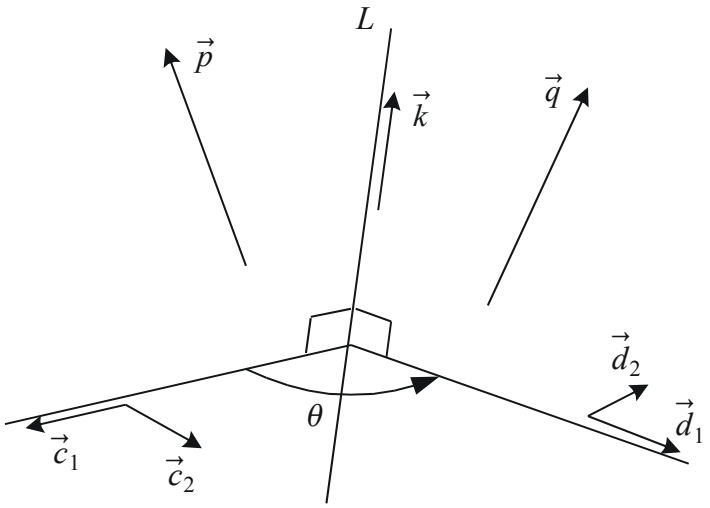

Figure 6.7: Rotation of the vector $\vec{p}$ by an angle θ around the $\vec{k}$ vector.

and

$$\vec{q} = x\vec{d_1} + y\vec{d_2} + z\vec{k} \tag{6.132}$$

The unit vectors $\vec{d_1}$ and $\vec{d_2}$ can be written

$$\begin{aligned} \vec{d_1} &= \cos\theta\vec{c}_1 + \sin\theta\vec{c}_2 & (6.133)\\ \vec{d_2} &= -\sin\theta\vec{c}_1 + \cos\theta\vec{c}_2 & (6.134) \end{aligned}$$

and insertion into (6.132) gives

$$\vec{q} = (x\cos\theta - y\sin\theta)\vec{c}_1 + (x\sin\theta + y\cos\theta)\vec{c}_2 + z\vec{k} \tag{6.135}$$

Consider the calculation

$$\begin{aligned} & \cos\theta\vec{p} + \sin\theta\vec{k}\times\vec{p} + (1-\cos\theta)\vec{k}\vec{k}\cdot\vec{p} \\ = \;& \cos\theta(x\vec{c}_1 + y\vec{c}_2 + z\vec{k}) - \sin\theta(-x\vec{c}_2 + y\vec{c}_1) + (1-\cos\theta)z\vec{k} \\ = \;& (x\cos\theta - y\sin\theta)\vec{c}_1 + (x\sin\theta + y\cos\theta)\vec{c}_2 + z\vec{k} \qquad (6.136) \end{aligned}$$

where it is used that $\vec{c}_1\times\vec{k} = -\vec{c}_2$, $\vec{c}_2\times\vec{k} = \vec{c}_1$, and that $\vec{p}\cdot\vec{k} = z$. It follows by comparison with (6.135) that

$$\begin{aligned} \vec{q} &= \cos\theta\vec{p} + \sin\theta\vec{k}\times\vec{p} + (1-\cos\theta)\vec{k}\vec{k}\cdot\vec{p} \\ &= \left(\cos\theta\vec{I} + \sin\theta\vec{k}^{\times} + (1-\cos\theta)\vec{k}\vec{k}\right)\cdot\vec{p} \qquad (6.137) \end{aligned}$$

6.6.4 The rotation dyadic

The rotation of the vector $\vec{p}$ to the vector $\vec{q}$ as given by (6.137) can be written in the dyadic form

$$\vec{q} = \vec{R}_{k,\theta}\cdot\vec{p} \tag{6.138}$$

where

$$\vec{R}_{k,\theta} = \cos\theta\vec{I} + \sin\theta\vec{k}^{\times} + (1-\cos\theta)\vec{k}\vec{k} \tag{6.139}$$

is the *rotation dyadic* of the angle-axis description. Now, recall that $\vec{q}$ is obtained by rotating $\vec{p}$ according to $\mathbf{q}^a = \mathbf{R}_a^b\mathbf{p}^a$, which in combination with (6.138) implies that the rotation matrix $\mathbf{R}_a^b$ is the matrix representation of the rotation dyadic $\vec{R}_{k,\theta}$ in a. This leads to the result

The rotation matrix $\mathbf{R}_b^a$ can be described as a rotation by an angle θ about a unit vector $\vec{k}$ where $\mathbf{R}_b^a$ is given by

$$\mathbf{R}_b^a = \cos\theta\mathbf{I} + \sin\theta(\mathbf{k}^a)^{\times} + (1-\cos\theta)\mathbf{k}^a(\mathbf{k}^a)^T \tag{6.140}$$

This is the angle-axis parameterization of the rotation matrix.

Using the standard transformation rule and the identities $(\mathbf{k}^a)^{\times}\mathbf{k}^a = \mathbf{0}$ and $(\mathbf{k}^a)^T\mathbf{k}^a = 1$ gives

$$\mathbf{k}^a = \mathbf{R}_b^a\mathbf{k}^b = \mathbf{k}^b \tag{6.141}$$

which shows that the rotation vector $\vec{k}$ has the same coordinates in a and b.

Example 91 *Inserting* $\mathbf{k}^a = (k_x \;\; k_y \;\; k_z)^T$ *we get*

$$\mathbf{R}_b^a = \begin{pmatrix} k_x^2v_\theta + c_\theta & k_xk_yv_\theta - k_zs_\theta & k_xk_zv_\theta + k_ys_\theta \\ k_xk_yv_\theta + k_zs_\theta & k_y^2v_\theta + c_\theta & k_yk_zv_\theta - k_xs_\theta \\ k_xk_zv_\theta - k_ys_\theta & k_yk_zv_\theta + k_xs_\theta & k_z^2v_\theta + c_\theta \end{pmatrix} \tag{6.142}$$

where the notation $s_\theta = \sin\theta$, $c_\theta = \cos\theta$ *and* $v_\theta = 1-c_\theta$ *is used to simplify the expression.*

Example 92 *Suppose that the rotation axis is given by* $\vec{k} = \vec{a}_3$, *which means that* $\mathbf{k}^a = (0,0,1)^T$. *Then the matrix representation of* $\vec{R}_{k,\theta}$ *in* a *is*

$$\mathbf{R}_b^a = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{R}_{z,\theta} \tag{6.143}$$

which is to be expected as this is a rotation by an angle θ *about the* z *axis.*

Example 93 *As* $\vec{R}_{k,\theta}$ *is a dyadic and* $\vec{k} = \vec{c}_3$, *it follows from the transformation rule (6.109) that*

$$\mathbf{R}_b^a = \mathbf{R}_c^a\mathbf{R}_{z,\theta}\mathbf{R}_a^c \tag{6.144}$$

6.6.5 Rotation matrix

We use the notation $\mathbf{R}_{k,\theta} = \mathbf{R}_b^a$ and $\mathbf{k} = \mathbf{k}^a$ so that the rotation matrix is written

$$\mathbf{R}_{k,\theta} := \cos\theta\mathbf{I} + \mathbf{k}^{\times}\sin\theta + \mathbf{k}\mathbf{k}^T(1-\cos\theta) \tag{6.145}$$

An alternative expression is found by inserting the identity

$$\mathbf{k}^{\times}\mathbf{k}^{\times} = \mathbf{k}\mathbf{k}^T - \mathbf{k}^T\mathbf{k}\mathbf{I} = \mathbf{k}\mathbf{k}^T - \mathbf{I}. \tag{6.146}$$

which gives

$$\mathbf{R}_{k,\theta} = \mathbf{I} + \mathbf{k}^{\times}\sin\theta + \mathbf{k}^{\times}\mathbf{k}^{\times}(1-\cos\theta) \tag{6.147}$$

The inverse to the rotation matrix is

$$(\mathbf{R}_b^a)^T = \mathbf{R}_a^b = \mathbf{R}_{k,-\theta}. \tag{6.148}$$

Example 94 *The derivative of* $\mathbf{R}_{k,\theta}$ *with respect to* θ *is found from (6.147) to be*

$$\frac{d\mathbf{R}_{k,\theta}}{d\theta} = \mathbf{k}^\times \cos\theta + \mathbf{k}^\times\mathbf{k}^\times \sin\theta \tag{6.149}$$

Using (6.27) we find that

$$\mathbf{k}^\times\mathbf{R}_{k,\theta} = \mathbf{k}^\times + \mathbf{k}^\times\mathbf{k}^\times \sin\theta - \mathbf{k}^\times(1-\cos\theta) \tag{6.150}$$

and we may conclude that

$$\frac{d}{d\theta}\mathbf{R}_{k,\theta} = \mathbf{k}^\times\mathbf{R}_{k,\theta} \tag{6.151}$$

Example 95 *It is known that the differential equation* $\frac{d}{dt}\mathbf{x} = \mathbf{A}\mathbf{x}$ *has the solution* $\mathbf{x}(t) = \mathbf{x}(0)\exp(\mathbf{A}t)$ *when* $\mathbf{A}$ *is a constant matrix. When* $\mathbf{k}$ *is a constant vector the solution of (6.151) is found in the same way to be*

$$\mathbf{R}_{k,\theta} = \exp[\mathbf{k}^\times\theta] \tag{6.152}$$

as $\mathbf{R}_{k,\theta}(\theta = 0) = \mathbf{I}$.

Example 96 *The matrix exponential* $\exp(A)$ *for a quadratic matrix* $\mathbf{A}$ *is defined by*

$$\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 \ldots \tag{6.153}$$

The result (6.152) can be derived directly from (6.147). First we use (6.27) to establish the identity

$$(\mathbf{k}^\times)^{2n+1} = (-1)^n\mathbf{k}^\times \tag{6.154}$$

by induction. This is done by noting that (6.27) implies that (6.154) is true for $n = 1$*, and moreover, for* $n = 1, 2, \ldots$ *we have*

$$(\mathbf{k}^\times)^{2n+1} = (-1)^n\mathbf{k}^\times \quad \Rightarrow \quad (\mathbf{k}^\times)^{2(n+1)+1} = (-1)^n\mathbf{k}^\times(\mathbf{k}^\times)^2 = (-1)^n(-1)\mathbf{k}^\times. \tag{6.155}$$

Then we may evaluate $\exp[\mathbf{k}^\times\theta]$ *directly from the definition (6.153), and find that*

$$\begin{aligned}
\exp[\mathbf{k}^\times\theta] &= \mathbf{I} + \mathbf{k}^\times\theta + (\mathbf{k}^\times)^2\frac{\theta^2}{2!} + (\mathbf{k}^\times)^3\frac{\theta^3}{3!} + (\mathbf{k}^\times)^4\frac{\theta^4}{4!} + (\mathbf{k}^\times)^5\frac{\theta^5}{5!} + (\mathbf{k}^\times)^6\frac{\theta^6}{6!}\ldots \\
&= \mathbf{I} + \mathbf{k}^\times[\theta + \mathbf{k}^\times\frac{\theta^2}{2!}] + (\mathbf{k}^\times)^3[\frac{\theta^3}{3!} + \mathbf{k}^\times\frac{\theta^4}{4!}] + (\mathbf{k}^\times)^5[\frac{\theta^5}{5!} + \mathbf{k}^\times\frac{\theta^6}{6!}]\ldots \\
&= \mathbf{I} + \mathbf{k}^\times[\theta + \mathbf{k}^\times\frac{\theta^2}{2!}] - \mathbf{k}^\times[\frac{\theta^3}{3!} + \mathbf{k}^\times\frac{\theta^4}{4!}] + \mathbf{k}^\times[\frac{\theta^5}{5!} + \mathbf{k}^\times\frac{\theta^6}{6!}]\ldots \\
&= \mathbf{I} + \mathbf{k}^\times[\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!}\ldots] - (\mathbf{k})^2[\frac{\theta^2}{2!} - \frac{\theta^4}{4!} + \frac{\theta^6}{6!}\ldots] \\
&= \mathbf{I} + \mathbf{k}^\times\sin\theta + \left(\mathbf{k}^\times\right)^2(1-\cos\theta)
\end{aligned} \tag{6.156}$$

This shows in view of (6.147) that

$$\mathbf{R}_{k,\theta} = \exp[\mathbf{k}^\times\theta] \tag{6.157}$$

which is in agreement with the previously derived result (6.152).

6.7 Euler parameters

6.7.1 Definition

The Euler parameters were introduced by Euler in 1770, and are essentially the same as the unit quaternions that were devised by Hamilton in on October 16, 1843, and which involved the definition of a complex number with one real part and three imaginary parts. The Euler parameters have no singularities, and give rational expressions for the rotation matrix as opposed to the angle/axis parameters, which lead to trigonometric terms in the expressions for the rotation matrix. The Euler parameters are of particular use in the numerical simulation of rotation, and in stability analysis of attitude control systems.

The Euler parameters are defined in terms of the angle-axis parameters θ and $\vec{k}$, and are given by the scalar η and the vector $\vec{\epsilon}$ defined by

$$\eta = \cos\frac{\theta}{2}, \qquad \vec{\epsilon} = \vec{k}\sin\frac{\theta}{2} \tag{6.158}$$

In coordinate form this is written

$$\eta = \cos\frac{\theta}{2}, \qquad \boldsymbol{\epsilon} = \mathbf{k}\sin\frac{\theta}{2} \tag{6.159}$$

We note that

$$\eta^2 + \vec{\epsilon}\cdot\vec{\epsilon} = \eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \cos^2\frac{\theta}{2} + \sin^2\frac{\theta}{2} = 1 \tag{6.160}$$

Insertion of the trigonometric identities

$$\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} \tag{6.161}$$

$$\cos\theta = \cos^2\frac{\theta}{2} - \sin^2\frac{\theta}{2} = 2\cos^2\frac{\theta}{2} - 1 = 1 - 2\sin^2\frac{\theta}{2} \tag{6.162}$$

into (6.145) makes it possible to express the rotation matrix $\mathbf{R}_{k,\theta}$ in terms of the Euler parameters $\mathbf{R}_{k,\theta} = \mathbf{R}_e(\eta,\boldsymbol{\epsilon})$, where

$$\mathbf{R}_e(\eta,\boldsymbol{\epsilon}) = (\eta^2 - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon})\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \tag{6.163}$$

$$= (2\eta^2 - 1)\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \tag{6.164}$$

$$= (1 - 2\boldsymbol{\epsilon}^T\boldsymbol{\epsilon})\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \tag{6.165}$$

Here it is used that $\eta^2 = \cos^2\frac{\theta}{2}$, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sin^2\frac{\theta}{2}$ and $\sin\frac{\theta}{2}\cos\frac{\theta}{2}\mathbf{k}^\times = \eta\boldsymbol{\epsilon}^\times$. From (6.165) and (6.146) an alternative form of the rotation matrix is found.

The rotation matrix is given by the corresponding Euler parameters according to

$$\mathbf{R}_e(\eta,\boldsymbol{\epsilon}) = \mathbf{I} + 2\eta\boldsymbol{\epsilon}^\times + 2\boldsymbol{\epsilon}^\times\boldsymbol{\epsilon}^\times \tag{6.166}$$

A given rotation will correspond to two sets of Euler parameters (η,ϵ) and $(-\eta,-\epsilon)$ with opposite signs as

$$\mathbf{R}_e(-\eta,-\boldsymbol{\epsilon}) = \mathbf{R}_e(\eta,\boldsymbol{\epsilon}) \tag{6.167}$$

The inverse of $\mathbf{R}_e(\eta,\epsilon)$ given by

$$\mathbf{R}_e(\eta,\boldsymbol{\epsilon})^T = \mathbf{R}_e(\eta,-\boldsymbol{\epsilon}) \tag{6.168}$$

corresponds to the Euler parameters $(\eta,-\epsilon)$.

Example 97 *From (6.165) and (6.160) we find that*

$$\text{Trace}\mathbf{R} = 3(\eta^2 - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}) + 2\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = 4\eta^2 - 1 \tag{6.169}$$

6.7.2 Quaternions

The vector

$$\mathbf{p} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \tag{6.170}$$

of Euler parameters can be treated as a *unit quaternion vector*. This makes it possible to introduce a wealth of techniques and analysis tool from the theory of quaternions. In the following, the necessary background on quaternions will be presented, and this will be specialized to unit quaternions representing a rotation matrix through its Euler parameters.

A *quaternion* is represented by a vector

$$\mathbf{q} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} \tag{6.171}$$

of dimension 4 where α is the scalar part and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ is the vector part.

The quaternion product between two quaternion vectors $\mathbf{q}_1 = (\alpha_1 \quad \boldsymbol{\beta}_1^T)^T$ and $\mathbf{q}_2 = (\alpha_2 \quad \boldsymbol{\beta}_2^T)^T$ in R^4 is defined by

$$\begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \alpha_1\alpha_2 - \boldsymbol{\beta}_1^T\boldsymbol{\beta}_2 \\ \alpha_1\boldsymbol{\beta}_2 + \alpha_2\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^\times\boldsymbol{\beta}_2 \end{pmatrix} \tag{6.172}$$

where $\alpha_1, \alpha_2 \in R$ and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in R^3$.

Example 98 *The commutator of the quaternion product is given by*

$$\begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} - \begin{pmatrix} \alpha_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} \otimes \begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_1 \end{pmatrix} = 2\begin{pmatrix} 0 \\ \boldsymbol{\beta}_1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \boldsymbol{\beta}_2 \end{pmatrix} \tag{6.173}$$

where it is used that $\boldsymbol{\beta}_1^\times\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_2^\times\boldsymbol{\beta}_1$.

Example 99 *Define the matrices*

$$\mathbf{F}(\mathbf{q}) = \begin{pmatrix} \alpha & -\boldsymbol{\beta}^T \\ \boldsymbol{\beta} & \alpha\mathbf{I} + \boldsymbol{\beta}^\times \end{pmatrix} \in R^{4\times 4} \tag{6.174}$$

$$\mathbf{E}(\mathbf{q}) = \begin{pmatrix} \alpha & -\boldsymbol{\beta}^T \\ \boldsymbol{\beta} & \alpha\mathbf{I} - \boldsymbol{\beta}^\times \end{pmatrix} \in R^{4\times 4} \tag{6.175}$$

The matrix $\mathbf{F}(\mathbf{q})$ *represents quaternion pre-multiplication with* $\mathbf{q}$ *in the sense that for any* $\mathbf{u} \in R^4$

$$\mathbf{q} \otimes \mathbf{u} = \mathbf{F}(\mathbf{q})\mathbf{u} \tag{6.176}$$

while $\mathbf{E}(\mathbf{q})$ *represents quaternion post-multiplication with* $\mathbf{q}$ *in the sense that*

$$\mathbf{u} \otimes \mathbf{q} = \mathbf{E}(\mathbf{q})\mathbf{u} \tag{6.177}$$

Example 100 *The concept of quaternions was introduced by Hamilton who got the idea on the 16th of October 1843 while he was walking with his wife to the Royal Irish Academy (der Waerden 1976). In Hamilton's formulation the quaternion was written*

$$\mathsf{q} = \alpha + i\beta_1 + j\beta_2 + k\beta_3 \tag{6.178}$$

where i, j and k are imaginary units satisfying

$$i^2 = j^2 = k^2 = -1 \tag{6.179}$$

$$ij = -ji = k, \quad jk = -kj = i, \quad ki = -ik = j \tag{6.180}$$

It is interesting to see that the quaternion is actually an extension of complex numbers $z = a + ib$ where $i^2 = -1$. The complex numbers form a division algebra, which means that if z_1 and z_2 are complex numbers, then the sum $z_1 + z_2$, the difference $z_1 - z_2$, and the product $z_1 z_2$ are complex numbers, and likewise z_1/z_2 is a complex number if $z_2 \neq 0$. In addition, the magnitudes or moduli satisfy $|z| = |z_1||z_2|$, which is referred to as the law of the moduli. Hamilton had for a long time attempted to extend the theory of complex numbers to triplets $a + ib + jc$ where i and j are imaginary units, but he was unable to achieve a division algebra that satisfy the law of the moduli. Hamilton's great idea was then to introduce one more complex unit, which resulted in the quaternion $a+ib+jc+dk$, which lead to a division algebra where the law of the moduli was satisfied. It is now established that this is possible for dimensions 1, 2, 4 and 8, so the attempt to do this for triplets could not have succeeded. Hamilton's formulation of the product of quaternions is based on the rules for the imaginary units i, j and k. In this setting the quaternion vectors $\mathbf{q}_1 = (\alpha_1, \beta_{11}, \beta_{12}, \beta_{13})^T$ and $\mathbf{q}_2 = (\alpha_2, \beta_{21}, \beta_{22}, \beta_{23})^T$ can be represented by the quaternion numbers (Samson, Borgne and Espiau 1991)

$$\mathsf{q}_1 = \alpha_1 + i\beta_{11} + j\beta_{12} + k\beta_{13} \tag{6.181}$$

$$\mathsf{q}_2 = \alpha_2 + i\beta_{21} + j\beta_{22} + k\beta_{23} \tag{6.182}$$

Then the product of q_1 and q_2 is found to be

$$\begin{aligned} \mathsf{q}_1\mathsf{q}_2 &= (\alpha_1 + i\beta_{11} + j\beta_{12} + k\beta_{13})(\alpha_2 + i\beta_{21} + j\beta_{22} + k\beta_{23}) \\ &= \alpha_1\alpha_2 - \beta_{11}\beta_{21} - \beta_{12}\beta_{22} - \beta_{13}\beta_{23} \\ &\quad + i(\alpha_1\beta_{21} + \alpha_2\beta_{11} + \beta_{12}\beta_{23} - \beta_{13}\beta_{22}) \\ &\quad + j(\alpha_1\beta_{22} + \alpha_2\beta_{12} + \beta_{13}\beta_{21} - \beta_{11}\beta_{23}) \\ &\quad + k(\alpha_1\beta_{23} + \alpha_2\beta_{13} + \beta_{11}\beta_{22} - \beta_{12}\beta_{21}) \end{aligned} \tag{6.183}$$

We see that this product of quaternion numbers satisfy

$$\mathsf{q} = \mathsf{q}_1\mathsf{q}_2 \tag{6.184}$$

where

$$\mathsf{q} = \alpha + i\beta_1 + j\beta_2 + k\beta_3 \tag{6.185}$$

corresponds to the quaternion vector $\mathbf{q} = \mathbf{q}_1 \otimes \mathbf{q}_2$ as defined in (6.172).

6.7.3 Unit quaternions

A *unit quaternion*

$$\mathbf{p} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \tag{6.186}$$

is a quaternion with unit length, that is, a quaternion that satisfies

$$\mathbf{p}^T\mathbf{p} = \eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = 1 \tag{6.187}$$

We see that if η and $\boldsymbol{\epsilon}$ are Euler parameters, then $\mathbf{p}$ is a unit quaternion corresponding to the rotation matrix $\mathbf{R}_{\eta,\epsilon}$. The unit quaternion corresponding to $\mathbf{R}_{\eta,\epsilon}^{-1} = \mathbf{R}_{\eta,-\epsilon}$ is the *inverse unit quaternion* $\bar{\mathbf{p}}$ defined by

$$\bar{\mathbf{p}} = \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} \tag{6.188}$$

The unit quaternion corresponding to the identity matrix $\mathbf{R}_{1,0} = \mathbf{I}$ is the identity quaternion $\mathbf{p}_{id}$ defined by

$$\mathbf{p}_{id} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} \tag{6.189}$$

6.7.4 The quaternion product for unit quaternions

The quaternion product of two unit quaternions $\mathbf{p}_1$ and $\mathbf{p}_2$ is a unit quaternion

$$\mathbf{p} := \mathbf{p}_1 \otimes \mathbf{p}_2 = \begin{pmatrix} \eta_1\eta_2 - \boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_2 \\ \eta_1\boldsymbol{\epsilon}_2 + \eta_2\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_2 \end{pmatrix} \tag{6.190}$$

This is shown by direct computation of $\mathbf{p}^T\mathbf{p}$ which gives

$$\begin{aligned} \mathbf{p}^T\mathbf{p} &= \eta_1^2\eta_2^2 - 2\eta_1\eta_2\boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_2 + (\boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_2)^2 + \eta_1^2\boldsymbol{\epsilon}_2^T\boldsymbol{\epsilon}_2 + 2\eta_1\eta_2\boldsymbol{\epsilon}_2^T\boldsymbol{\epsilon}_1 + \eta_2^2\boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2^T\boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_2 \\ &= (\eta_1^2 + \boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_1)(\eta_2^2 + \boldsymbol{\epsilon}_2^T\boldsymbol{\epsilon}_2) = 1 \end{aligned}$$

where it is used that $\boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_2$ is orthogonal to $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, $\boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_1^\times = \boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_1^T - \boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_1\mathbf{I}$, $\eta_1^2 + \boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_1 = 1$ and $\eta_2^2 + \boldsymbol{\epsilon}_2^T\boldsymbol{\epsilon}_1 = 1$.

It is straightforward to check that the quaternion product of $\mathbf{p}$ and the inverse $\bar{\mathbf{p}}$ is the identity quaternion, that is,

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \bar{\mathbf{p}} \otimes \mathbf{p} = \mathbf{p}_{id} \tag{6.191}$$

This follows from

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ \eta\boldsymbol{\epsilon} - \eta\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\times\boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} = \mathbf{p}_{id} \tag{6.192}$$

In the same way it can be shown that $\bar{\mathbf{p}} \otimes \mathbf{p} = \mathbf{p}_{id}$.

We may also verify that

$$\mathbf{p} \otimes \mathbf{p}_{id} = \mathbf{p}_{id} \otimes \mathbf{p} = \mathbf{p} \tag{6.193}$$

Example 101 *The unit quaternion satisfies*

$$\dot{\mathbf{p}} \otimes \bar{\mathbf{p}} = \begin{pmatrix} \dot{\eta} \\ \dot{\boldsymbol{\epsilon}} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} 0 \\ -\dot{\eta}\boldsymbol{\epsilon} + \eta\dot{\boldsymbol{\epsilon}} + \boldsymbol{\epsilon}^\times\dot{\boldsymbol{\epsilon}} \end{pmatrix} \tag{6.194}$$

where we have used

$$\dot{\eta}\eta + \dot{\boldsymbol{\epsilon}}^T\boldsymbol{\epsilon} = \frac{1}{2}\frac{d}{dt}\left(\eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right) = 0 \tag{6.195}$$

This result is used in the derivation of the kinematic differential equation for unit quaternions.

Example 102 *The time derivative of the inverse unit quaternion is found from*

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \mathbf{p}_{id} \quad \Rightarrow \quad \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \dot{\bar{\mathbf{p}}} = \mathbf{0} \tag{6.196}$$

which implies

$$\dot{\bar{\mathbf{p}}} = -\bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \tag{6.197}$$

6.7.5 Rotation by the quaternion product

Let $\mathbf{R} := \mathbf{R}_e(\eta, \boldsymbol{\epsilon})$ be the rotation matrix corresponding to the Euler parameters η and $\boldsymbol{\epsilon}$. Let $\mathbf{v} \in R^3$ be an arbitrary vector. We are already familiar with the notion that $\mathbf{Rv}$ is either the coordinate vector of the vector $\mathbf{v}$ in some other frame, or it is a rotation of the vector $\mathbf{v}$.

The transformation $\mathbf{Rv}$ can be achieved with the Euler parameters and the quaternion product according to

$$\begin{pmatrix} 0 \\ \mathbf{Rv} \end{pmatrix} = \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} \tag{6.198}$$

This is shown by direct computation of the quaternion products:

$$\begin{aligned} \begin{pmatrix} \eta \\ \boldsymbol{\epsilon} \end{pmatrix} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \begin{pmatrix} \eta \\ -\boldsymbol{\epsilon} \end{pmatrix} &= \begin{pmatrix} \eta\boldsymbol{\epsilon}^T\mathbf{v} - \eta\boldsymbol{\epsilon}^T\mathbf{v} - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}^\times\mathbf{v} \\ \eta^2\mathbf{v} + 2\eta\boldsymbol{\epsilon}^\times\mathbf{v} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{v} + \boldsymbol{\epsilon}^\times\boldsymbol{\epsilon}^\times\mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ (\mathbf{I} + 2\eta\boldsymbol{\epsilon}^\times + 2\boldsymbol{\epsilon}^\times\boldsymbol{\epsilon}^\times)\,\mathbf{v} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \mathbf{Rv} \end{pmatrix} \end{aligned} \tag{6.199}$$

where we have used $(\boldsymbol{\epsilon}^\times)^2 = \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\mathbf{I}$ and $\eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = 1$.

We will now see how composite rotations can be expressed in terms of unit quaternions. Let

$$\mathbf{R}_1 = \mathbf{R}_e(\eta_1, \boldsymbol{\epsilon}_1) \quad \text{and} \quad \mathbf{R}_2 = \mathbf{R}_e(\eta_2, \boldsymbol{\epsilon}_2) \tag{6.200}$$

and let

$$\mathbf{R} = \mathbf{R}_1\mathbf{R}_2 = \mathbf{R}_e(\eta, \boldsymbol{\epsilon}). \tag{6.201}$$

be the composite rotation where $\mathbf{p} = (\eta \;\; \boldsymbol{\epsilon}^T)^T$, $\mathbf{p}_1 = (\eta_1 \;\; \boldsymbol{\epsilon}_1^T)^T$ and $\mathbf{p}_2 = (\eta_2 \;\; \boldsymbol{\epsilon}_2^T)^T$.

Let $\mathbf{u}$ be an arbitrary vector, and define $\mathbf{v} := \mathbf{R}_2\mathbf{u}$ and $\mathbf{w} := \mathbf{R}_1\mathbf{v} = \mathbf{Ru}$. Then

$$\begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} = \mathbf{p}_2 \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \overline{\mathbf{p}}_2 \tag{6.202}$$

$$\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix} = \mathbf{p}_1 \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \overline{\mathbf{p}}_1 = \mathbf{p}_1 \otimes \mathbf{p}_2 \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \overline{\mathbf{p}}_2 \otimes \overline{\mathbf{p}}_1 \tag{6.203}$$

At the same time we have $\mathbf{w} = \mathbf{R}_1\mathbf{R}_2\mathbf{u} = \mathbf{Ru}$ which gives

$$\begin{pmatrix} 0 \\ \mathbf{w} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \overline{\mathbf{p}} \tag{6.204}$$

Comparing these results we find that the Euler parameters η, $\boldsymbol{\epsilon}$ corresponding to the composite rotation $\mathbf{R}$ is given by

$$\mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2 \tag{6.205}$$

which can also be written

$$\eta = \eta_1\eta_2 - \boldsymbol{\epsilon}_1^T\boldsymbol{\epsilon}_2 \tag{6.206}$$

$$\boldsymbol{\epsilon} = \eta_1\boldsymbol{\epsilon}_2 + \eta_2\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_1^\times\boldsymbol{\epsilon}_2 \tag{6.207}$$

Example 103 *We see that*

$$\mathbf{R}\mathbf{R}^T = \mathbf{I} \tag{6.208}$$

is consistent with

$$\mathbf{p} \otimes \bar{\mathbf{p}} = \mathbf{p}_{id} \tag{6.209}$$

Moreover,

$$\mathbf{R}\mathbf{I} = \mathbf{I}\mathbf{R} = \mathbf{R} \tag{6.210}$$

is seen to be consistent with

$$\mathbf{p} \otimes \mathbf{p}_{id} = \mathbf{p}_{id} \otimes \mathbf{p} = \mathbf{p}_{id} \tag{6.211}$$

Example 104 *Let* $\mathbf{F}(\cdot)$ *and* $\mathbf{E}(\cdot)$ *be the matrices corresponding to pre-multiplication and post-multiplication, respectively, as defined in (6.174) and (6.175) Then*

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \bar{\mathbf{p}} \tag{6.212}$$

can be written

$$\begin{pmatrix} 0 \\ \mathbf{R}\mathbf{v} \end{pmatrix} = \mathbf{F}(\mathbf{p}) \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \otimes \bar{\mathbf{p}} = \mathbf{F}(\mathbf{p})\mathbf{E}(\bar{\mathbf{p}}) \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix} \tag{6.213}$$

This leads to one more formula for the rotation matrix:

$$\mathbf{R} = \begin{pmatrix} -\boldsymbol{\epsilon} & \eta\mathbf{I} + \boldsymbol{\epsilon}^{\times} \end{pmatrix} \begin{pmatrix} -\boldsymbol{\epsilon} & \eta\mathbf{I} - \boldsymbol{\epsilon}^{\times} \end{pmatrix}^T \tag{6.214}$$

6.7.6 Euler parameters from the rotation matrix

The problem to be solved in this section is how to find the Euler parameters η, $\boldsymbol{\epsilon}$ when the rotation matrix $\mathbf{R} = \{r_{ij}\}$ is given. This is done using a method due to Shepperd (Shepperd 1978).

The rotation matrix is given in terms of the Euler parameters by (6.165):

$$\mathbf{R} = \mathbf{R}_e(\eta, \boldsymbol{\epsilon}) = \begin{pmatrix} \eta^2 + \epsilon_1^2 - \epsilon_2^2 - \epsilon_3^2 & 2(\epsilon_1\epsilon_2 - \eta\epsilon_3) & 2(\epsilon_1\epsilon_3 + \eta\epsilon_2) \\ 2(\epsilon_1\epsilon_2 + \eta\epsilon_3) & \eta^2 - \epsilon_1^2 + \epsilon_2^2 - \epsilon_3^2 & 2(\epsilon_2\epsilon_3 - \eta\epsilon_1) \\ 2(\epsilon_1\epsilon_3 - \eta\epsilon_2) & 2(\epsilon_2\epsilon_3 + \eta\epsilon_1) & \eta^2 - \epsilon_1^2 - \epsilon_2^2 + \epsilon_3^2 \end{pmatrix} \tag{6.215}$$

In addition, $\eta^2 + \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 = 1$. The following notation is introduced to simplify the algorithms:

$$\mathbf{z} = \begin{pmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{pmatrix} := 2 \begin{pmatrix} \eta \\ \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} \tag{6.216}$$

$$T := r_{11} + r_{22} + r_{33} = \text{Trace}\mathbf{R} \tag{6.217}$$

and

$$r_{00} := T \tag{6.218}$$

This gives the symmetric set of equations

$$z_0^2 = 1 + 2r_{00} - T \tag{6.219}$$

$$z_1^2 = 1 + 2r_{11} - T \tag{6.220}$$

$$z_2^2 = 1 + 2r_{22} - T \tag{6.221}$$

$$z_3^2 = 1 + 2r_{33} - T \tag{6.222}$$

that appear from the diagonal elements of $\mathbf{R}$, while the off-diagonal terms give the equations

$$z_0 z_1 = r_{32} - r_{23} \qquad z_2 z_3 = r_{32} + r_{23} \tag{6.223}$$
$$z_0 z_2 = r_{13} - r_{31} \qquad z_3 z_1 = r_{13} + r_{31} \tag{6.224}$$
$$z_0 z_3 = r_{21} - r_{12} \qquad z_1 z_2 = r_{21} + r_{12} \tag{6.225}$$

The algorithm is as follows:

1. Find the largest element in $\{r_{00}, r_{11}, r_{22}, r_{33}\}$. This element is denoted r_{ii}.
2. Compute
$$|z_i| = \sqrt{1 + 2r_{ii} - T} \tag{6.226}$$
3. Determine the sign of z_i from some criterion, like continuity of solution, or $\eta > 0$.
4. Find the remaining z_j from the three equations out of (6.223–6.225) that have as the left side $z_j z_i$ for all $j \neq i$. For example, if z_0 was found under step 2 and 3, then the remaining z_j are found from
$$z_1 = (r_{32} - r_{23})/z_0 \tag{6.227}$$
$$z_2 = (r_{13} - r_{31})/z_0 \tag{6.228}$$
$$z_3 = (r_{21} - r_{12})/z_0 \tag{6.229}$$
5. Compute $\eta = z_0/2$ and $\epsilon_i = z_i/2$.

Note that this algorithm avoids division by zero as the division is done with the z_i that has the largest absolute value.

6.7.7 The Euler rotation vector

The Euler rotation vector

$$\mathbf{e} = \mathbf{k} \sin\theta \in R^3 \tag{6.230}$$

is defined from the angle-axis parameters ($\mathbf{k}$,θ) . From (6.145) it is seen that the rotation matrix $\mathbf{R}_{k,\theta}$ and its transpose $\mathbf{R}^T_{k,\theta}$ are given by

$$\mathbf{R}_{k,\theta} = \mathbf{e}^\times + \cos\theta \mathbf{I} + \mathbf{k}\mathbf{k}^T(1 - \cos\theta) \tag{6.231}$$
$$\mathbf{R}^T_{k,\theta} = -\mathbf{e}^\times + \cos\theta \mathbf{I} + \mathbf{k}\mathbf{k}^T(1 - \cos\theta) \tag{6.232}$$

which implies that

$$\mathbf{e}^\times = \frac{1}{2}\left(\mathbf{R}_{k,\theta} - \mathbf{R}^T_{k,\theta}\right) \tag{6.233}$$

From this we see that if $\mathbf{R}_{k,\theta} = \{r_{ij}\}$, then the Euler rotation vector can be found from

$$\mathbf{e} = \frac{1}{2}\begin{pmatrix} r_{32} - r_{23} \\ r_{13} - r_{31} \\ r_{21} - r_{12} \end{pmatrix} \tag{6.234}$$

We note that if $\mathbf{R}_{k,\theta} = \mathbf{R}^a_b$, then

$$\mathbf{e} = \mathbf{e}^a = \mathbf{e}^b \tag{6.235}$$

as $\mathbf{R}_{k,\theta}\mathbf{k} = \mathbf{k}$.

Example 105 *In robot control the desired orientation of the robot hand may be specified to be*

$$\mathbf{R}_d = \begin{pmatrix} \mathbf{n}_d & \mathbf{s}_d & \mathbf{a}_d \end{pmatrix} \in SO(3) \tag{6.236}$$

Suppose that the actual orientation of the robot hand is

$$\mathbf{R} = \begin{pmatrix} \mathbf{n} & \mathbf{s} & \mathbf{a} \end{pmatrix} \in SO(3) \tag{6.237}$$

where $\mathbf{n}$ *is the normal vector,* $\mathbf{s}$ *is the slide vector and* $\mathbf{a}$ *is the approach vector of the hand (Spong and Vidyasagar 1989), (Sciavicco and Siciliano 2000). Then the deviation of* $\mathbf{R}$ *from* $\mathbf{R}_d$ *is given by the rotation matrix* $\widetilde{\mathbf{R}} = \{\tilde{r}_{ij}\}$ *which is defined by*

$$\widetilde{\mathbf{R}} := \mathbf{R}\mathbf{R}_d^T \quad \Rightarrow \quad \mathbf{R} = \widetilde{\mathbf{R}}\mathbf{R}_d \tag{6.238}$$

The component form of this equation is

$$\tilde{r}_{ij} = n_i n_{dj} + s_i s_{dj} + a_i a_{dj} \tag{6.239}$$

If an angle-axis parameters of $\widetilde{\mathbf{R}}$ *are* $\left(\widetilde{\mathbf{k}},\tilde{\theta}\right)$ *and* $\widetilde{\mathbf{e}} = \widetilde{\mathbf{k}}\sin\tilde{\theta}$ *is the associated Euler rotation vector, then (6.234) gives*

$$\begin{aligned} \widetilde{\mathbf{e}} &= \frac{1}{2}\begin{pmatrix} \tilde{r}_{32} - \tilde{r}_{23} \\ \tilde{r}_{13} - \tilde{r}_{31} \\ \tilde{r}_{21} - \tilde{r}_{12} \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} n_3 n_{d2} - n_2 n_{d3} \\ n_1 n_{d3} - n_3 n_{d1} \\ n_2 n_{d1} - n_1 n_{d2} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} s_3 s_{d2} - s_2 s_{d3} \\ s_1 s_{d3} - s_3 s_{d1} \\ s_2 s_{d1} - s_1 s_{d2} \end{pmatrix} + \frac{1}{2}\begin{pmatrix} a_3 a_{d2} - a_2 a_{d3} \\ a_1 a_{d3} - a_3 a_{d1} \\ a_2 a_{d1} - a_1 a_{d2} \end{pmatrix} \end{aligned}$$

Using the definition of the vector cross product the Euler rotation vector $\widetilde{\mathbf{e}}$ *corresponding to the deviation* $\widetilde{\mathbf{R}}$ *can be written*

$$\widetilde{\mathbf{e}} = \frac{1}{2}(\mathbf{n}_d^\times \mathbf{n} + \mathbf{s}_d^\times \mathbf{s} + \mathbf{a}_d^\times \mathbf{a}) \tag{6.240}$$

6.7.8 Euler-Rodrigues parameters

The *Euler-Rodrigues parameters* are defined by (Hughes 1986)

$$\boldsymbol{\rho} = \mathbf{k}\tan\frac{\theta}{2} \tag{6.241}$$

This can be expressed in terms of the Euler parameters according to

$$\boldsymbol{\rho} = \frac{\boldsymbol{\epsilon}}{\eta} \tag{6.242}$$

It is evident that the Euler-Rodrigues parameters are undefined when $\eta = 0 \Leftrightarrow \theta = \pi + 2k\pi$ where $k \in \{\ldots -1, 0, 1 \ldots\}$.

The derivations that follows use the relation

$$1 = \eta^2 + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \eta^2(1 + \boldsymbol{\rho}^T\boldsymbol{\rho}) \tag{6.243}$$

which implies

$$\eta^2 = \frac{1}{1 + \boldsymbol{\rho}^T\boldsymbol{\rho}} \tag{6.244}$$

Then

$$\mathbf{R} = \mathbf{I} + \frac{2}{1+\boldsymbol{\rho}^T\boldsymbol{\rho}}[\boldsymbol{\rho}^\times + \boldsymbol{\rho}^\times\boldsymbol{\rho}^\times] \tag{6.245}$$

is found from (6.166). We note that there are no trigonometric terms in (6.245).

The Euler-Rodrigues parameters can be found from the rotation matrix using

$$\boldsymbol{\rho} = \frac{\boldsymbol{\epsilon}}{\eta} = \frac{\mathbf{e}}{2\eta^2} = \frac{1}{\mathrm{Trace}\mathbf{R}+1}\begin{pmatrix} r_{32}-r_{23} \\ r_{13}-r_{31} \\ r_{21}-r_{12} \end{pmatrix} \tag{6.246}$$

where (6.234) and (6.169) are used.

Next we will derive *Cayley's formula* (Angeles 1988) from equation (6.245). To do this we need some algebraic manipulations. First we observe that $\boldsymbol{\rho}^\times\boldsymbol{\rho}^\times = \boldsymbol{\rho}\boldsymbol{\rho}^T - \boldsymbol{\rho}^T\boldsymbol{\rho}\mathbf{I}$ implies that

$$\boldsymbol{\rho}^\times\boldsymbol{\rho}^\times\boldsymbol{\rho}^\times = -\left(\boldsymbol{\rho}^T\boldsymbol{\rho}\right)\boldsymbol{\rho}^\times \tag{6.247}$$

Using this result and (6.245) we find that

$$\mathbf{R}\left(\mathbf{I}-\boldsymbol{\rho}^\times\right) = \left[\mathbf{I} + \frac{2}{1+\boldsymbol{\rho}^T\boldsymbol{\rho}}\left(\boldsymbol{\rho}^\times + \boldsymbol{\rho}^\times\boldsymbol{\rho}^\times\right)\right]\left(\mathbf{I}-\boldsymbol{\rho}^\times\right) = \mathbf{I} + \boldsymbol{\rho}^\times \tag{6.248}$$

This leads to the following result:

The rotation matrix can be given by Cayley's formula

$$\mathbf{R} = \left(\mathbf{I}+\boldsymbol{\rho}^\times\right)\left(\mathbf{I}-\boldsymbol{\rho}^\times\right)^{-1} \tag{6.249}$$

where $\boldsymbol{\rho}$ is the vector of Euler-Rodrigues parameters corresponding to $\mathbf{R}$.

The *Cayley transformation* $\mathrm{cay}(\mathbf{u}) \in SO(3)$ maps a three-dimensional vector $\mathbf{u}$ into a rotation matrix according to

$$\mathrm{cay}(\mathbf{u}) := \left[\mathbf{I} + \frac{1}{2}\mathbf{u}^\times\right]\left[\mathbf{I} - \frac{1}{2}\mathbf{u}^\times\right]^{-1} \in SO(3) \tag{6.250}$$

This transformation is used in numerical integrators in attitude problems (Lewis and Simo 1994). In particular it is well suited for the implementation of the implicit midpoint rule for the integration of the rotation matrix. We note that

$$\mathbf{R} = \mathrm{cay}\left(2\boldsymbol{\rho}\right) \tag{6.251}$$

and that

$$\mathrm{cay}(\mathbf{k}\theta) \approx \mathbf{R}_{k,\theta}, \quad \theta \text{ small} \tag{6.252}$$

6.8 Angular velocity

6.8.1 Introduction

If the position vector $\mathbf{r}$ is given in an inertial frame, then the velocity vector $\mathbf{v} = \dot{\mathbf{r}}$ is known to be the rate of change of the position vector $\mathbf{r}$. In the same way we would like to have some physical entity that describes the rate of change of a rotation matrix $\mathbf{R}_b^a$. This is not quite as simple as for the case of position and velocity. However, the

rotation matrix can be described by three independent variables, and this indicates that there might be some entity that represents the time derivative of the rotation matrix using three parameters. We will in the following analyze this problem, and arrive at the definition of the angular velocity vector $\vec{\omega}$, which represents the time derivative of the rotation matrix.

6.8.2 Definition

The rotation matrix $\mathbf{R}_b^a$ is orthogonal and satisfies

$$\mathbf{R}_b^a(\mathbf{R}_b^a)^T = \mathbf{I} \tag{6.253}$$

Time differentiation of the matrix product gives

$$\frac{d}{dt}\left[\mathbf{R}_b^a(\mathbf{R}_b^a)^T\right] = \dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T + \mathbf{R}_b^a(\dot{\mathbf{R}}_b^a)^T = \mathbf{0} \tag{6.254}$$

From this equation it is seen that the matrix $\dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T$ is skew symmetric. Now, any skew symmetric 3×3 matrix can be seen as the skew symmetric form of a column vector. This means that it is possible to define a vector so that its skew symmetric form is equal to $\dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T$. It is quite remarkable that this vector can be given a physical interpretation, and that it is of fundamental importance in dynamics.

Let the vector $\vec{\omega}_{ab}$ be defined by requiring that its coordinate form $\boldsymbol{\omega}_{ab}^a$ in frame a satisfies

$$(\boldsymbol{\omega}_{ab}^a)^\times = \dot{\mathbf{R}}_b^a(\mathbf{R}_b^a)^T \tag{6.255}$$

The vector $\vec{\omega}_{ab}$ is said to be the *angular velocity* vector of frame b relative to frame a.

A kinematic differential equation for the rotation matrix $\mathbf{R}_b^a$ appears from the definition of the angular velocity by post-multiplication of (6.255) with $\mathbf{R}_b^a$. Moreover, using the coordinate transformation rule $(\boldsymbol{\omega}_{ab}^a)^\times = \mathbf{R}_b^a(\boldsymbol{\omega}_{ab}^b)^\times\mathbf{R}_a^b$ for the skew symmetric form of a vector an alternative formulation of the kinematic differential equation is found.

The kinematic differential equation of the rotation matrix is given by the two alternative forms

$$\dot{\mathbf{R}}_b^a = (\boldsymbol{\omega}_{ab}^a)^\times\mathbf{R}_b^a \tag{6.256}$$

$$\dot{\mathbf{R}}_b^a = \mathbf{R}_b^a(\boldsymbol{\omega}_{ab}^b)^\times \tag{6.257}$$

6.8.3 Simple rotations

Using the rotation matrices $\mathbf{R}_x(\phi)$, $\mathbf{R}_y(\theta)$ and $\mathbf{R}_z(\psi)$ of the simple rotations about the x, y and z axes, we define the angular velocities

$$\left[\boldsymbol{\omega}_x(\dot{\phi})\right]^\times \quad : \quad = \dot{\mathbf{R}}_x(\phi)\mathbf{R}_x^T(\phi) \tag{6.258}$$

$$\left[\boldsymbol{\omega}_y(\dot{\theta})\right]^\times \quad : \quad = \dot{\mathbf{R}}_y(\theta)\mathbf{R}_y^T(\theta) \tag{6.259}$$

$$\left[\boldsymbol{\omega}_z(\dot{\psi})\right]^\times \quad : \quad = \dot{\mathbf{R}}_z(\psi)\mathbf{R}_z^T(\psi) \tag{6.260}$$

From the definitions it is clear that $\boldsymbol{\omega}_x(\dot{\phi})$ is the angular velocity of a rotation by an angular rate $\dot{\phi}$ about the x axis, $\boldsymbol{\omega}_y(\dot{\theta})$ is the angular velocity of a rotation by an angular rate $\dot{\theta}$ about the y axis, and $\boldsymbol{\omega}_z(\dot{\psi})$ is the angular velocity of a rotation by an angular rate $\dot{\psi}$ about the z axis. From (6.101) we find that

$$\begin{aligned}\left[\boldsymbol{\omega}_x(\dot{\phi})\right]^\times &= \dot{\phi}\begin{pmatrix}0&0&0\\0&-\sin\phi&-\cos\phi\\0&\cos\phi&-\sin\phi\end{pmatrix}\begin{pmatrix}1&0&0\\0&\cos\phi&\sin\phi\\0&-\sin\phi&\cos\phi\end{pmatrix} &(6.261)\\ &= \begin{pmatrix}0&0&0\\0&0&-\dot{\phi}\\0&\dot{\phi}&0\end{pmatrix} &(6.262)\end{aligned}$$

In the same way we may compute $\left[\boldsymbol{\omega}_y(\dot{\theta})\right]^\times$ and $\left[\boldsymbol{\omega}_z(\dot{\psi})\right]^\times$. This results in

$$\boldsymbol{\omega}_x(\dot{\phi}) = \begin{pmatrix}\dot{\phi}\\0\\0\end{pmatrix}, \quad \boldsymbol{\omega}_y(\dot{\theta}) = \begin{pmatrix}0\\\dot{\theta}\\0\end{pmatrix}, \quad \text{and} \quad \boldsymbol{\omega}_z(\dot{\psi}) = \begin{pmatrix}0\\0\\\dot{\psi}\end{pmatrix} \tag{6.263}$$

Consider the angle-axis parameterization when $\mathbf{k} = \mathbf{k}^a$ is a constant vector and

$$\mathbf{R}^a_b = \mathbf{R}_{k,\theta} = \mathbf{I} + \mathbf{k}^\times \sin\theta + \mathbf{k}^\times\mathbf{k}^\times(1-\cos\theta) \tag{6.264}$$

Then the angular velocity is

$$\begin{aligned}(\boldsymbol{\omega}^a_{ab})^\times &= \dot{\theta}\left(\mathbf{k}^\times\cos\theta + \mathbf{k}^\times\mathbf{k}^\times\sin\theta\right)\left(\mathbf{I} - \mathbf{k}^\times\sin\theta + \mathbf{k}^\times\mathbf{k}^\times(1-\cos\theta)\right)\\ &= \dot{\theta}\left[\mathbf{k}^\times\cos\theta + \mathbf{k}^\times\mathbf{k}^\times\sin\theta - \mathbf{k}^\times\mathbf{k}^\times\cos\theta\sin\theta + \mathbf{k}^\times\sin^2\theta\right.\\ &\quad \left.-\mathbf{k}^\times\left(\cos\theta - \cos^2\theta\right) + \mathbf{k}^\times\mathbf{k}^\times\cos\theta\sin\theta - \mathbf{k}^\times\mathbf{k}^\times\sin\theta\right]\\ &= \dot{\theta}\mathbf{k}^\times &(6.265)\end{aligned}$$

This shows that:

For a simple rotation, the angular velocity vector $\vec{\omega}_{ab}$ is along the axis of rotation $\vec{k}$, and is given by

$$\vec{\omega}_{ab} = \dot{\theta}\vec{k} \tag{6.266}$$

This gives an intuitively appealing interpretation of the angular velocity. If the axis of rotation is not constant, then the expressions become somewhat more involved.

6.8.4 Composite rotations

Consider the composite rotation $\mathbf{R}^a_d = \mathbf{R}^a_b\mathbf{R}^b_c\mathbf{R}^c_d$. The time derivative of $\mathbf{R}^a_d$ is, according to the product rule,

$$\dot{\mathbf{R}}^a_d = \dot{\mathbf{R}}^a_b\mathbf{R}^b_c\mathbf{R}^c_d + \mathbf{R}^a_b\dot{\mathbf{R}}^b_c\mathbf{R}^c_d + \mathbf{R}^a_b\mathbf{R}^b_c\dot{\mathbf{R}}^c_d \tag{6.267}$$

and the transpose is $\left(\mathbf{R}^a_d\right)^T = \left(\mathbf{R}^c_d\right)^T\left(\mathbf{R}^b_c\right)^T\left(\mathbf{R}^a_b\right)^T$. Then the angular velocity of the composite rotation is

$$\begin{aligned}(\boldsymbol{\omega}^a_{ad})^\times &= \dot{\mathbf{R}}^a_d\left(\mathbf{R}^a_d\right)^T = \left(\dot{\mathbf{R}}^a_b\mathbf{R}^b_c\mathbf{R}^c_d + \mathbf{R}^a_b\dot{\mathbf{R}}^b_c\mathbf{R}^c_d + \mathbf{R}^a_b\mathbf{R}^b_c\dot{\mathbf{R}}^c_d\right)\left(\mathbf{R}^c_d\right)^T\left(\mathbf{R}^b_c\right)^T\left(\mathbf{R}^a_b\right)^T\\ &= \dot{\mathbf{R}}^a_b\left(\mathbf{R}^a_b\right)^T + \mathbf{R}^a_b\dot{\mathbf{R}}^b_c\left(\mathbf{R}^b_c\right)^T\left(\mathbf{R}^a_b\right)^T + \mathbf{R}^a_c\dot{\mathbf{R}}^c_d\left(\mathbf{R}^c_d\right)^T\left(\mathbf{R}^a_c\right)^T\\ &= (\boldsymbol{\omega}^a_{ab})^\times + \mathbf{R}^a_b(\boldsymbol{\omega}^b_{bc})^\times\left(\mathbf{R}^a_b\right)^T + \mathbf{R}^a_c(\boldsymbol{\omega}^c_{cd})^\times\left(\mathbf{R}^a_c\right)^T\\ &= (\boldsymbol{\omega}^a_{ab})^\times + (\boldsymbol{\omega}^a_{bc})^\times + (\boldsymbol{\omega}^a_{cd})^\times &(6.268)\end{aligned}$$

This implies that the angular velocities $\vec{\omega}_{ab}$, $\vec{\omega}_{bc}$ and $\vec{\omega}_{cd}$ can be added vectorially.

The angular velocity of the composite rotation matrix $\mathbf{R}_d^a = \mathbf{R}_b^a \mathbf{R}_c^b \mathbf{R}_d^c$ is the sum of the angular velocities according to

$$\vec{\omega}_{ad} = \vec{\omega}_{ab} + \vec{\omega}_{bc} + \vec{\omega}_{cd} \tag{6.269}$$

Example 106 *In a gimbal system for inertial navigation the rotation matrix from the vehicle frame b to the instrumented platform frame p will be*

$$\mathbf{R}_p^b = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_x(\phi) \tag{6.270}$$

which corresponds to the angular velocity vector

$$\boldsymbol{\omega}_{bp}^b = \boldsymbol{\omega}_z(\dot{\psi}) + \mathbf{R}_z(\psi)\boldsymbol{\omega}_y(\dot{\theta}) + \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\boldsymbol{\omega}_x(\dot{\phi}) \tag{6.271}$$

6.8.5 Differentiation of coordinate vectors

A coordinate vector is differentiated with respect to time by differentiating the components of the vector with respect to time:

$$\dot{\mathbf{u}}^a := \frac{d}{dt}(\mathbf{u}^a) = \frac{d}{dt}\begin{pmatrix} u_1^a \\ u_2^a \\ u_3^a \end{pmatrix} = \begin{pmatrix} \dot{u}_1^a \\ \dot{u}_2^a \\ \dot{u}_3^a \end{pmatrix} \tag{6.272}$$

The relation between the time derivative in frame a and the time derivative in frame b is found by differentiating the equation

$$\mathbf{u}^a = \mathbf{R}_b^a \mathbf{u}^b \tag{6.273}$$

which gives

$$\dot{\mathbf{u}}^a = \mathbf{R}_b^a \dot{\mathbf{u}}^b + \dot{\mathbf{R}}_b^a \mathbf{u}^b \tag{6.274}$$

Insertion of $\dot{\mathbf{R}}_b^a = \mathbf{R}_b^a (\boldsymbol{\omega}_{ab}^b)^\times$ gives the relation

$$\dot{\mathbf{u}}^a = \mathbf{R}_b^a \left[\dot{\mathbf{u}}^b + (\boldsymbol{\omega}_{ab}^b)^\times \mathbf{u}^b\right] \tag{6.275}$$

6.8.6 Differentiation of vectors

Differentiation of a vector $\vec{u}$ must be done with reference to some reference frame. The time derivative of the vector $\vec{u} = u_1^a \vec{a}_1 + u_2^a \vec{a}_2 + u_3^a \vec{a}_3$ referenced to frame a is defined by

$$\frac{{}^a d}{dt}\vec{u} := \dot{u}_1^a \vec{a}_1 + \dot{u}_2^a \vec{a}_2 + \dot{u}_3^a \vec{a}_3 \tag{6.276}$$

where the leading superscript a on the time differentiation operator denotes that the differentiation is taken with reference to frame a. The time derivative referenced to frame b is

$$\frac{{}^b d}{dt}\vec{u} = \dot{u}_1^b \vec{b}_1 + \dot{u}_2^b \vec{b}_2 + \dot{u}_3^b \vec{b}_3 \tag{6.277}$$

where it is assumed that $\vec{u} = u_1^b\vec{b}_1 + u_2^b\vec{b}_2 + u_3^b\vec{b}_3$. The corresponding column vector representation is

$$\dot{\mathbf{u}}^a = \begin{pmatrix} \dot{u}_1^a \\ \dot{u}_2^a \\ \dot{u}_3^a \end{pmatrix}, \quad \dot{\mathbf{u}}^b = \begin{pmatrix} \dot{u}_1^b \\ \dot{u}_2^b \\ \dot{u}_3^b \end{pmatrix} \tag{6.278}$$

From (6.275) we find that

$$\frac{{}^a d}{dt}\vec{u} = \frac{{}^b d}{dt}\vec{u} + \vec{\omega}_{ab} \times \vec{u} \tag{6.279}$$

where

$$\frac{{}^a d}{dt}\vec{u} = \dot{u}_1^a\vec{a}_1 + \dot{u}_2^a\vec{a}_2 + \dot{u}_3^a\vec{a}_3 \tag{6.280}$$

$$\frac{{}^b d}{dt}\vec{u} = \dot{u}_1^b\vec{b}_1 + \dot{u}_2^b\vec{b}_2 + \dot{u}_3^b\vec{b}_3 \tag{6.281}$$

Example 107 *In the same way partial differentiation with respect to some variable q in frame a and b is defined by*

$$\frac{{}^a\partial\vec{u}}{\partial q} := \frac{\partial u_1^a}{\partial q}\vec{a}_1 + \frac{\partial u_2^a}{\partial q}\vec{a}_2 + \frac{\partial u_3^a}{\partial q}\vec{a}_3 \tag{6.282}$$

$$\frac{{}^b\partial\vec{u}}{\partial q} := \frac{\partial u_1^b}{\partial q}\vec{b}_1 + \frac{\partial u_2^b}{\partial q}\vec{b}_2 + \frac{\partial u_3^b}{\partial q}\vec{b}_3 \tag{6.283}$$

Example 108 *An alternative definition of the angular velocity vector is used in (Kane and Levinson 1985):*

$$\vec{\omega}_{ab} = \vec{b}_1\left(\frac{{}^a d\vec{b}_2}{dt}\cdot\vec{b}_3\right) + \vec{b}_2\left(\frac{{}^a d\vec{b}_3}{dt}\cdot\vec{b}_1\right) + \vec{b}_3\left(\frac{{}^a d\vec{b}_1}{dt}\cdot\vec{b}_2\right) \tag{6.284}$$

Here $\vec{b}_1, \vec{b}_2, \vec{b}_3$ are the orthogonal unit vectors of the frame b. We will now show that this is in agreement by our definition (6.255). From (6.94) we have

$$\mathbf{R}_b^a = \begin{pmatrix} \mathbf{b}_1^a & \mathbf{b}_2^a & \mathbf{b}_3^a \end{pmatrix} \tag{6.285}$$

and from the definition of $\boldsymbol{\omega}_{ab}^b$ in (6.255) we get

$$(\boldsymbol{\omega}_{ab}^b)^\times = \mathbf{R}^T\dot{\mathbf{R}} = \begin{pmatrix} 0 & \mathbf{b}_1^{aT}\dot{\mathbf{b}}_2^a & \mathbf{b}_1^{aT}\dot{\mathbf{b}}_3^a \\ \mathbf{b}_2^{aT}\dot{\mathbf{b}}_1^a & 0 & \mathbf{b}_2^{aT}\dot{\mathbf{b}}_3^a \\ \mathbf{b}_3^{aT}\dot{\mathbf{b}}_1^a & \mathbf{b}_3^{aT}\dot{\mathbf{b}}_2^a & 0 \end{pmatrix} \tag{6.286}$$

Before proceeding we show that this matrix is skew symmetric. We note that $\mathbf{b}_1^a$, $\mathbf{b}_2^a$ and $\mathbf{b}_3^a$ are orthogonal. Then $\mathbf{b}_1^{aT}\mathbf{b}_2^a = 0$ and $\frac{d}{dt}(\mathbf{b}_1^{aT}\mathbf{b}_2^a) = \dot{\mathbf{b}}_1^{aT}\mathbf{b}_2^a + \mathbf{b}_1^{aT}\dot{\mathbf{b}}_2^a = 0$, which implies that $\mathbf{b}_1^{aT}\dot{\mathbf{b}}_2^a = -\mathbf{b}_2^{aT}\dot{\mathbf{b}}_1^a$. In the same way it is found that $\mathbf{b}_2^{aT}\dot{\mathbf{b}}_3^a = -\mathbf{b}_3^{aT}\dot{\mathbf{b}}_2^a$ and $\mathbf{b}_3^{aT}\dot{\mathbf{b}}_1^a = -\mathbf{b}_1^{aT}\dot{\mathbf{b}}_3^a$, and the right side in (6.286) is seen to be skew symmetric. We write $\boldsymbol{\omega}_{ab}^b$ in its vector form, and express the scalar products in terms of coordinate-free vectors to get

$$\boldsymbol{\omega}_{ab}^b = \begin{pmatrix} \mathbf{b}_3^{aT}\dot{\mathbf{b}}_2^a \\ \mathbf{b}_1^{aT}\dot{\mathbf{b}}_3^a \\ \mathbf{b}_2^{aT}\dot{\mathbf{b}}_1^a \end{pmatrix} = \begin{pmatrix} \vec{b}_3\cdot\frac{{}^a d\vec{b}_2}{dt} \\ \vec{b}_1\cdot\frac{{}^a d\vec{b}_3}{dt} \\ \vec{b}_2\cdot\frac{{}^a d\vec{b}_1}{dt} \end{pmatrix} \tag{6.287}$$

We see that this is indeed the coordinate form of the definition (6.284) of (Kane and Levinson 1985).

6.9 Kinematic differential equations

6.9.1 Introduction

A model describing the rotation of a rigid body can be separated into the equation of motion, which is a differential equation for the angular velocity, and a kinematic differential equation which gives the time derivative of some parameterization of the rotation matrix as a function of the angular velocity. From a modeling perspective it is interesting to note that kinematic differential equations are exact models with no uncertainty and no approximations involved. In the following we will derive kinematic differential equations for the different parametrizations of rotation that have been presented in the previous sections.

6.9.2 Attitude deviation

We consider the problem where the rotation of a rigid body is to be controlled. This will be the case in the attitude control of a satellite, or in the control of a robotic hand. Let the frame a define a reference orientation, let the frame b be a frame fixed in the body. Then the rotation matrix $\mathbf{R} := \mathbf{R}_b^a$ will describe the orientation of the body. Suppose that it is specified that the desired rotation of the body is given by a rotation matrix $\mathbf{R}_d$. The question is then how to represent the control deviation between the actual value $\mathbf{R}$ and the desired value $\mathbf{R}_d$. In a typical control setting we control some output vector $\mathbf{y}$ to its desired value $\mathbf{y}_d$, and the control deviation $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}_d$ is simply obtained by subtraction. In the case of rotation matrices it does not make sense to subtract $\mathbf{R}_d$ from $\mathbf{R}$ as the result would not be a rotation matrix. Instead the deviation between the two rotation matrices is described by the rotation matrix $\widetilde{\mathbf{R}}_a \in SO(3)$ defined by

$$\widetilde{\mathbf{R}}_a := \mathbf{R}\mathbf{R}_d^T \quad \Rightarrow \quad \mathbf{R} = \widetilde{\mathbf{R}}_a\mathbf{R}_d \tag{6.288}$$

We see that the rotation matrix $\mathbf{R}$ is described as the composite rotation defined by the rotation matrices $\widetilde{\mathbf{R}}_a$ and $\mathbf{R}_d$. To make this clear we introduce the intermediate frame c so that

$$\widetilde{\mathbf{R}}_a = \mathbf{R}_c^a, \;\; \mathbf{R}_d = \mathbf{R}_b^c \tag{6.289}$$

and

$$\frac{d}{dt}\widetilde{\mathbf{R}}_a = \dot{\mathbf{R}}_c^a = (\boldsymbol{\omega}_{ac}^a)^\times \mathbf{R}_c^a \tag{6.290}$$

$$\dot{\mathbf{R}}_d = \dot{\mathbf{R}}_b^c = \mathbf{R}_b^c(\boldsymbol{\omega}_{cb}^b)^\times \tag{6.291}$$

Then, if we define the angular velocity vectors

$$\boldsymbol{\omega}^a = \boldsymbol{\omega}_{ab}^a, \;\; \widetilde{\boldsymbol{\omega}}^a := \boldsymbol{\omega}_{ac}^a, \;\; \boldsymbol{\omega}_d^b := \boldsymbol{\omega}_{cb}^b \tag{6.292}$$

we find that the kinematic differential equations for $\mathbf{R}$, $\widetilde{\mathbf{R}}_a$ and $\mathbf{R}_d$ are given by

$$\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^\times \mathbf{R}, \;\; \frac{d}{dt}\widetilde{\mathbf{R}}_a = (\widetilde{\boldsymbol{\omega}}^a)^\times \widetilde{\mathbf{R}}_a, \;\; \dot{\mathbf{R}}_d = \mathbf{R}_d(\boldsymbol{\omega}_d^b)^\times \tag{6.293}$$

It follows from

$$\boldsymbol{\omega}_{ab}^a = \boldsymbol{\omega}_{ac}^a + \boldsymbol{\omega}_{cb}^a \tag{6.294}$$

that $\widetilde{\boldsymbol{\omega}}^a = \boldsymbol{\omega}^a - \boldsymbol{\omega}_d^a$, and we may sum up that the kinematic differential equations for the attitude deviation is

$$\widetilde{\mathbf{R}}_a := \mathbf{R}\mathbf{R}_d^T \tag{6.295}$$

$$\begin{aligned} \widetilde{\boldsymbol{\omega}}^a &= \boldsymbol{\omega}^a - \boldsymbol{\omega}_d^a & (6.296)\\ \frac{d}{dt}\widetilde{\mathbf{R}}_a &= (\widetilde{\boldsymbol{\omega}}^a)^\times \widetilde{\mathbf{R}}_a & (6.297) \end{aligned}$$

We define an alternative representation of the deviation between the two rotations using the rotation matrix $\widetilde{\mathbf{R}}_b \in SO(3)$ defined by

$$\widetilde{\mathbf{R}}_b := \mathbf{R}_d^T\mathbf{R} \quad \Rightarrow \quad \mathbf{R} = \mathbf{R}_d\widetilde{\mathbf{R}}_b \tag{6.298}$$

The desired angular velocity $\boldsymbol{\omega}_d^a$ is in this case defined by

$$\dot{\mathbf{R}}_d = (\boldsymbol{\omega}_d^a)^\times \mathbf{R}_d \tag{6.299}$$

Then, by introducing an intermediate frame as in the case above, we find that the kinematic differential equations referred to the b frame is

$$\widetilde{\mathbf{R}}_b := \mathbf{R}_d^T\mathbf{R} \tag{6.300}$$

$$\begin{aligned} \widetilde{\boldsymbol{\omega}}^b &= \boldsymbol{\omega}^b - \boldsymbol{\omega}_d^b & (6.301)\\ \frac{d}{dt}\widetilde{\mathbf{R}}_b &= \widetilde{\mathbf{R}}_b(\widetilde{\boldsymbol{\omega}}^b)^\times & (6.302) \end{aligned}$$

6.9.3 Homogeneous transformation matrices

The time derivative of the homogeneous transformation matrix

$$\mathbf{T}_b^a = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \in SE(3) \tag{6.303}$$

is found to be

$$\begin{aligned} \dot{\mathbf{T}}_b^a &= \begin{pmatrix} \mathbf{R}_b^a(\boldsymbol{\omega}_{ab}^b)^\times & \dot{\mathbf{r}}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} (\boldsymbol{\omega}_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \\ &= \mathbf{T}_b^a \begin{pmatrix} (\boldsymbol{\omega}_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \end{aligned} \tag{6.304}$$

We see that the time derivative of a homogeneous transformation matrix has certain similarities with the time derivative of a rotation matrix as expressed in (6.257). This similarity becomes more evident if we introduce the vector

$$\mathbf{w} = \begin{pmatrix} \mathbf{v}_{ab}^b \\ \boldsymbol{\omega}_{ab}^b \end{pmatrix} \tag{6.305}$$

which is the *twist vector* in the b frame. The twist vector $\mathbf{w}$ is a six-dimensional vector containing the velocity and the angular velocity. In analogy with the angular velocity $\boldsymbol{\omega}_{ab}^b$ and its matrix form $\left(\boldsymbol{\omega}_{ab}^b\right)^\times$, the twist vector $\mathbf{w}$ has a matrix form in the set $se(3)$ which is

$$\hat{\mathbf{w}} = \begin{pmatrix} (\boldsymbol{\omega}_{ab}^b)^\times & \mathbf{v}_{ab}^b \\ \mathbf{0}^T & 0 \end{pmatrix} \in se(3) \tag{6.306}$$

The time derivative of the homogeneous transformation matrix is given by

$$\dot{\mathbf{T}}_b^a = \mathbf{T}_b^a \hat{\mathbf{w}} \tag{6.307}$$

This topic is treated in great detail in (Murray, Li and Sastry 1994).

Example 109 *The transformation rule for a twist vector is not as straightforward as for the angular velocity vector. This is seen in the time derivative of* $\mathbf{T}_b^a$ *when it expressed in the a frame:*

$$\begin{aligned} \dot{\mathbf{T}}_b^a &= \begin{pmatrix} (\boldsymbol{\omega}_{ab}^a)^\times \mathbf{R}_b^a & \mathbf{v}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{\omega}_{ab}^a)^\times & \mathbf{v}_{ab}^a - (\boldsymbol{\omega}_{ab}^a)^\times \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R}_b^a & \mathbf{r}_{ab}^a \\ \mathbf{0}^T & 1 \end{pmatrix} \end{aligned} \tag{6.308}$$

The physical interpretation of the velocity term $\mathbf{v}_{ab}^a - (\boldsymbol{\omega}_{ab}^a)^\times \mathbf{r}_{ab}^a$ *is not as obvious as when the coordinates of the b frame is used. A geometric interpretation is given in (Murray et al. 1994).*

6.9.4 Euler angles

When Euler angles are used the rotation matrix $\mathbf{R}_d^a$ from frame a to frame d is a composite rotation involving three simple rotations. In the roll-pitch-yaw case the simple rotations are

$$\mathbf{R}_b^a = \mathbf{R}_z(\psi), \quad \mathbf{R}_c^b = \mathbf{R}_y(\theta) \quad \text{and} \quad \mathbf{R}_d^c = \mathbf{R}_x(\phi) \tag{6.309}$$

We see that the angular velocities associated with the simple rotations are

$$\boldsymbol{\omega}_{ab}^a = \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix}, \quad \boldsymbol{\omega}_{bc}^b = \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\omega}_{cd}^c = \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \tag{6.310}$$

From (6.269) we have that the angular velocity of d relative to a is the sum of the angular velocities resulting from each of the three simple rotations due to ψ, θ and ϕ:

$$\vec{\omega}_{ad} = \vec{\omega}_{ab} + \vec{\omega}_{bc} + \vec{\omega}_{cd} \tag{6.311}$$

In the a frame this gives:

$$\begin{aligned} \boldsymbol{\omega}_{ad}^a &= \begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix} + \mathbf{R}_{z,\psi} \begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} + \mathbf{R}_{z,\psi} \mathbf{R}_{y,\theta} \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -\sin\psi\dot{\theta} + \cos\psi\cos\theta\dot{\phi} \\ \cos\psi\dot{\theta} + \sin\psi\cos\theta\dot{\phi} \\ \dot{\psi} - \sin\theta\dot{\phi} \end{pmatrix} \end{aligned} \tag{6.312}$$

In the d frame we find

$$\boldsymbol{\omega}_{ad}^{d} = \mathbf{R}_{x,-\phi}\mathbf{R}_{y,-\theta}\begin{pmatrix} 0 \\ 0 \\ \dot{\psi} \end{pmatrix} + \mathbf{R}_{x,-\phi}\begin{pmatrix} 0 \\ \dot{\theta} \\ 0 \end{pmatrix} + \begin{pmatrix} \dot{\phi} \\ 0 \\ 0 \end{pmatrix} \tag{6.313}$$

$$= \begin{pmatrix} -\sin\theta\dot{\psi} + \dot{\phi} \\ \sin\phi\cos\theta\dot{\psi} + \cos\phi\dot{\theta} \\ \cos\phi\cos\theta\dot{\psi} - \sin\phi\dot{\theta} \end{pmatrix} \tag{6.314}$$

Define the vector $\boldsymbol{\phi} = (\phi, \theta, \psi)^T$. We can then write

$$\boldsymbol{\omega}_{ad}^{a} = \mathbf{E}_a(\boldsymbol{\phi})\dot{\boldsymbol{\phi}} = \begin{pmatrix} \cos\psi\cos\theta & -\sin\psi & 0 \\ \sin\psi\cos\theta & \cos\psi & 0 \\ -\sin\theta & 0 & 1 \end{pmatrix}\dot{\boldsymbol{\phi}} \tag{6.315}$$

and

$$\boldsymbol{\omega}_{ad}^{d} = \mathbf{E}_d(\boldsymbol{\phi})\dot{\boldsymbol{\phi}} = \begin{pmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\phi & \sin\phi\cos\theta \\ 0 & -\sin\phi & \cos\phi\cos\theta \end{pmatrix}\dot{\boldsymbol{\phi}}. \tag{6.316}$$

We note that $\det[\mathbf{E}_a(\boldsymbol{\phi})] = \det[\mathbf{E}_d(\boldsymbol{\phi})] = \cos\theta$ which implies that the matrices are singular for $\cos\theta = 0$.

We can solve for $\dot{\boldsymbol{\phi}}$, and find that

$$\dot{\boldsymbol{\phi}} = \mathbf{E}_a(\boldsymbol{\phi})^{-1}\boldsymbol{\omega}_{ad}^{a} = \frac{1}{\cos\theta}\begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi\cos\theta & \cos\psi\cos\theta & 0 \\ \cos\psi\sin\theta & \sin\psi\sin\theta & \cos\theta \end{pmatrix}\boldsymbol{\omega}_{ad}^{a} \tag{6.317}$$

and

$$\dot{\boldsymbol{\phi}} = \mathbf{E}_d(\boldsymbol{\phi})^{-1}\boldsymbol{\omega}_{ad}^{d} = \frac{1}{\cos\theta}\begin{pmatrix} \cos\theta & \sin\phi\sin\theta & \cos\phi\sin\theta \\ 0 & \cos\phi\cos\theta & -\sin\phi\cos\theta \\ 0 & \sin\phi & \cos\phi \end{pmatrix}\boldsymbol{\omega}_{ad}^{d} \tag{6.318}$$

Let $\vec{a}_i$ be the orthogonal unit vectors of the a frame, $\vec{b}_i$ be the unit vectors of the b frame, and let $\vec{c}_i$ be the orthogonal unit vectors of the c frame. Then the roll-pitch-yaw description gives the angular velocity $\vec{\omega}_{ad}$ as a sum of an angular velocity $\vec{\omega}_{ab}$ along $\vec{a}_3$, an angular velocity $\vec{\omega}_{bc}$ along $\vec{b}_2$, and an angular velocity $\vec{\omega}_{cd}$ along $\vec{c}_1$. The physical interpretation of the singularity of $\mathbf{E}_a(\boldsymbol{\phi})$ and $\mathbf{E}_d(\boldsymbol{\phi})$ at $\cos\theta = 0$ is due to the fact that when $\cos\theta = 0$, then the rotation vectors $\vec{a}_3$ and $\vec{c}_1$ align so that both $\vec{\omega}_{ab}$ and $\vec{\omega}_{cd}$ are along the $\vec{a}_3$ vector while $\vec{\omega}_{bc}$ is along the $\vec{b}_2$ vector. This means that it is not possible to describe an angular velocity along $\vec{a}_3 \times \vec{b}_2$ when $\cos\theta = 0$. This is the *Euler-angle singularity*, which is a singularity due to the mathematical representation of the rotation matrix.

6.9.5 Euler parameters

In this section we will derive the kinematic differential equations for the Euler parameters. These differential equations give the time derivatives of the Euler parameters as functions of the angular velocity. We let $\mathbf{R} := \mathbf{R}_b^a$ and $\boldsymbol{\omega} := \boldsymbol{\omega}_{ab}$ so that $\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^\times\mathbf{R}$. Moreover we let $\mathbf{R} = \mathbf{R}_e(\eta, \boldsymbol{\epsilon})$ and $\mathbf{p} = (\eta \quad \boldsymbol{\epsilon}^T)^T$. We then have

$$\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^\times\mathbf{R} = \mathbf{R}(\boldsymbol{\omega}^b)^\times \tag{6.319}$$

The derivation is based on the coordinate transformation rule using the quaternion product. For an arbitrary vector $\mathbf{u} \in R^3$ we have

$$\begin{pmatrix} 0 \\ \mathbf{Ru} \end{pmatrix} = \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} \tag{6.320}$$

We take the time derivative of both sides and get

$$\begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{R}\dot{\mathbf{u}} \end{pmatrix} = \dot{\mathbf{p}} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \dot{\mathbf{u}} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \dot{\bar{\mathbf{p}}} \tag{6.321}$$

Then, because the transformation rule in (6.320) is valid for any vector it is also valid for $\dot{\mathbf{u}}$. This implies that

$$\begin{aligned} \begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} &= \dot{\mathbf{p}} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} + \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \dot{\bar{\mathbf{p}}} \\ &= \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \otimes \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} - \mathbf{p} \otimes \begin{pmatrix} 0 \\ \mathbf{u} \end{pmatrix} \otimes \bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \otimes \bar{\mathbf{p}} \\ &= (\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}) \otimes \begin{pmatrix} 0 \\ \mathbf{Ru} \end{pmatrix} - \begin{pmatrix} 0 \\ \mathbf{Ru} \end{pmatrix} \otimes (\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}) \\ &= 2 \begin{pmatrix} 0 \\ (\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\times}\dot{\boldsymbol{\epsilon}})^{\times} \mathbf{Ru} \end{pmatrix} \end{aligned} \tag{6.322}$$

where we have used (6.197), (6.191), (6.193), (6.194) and (6.173). From $\dot{\mathbf{R}} = (\boldsymbol{\omega}^a)^{\times}\mathbf{R}$ we have

$$\begin{pmatrix} 0 \\ \dot{\mathbf{R}}\mathbf{u} \end{pmatrix} = \begin{pmatrix} 0 \\ (\boldsymbol{\omega}^a)^{\times}\mathbf{Ru} \end{pmatrix} \tag{6.323}$$

Comparing this with (6.322) we find that the angular velocity $\boldsymbol{\omega}^a$ is given by

$$\boldsymbol{\omega}^a = 2[\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\times}\dot{\boldsymbol{\epsilon}}] \tag{6.324}$$

From (6.194) it is seen that this can be written in quaternion form, and this leads to the result

The angular velocity is given in frames a and b by

$$\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = 2\dot{\mathbf{p}} \otimes \bar{\mathbf{p}}, \quad \begin{pmatrix} 0 \\ \boldsymbol{\omega}^b \end{pmatrix} = 2\bar{\mathbf{p}} \otimes \dot{\mathbf{p}} \tag{6.325}$$

and the kinematic differential equation for the quaternion vector is

$$\dot{\mathbf{p}} = \frac{1}{2} \begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} \otimes \mathbf{p}, \quad \dot{\mathbf{p}} = \frac{1}{2}\mathbf{p} \otimes \begin{pmatrix} 0 \\ \boldsymbol{\omega}^b \end{pmatrix} \tag{6.326}$$

Here the transformation rule (6.320) has been used, and the kinematic differential equations appear by postmultiplication with $\mathbf{p}$ for the expression in the a frame, and by premultiplication with $\mathbf{p}$ for the expression in the b frame.

The component form of these last four equations gives the result

$$\boldsymbol{\omega}^b = 2[\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^{\times}\dot{\boldsymbol{\epsilon}}] \tag{6.327}$$

$$\boldsymbol{\omega}^a = 2[\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\times}\dot{\boldsymbol{\epsilon}}] \tag{6.328}$$

$$\dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\omega}^b \tag{6.329}$$

$$\dot{\boldsymbol{\epsilon}} = \frac{1}{2}[\eta\mathbf{I} + \boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^b \tag{6.330}$$

and

$$\dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\omega}^a \tag{6.331}$$

$$\dot{\boldsymbol{\epsilon}} = \frac{1}{2}[\eta\mathbf{I} - \boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^a \tag{6.332}$$

Example 110 *From (6.174), (6.175) and (6.329–6.332) it is seen that the kinematic differential equations can be written in vector form as*

$$\dot{\mathbf{p}} = \frac{1}{2}\begin{pmatrix} 0 & -(\boldsymbol{\omega}^a)^T \\ \boldsymbol{\omega}^a & (\boldsymbol{\omega}^a)^\times \end{pmatrix}\mathbf{p} = \frac{1}{2}\begin{pmatrix} 0 & -(\boldsymbol{\omega}^b)^T \\ \boldsymbol{\omega}^b & -(\boldsymbol{\omega}^b)^\times \end{pmatrix}\mathbf{p} \tag{6.333}$$

or

$$\dot{\mathbf{p}} = \frac{1}{2}\begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta\mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta\mathbf{I} - \boldsymbol{\epsilon}^\times \end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} \tag{6.334}$$

6.9.6 Normalization for numerical integration

From (6.334) it is seen that

$$\frac{d}{dt}\left(\mathbf{p}^T\mathbf{p}\right) = \mathbf{p}^T\begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta\mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} = 0 \tag{6.335}$$

This shows that if $\mathbf{p}$ is initialized as a unit vector, then it will remain a unit vector, as should be expected. Numerical integration of the quaternion vector $\mathbf{p}$ from the kinematic differential equation will introduce numerical errors that will cause the length of $\mathbf{p}$ to deviate from unity. To compensate for such errors a normalization term is added to the kinematic differential equation. This can be done with the following modification of the kinematic differential equation, which should be used in numerical integration:

$$\dot{\mathbf{p}} = \frac{1}{2}\begin{pmatrix} \eta & -\boldsymbol{\epsilon}^T \\ \boldsymbol{\epsilon} & \eta\mathbf{I} + \boldsymbol{\epsilon}^\times \end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{\omega}^a \end{pmatrix} + \frac{\lambda}{2}(1 - \mathbf{p}^T\mathbf{p})\mathbf{p} \tag{6.336}$$

Here λ is a positive gain. Then

$$\frac{d}{dt}\left(\mathbf{p}^T\mathbf{p}\right) = \frac{\lambda}{2}(1 - \mathbf{p}^T\mathbf{p})\mathbf{p}^T\mathbf{p} \tag{6.337}$$

We see that this will give the desired result as $\mathbf{p}^T\mathbf{p}$ will increase whenever $\mathbf{p}^T\mathbf{p} < 1$, and $\mathbf{p}^T\mathbf{p}$ will decrease whenever $\mathbf{p}^T\mathbf{p} > 1$. When $\mathbf{p}^T\mathbf{p} = 1$ the usual kinematic differential equations are recovered. Linearization about $\mathbf{p}^T\mathbf{p} = 1$ gives $\dot{e} = -\lambda e$ where $e = 1 - \mathbf{p}^T\mathbf{p}$. This means that the normalization converges with a time constant $T = \lambda^{-1}$. A Simulink toolbox has implemented this algorithm with $\lambda = 100$, which means that the normalization converges with a time constant of 0.01 s.

Another alternative is to normalize directly after each time step using the normalization assignment

$$\mathbf{p} := \frac{\mathbf{p}}{\sqrt{\mathbf{p}^T\mathbf{p}}} \tag{6.338}$$

6.9.7 Euler rotation

As $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$, it is seen that $\mathbf{e} = \mathbf{k}\sin\theta$ can be expressed by the Euler parameters according to

$$\mathbf{e} = 2\eta\boldsymbol{\epsilon}. \tag{6.339}$$

The kinematic differential equation is then found from

$$\dot{\mathbf{e}} = 2(\dot{\eta}\boldsymbol{\epsilon} + \eta\dot{\boldsymbol{\epsilon}}). \tag{6.340}$$

This gives

$$\begin{aligned} \dot{\mathbf{e}} &= [\eta^2\mathbf{I} - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - \eta\boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^a & (6.341)\\ &= [\eta^2\mathbf{I} - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + \eta\boldsymbol{\epsilon}^\times]\boldsymbol{\omega}^b & (6.342) \end{aligned}$$

where (6.331) and (6.332) are used.

Alternative expressions are found from

$$\mathbf{R} = (2\eta^2 - 1)\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T + 2\eta\boldsymbol{\epsilon}^\times \tag{6.343}$$

and

$$\mathbf{R}^T = (2\eta^2 - 1)\mathbf{I} + 2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T - 2\eta\boldsymbol{\epsilon}^\times \tag{6.344}$$

which leads to

$$\begin{aligned} \dot{\mathbf{e}} &= \frac{1}{2}[\mathrm{Trace}(\mathbf{R})\mathbf{I} - \mathbf{R}]\boldsymbol{\omega}^a & (6.345)\\ \dot{\mathbf{e}} &= \frac{1}{2}[\mathrm{Trace}(\mathbf{R}^T)\mathbf{I} - \mathbf{R}^T]\boldsymbol{\omega}^b & (6.346) \end{aligned}$$

where we have used that $4\eta^2 - 1 = \mathrm{Trace}(\mathbf{R})$.

Note that for $\theta = 0$ we have

$$\dot{\mathbf{e}}|_{\theta=0} = \boldsymbol{\omega}^a = \boldsymbol{\omega}^b \tag{6.347}$$

6.9.8 Euler-Rodrigues parameters

The kinematic differential equation is derived from

$$\dot{\boldsymbol{\rho}} = \frac{d}{dt}\frac{\boldsymbol{\epsilon}}{\eta} = \frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} \tag{6.348}$$

which gives

$$\dot{\boldsymbol{\rho}} = \frac{1}{2}[\mathbf{I} + \boldsymbol{\rho}^\times + \boldsymbol{\rho}\boldsymbol{\rho}^T]\,\boldsymbol{\omega}^b. \tag{6.349}$$

An equation for the angular velocity is found from

$$\boldsymbol{\rho}^\times\dot{\boldsymbol{\rho}} = \frac{1}{\eta}\boldsymbol{\epsilon}^\times\frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} = \frac{1}{\eta^2}\boldsymbol{\epsilon}^\times\dot{\boldsymbol{\epsilon}} \tag{6.350}$$

as $\boldsymbol{\epsilon}^\times\boldsymbol{\epsilon} = \mathbf{0}$. From (6.327) it is seen that

$$\boldsymbol{\omega}^b = 2\eta^2\left(\frac{\eta\dot{\boldsymbol{\epsilon}} - \dot{\eta}\boldsymbol{\epsilon}}{\eta^2} - \frac{1}{\eta^2}\boldsymbol{\epsilon}^\times\dot{\boldsymbol{\epsilon}}\right) \tag{6.351}$$

Insertion of (6.244), (6.348) and (6.350) gives

$$\boldsymbol{\omega}^b = \frac{2}{1 + \boldsymbol{\rho}^T\boldsymbol{\rho}}[\mathbf{I} - \boldsymbol{\rho}^\times]\dot{\boldsymbol{\rho}} \tag{6.352}$$

Example 111 *Equation (6.245) can be written*

$$\mathbf{R} = \frac{1}{1+\boldsymbol{\rho}^T\boldsymbol{\rho}}[2\mathbf{I} + 2\boldsymbol{\rho}^\times + 2\boldsymbol{\rho}\boldsymbol{\rho}^T - (1+\boldsymbol{\rho}^T\boldsymbol{\rho})\mathbf{I}] \tag{6.353}$$

which implies that

$$\mathbf{I} + \boldsymbol{\rho}^\times + \boldsymbol{\rho}\boldsymbol{\rho}^T = \frac{1+\boldsymbol{\rho}^T\boldsymbol{\rho}}{2}(\mathbf{R}+\mathbf{I}) \tag{6.354}$$

Then (6.349) can be written

$$\dot{\boldsymbol{\rho}} = \frac{1+\boldsymbol{\rho}^T\boldsymbol{\rho}}{4}(\mathbf{R}+\mathbf{I})\,\boldsymbol{\omega}^b \tag{6.355}$$

We recall from (6.169) that

$$\text{Trace}\mathbf{R} + 1 = 4\eta^2 \tag{6.356}$$

and, using (6.244), we arrive at

$$\dot{\boldsymbol{\rho}} = \frac{1}{\text{Trace}\mathbf{R}+1}(\mathbf{R}+\mathbf{I})\,\boldsymbol{\omega}^b \tag{6.357}$$

6.9.9 Passivity of kinematic differential equations

In translational dynamics the integration from velocity $\mathbf{v} = \dot{\mathbf{x}}$ to position $\mathbf{x}$ is a passive dynamic system. In fact, the function $V_x = \frac{1}{2}\mathbf{x}^T\mathbf{x} \geq 0$ has time derivative

$$\dot{V}_x = \frac{\partial V_x}{\partial \mathbf{x}}\dot{\mathbf{x}} = \mathbf{x}^T\mathbf{v} \tag{6.358}$$

so that the system with input $\mathbf{v}$ and output $\mathbf{x}$ is clearly passive. It is interesting to investigate if similar results can be established for rotational dynamics.

The starting point for such an investigation (Egeland and Godhavn 1994) is the differential equation

$$\dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\omega} \tag{6.359}$$

where $|\eta| = \left|\cos\frac{\theta}{2}\right| \leq 1$. Define

$$V_\epsilon = 2\,(1-\eta) \geq 0 \tag{6.360}$$

The time derivative for solutions of the kinematic differential equations of the Euler parameters is

$$\dot{V}_\epsilon = -2\dot{\eta} = \boldsymbol{\epsilon}^T\boldsymbol{\omega} \tag{6.361}$$

It follows that the kinematic system with input $\boldsymbol{\omega}$ and output $\boldsymbol{\epsilon}$ is passive.

At this stage it is not very difficult to extend this result to other kinematic representations based on the Euler parameters. First we note that if we multiply the equation for $\dot{\eta}$ by η, we get

$$\eta\dot{\eta} = -\frac{1}{2}\eta\boldsymbol{\epsilon}^T\boldsymbol{\omega} = -\frac{1}{4}\mathbf{e}^T\boldsymbol{\omega} \tag{6.362}$$

where $\mathbf{e} = 2\eta\boldsymbol{\epsilon}$ is the Euler rotation vector. We are then lead to the function

$$V_e = 2\left(1-\eta^2\right) \geq 0 \tag{6.363}$$

which has time derivative

$$\dot{V}_e = -4\eta\dot{\eta} = \mathbf{e}^T\boldsymbol{\omega} \tag{6.364}$$

and we have shown that the kinematic system with input $\boldsymbol{\omega}$ and output $\mathbf{e}$ is passive.

Finally we note that

$$\frac{\dot{\eta}}{\eta} = -\frac{1}{2}\frac{\boldsymbol{\epsilon}^T}{\eta}\boldsymbol{\omega} = -\frac{1}{2}\boldsymbol{\rho}^T\boldsymbol{\omega} \tag{6.365}$$

Define

$$V_\rho = -2\ln|\eta| \geq 0, \quad \eta \neq 0 \tag{6.366}$$

which is defined for all η except for $\eta = 0$ where also $\boldsymbol{\rho}$ is undefined. Then

$$\dot{V}_\rho = -2\frac{\dot{\eta}}{\eta} = \boldsymbol{\rho}^T\boldsymbol{\omega} \tag{6.367}$$

and it is seen that the kinematic system with input $\boldsymbol{\omega}$ and output $\boldsymbol{\rho}$ is passive.

Example 112 *It is interesting to note that*

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + (1-\eta)^2 = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \eta^2 - 2\eta + 1 = 2(1-\eta) = V_\epsilon \tag{6.368}$$

and that

$$2\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = 2\left(1-\eta^2\right) = V_e \tag{6.369}$$

Example 113 *Consider a system with equation of motion*

$$\mathbf{M}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^\times\mathbf{M}\boldsymbol{\omega} = \boldsymbol{\tau} \tag{6.370}$$

where $\mathbf{M}$ *is a constant, symmetric and positive definite matrix, and where the input is selected to be*

$$\boldsymbol{\tau} = -\mathbf{K}_d\boldsymbol{\omega} - k_p\boldsymbol{\epsilon} \tag{6.371}$$

where $\mathbf{K}_d$ *is a constant, symmetric and positive definite matrix, and* k_p *is a positive constant. The energy function*

$$V = \frac{1}{2}\boldsymbol{\omega}^T\mathbf{M}\boldsymbol{\omega} + 2k_p(1-\eta) \geq 0 \tag{6.372}$$

has time derivative

$$\begin{aligned}\dot{V} &= \boldsymbol{\omega}^T\left(-\boldsymbol{\omega}^\times\mathbf{M}\boldsymbol{\omega} - \mathbf{K}_d\boldsymbol{\omega} - k_p\epsilon\right) + k_p\boldsymbol{\epsilon}^T\boldsymbol{\omega} \\ &= -\boldsymbol{\omega}^T\mathbf{K}_d\boldsymbol{\omega}\end{aligned} \tag{6.373}$$

along the solutions of the system. This means that the energy of the system decreases whenever $\boldsymbol{\omega} \neq \mathbf{0}$*. For further details see (Wen and Kreutz-Delgado 1991), where a cross-term was added to the energy function, and (Egeland and Godhavn 1994).*

6.9.10 Angle-axis representation

The kinematic differential equations for the angle θ and the unit vector $\mathbf{k}$ in the angle axis representation of the rotation matrix are derived in this section. The derivation is based on differentiation of the Euler parameters. To find the equation for $\dot{\theta}$ we observe that

$$-\frac{1}{2}\sin\left(\frac{\theta}{2}\right)\dot{\theta} = \dot{\eta} = -\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\omega} = -\frac{1}{2}\sin\left(\frac{\theta}{2}\right)\mathbf{k}^T\boldsymbol{\omega}$$

Whenever $\sin(\theta/2) \neq 0$, this implies

$$\dot{\theta} = \mathbf{k}^T \boldsymbol{\omega} \tag{6.374}$$

To find the equation for $\dot{\mathbf{k}}$ we note that

$$\dot{\boldsymbol{\epsilon}} = \left(\frac{d}{dt}\sin\frac{\theta}{2}\right)\mathbf{k} + \sin\frac{\theta}{2}\dot{\mathbf{k}} = \frac{1}{2}\cos\left(\frac{\theta}{2}\right)\dot{\theta}\mathbf{k} + \sin\frac{\theta}{2}\dot{\mathbf{k}}$$

Combining this with the kinematic differential equations of the Euler parameters, we get

$$\frac{1}{2}\left[\eta\mathbf{I} + \boldsymbol{\epsilon}^\times\right]\boldsymbol{\omega} = \frac{1}{2}\cos\left(\frac{\theta}{2}\right)\dot{\theta}\mathbf{k} + \sin\frac{\theta}{2}\dot{\mathbf{k}}$$

which gives

$$\begin{aligned} 2\sin\frac{\theta}{2}\dot{\mathbf{k}} &= \eta\left(\mathbf{I} - \mathbf{k}\mathbf{k}^T\right)\boldsymbol{\omega} + \boldsymbol{\epsilon}^\times\boldsymbol{\omega} \\ &= \cos\frac{\theta}{2}\left(\mathbf{I} - \mathbf{k}\mathbf{k}^T\right)\boldsymbol{\omega} + \sin\frac{\theta}{2}\mathbf{k}^\times\boldsymbol{\omega} \end{aligned}$$

Then the kinematic differential equation for $\mathbf{k}$ is found using $\mathbf{k}^\times\mathbf{k}^\times = \mathbf{k}\mathbf{k}^T - \mathbf{I}$. Whenever $\sin(\theta/2) \neq 0$ the result is

$$\dot{\mathbf{k}} = \frac{1}{2}\left[\mathbf{k}^\times - \mathbf{k}^\times\mathbf{k}^\times\cot\frac{\theta}{2}\right]\boldsymbol{\omega} \tag{6.375}$$

The equations (6.374) and (6.375) have a singularity at $\theta = 0$, which is in agreement with the fact that $\mathbf{k}$ is undefined for a zero rotation $\theta = 0$.

6.10 The Serret-Frenet frame

6.10.1 Kinematics

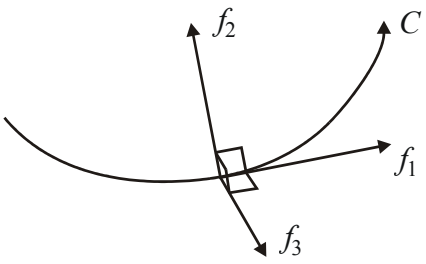

Figure 6.8: The Serret-Frenet frame for a curve C.

In aerospace, automotive steering and ship control the desired trajectory of the system may be given as a curve in a fixed frame i. The control deviations from the desired curve to the actual configuration of the system can then be calculated in the *Serret-Frenet frame* f. This frame has axes along the tangent, the normal and the binormal of the curve as shown in Figure 6.8. We will develop the equations for this frame in this section.

6.12 Kinematics of a rigid body

6.12.1 Configuration

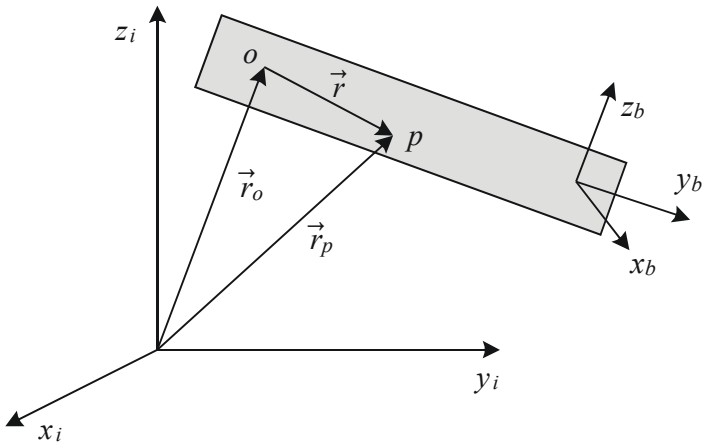

Figure 6.11: Rigid body b with the fixed frame b and the fixed points o and p.

The configuration of a rigid body defines the position of all points in the rigid body. For a rigid body the configuration can be specified in terms of the position $\vec{r}_o$ of one fixed point in the rigid body, and the rotation matrix $\mathbf{R}_b^i$ from a reference frame i to a body-fixed frame b. Then the position of any point p in the rigid body, which is not necessarily fixed in the rigid body, is given by

$$\vec{r}_p = \vec{r}_o + \vec{r} \tag{6.397}$$

as shown in Figure 6.11.

Here $\vec{r}$ is the vector from o to p with coordinate vector $\mathbf{r}^b$ in the b frame. This vector is given in the i frame by

$$\mathbf{r}^i = \mathbf{R}_b^i \mathbf{r}^b \tag{6.398}$$

6.12.2 Velocity

The frame i is assumed to be an *inertial frame* which is also referred to as a *Newtonian frame*. The velocities of o and p are given by

$$\vec{v}_o := \frac{{}^i d}{dt}\vec{r}_o, \quad \vec{v}_p := \frac{{}^i d}{dt}\vec{r}_p \tag{6.399}$$

From (6.397) and the rule for differentiation in moving frames it is seen that the velocity of p can be expressed as

$$\vec{v}_p = \vec{v}_o + \frac{{}^b d}{dt}\vec{r} + \vec{\omega}_{ib} \times \vec{r} \tag{6.400}$$

6.12.3 Acceleration

The acceleration vectors are defined by

$$\vec{a}_p := \frac{{}^i d^2}{dt^2}\vec{r}_p, \quad \vec{a}_o := \frac{{}^i d^2}{dt^2}\vec{r}_o \tag{6.401}$$

while the angular acceleration vector is defined by

$$\vec{\alpha}_{ib} := \frac{^i d}{dt}\vec{\omega}_{ib} = \frac{^b d}{dt}\vec{\omega}_{ib} \tag{6.402}$$

where the second equality is a consequence of

$$\frac{^i d}{dt}\vec{\omega}_{ib} = \frac{^b d}{dt}\vec{\omega}_{ib} + \vec{\omega}_{ib} \times \vec{\omega}_{ib} = \frac{^b d}{dt}\vec{\omega}_{ib} \tag{6.403}$$

In the formulation of the equations of motion for rigid bodies, we need the following result:

$$\begin{aligned}
\frac{^i d^2}{dt^2}\vec{r}_p &= \frac{^i d^2}{dt^2}\vec{r}_o + \frac{^i d}{dt}\left(\frac{^i d}{dt}\vec{r}\right) = \frac{^i d^2}{dt^2}\vec{r}_o + \frac{^i d}{dt}\left(\frac{^b d}{dt}\vec{r} + \vec{\omega}_{ib} \times \vec{r}\right) \\
&= \frac{^i d^2}{dt^2}\vec{r}_o + \frac{^b d}{dt}\left(\frac{^b d}{dt}\vec{r} + \vec{\omega}_{ib} \times \vec{r}\right) + \vec{\omega}_{ib} \times \left(\frac{^b d}{dt}\vec{r} + \vec{\omega}_{ib} \times \vec{r}\right) \\
&= \frac{^i d^2}{dt^2}\vec{r}_o + \frac{^b d^2}{dt^2}\vec{r} + 2\vec{\omega}_{ib} \times \frac{^b d}{dt}\vec{r} + \left(\frac{^b d}{dt}\vec{\omega}_{ib}\right) \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r})
\end{aligned} \tag{6.404}$$

In terms of acceleration, angular acceleration and velocities this is written

$$\begin{aligned}
\underbrace{\vec{a}_p}_{\substack{\text{Acceleration} \\ \text{of } p}} &= \underbrace{\vec{a}_o}_{\substack{\text{Acceleration} \\ \text{of } o}} + \underbrace{\frac{^b d^2}{dt^2}\vec{r}}_{\substack{\text{Second derivative} \\ \text{of } \vec{r} \text{ in } b}} \\
&+ \underbrace{2\vec{\omega}_{ib} \times \frac{^b d}{dt}\vec{r}}_{\substack{\text{Coriolis} \\ \text{acceleration}}} + \underbrace{\vec{\alpha}_{ib} \times \vec{r}}_{\substack{\text{Transversal} \\ \text{acceleration}}} + \underbrace{\vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r})}_{\substack{\text{Centripetal} \\ \text{acceleration}}}
\end{aligned} \tag{6.405}$$

An alternative formulation is obtained by inserting the expression

$$\vec{a}_o = \frac{^i d}{dt}\vec{v}_o = \frac{^b d}{dt}\vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o \tag{6.406}$$

which gives

$$\vec{a}_p = \frac{^b d}{dt}\vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \frac{^b d^2}{dt^2}\vec{r} + 2\vec{\omega}_{ib} \times \frac{^b d}{dt}\vec{r} + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}) \tag{6.407}$$

Note the difference between the term $\vec{\omega}_{ib} \times \vec{v}_o$ which is related to the velocity of o, and the Coriolis acceleration $2\vec{\omega}_{ib} \times \frac{^b d}{dt}\vec{r}$ which is related to the motion of p in the b frame relative to o.

If the point p is fixed in the body b, then the vector $\vec{r}$ is constant in frame b so that

$$\frac{^b d}{dt}\vec{r} = \vec{0} \Rightarrow \frac{^i d}{dt}\vec{r} = \vec{\omega}_{ib} \times \vec{r}, \quad \vec{r} \text{ fixed in } b \tag{6.408}$$

For $\vec{v}_p$ this gives

$$\vec{v}_p = \vec{v}_o + \vec{\omega}_{ib} \times \vec{r}, \quad \vec{r} \text{ fixed in } b \tag{6.409}$$

The acceleration is found to be

$$\vec{a}_p = \vec{a}_o + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}), \quad \vec{r} \text{ fixed in } b \tag{6.410}$$

We see that in this case there is no Coriolis acceleration. Using (6.406) the acceleration can be written

$$\vec{a}_p = \frac{^b d}{dt}\vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \vec{\alpha}_{ib} \times \vec{r} + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}), \quad \vec{r} \text{ fixed in } b \tag{6.411}$$

6.13 The center of mass

6.13.1 System of particles

Consider a system of N particles each of mass m_k and with position $\vec{r}_k$ relative to the origin of the inertial frame i. The *center of mass* is the point with position $\vec{r}_c$ defined by

$$m\vec{r}_c = \sum_{k=1}^{N} m_k \vec{r}_k \tag{6.412}$$

where $m = \sum_{k=1}^{N} m_k$ is the sum of the mass of the particles. The velocity $\vec{v}_c$ and the acceleration $\vec{a}_c$ of the center of mass are defined by

$$m\vec{v}_c = m\frac{^i d\vec{r}_c}{dt} = \sum_{k=1}^{N} m_k \vec{v}_k \tag{6.413}$$

$$m\vec{a}_c = m\frac{^i d^2\vec{r}_c}{dt^2} = \sum_{k=1}^{N} m_k \vec{a}_k \tag{6.414}$$

6.13.2 Rigid body

The position $\vec{r}_c$ of the center of mass of a rigid body b is defined by

$$m\vec{r}_c = \int_b \vec{r}_p dm \tag{6.415}$$

where $m = \int_b dm$ is the mass of the rigid body, and $\vec{r}_p$ is the position of a mass element dm which is fixed in frame b. The position of the mass element relative to the center of mass is given by $\vec{r}$ so that

$$\vec{r}_p = \vec{r}_c + \vec{r} \tag{6.416}$$

as shown in Figure 6.12. From the definition of the center of mass we see that

$$\int_b \vec{r} dm = \int_b \vec{r}_p dm - m\vec{r}_c = \vec{0} \tag{6.417}$$

The velocity $\vec{v}_c$ of the center of mass is given by

$$m\vec{v}_c = m\frac{^i d\vec{r}_c}{dt} = \frac{^i d}{dt}\int_b \vec{r}_p dm = \int_b \frac{^i d\vec{r}_p}{dt} dm = \int_b \vec{v}_p dm \tag{6.418}$$

while the acceleration $\vec{a}_c$ of the center of mass is found in the same way. We conclude that

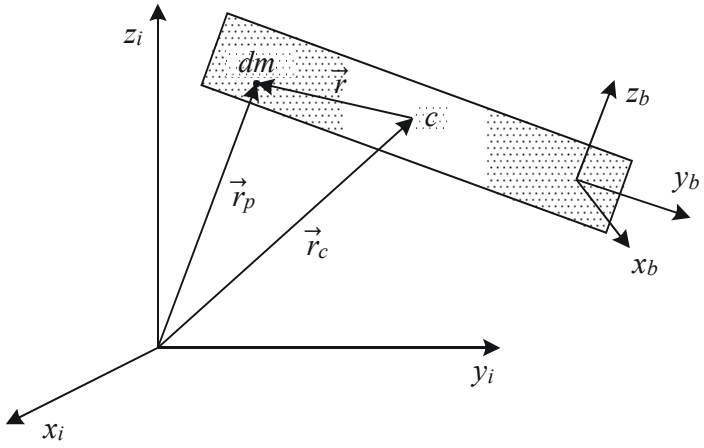

Figure 6.12: Mass element dm relative to the center of mass c.

The motion of the mass center in the rigid body b satisfies the equations

$$m\vec{r}_c = \int_b \vec{r}_p dm, \quad m\vec{v}_c = \int_b \vec{v}_p dm, \quad m\vec{a}_c = \int_b \vec{a}_p dm \tag{6.419}$$

Chapter 7

Newton-Euler equations of motion

7.1 Introduction

The development of the equations of motions for rigid bodies and systems of rigid bodies is the topic of this chapter. The equations of motion are differential equations for the velocity and angular velocity. The derivations in this chapter are based on Newton's law and its extension to rotational dynamics, which is usually attributed to Euler. This provides the motivation for the term Newton-Euler equation of motion. The derivations rely on vector operations. The presentation starts with some results on forces and torques on rigid bodies. Then the basic Newton-Euler equations of motion are presented and used to derive the equations of motion for the ball-and-beam system, the Furuta pendulum and the inverted pendulum. Then the principle of virtual work is presented, and its use is demonstrated for multi-body systems. The use of recursive computations in manipulator dynamics is also discussed.

7.2 Forces and torques

To derive the equations of motions for rigid bodies we need some results on resultant forces and moments, which are presented in this section. The material is taken from (Kane and Levinson 1985).

7.2.1 Resultant force

A force vector $\vec{F}$ will have a *line of action*, which means that the moment of $\vec{F}$ about a point P is $\vec{r}\times\vec{F}$ where $\vec{r}$ is the position vector from P to some arbitrary point on the line of action as shown in Figure 7.1. A vector with a line of action is called a *bound vector*.

Consider a set S of n_F forces $\vec{F}_j$. The resultant force $\vec{F}_S^{(r)}$ of the set S is the vector

$$\vec{F}_S^{(r)} = \sum_{j=1}^{n_F} \vec{F}_j \tag{7.1}$$

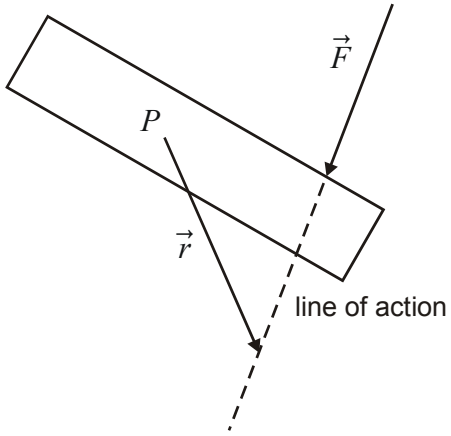

Figure 7.1: A force $\vec{F}$ acting on a rigid body. The line of action of the force is indicated as a dashed line, and the distance $\vec{r}$ from a point P is shown.

while the moment about P of the set S of forces is

$$\vec{N}_{S/P} = \sum_{j=1}^{n_F} \vec{r}_{Pj} \times \vec{F}_j \tag{7.2}$$

where $\vec{r}_{Pj}$ is the position vector from P to an arbitrary point on the line of action of $\vec{F}_j$. Note that in this description the resultant $\vec{F}_S^{(r)}$ is a sum of forces, and can not be considered to be a force with a line of action. Note in particular that the resultant force does not appear in the expression for the moment $\vec{N}_{S/P}$. The moment $\vec{N}_{S/Q}$ about some other point Q is found from

$$\begin{aligned} \vec{N}_{S/Q} &= \sum_{j=1}^{n_F} \vec{r}_{Qj} \times \vec{F}_j = \sum_{j=1}^{n_F} (\vec{r}_{Pj} + \vec{r}_{QP}) \times \vec{F}_j \\ &= \sum_{j=1}^{n_F} \vec{r}_{Pj} \times \vec{F}_j + \vec{r}_{QP} \times \sum_{j=1}^{n_F} \vec{F}_j \end{aligned} \tag{7.3}$$

The moment $\vec{N}_{S/Q}$ of the set S about a point Q is the moment $\vec{N}_{S/P}$ of the set S about the point P plus the moment about Q that would have resulted if the resultant $\vec{F}_S^{(r)}$ had line of action through P:

$$\vec{N}_{S/Q} = \vec{N}_{S/P} + \vec{r}_{QP} \times \vec{F}_S^{(r)} \tag{7.4}$$

This result is straightforward to apply, however, the resultant force does not have a line of action, so the procedure of pretending that $\vec{F}_S^{(r)}$ has its line of action through P is not completely satisfying. Therefore we will follow the approach of (Kane and Levinson 1985) and introduce an equivalent representation with a bound vector and a torque. To do this it is necessary to introduce the concept of a torque.

7.2.2 Torque

A *couple* is a set C of forces with zero resultant force, that is $\vec{F}_C^{(r)} = \vec{0}$. From (7.4) it is seen that this implies that $\vec{N}_{C/P} = \vec{N}_{C/Q}$, which means that the moment of a couple will be the same about any point, and it is therefore meaningful to define the moment of the couple without reference to any point.

The *torque* $\vec{T}_C$ is defined as the moment of the couple C. The resultant $\vec{F}_C^{(r)}$ of a couple is by definition zero. Therefore, the moment of the couple C is the same about any point, which means that

$$\vec{T}_C := \vec{N}_{C/P} = \vec{N}_{C/Q} \tag{7.5}$$

for arbitrary points Q and P.

Example 114 *Consider a couple with two forces $\vec{F}_1$ and $\vec{F}_2$ that have zero resultant force, which implies $\vec{F}_2 = -\vec{F}_1$. Define the position vector $\vec{r}_{21}$ between an arbitrary point on the line of action of $\vec{F}_2$ and the line of action of $\vec{F}_1$. The torque $\vec{T}$ of the couple, which is the moment of the two forces about an arbitrary point P, is then found from*

$$\vec{T} = \vec{r}_1 \times \vec{F}_1 + \vec{r}_2 \times \vec{F}_2 = \vec{r}_1 \times \vec{F}_1 - (\vec{r}_1 - \vec{r}_{21}) \times \vec{F}_1 = \vec{r}_{21} \times \vec{F}_1 \tag{7.6}$$

We see that the torque does not depend on the selection of the point P.

Example 115 *In this example we will derive force and torque expressions for a satellite with six gas jet actuators and three momentum wheels. The gas jet actuators set up forces $\vec{F}_j$, and the momentum wheels set up torques $\vec{T}_j$. The resultant force and the total moment about the center of mass are then*

$$\vec{F}^{(r)} = \sum_{j=1}^{6} \vec{F}_j, \quad \vec{N}_c = \sum_{j=1}^{3} \vec{T}_j + \sum_{j=1}^{6} \vec{r}_j \times \vec{F}_j \tag{7.7}$$

In the control of the attitude of the satellite it would make sense to arrange the gas jet actuators in pairs that produce torques in the form of couples. This is done by requiring $\vec{F}_1 = -\vec{F}_4$, $\vec{F}_2 = -\vec{F}_5$, $\vec{F}_3 = -\vec{F}_6$, $\vec{r}_1 = -\vec{r}_4$, $\vec{r}_2 = -\vec{r}_5$ and $\vec{r}_3 = -\vec{r}_6$. This implies that the resultant force is zero, that is, $\vec{F}^{(r)} = \vec{0}$. Therefore, the set of forces constitute a couple, and because of this the moment about the center of mass is actually a torque $\vec{T}_c = \vec{N}_c$ given by

$$\vec{T}_c = \sum_{j=1}^{3} \vec{T}_j + \sum_{j=1}^{3} 2\vec{r}_j \times \vec{F}_j \tag{7.8}$$

7.2.3 Equivalent force and torque

Two sets S and Σ of force vectors are said to be *equivalent* if they have equal resultant and equal moment about any point. Consider a set S of n_F forces with resultant force $\vec{F}_S^{(r)}$ and moment $\vec{N}_{S/P}$ about a point P. An equivalent set Σ of forces can then be defined with a single force and a torque due to a couple. To do this we let the set Σ to be the force $\vec{F}_\Sigma$ with line of action through the point P, and the torque $\vec{T}_\Sigma$ so that

$$\vec{F}_\Sigma = \vec{F}_S^{(r)}, \quad \vec{T}_\Sigma = \vec{N}_{S/P} \tag{7.9}$$

To see that the sets S and Σ will be equivalent we observe that the set Σ will have resultant $\vec{F}_\Sigma^{(r)} = \vec{F}_\Sigma = \vec{F}_S^{(r)}$, and the moments about an arbitrary point Q will be equal, which is confirmed by comparing the expression for $\vec{N}_{S/Q}$ in (7.4) with the moment $\vec{N}_{\Sigma/Q}$, which is the torque $\vec{T}_\Sigma$ plus the moment of $\vec{F}_\Sigma$ about Q, that is,

$$\vec{N}_{\Sigma/Q} = \vec{T}_\Sigma + \vec{r}_{QP} \times \vec{F}_\Sigma \tag{7.10}$$

that result from (7.4).

We conclude that the following sets of forces are equivalent:

1. A set S with resultant $\vec{F}_S^{(r)}$ and with moment $\vec{N}_{S/P}$ about P, where the resultant $\vec{F}_S^{(r)}$ does not have a line of action, and where the moment $\vec{N}_{S/Q}$ about a point Q is found from the rule (7.4).
2. A force $\vec{F}_\Sigma = \vec{F}_S^{(r)}$ with line of action through P in combination with a torque $\vec{T}_\Sigma = \vec{N}_{S/P}$. Then the moment about a point Q is found from $\vec{F}_\Sigma$ and $\vec{T}_\Sigma$ according to equation (7.10).

The main difference between the two equivalent representations S and Σ is that when S is used the resultant is not a true force vector as it is not a bound vector, and the additional rule (7.4) must be used to find the moment about some other point Q. In contrast to this, when the set Σ is used, the force $\vec{F}_\Sigma$ can be treated as a force vector and the torque $\vec{T}_\Sigma$ can be treated as a torque, and hence the usual definition of a moment about a point can be used to calculate the moment about a point Q.

7.2.4 Forces and torques on a rigid body

A mass force is a force $\vec{f}dm$ that acts on a mass element $dm = \rho dV$ at position $\vec{r}_p$. An example of this is the gravity force $\vec{g}dm$ acting on dm. The resultant gravity force on a body is

$$\vec{G} = \int_b \vec{g}dm = m\vec{g} \tag{7.11}$$

The moment of the gravity force about the origin of frame i is

$$\vec{N}_{G/i} = \int_b \vec{r}_p \times \vec{g}dm = \int_b \vec{r}_p dm \times \vec{g} = \vec{r}_c \times m\vec{g} = \vec{r}_c \times \vec{G} \tag{7.12}$$

The interpretation of this is that the gravity forces $\vec{g}dm$ will set up a moment equal to the moment of the resultant gravity force $\vec{G}$ would give if $\vec{G}$ had line of action through the center of mass. For this reason the center of mass is also called the *center of gravity*. In this connection it may be argued that the concept of a center of mass is more fundamental than a center of gravity which requires the presence of a field of gravity. From (7.4) it follows that the moment of gravity about the center of mass is zero, that is, $\vec{N}_{G/c} = \vec{0}$.

The resultant forces acting on a body b will be

$$\vec{F}_b^{(r)} = \vec{G} + \sum_{j=1}^{n_F} \vec{F}_j \tag{7.13}$$

where $\vec{F}_j$ are n_F contact forces acting on the body. The moment on the body b about its center of mass c is

$$\vec{N}_{b/c} = \vec{T}_b + \sum_{j=1}^{n_F} \vec{r}_{cj} \times \vec{F}_j \tag{7.14}$$

where $\vec{r}_{cj} = \vec{r}_{Fj} - \vec{r}_c$ is the vector from the center of mass c to the line of action of the forces $\vec{F}_j$, and $\vec{T}_b$ is the contact torque due to couples acting the body. Typically this would be motor torques. There is no moment from gravity as the moment is about the center of mass. The moment about some other point o is found from the rule (7.4), which gives

$$\vec{N}_{b/o} = \vec{N}_{b/c} + \vec{r}_g \times \vec{F}_b^{(r)} \tag{7.15}$$

where

$$\vec{r}_g := \vec{r}_{oc} \tag{7.16}$$

is the vector from o to the center of mass c.

Equivalent descriptions of the forces and moments on a body b are (Kane and Levinson 1985)

1. The resultant force $\vec{F}_b^{(r)}$ without specification of line of action, the moment $\vec{N}_{b/c}$ about the center of mass, and, in addition, the rule (7.15) for calculating the moment about some other point o.

2. The force $\vec{F}_{bc} = \vec{F}_b^{(r)}$ with line of action through the center of mass c in combination with the torque $\vec{T}_{bc} = \vec{N}_{b/c}$.

3. The force $\vec{F}_{bo} = \vec{F}_b^{(r)}$ with line of action through the point o in combination with the torque $\vec{T}_{bo} = \vec{N}_{b/o}$. Then $\vec{F}_{bo}$ and $\vec{T}_{bo}$ can be found from $\vec{F}_{bc}$ and $\vec{T}_{bc}$ with

$$\vec{F}_{bo} = \vec{F}_{bc} \tag{7.17}$$
$$\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \tag{7.18}$$

The resultant force and the moment are represented using Descriptions 2 and 3 is used in the software package Autolev for multibody simulation based on Kane's formulation of the equations of motion.

Example 116 *Suppose that Description 2 is used, and that $\vec{F}_{bc}$ and $\vec{T}_{bc}$ are given. Then the moment on b about a point o is found from*

$$\vec{N}_{b/o} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \tag{7.19}$$

If Description 3 is used and $\vec{F}_{bo}$ and $\vec{T}_{bo}$ are given, then the moment on b about c is found from

$$\vec{N}_{b/c} = \vec{T}_{bo} - \vec{r}_g \times \vec{F}_{bo} \tag{7.20}$$

7.2.5 Example: Robotic link

Consider a robot manipulator with 6 rigid bodies, called links, which are connected with rotary joints. The forces acting on a link k are the contact force $\vec{F}_{k-1,k}$ on link k from link $k-1$, the contact force $\vec{F}_{k+1,k}$ from link $k+1$ on link k, and the gravity force $\vec{G}_k$. The line of action of $\vec{F}_{k-1,k}$ passes through a point of position $\vec{r}_{k-1}$, and the line of action of $\vec{F}_{k+1,k}$ goes through a point of position $\vec{r}_k$. The center of mass has position $\vec{r}_{k_c}$. We note that due to the principle of action and reaction $\vec{F}_{k+1,k} = -\vec{F}_{k,k+1}$ where $\vec{F}_{k,k+1}$ is the force acting on link $k+1$ from link k. The torques in the form of couples that act on the link are the contact torque $\vec{T}_{k-1,k}$ on link k from link $k-1$, and the contact force $\vec{T}_{k+1,k} = -\vec{T}_{k,k+1}$ from link $k+1$ on link k. This gives the following expression for the resultant forces on link k

$$\vec{F}_k^{(r)} = \vec{F}_{k-1,k} + \vec{F}_{k+1,k} + \vec{G}_k \tag{7.21}$$

The moment about the center of mass k_c with position $\vec{r}_{k_c}$ is

$$\vec{N}_{k/k_c} = \vec{T}_{k-1,k} + \vec{T}_{k+1,k} + (\vec{r}_{k-1} - \vec{r}_{k_c}) \times \vec{F}_{k-1,k} + (\vec{r}_k - \vec{r}_{k_c}) \times \vec{F}_{k+1,k} \tag{7.22}$$

An equivalent description is possible with the force

$$\vec{F}_{kc} := \vec{F}_k^{(r)} \tag{7.23}$$

with line of action through the center of mass, and the torque

$$\vec{T}_{kc} := \vec{N}_{k/k_c} \tag{7.24}$$

To calculate the moment $\vec{N}_{k/k-1}$ on link k about the point $k-1$ with position vector $\vec{r}_{k-1}$, it is used that the force $\vec{F}_{kc}$ has line of action through c, and the the moment is found to be the torque $\vec{T}_{kc}$ plus the moment of $\vec{F}_{kc}$ about $k-1$, which gives

$$\vec{N}_{k/k-1} = \vec{T}_{kc} + (\vec{r}_{k_c} - \vec{r}_{k-1}) \times \vec{F}_{kc}$$

We may check that this makes sense by inserting of the expressions for $\vec{T}_{kc}$ and $\vec{F}_{kc}$, which gives

$$\vec{N}_{k/k-1} = \vec{T}_{k-1,k} + \vec{T}_{k+1,k} + (\vec{r}_k - \vec{r}_{k-1}) \times \vec{F}_{k+1,k} + (\vec{r}_{k_c} - \vec{r}_{k-1}) \times \vec{G}_k \tag{7.25}$$

7.3 Newton-Euler equations for rigid bodies

7.3.1 Equations of motion for a system of particles

Consider a system of N particles each of mass m_k and with position $\vec{r}_k$ relative to the origin of the inertial frame i. By setting up Newton's law for each particle and summing up we get the result (Goldstein 1980), (Kane and Levinson 1985)

$$m\vec{a}_c = \vec{F}^{(r)} \tag{7.26}$$

where $\vec{F}^{(r)}$ is the resultant force on the system of particles, and $\vec{a}_c$ is the acceleration of the center of mass.

The angular momentum of particle k about the center of mass is

$$\vec{h}_{k/c} = (\vec{r}_k - \vec{r}_c) \times m_k \vec{v}_k \tag{7.27}$$

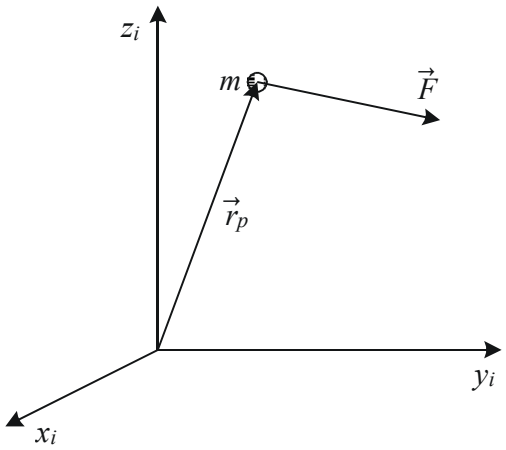

Figure 7.2: Mass point subject to a force $\vec{F}$.

where $\vec{v}_k$ is the velocity of particle k and $\vec{r}_c$ is the position of the center of mass. With reference to frame i, the time derivative of $\vec{h}_{k/c}$ is

$$\begin{aligned} \frac{^i d}{dt}\vec{h}_{k/c} &= (\vec{v}_k - \vec{v}_c) \times m_k \vec{v}_k + \vec{r}_{ck} \times m_k \vec{a}_k \\ &= -\vec{v}_c \times m_k \vec{v}_k + \vec{r}_{ck} \times m_k \vec{a}_k \end{aligned} \tag{7.28}$$

Summation over all particles leads to

$$\frac{^i d}{dt}\vec{h}_c = \vec{N}_c \tag{7.29}$$

where (6.413) and (7.26) is used, and where

$$\vec{h}_c = \sum_{k=1}^{N} \vec{r}_{ck} \times m_k \vec{v}_k \tag{7.30}$$

is the angular momentum of the system about the center of mass, and

$$\vec{N}_c = \sum_{k=1}^{N} \vec{r}_{ck} \times \vec{F}_k \tag{7.31}$$

is the moment of the forces about the center of mass. This means that the time derivative of the angular momentum about the center of mass is equal to the moment of the forces about the center of mass.

7.3.2 Equations of motion for a rigid body

This result (7.26, 7.29) in the previous section was derived for a system of N particles. The result can be generalized to a rigid body b by summing up the equations of motion for mass elements dm of position $\vec{r}_p$, velocity $\vec{v}_p$ and acceleration $\vec{a}_p$. To simplify expressions the set of forces and torques acting on the rigid body is represented by the equivalent

set with a force $\vec{F}_{bc}$ with line of action through the center of mass and magnitude equal to the resultant force, and a torque $\vec{T}_{bc} = \vec{N}_{b/c}$ that equals the moment about the center of the mass. The equations of motion for a rigid body are then found from (7.26, 7.29) to be

$$\vec{F}_{bc} = m\vec{a}_c \tag{7.32}$$

$$\vec{T}_{bc} = {}^i\frac{d}{dt}\vec{h}_{b/c} \tag{7.33}$$

where

$$\vec{h}_{b/c} = \int_b \vec{r} \times \vec{v}_p dm \tag{7.34}$$

is the angular momentum of the body b about the center of mass, and

$$\vec{r} = \vec{r}_p - \vec{r}_c \tag{7.35}$$

is the position of the mass element relative to the center of mass. Using $\vec{v}_p = \vec{v}_c + \vec{\omega}_{ib} \times \vec{r}$, this can be written

$$\begin{aligned} \vec{h}_{b/c} &= \int_b \vec{r} dm \times \vec{v}_c + \int_b \vec{r} \times (\vec{\omega}_{ib} \times \vec{r}) dm \\ &= \int_b \vec{r} \times (\vec{\omega}_{ib} \times \vec{r}) dm \\ &= -\int_b \vec{r} \times (\vec{r} \times \vec{\omega}_{ib}) dm \end{aligned} \tag{7.36}$$

where we have used (6.417). By introducing the dyadic representation of the vector cross product we may write this in the form

$$\vec{h}_{b/c} = -\int_b \vec{r}^{\times} \cdot (\vec{r}^{\times} \cdot \vec{\omega}_{ib}) dm = -\int_b \vec{r}^{\times} \cdot \vec{r}^{\times} dm \cdot \vec{\omega}_{ib} \tag{7.37}$$

This expression motivates the definition of the *inertia dyadic* of b about c as

$$\vec{M}_{b/c} = -\int_b \vec{r}^{\times} \cdot \vec{r}^{\times} dm \tag{7.38}$$

The angular momentum about c can then be written

$$\vec{h}_{b/c} = \vec{M}_{b/c} \cdot \vec{\omega}_{ib} \tag{7.39}$$

Insertion in (7.33) gives

$$\vec{T}_{bc} = {}^i\frac{d}{dt}\left(\vec{M}_{b/c} \cdot \vec{\omega}_{ib}\right) = {}^b\frac{d}{dt}(\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \tag{7.40}$$

Finally it is used that $\vec{M}_{b/c}$ is constant in b. This leads to the following result:

When referenced to the center of mass the equation of motion for a rigid body can be written

$$\vec{F}_{bc} = m\vec{a}_c \tag{7.41}$$

$$\vec{T}_{bc} = \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \tag{7.42}$$

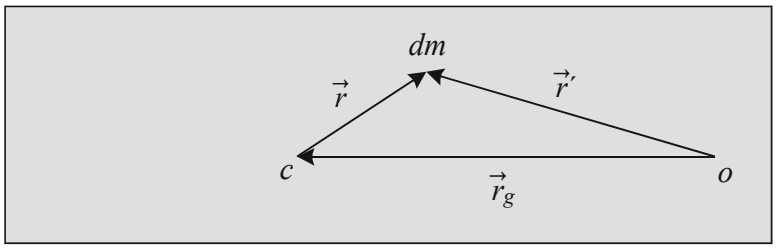

Figure 7.3: Definition of the vectors $\vec{r}$, $\vec{r}'$ and $\vec{r}_g$.

7.3.3 Equations of motion about a point

In important applications like ship dynamics and aeroplane dynamics the motion of the body b is described in terms of translation of a fixed point o which is not the mass center, and the rotation about the point o. In this case it is convenient to represent the forces and the moments acting on a rigid body by an equivalent set with a force $\vec{F}_{bo} = \vec{F}_{bc}$ with line of action through the point o and magnitude equal to the resultant force, and a torque

$$\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc} \tag{7.43}$$

where $\vec{r}_g = \vec{r}_c - \vec{r}_o$ is the vector from o to c as shown in Figure 7.3. We mention the following result before proceeding towards a description where the dynamics are referenced to a point o:

A mixed formulation of the equations of motion where the force and torque are referenced to the point o and the acceleration and inertia dyadics are referenced to the mass center is given by

$$\vec{F}_{bo} = m\vec{a}_c \tag{7.44}$$
$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_c + \vec{M}_{b/c} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/c} \cdot \vec{\omega}_{ib}) \tag{7.45}$$

Then the force equation can be referenced to the point o by combining (7.32) and (6.410). This gives

$$\vec{F}_{bo} = m\left[\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)\right]. \tag{7.46}$$

The torque equation involves the angular momentum about c, while we would like to have an expression involving the angular momentum about o, which is given by

$$\vec{h}_{b/o} = \int_b \vec{r}' \times \vec{v}_p dm \tag{7.47}$$

where

$$\vec{r}_p = \vec{r}_o + \vec{r}' \tag{7.48}$$
$$\vec{v}_p = \vec{v}_o + \vec{\omega}_{ib} \times \vec{r}' \tag{7.49}$$

Insertion of (7.49) into (7.47) leads to

$$\begin{aligned}\vec{h}_{b/o} &= \int_b \vec{r}'dm \times \vec{v}_o + \int_b \vec{r}' \times (\vec{\omega}_{ib} \times \vec{r}')\, dm \\ &= \int_b (\vec{r}_p - \vec{r}_o)\, dm \times \vec{v}_o - \int_b \vec{r}' \times (\vec{r}' \times \vec{\omega}_{ib})\, dm \\ &= \vec{r}_g \times m\vec{v}_o + \vec{M}_{b/o} \cdot \vec{\omega}_{ib} \end{aligned} \tag{7.50}$$

where the inertia dyadic $\vec{M}_{b/o}$ of b about o is defined by

$$\vec{M}_{b/o} = -\int_b (\vec{r}')^{\times} \cdot (\vec{r}')^{\times} dm \tag{7.51}$$

Note that $\vec{M}_{b/o}$ is constant in frame b. Time differentiation with respect to reference to frame i gives

$$\begin{aligned}\frac{^i d}{dt}\vec{h}_{b/o} &= (\vec{v}_c - \vec{v}_o) \times m\vec{v}_o + \vec{r}_g \times m\vec{a}_o + \frac{^b d}{dt}\left(\vec{M}_{b/o} \cdot \vec{\omega}_{ib}\right) + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \\ &= \vec{v}_c \times m\vec{v}_o + \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \end{aligned} \tag{7.52}$$

We may also express the angular momentum about o with the angular momentum about c by combining (7.34), (7.47) and $\vec{r}' = \vec{r} + \vec{r}_g$. This gives

$$\vec{h}_{b/o} = \int_b (\vec{r} + \vec{r}_g) \times \vec{v}_p dm = \vec{h}_{b/c} + \int_b \vec{r}_g \times \vec{v}_p dm \tag{7.53}$$

$$= \vec{h}_{b/c} + \vec{r}_g \times m\vec{v}_c \tag{7.54}$$

which implies that

$$\frac{^i d}{dt}\vec{h}_{b/o} = \frac{^i d}{dt}\vec{h}_{b/c} + \vec{r}_g \times m\vec{a}_c - \vec{v}_o \times m\vec{v}_c \tag{7.55}$$

From this equation and (7.52) it follows that

$$\frac{^i d}{dt}\vec{h}_{b/c} = \vec{r}_g \times m\,(\vec{a}_o - \vec{a}_c) + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \tag{7.56}$$

This result in combination with equations (7.32), (7.33) and (7.43) gives

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \tag{7.57}$$

With reference to a point o the equation of motion for a rigid body can be written

$$\vec{F}_{bo} = m\left[\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)\right] \tag{7.58}$$

$$\vec{T}_{bo} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \tag{7.59}$$

7.3.4 The inertia dyadic

The inertia dyadic of b about the center of the mass was defined as

$$\vec{M}_{b/c} = -\int_b \vec{r}^{\times} \cdot \vec{r}^{\times} dm \tag{7.60}$$

From (6.73) the alternative expression

$$\vec{M}_{b/c} = \int_b (\vec{r}^{\,2}\vec{I} - \vec{r}\vec{r})dm \tag{7.61}$$

is found. The dyadic can be evaluated in the b frame and is written

$$\vec{M}_{b/c} = \sum_{i=1}^{3}\sum_{j=1}^{3} m_{ij}^b \vec{b}_i\vec{b}_j \tag{7.62}$$

Example 117 *Consider a rigid body with a fixed coordinate frame b with orthogonal unit vectors $\vec{b}_1$, $\vec{b}_2$ and $\vec{b}_3$ that coincide with the main axes of inertia of the body b. Then the inertia dyadic is*

$$\vec{M}_{b/c} = m_{11}\vec{b}_1\vec{b}_1 + m_{22}\vec{b}_2\vec{b}_2 + m_{33}\vec{b}_3\vec{b}_3. \tag{7.63}$$

where m_{11}, m_{22} and m_{33} are constants. The angular velocity is written

$$\vec{\omega}_{ib} = \omega_1\vec{b}_1 + \omega_2\vec{b}_2 + \omega_3\vec{b}_3 \tag{7.64}$$

where $\omega_i = \vec{\omega}_{ib}\cdot\vec{b}_i$ is the component of $\vec{\omega}_{ib}$ along $\vec{b}_i$. The angular momentum is then

$$\vec{h}_{b/c} = (m_{11}\vec{b}_1\vec{b}_1 + m_{22}\vec{b}_2\vec{b}_2 + m_{33}\vec{b}_3\vec{b}_3)\cdot(\omega_1\vec{b}_1 + \omega_2\vec{b}_2 + \omega_3\vec{b}_3). \tag{7.65}$$

As the unit vectors are orthogonal, it follows that $\vec{b}_i\cdot\vec{b}_j = 0$ for $i \neq j$, and $\vec{b}_i\cdot\vec{b}_i = 1$. This gives

$$\vec{h}_{b/c} = m_{11}\omega_1\vec{b}_1 + m_{22}\omega_2\vec{b}_2 + m_{33}\omega_3\vec{b}_3 \tag{7.66}$$

Example 118 *The kinetic energy of a rigid body is*

$$K = \frac{1}{2}\int_b \vec{v}_p\cdot\vec{v}_p dm. \tag{7.67}$$

Insertion of $\vec{v}_p = \vec{v}_c + \vec{\omega}_{ib}\times\vec{r}$ gives

$$K = \frac{1}{2}m\vec{v}_c^{\,2} + \frac{1}{2}\int_b (\vec{\omega}_{ib}\times\vec{r})\cdot(\vec{\omega}_{ib}\times\vec{r})dm \tag{7.68}$$

as $\vec{v}_c\cdot\vec{\omega}_{ib}\times\int_b\vec{r}dm = 0$. The last term on the right side is simplified using

$$\begin{aligned}
\frac{1}{2}\int_b (\vec{\omega}_{ib}\times\vec{r})\cdot(\vec{\omega}_{ib}\times\vec{r})dm &= -\frac{1}{2}\int_b (\vec{\omega}_{ib}\times\vec{r})\cdot(\vec{r}\times\vec{\omega}_{ib})dm \\
&= -\frac{1}{2}\int_b \vec{\omega}_{ib}\cdot\vec{r}^{\times}\cdot\vec{r}^{\times}\cdot\vec{\omega}_{ib}dm \\
&= -\frac{1}{2}\vec{\omega}_{ib}\cdot\int_b \vec{r}^{\times}\cdot\vec{r}^{\times}dm\cdot\vec{\omega}_{ib} \\
&= \frac{1}{2}\vec{\omega}_{ib}\cdot\vec{M}_{b/c}\cdot\vec{\omega}_{ib}
\end{aligned} \tag{7.69}$$

This leads to the following expression for the kinetic energy:

$$K = \frac{1}{2}m\vec{v}_c^{\,2} + \frac{1}{2}\vec{\omega}_{ib}\cdot\vec{M}_{b/c}\cdot\vec{\omega}_{ib} \tag{7.70}$$

Example 119 *Using (7.49) the kinetic energy is found from the computation*

$$\begin{aligned} K &= \frac{1}{2}\int_b \vec{v}_p \cdot \vec{v}_p dm \\ &= \frac{1}{2} m\vec{v}_o \cdot \vec{v}_o + \vec{v}_o \cdot \left(\vec{\omega}_{ib} \times \int_b \vec{r}\,' dm\right) + \frac{1}{2}\int_b (\vec{\omega}_{ib} \times \vec{r}\,') \cdot (\vec{\omega}_{ib} \times \vec{r}\,') dm \\ &= \frac{1}{2} m\vec{v}_o \cdot \vec{v}_o + \vec{v}_o \cdot (\vec{\omega}_{ib} \times m\vec{r}_g) + \frac{1}{2}\vec{\omega}_{ib} \cdot \vec{M}_{b/o} \cdot \vec{\omega}_{ib} \end{aligned} \tag{7.71}$$

to have the form

$$K = \frac{1}{2} m\vec{v}_o \cdot \vec{v}_o - \vec{v}_o \cdot m\vec{r}_g^{\times} \cdot \vec{\omega}_{ib} + \frac{1}{2}\vec{\omega}_{ib} \cdot \vec{M}_{b/o} \cdot \vec{\omega}_{ib} \tag{7.72}$$

7.3.5 The inertia matrix

The matrix representation of the inertia dyadic $\vec{M}_{b/c}$ in frame b is the *inertia matrix*

$$\mathbf{M}_{b/c}^b = -\int_b \left(\mathbf{r}^b\right)^{\times} \left(\mathbf{r}^b\right)^{\times} dm \tag{7.73}$$

From (6.25) the more usual expression

$$\mathbf{M}_{b/c}^b = \int_b \left[\left(\mathbf{r}^b\right)^T \mathbf{r}^b \mathbf{I} - \mathbf{r}^b \left(\mathbf{r}^b\right)^T\right] dm \tag{7.74}$$

is found. The angular momentum of b about c can then be written in column vector form as

$$\mathbf{h}_{b/c}^b = \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \tag{7.75}$$

Equation (7.75) can be transformed to frame i using the transformation rule:

$$\mathbf{R}_b^i \mathbf{h}_{ib}^b = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \tag{7.76}$$

Insertion of $\boldsymbol{\omega}_{ib}^b = \mathbf{R}_i^b \boldsymbol{\omega}_{ib}^i$ gives

$$\mathbf{h}_{ib}^i = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \mathbf{R}_i^b \boldsymbol{\omega}_{ib}^i \tag{7.77}$$

Moreover, the inertia dyadic can also be represented by a matrix frame i using

$$\mathbf{M}_{b/c}^i = \int_b \left[\left(\mathbf{r}^i\right)^T \mathbf{r}^i \mathbf{I} - \mathbf{r}^i \left(\mathbf{r}^i\right)^T\right] dm \tag{7.78}$$

which satisfies

$$\mathbf{h}_{b/c}^i = \mathbf{M}_{b/c}^i \boldsymbol{\omega}_{ib}^i \tag{7.79}$$

Comparison of the two expressions (7.77) and (7.79) leads to the conclusion

$$\mathbf{M}_{b/c}^i = \mathbf{R}_b^i \mathbf{M}_{b/c}^b \mathbf{R}_i^b \tag{7.80}$$

We see that the inertia matrix transforms from frame b to frame i by a similarity transformation. This is to be expected as it is the matrix representation of a dyadic. The generic element m_{ij}^b of the dyadic is a second order tensor. Because of this the inertia matrix is often referred to as the *inertia tensor*.

Example 120 *The kinetic energy as given by (7.70) can be expressed in coordinate form as*

$$K = \frac{1}{2} m \left(\mathbf{v}_c^b\right)^T \mathbf{v}_c^b + \frac{1}{2} \left(\boldsymbol{\omega}_{ib}^b\right)^T \mathbf{M}_{b/c}^b \boldsymbol{\omega}_{ib}^b \tag{7.81}$$

The second term is the kinetic energy due to rotation. As the kinetic energy is greater or equal to zero, it follows that $\mathbf{M}_{b/c}^b$ *is a positive definite matrix.*

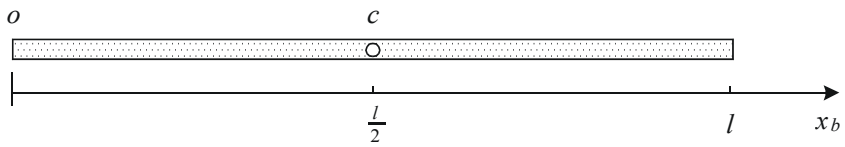

Figure 7.4: Slender beam of length l.

7.3.6 Expressions for the inertia matrix

The inertia matrix is given by

$$\mathbf{M}_{b/c}^{b} = \int_b \left[(\mathbf{r}^b)^2 \mathbf{I} - \mathbf{r}^b (\mathbf{r}^b)^T \right] dm \tag{7.82}$$

in the body-fixed frame b. Let $\mathbf{r}^b = (x, y, z)^T$. The the inertia matrix is then found to be

$$\mathbf{M}_{b/c}^{b} = \int_b \begin{pmatrix} y^2 + z^2 & -xy & -xz \\ -xy & x^2 + z^2 & -yz \\ -xz & -yz & x^2 + y^2 \end{pmatrix} dm \tag{7.83}$$

Under the assumption that the b frame is fixed in the body b, the inertia matrix $\mathbf{M}_{b/c}^{b}$ in the b frame is a constant matrix. In frame i we have $\mathbf{M}_{b/c}^{i} = \mathbf{R}_b^i \, \mathbf{M}_{b/c}^{b} \mathbf{R}_i^b$ which will not be constant if frame b is rotating relative to frame i.

7.3.7 The parallel axes theorem

The inertia dyadic about the point o is

$$\begin{aligned} \vec{M}_{b/o} &= -\int_b (\vec{r}\,')^{\times} \cdot (\vec{r}\,')^{\times} dm = -\int_b (\vec{r} + \vec{r}_g)^{\times} \cdot (\vec{r} + \vec{r}_g)^{\times} dm \\ &= -\int_b \vec{r}^{\times} \cdot \vec{r}^{\times} dm - \vec{r}_g^{\times} \cdot \left(\int_b \vec{r} dm \right)^{\times} - \left(\int_b \vec{r} dm \right)^{\times} \cdot \vec{r}_g^{\times} - \vec{r}_g^{\times} \cdot \vec{r}_g^{\times} \int_b dm \end{aligned}$$

Using (6.417) we find the following result:

The inertia dyadic of b about o is related to the inertia dyadic of b about c according to

$$\vec{M}_{b/o} = \vec{M}_{b/c} - m\vec{r}_g^{\times}\vec{r}_g^{\times} = \vec{M}_{b/c} + m\left[(\vec{r}_g \cdot \vec{r}_g)\vec{I} - \vec{r}_g\vec{r}_g \right] \tag{7.84}$$

The corresponding matrix expressions is

$$\mathbf{M}_{b/o}^{b} = \mathbf{M}_{b/c}^{b} - m\left(\mathbf{r}_g^b\right)^{\times}\left(\mathbf{r}_g^b\right)^{\times} = \mathbf{M}_{b/c}^{b} + m\left[(\mathbf{r}_g^b)^2 \mathbf{I} - \mathbf{r}_g^b (\mathbf{r}_g^b)^T \right] \tag{7.85}$$

This is *the parallel axes theorem*.

Example 121 *A slender beam has length ℓ and mass m. A coordinate frame b is fixed in the beam with the x axis in the length axis of the beam as shown in Figure 7.4. The mass element is set to be $dm = (m/\ell)dx$. The inertia matrix about the center of mass is*

$$\mathbf{M}_{b/c}^{b} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{12} & 0 \\ 0 & 0 & \frac{m\ell^2}{12} \end{pmatrix} \tag{7.86}$$

The inertia matrix about the endpoint o of the beam is found from the parallel axes theorem to be

$$\mathbf{M}^b_{b/o} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{12} & 0 \\ 0 & 0 & \frac{m\ell^2}{12} \end{pmatrix} + m\left(\frac{\ell}{2}\right)^2 \mathbf{I} - \begin{pmatrix} m(\frac{\ell}{2})^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{m\ell^2}{3} & 0 \\ 0 & 0 & \frac{m\ell^2}{3} \end{pmatrix}$$

Example 122 *The kinetic energy of a rigid body can be expressed with reference to a point o by*

$$\begin{aligned} K &= \frac{1}{2}m\left(\mathbf{v}^b_c\right)^T \mathbf{v}^b_c + \frac{1}{2}\left(\boldsymbol{\omega}^b_{ib}\right)^T \mathbf{M}^b_{b/c}\boldsymbol{\omega}^b_{ib} \\ &= \frac{1}{2}m\left(\mathbf{v}^b_o + (\boldsymbol{\omega}^b_{ib})^\times \mathbf{r}^b_g\right)^T \left(\mathbf{v}^b_o + (\boldsymbol{\omega}^b_{ib})^\times \mathbf{r}^b_g\right) + \frac{1}{2}\left(\boldsymbol{\omega}^b_{ib}\right)^T \mathbf{M}^b_{b/c}\boldsymbol{\omega}^b_{ib} \\ &= \frac{1}{2}m\left(\mathbf{v}^b_c\right)^T \mathbf{v}^b_c + m\left(\mathbf{v}^b_o\right)^T (\mathbf{r}^b_g)^{\times T}\boldsymbol{\omega}^b_{ib} + m\left((\boldsymbol{\omega}^b_{ib})^\times \mathbf{r}^b_g\right)^T \mathbf{v}^b_o \\ &\quad + \frac{1}{2}\left(\boldsymbol{\omega}^b_{ib}\right)^T \left(\mathbf{M}^b_{b/c} - m\left(\mathbf{r}^b_g\right)^\times \left(\mathbf{r}^b_g\right)^\times\right)\boldsymbol{\omega}^b_{ib} \\ &= \frac{1}{2}\begin{pmatrix} \mathbf{v}^b_o \\ \boldsymbol{\omega}^b_{ib} \end{pmatrix}^T \begin{pmatrix} m\mathbf{I} & m\left(\mathbf{r}^b_g\right)^{\times T} \\ m\left(\mathbf{r}^b_g\right)^\times & \mathbf{M}^b_{b/o} \end{pmatrix} \begin{pmatrix} \mathbf{v}^b_o \\ \boldsymbol{\omega}^b_{ib} \end{pmatrix} \end{aligned} \tag{7.87}$$

when the description is referenced to a fixed point o in the body. In the derivation the rules $\mathbf{a}^\times\mathbf{b} = -\mathbf{b}^\times\mathbf{a}$ *and* $(\mathbf{a}^\times\mathbf{b})^T\mathbf{c} = (\mathbf{c}^\times\mathbf{b})^T\mathbf{a}$ *are used.*

7.3.8 The equations of motion for a rigid body

In this section we will sum up with different versions of the equations of motion for a rigid body b where the resultant force is $\vec{F}^{(r)}_b$ and the total moment on b about the center of mass is $\vec{N}_{b/c}$. First we represent the forces and the moments by the equivalent representation with a force $\vec{F}_{bc} = \vec{F}^{(r)}_b$ with line of action through the center of mass c in combination with a torque $\vec{T}_{bc} = \vec{N}_{b/c}$. The equations of motion are

$$\vec{F}_{bc} = m\vec{a}_c \tag{7.88}$$
$$\vec{T}_{bc} = \vec{M}_{b/c}\cdot\vec{\alpha}_{ib} + \vec{\omega}_{ib}\times(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}) \tag{7.89}$$

In the b frame the coordinate form is written in matrix form as

$$\begin{pmatrix} m\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^b_{b/c} \end{pmatrix}\begin{pmatrix} \mathbf{a}^b_c \\ \boldsymbol{\alpha}^b_{ib} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \left(\boldsymbol{\omega}^b_{ib}\right)^\times \mathbf{M}^b_{b/c}\,\boldsymbol{\omega}^b_{ib} \end{pmatrix} = \begin{pmatrix} \mathbf{F}^b_{bc} \\ \mathbf{T}^b_{b/c} \end{pmatrix} \tag{7.90}$$

In view of $\vec{a}_c = \frac{{}^b d}{dt}\vec{v}_c + \vec{\omega}_{ib}\times\vec{v}_c$ the equations of motion can be written

$$\vec{F}_{bc} = m\frac{{}^b d}{dt}\vec{v}_c + m\vec{\omega}_{ib}\times\vec{v}_c \tag{7.91}$$
$$\vec{T}_{bc} = \vec{M}_{b/c}\cdot\vec{\alpha}_{ib} + \vec{\omega}_{ib}\times(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}) \tag{7.92}$$

or

$$\begin{pmatrix} m\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^b_{b/c} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{v}}^b_c \\ \boldsymbol{\alpha}^b_{ib} \end{pmatrix} + \begin{pmatrix} m\left(\boldsymbol{\omega}^b_{ib}\right)^\times \mathbf{v}^b_c \\ \left(\boldsymbol{\omega}^b_{ib}\right)^\times \mathbf{M}^b_{b/c}\boldsymbol{\omega}^b_{ib} \end{pmatrix} = \begin{pmatrix} \mathbf{F}^b_{bc} \\ \mathbf{T}^b_{bc} \end{pmatrix} \tag{7.93}$$

The representation of the forces and torques is changed to an equivalent representation with a force $\vec{F}_{bo} = \vec{F}_b^{(r)}$ with line of action through o in combination with a torque $\vec{T}_{bo} = \vec{T}_{bc} + \vec{r}_g \times \vec{F}_{bc}$, where $\vec{r}_g$ is the vector from o to c. The equations of motion are then

$$\vec{F}_{bo} = m[\vec{a}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)] \tag{7.94}$$

$$\vec{T}_{b/o} = \vec{r}_g \times m\vec{a}_o + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \tag{7.95}$$

In the special case where o is the center of mass, then $\vec{r}_g = \vec{0}$, and the result is the same as in (7.91, 7.92).

The coordinate form in the b frame is written in matrix form as

$$\begin{pmatrix} m\mathbf{I} & m(\mathbf{r}^b_g)^{\times T} \\ m(\mathbf{r}^b_g)^\times & \mathbf{M}^b_{b/o} \end{pmatrix} \begin{pmatrix} \mathbf{a}^b_o \\ \boldsymbol{\alpha}^b_{ib} \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{\omega}^b_{ib})^\times \left(\boldsymbol{\omega}^b_{ib}\right)^\times \mathbf{r}^b_g \\ (\boldsymbol{\omega}^b_{ib})^\times \mathbf{M}^b_{b/o}\boldsymbol{\omega}^b_{ib} \end{pmatrix} = \begin{pmatrix} \mathbf{F}^b_{bo} \\ \mathbf{T}^b_{bo} \end{pmatrix} \tag{7.96}$$

Here it is used that $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$ for any two vectors $\vec{a}$ and $\vec{b}$, and that $(\cdot)^\times = -\left[(\cdot)^\times\right]^T$. Note that the leading matrix on the left side is symmetric and positive definite. This matrix can be regarded as a mass matrix.

An alternative formulation is

$$\vec{F}_{bo} = m\left[\frac{{}^b d}{dt}\vec{v}_o + \vec{\omega}_{ib} \times \vec{v}_o + \vec{\alpha}_{ib} \times \vec{r}_g + \vec{\omega}_{ib} \times (\vec{\omega}_{ib} \times \vec{r}_g)\right] \tag{7.97}$$

$$\vec{T}_{bo} = m\vec{r}_g \times \frac{{}^b d}{dt}\vec{v}_o + m\vec{r}_g \times (\vec{\omega}_{ib} \times \vec{v}_o) + \vec{M}_{b/o} \cdot \vec{\alpha}_{ib} + \vec{\omega}_{ib} \times (\vec{M}_{b/o} \cdot \vec{\omega}_{ib}) \tag{7.98}$$

with matrix form

$$\begin{pmatrix} m\mathbf{I} & m(\mathbf{r}^b_g)^{\times T} \\ m(\mathbf{r}^b_g)^\times & \mathbf{M}^b_{b/o} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{v}}^b_o \\ \boldsymbol{\alpha}^b_{ib} \end{pmatrix} + \begin{pmatrix} m(\boldsymbol{\omega}^b_{ib})^\times \left[(\boldsymbol{\omega}^b_{ib})^\times \mathbf{r}^b_g + \mathbf{v}^b_o\right] \\ (\boldsymbol{\omega}^b_{ib})^\times \mathbf{M}^b_{b/o}\boldsymbol{\omega}^b_{ib} + m(\mathbf{r}^b_g)^\times (\boldsymbol{\omega}^b_{ib})^\times \mathbf{v}^b_o \end{pmatrix} = \begin{pmatrix} \mathbf{F}^b_{bo} \\ \mathbf{T}^b_{bo} \end{pmatrix}$$

7.3.9 Satellite attitude dynamics

Suppose that the inertia matrix in the body-fixed frame b is

$$\mathbf{M}^b_{b/c} = \mathrm{diag}(m_{11}, m_{22}, m_{33}). \tag{7.99}$$

Then the angular momentum is

$$\mathbf{h}^b_{b/c} = \mathbf{M}^b_{b/c}\boldsymbol{\omega}^b_{ib} = \begin{pmatrix} m_{11}\omega_1 \\ m_{22}\omega_2 \\ m_{33}\omega_3 \end{pmatrix}. \tag{7.100}$$

The torque $\mathbf{T}_{bc}^b = (T_1, T_2, T_3)^T$ is acting on the body. The torque law is then

$$\vec{T}_{bc} = \frac{^i d}{dt}\vec{h}_{b/c} = \frac{^i d}{dt}\left(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}\right) = \frac{^b d}{dt}\left(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}\right) + \vec{\omega}_{ib}\times\left(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}\right) \tag{7.101}$$

The inertia dyadic is constant in the b frame, therefore

$$\vec{M}_{b/c}\cdot\vec{\alpha}_{ib} + \vec{\omega}_{ib}\times\left(\vec{M}_{b/c}\cdot\vec{\omega}_{ib}\right) = \vec{T}_{bc} \tag{7.102}$$

where (6.402) is used. With coordinate vectors this is written

$$\mathbf{M}_{b/c}^b\dot{\boldsymbol{\omega}}_{ib}^b + (\boldsymbol{\omega}_{ib}^b)^\times\mathbf{M}_{b/c}^b\boldsymbol{\omega}_{ib}^b = \mathbf{T}_{bc}^b. \tag{7.103}$$

Written out in components the model is

$$\begin{aligned} m_{11}\dot{\omega}_1 + (m_{33}-m_{22})\omega_2\omega_3 &= T_1 && (7.104)\\ m_{22}\dot{\omega}_2 + (m_{11}-m_{33})\omega_3\omega_1 &= T_2 && (7.105)\\ m_{33}\dot{\omega}_3 + (m_{22}-m_{11})\omega_1\omega_2 &= T_3 && (7.106) \end{aligned}$$

7.4 Example: Ball and beam dynamics

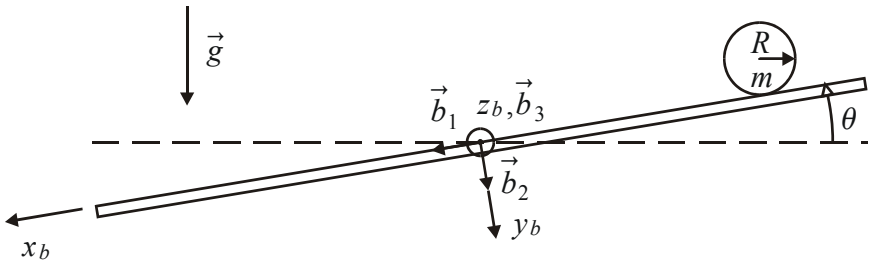

Figure 7.5: Ball and beam system.

In this section we will derive the equations of motion for a ball-and-beam system with a ball that rolls in a track on a beam. To do this we start with the kinematics of the system, and then combine the equations of motion for the ball and for the beam. We fix a coordinate system b in the beam so that the x_b axis is along the track, and the z_b axis is along the motor shaft. The orthogonal unit vectors $\vec{b}_1, \vec{b}_2, \vec{b}_3$ are placed along the x_b, y_b, z_b axes. According to (6.103) the scalar products of the unit vectors of frames n and b are related by

$$\begin{aligned} \vec{n}_1\cdot\vec{b}_1 &= \cos\theta, \quad \vec{n}_1\cdot\vec{b}_2 = -\sin\theta && (7.107)\\ \vec{n}_2\cdot\vec{b}_1 &= \sin\theta, \quad \vec{n}_2\cdot\vec{b}_2 = \cos\theta && (7.108)\\ \vec{n}_3\cdot\vec{b}_1 &= \vec{n}_3\cdot\vec{b}_2 = 0, \quad \vec{n}_3\cdot\vec{b}_3 = 1 && (7.109) \end{aligned}$$

which implies that

$$\begin{aligned} \vec{n}_1 &= \cos\theta\vec{b}_1 - \sin\theta\vec{b}_2 && (7.110)\\ \vec{n}_2 &= \sin\theta\vec{b}_1 + \cos\theta\vec{b}_2, && (7.111)\\ \vec{n}_3 &= \vec{b}_3 && (7.112) \end{aligned}$$

We note that

$$\vec{n}_1 \times \vec{b}_1 = \sin\theta \vec{b}_3, \quad \vec{n}_1 \times \vec{b}_2 = \cos\theta \vec{b}_3 \tag{7.113}$$

$$\vec{n}_2 \times \vec{b}_1 = -\cos\theta \vec{b}_3, \quad \vec{n}_2 \times \vec{b}_2 = \sin\theta \vec{b}_3 \tag{7.114}$$

Note that $\vec{n}_2$ is pointing vertically downwards so that the acceleration of gravity is $\vec{g} = g\vec{n}_2$. The beam is rotated with angular velocity $\vec{\omega}_1 = \dot{\theta}\vec{b}_3$ by a motor so that the track can be given an angle θ relative to the horizontal line, and the ball can be made to roll along the beam. The system is shown in Figure 7.5.

The radius of the ball is R, and the position of the ball along the track is denoted by x. The position of the center of the ball is

$$\vec{r}_2 = x\vec{b}_1 - R\vec{b}_2.$$

The velocity is

$$\vec{v}_2 = \frac{{}^b d}{dt}\vec{r} + \vec{\omega}_1 \times \vec{r} = \dot{x}\vec{b}_1 + \dot{\theta}\vec{b}_3 \times \left(x\vec{b}_1 - R\vec{b}_2\right) = \left(\dot{x} + \dot{\theta}R\right)\vec{b}_1 + \dot{\theta}x\vec{b}_2, \tag{7.115}$$

and the acceleration is

$$\begin{aligned} \vec{a}_2 &= \frac{{}^b d}{dt}\vec{v}_2 + \vec{\omega}_1 \times \vec{v}_2 \\ &= \left(\ddot{x} + \ddot{\theta}R\right)\vec{b}_1 + \left(\ddot{\theta}x + \dot{\theta}\dot{x}\right)\vec{b}_2 + \dot{\theta}\vec{b}_3 \times \left[\left(\dot{x} + \dot{\theta}R\right)\vec{b}_1 + \dot{\theta}x\vec{b}_2\right] \\ &= \left(\ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x\right)\vec{b}_1 + \left(\ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R\right)\vec{b}_2. \end{aligned} \tag{7.116}$$

It follows that the ball rolls along the track with an angular velocity given by

$$\vec{\omega}_2 = \left(\dot{\theta} + \frac{\dot{x}}{R}\right)\vec{b}_3 \tag{7.117}$$

as it is assumed that the ball does not slide.

The kinematic equations have now been established, and we will now develop the equations of motion. The mass of the ball is m, and the moment of inertia of the ball about its center of inertia is

$$J_2 = \frac{2}{5}m_2 R^2 \tag{7.118}$$

which is tabulated in textbooks on dynamics. The contact force acting from the beam on the ball is

$$\vec{F} = F_x\vec{b}_1 + F_y\vec{b}_2 \tag{7.119}$$

while the gravitational force on the ball is

$$\vec{G} = m_2 g\vec{n}_2 = m_2 g\left(\sin\theta\vec{b}_1 + \cos\theta\vec{b}_2\right). \tag{7.120}$$

It is noted that the contact torque between the ball and the beam is zero.

The angular momentum equation for the ball is (7.32)

$$\vec{T}_{2c} = J_2\frac{{}^n d}{dt}\vec{\omega}_2 = J_2\left(\ddot{\theta} + \frac{\ddot{x}}{R}\right)\vec{b}_3 \tag{7.121}$$

It is convenient to use the moment

$$\vec{N}_{2/o} = \left(-R\vec{b}_2\right) \times \vec{G}_2 = m_2 Rg \sin\theta \vec{b}_3 \tag{7.122}$$

about the contact point between the ball and the beam in the equation of motion. The reason for this is that the unknown constraint force F_x will not show up in the torque in this case. From (7.45) the moment $\vec{N}_{2/o}$ and the torque $\vec{T}_{2c}$ are related through the expression

$$\vec{N}_{2/o} = \vec{T}_{2c} + \left(-R\vec{b}_2\right) \times m_2\vec{a} \tag{7.123}$$

which gives

$$\begin{aligned} m_2 Rg \sin\theta &= J_2\left(\ddot{\theta} + \frac{\ddot{x}}{R}\right) + m_2 R\left(\ddot{x} + \ddot{\theta}R - \dot{\theta}^2 x\right) && (7.124)\\ &= \left(J_2 + m_2R^2\right)\ddot{\theta} + \left(\frac{J_2}{R} + m_2 R\right)\ddot{x} - m_2 Rx\dot{\theta}^2 && (7.125) \end{aligned}$$

This is written

$$\left(J_2 + m_2R^2\right)\ddot{\theta} + \frac{1}{R}\left(J_2 + m_2R^2\right)\ddot{x} = m_2 Rx\dot{\theta}^2 + Rm_2 g\sin\theta \tag{7.126}$$

By inserting the value of J_2 from (7.118), we get

$$\left(J_2 + m_2R^2\right) = \frac{7}{5}m_2R^2 \tag{7.127}$$

The Newton's law for the ball is

$$m_2\vec{a}_2 = \vec{F} + \vec{G}_2 \tag{7.128}$$

In the y_b direction this gives

$$m_2\left(\ddot{\theta}x + 2\dot{\theta}\dot{x} + \dot{\theta}^2 R\right) = F_y + m_2 g\cos\theta \tag{7.129}$$

To proceed an expression for the contact force F_y from the beam on the ball is needed. This can be found from the equation of motion for the beam. The contact force from the ball on the beam in the y_b direction is $-F_y$. In the equation of motion for the beam this gives

$$J_1\ddot{\theta}\vec{b}_3 = x\vec{b}_1 \times (-F_y\vec{b}_2) + T\vec{b}_3 \tag{7.130}$$

which leads to

$$J_1\ddot{\theta} = -xF_y + T \tag{7.131}$$

This equation is combined with (7.129), and the result is

$$\left(J_1 + m_2x^2\right)\ddot{\theta} = T + m_2 gx\cos\theta - 2m_2 x\dot{\theta}\dot{x} - m_2\dot{\theta}^2 xR \tag{7.132}$$

where T is the motor torque and F_y is the contact force in the y_b direction.

The model of the ball and beam is given by

$$\left(J_1 + m_2x^2\right)\ddot{\theta} = T + m_2 gx\cos\theta - 2m_2 x\dot{\theta}\dot{x} - m_2\dot{\theta}^2 xR \tag{7.133}$$

$$\left(J_2 + m_2R^2\right)\ddot{\theta} + \frac{1}{R}\left(J_2 + m_2R^2\right)\ddot{x} = m_2 Rx\dot{\theta}^2 + Rm_2 g\sin\theta \tag{7.134}$$

Example 123 *The rate of change of the energy of the ball and beam system will be equal to the power $\dot{\theta}T$ supplied by the torque T. If the model does not satisfy this condition, then the model is not correct, which provides us with a method to check the validity of the model. The total energy of the system is*

$$\begin{aligned} V &= \frac{1}{2}J_1\vec{\omega}_1\cdot\vec{\omega}_1+\frac{1}{2}J_2\vec{\omega}_2\cdot\vec{\omega}_2+\frac{1}{2}m_2\vec{v}_2\cdot\vec{v}_2+m_2g\left(-x\sin\theta+R\cos\theta\right) \\ &= \frac{1}{2}J_1\dot{\theta}^2+\frac{1}{2}J_2\left(\dot{\theta}+\frac{\dot{x}}{R}\right)^2+\frac{1}{2}m_2\left(\left(\dot{x}+\dot{\theta}R\right)^2+\left(\dot{\theta}x\right)^2\right) \\ &\quad +m_2g\left(-x\sin\theta+R\cos\theta\right) \end{aligned} \tag{7.135}$$

The time derivative along the solutions of the system is

$$\begin{aligned} \dot{V} &= \dot{\theta}J_1\ddot{\theta}+\left(\dot{\theta}+\frac{\dot{x}}{R}\right)J_2\left(\ddot{\theta}+\frac{\ddot{x}}{R}\right)+\left(\dot{x}+\dot{\theta}R\right)m_2\left(\ddot{x}+\ddot{\theta}R\right) \\ &\quad +\dot{\theta}xm_2\ddot{\theta}x+\dot{\theta}xm_2\dot{\theta}\dot{x}-m_2g\left(\dot{x}\sin\theta+x\dot{\theta}\cos\theta+R\dot{\theta}\sin\theta\right) \\ &= \dot{\theta}\left(J_1+m_2x^2\right)\ddot{\theta}+\left(\dot{\theta}+\frac{\dot{x}}{R}\right)\left(J_2+m_2R^2\right)\left(\ddot{\theta}+\frac{\ddot{x}}{R}\right)+m_2x\dot{x}\dot{\theta}^2 \\ &\quad -m_2g\left(\dot{x}\sin\theta+x\dot{\theta}\cos\theta+R\dot{\theta}\sin\theta\right) \\ &= \dot{\theta}T \end{aligned} \tag{7.136}$$

This result shows that the model is consistent with the energy flow in the system.

Example 124 *Insertion of $\omega_2=\dot{\theta}+\dot{x}/R$ gives a diagonal mass matrix:*

$$\left(J_1+m_2x^2\right)\ddot{\theta} = T+m_2gx\cos\theta-2m_2x\dot{\theta}\dot{x}-m_2\dot{\theta}^2xR \tag{7.137}$$

$$\frac{7}{5}m_2R^2\dot{\omega}_2 = Rm_2\dot{\theta}^2x+Rm_2g\sin\theta \tag{7.138}$$

Example 125 *If the radius of the ball becomes small, that is, when $R\to 0$, then the model becomes*

$$\left(J_1+mvx^2\right)\ddot{\theta} = T+m_2gx\cos\theta-2mvx\dot{\theta}\dot{x} \tag{7.139}$$

$$\frac{7}{5}m_2\ddot{x} = m_2\dot{\theta}^2x+m_2g\sin\theta \tag{7.140}$$

Example 126 *Linearization about $\dot{\theta}=0$, $\theta=0$, $\dot{x}=0$ and $x=0$ gives*

$$J_1\ddot{\theta} = T+m_2gx \tag{7.141}$$

$$\ddot{x} = \frac{5}{7}g\theta \tag{7.142}$$

which gives

$$\frac{d^4}{dt^4}x=\frac{1}{J_1}m_2gx+\frac{5}{7}\frac{g}{J_1}T \tag{7.143}$$

7.5 Example: Inverted pendulum

7.5.1 Equations of motion

Consider a pendulum on a cart as shown in Figure 7.6. The mass of the cart is m_v, the position of the cart is x, and the force on the cart is F. The pendulum is a point mass

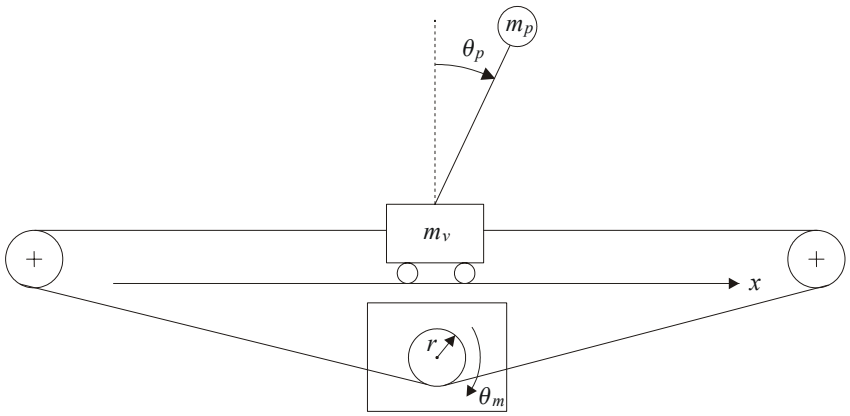

Figure 7.6: Inverted pendulum.

m_b at the end of a massless rod of length L_b. The angle of the pendulum is denoted θ_b, which is zero at the upright position.

To derive the equations of motion for the system we will first describe the kinematics of the system by assigning coordinate frames n and b to describe the motion of the cart and the pendulum, and then derive kinematic equations for the unit vectors of frames n and b. A non-moving coordinate frame n is defined with unit vector $\vec{n}_1$ along the motion of the cart, with $\vec{n}_2$ in the vertical downwards direction and $\vec{n}_3$ along the axis of rotation. A frame b is fixed to the pendulum. The rotation matrix is $\mathbf{R}_b^n = \mathbf{R}_{z,\theta_b}$ which is a rotation by an angle θ_b about the axis defined by $\vec{n}_3 = \vec{b}_3$. The angular velocity of the pendulum is $\vec{\omega}_b = \dot{\theta}_b\vec{n}_3 = \dot{\theta}_b\vec{b}_3$. The relation between the unit vectors in frames n and b is as given in Section 7.4.

The next step in the derivation of the equations of motion is to derive kinematic equations for position, velocity and acceleration. The position of the point mass m_1 is

$$\vec{r}_b = x\vec{n}_1 - L_b\vec{b}_2 \tag{7.144}$$

The velocity is found to be

$$\vec{v}_b = \frac{^n d\vec{r}_b}{dt} = \dot{x}\vec{n}_1 - \dot{\theta}_b\vec{b}_3 \times L_b\vec{b}_2 = \dot{x}\vec{n}_1 + \dot{\theta}_b L_b\vec{b}_1 \tag{7.145}$$

and acceleration is

$$\vec{a}_b = \ddot{x}\vec{n}_1 + \ddot{\theta}_b L_b\vec{b}_1 + \dot{\theta}_b\vec{b}_3 \times \dot{\theta}_b L\vec{b}_1 = \ddot{x}\vec{n}_1 + \ddot{\theta}_b L_b\vec{b}_1 + \dot{\theta}_b^2 L_b\vec{b}_2 \tag{7.146}$$

At this stage, the kinematic model has been established, and the equations of motion can be derived. This is done by combining Newton's law for the point mass and for the cart, and with the torque law for the pendulum. Newton's law for the point mass gives

$$m_b\vec{a}_b = \vec{F}_b + mg\vec{n}_2 \tag{7.147}$$

where g is the acceleration of gravity. In the $\vec{n}_1$ direction this gives

$$m_b\left(\ddot{x} + \ddot{\theta}_b L_b\cos\theta_b - \dot{\theta}_b^2 L_b\sin\theta_b\right) = \vec{F}_b \cdot \vec{n}_1 \tag{7.148}$$

Newton's law for the cart gives

$$m_v\ddot{x} = F - \vec{F}_b \cdot \vec{n}_1 \tag{7.149}$$

Combination of the two equations gives

$$(m_v + m_b)\ddot{x} + m_b L_b \ddot{\theta}_b \cos\theta_b - m_b \dot{\theta}_b^2 L_b \sin\theta_b = F \tag{7.150}$$

The torque law for the pendulum about the connection point is according to (7.45)

$$-L_b\vec{b}_2 \times m_b g\vec{n}_2 = -L_b\vec{b}_2 \times m_b\left(\ddot{x}\vec{n}_1 + \ddot{\theta}_b L_b \vec{b}_1 - \dot{\theta}_b L_b \vec{b}_2\right) \tag{7.151}$$

where $\vec{r}_g = -L_b\vec{b}_2$ in the notation of (7.45). The component of this equation in the $\vec{n}_3$ direction is

$$m_b L_b g \sin\theta_b = m_b L_b \ddot{x}\cos\theta_b + m_b L_b^2 \ddot{\theta}_b \tag{7.152}$$

The model for the cart and pendulum has then been found to be

$$\begin{aligned}(m_v + m_b)\ddot{x} + m_b L_b \ddot{\theta}_b \cos\theta_b &= m_b \dot{\theta}_b^2 L_b \sin\theta_b + F &(7.153)\\ m_b L_b^2 \ddot{\theta}_b + m_b L_b \ddot{x}\cos\theta_b &= m_b L_b g \sin\theta_b &(7.154)\end{aligned}$$

where F is the external force acting on the cart.

Example 127 *The rate of change of the energy in the system is equal to the power $F\dot{x}$ supplied by the external force F. We will now check if the model is consistent with this observation. The total energy of the system is*

$$\begin{aligned}V &= \frac{1}{2}m_v\dot{x}^2 + \frac{1}{2}m_b\vec{v}_b \cdot \vec{v}_b + m_b g L_b \cos\theta_b\\ &= \frac{1}{2}m_v\dot{x}^2 + \frac{1}{2}m_b(\dot{x}^2 + 2L_b\cos\theta_b\dot{x}\dot{\theta}_b + \dot{\theta}_b^2 L_b^2) + m_b g L_b\cos\theta_b \qquad (7.155)\end{aligned}$$

The time derivative of the energy along the solutions of the system is

$$\begin{aligned}\dot{V} &= \dot{x}\left[(m_v + m_b)\ddot{x} + m_b L_b \cos\theta_b \ddot{\theta}_b\right] + \dot{\theta}_b\left(m_b L_b^2\ddot{\theta}_b + m_b L_b\cos\theta_b\ddot{x}\right)\\ &\quad -L_b m_b \sin\theta_b \dot{x}\dot{\theta}_b^2 - m_b g L_b \dot{\theta}_b \sin\theta_b\\ &= \dot{x}\left(m_b\dot{\theta}_b^2 L_b\sin\theta_b + F\right) + \dot{\theta}_b m_b L_b g\sin\theta_b - L_b m_b\sin\theta_b\dot{x}\dot{\theta}_b^2 - m_b g L_b\dot{\theta}_b\sin\theta_b\\ &= F\dot{x} \qquad (7.156)\end{aligned}$$

This shows that the model is consistent with the energy flow of the system.

Next we combine the cart and pendulum model with the motor model. The cart is controlled with a current controlled DC motor with dynamics given by

$$J_m\ddot{\theta}_m = K_T u - T_L \tag{7.157}$$

where θ_m is the motor angle, u is the input, K_T is the torque constant, J_m is the inertial of the motor, and T_L is the load torque from the cart. The motor is connected to the cart with a string that runs over a pulley fixed to the motor axis. The radius of the pulley is r, and it follows that

$$T_L = rF, \quad \dot{x} = r\dot{\theta}_m \tag{7.158}$$

which gives

$$\frac{J_m}{r^2}\ddot{x} = \frac{K_T}{r}u - F \tag{7.159}$$

It is observed that the cart and pendulum is driven by the motor through a port with effort F and and flow $\dot{x}$. The effort F is input to the cart and pendulum model, and the flow $\dot{x}$ is output. At the same time the motor model has input F and output $\dot{x}$. This means that the inputs and the outputs of the port interconnection are incompatible, so that the equations must be combined by adding equations (7.153) and (7.159). This gives

$$(m + m_b)\ddot{x} + m_b L_b \cos\theta_b \ddot{\theta}_b - m_b \dot{\theta}_b^2 L_b \sin\theta_b = \frac{K_T}{r}u \tag{7.160}$$

where $m = m_v + J_m/r^2$.

The model of cart, pendulum and motor is

$$(m + m_b)\ddot{x} + m_b L_b \cos\theta_b \ddot{\theta}_b - m_b \dot{\theta}_b^2 L_b \sin\theta_b = \frac{K_T}{r}u \tag{7.161}$$

$$m_b L_b^2 \ddot{\theta}_b + m_b L_b \ddot{x} \cos\theta_b = m_b L_b g \sin\theta_b \tag{7.162}$$

7.5.2 Double inverted pendulum

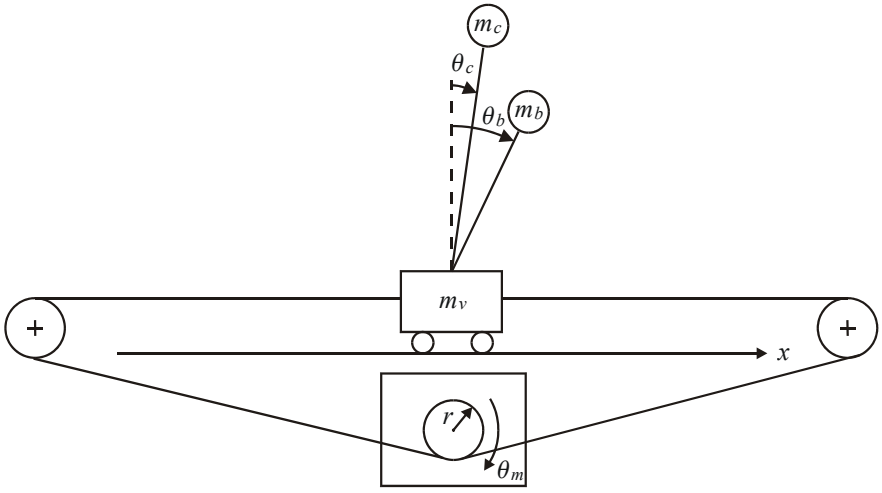

Figure 7.7: Double inverted pendulum.

A double pendulum system is obtained by adding one more pendulum to the cart and pendulum system as shown in Figure 7.7. The variables of the second pendulum are denoted with a subscript c. The position of the point mass m_c of the second pendulum is

$$\vec{r}_c = x\vec{n}_1 - L_c\vec{c}_2 \tag{7.163}$$

The velocity is

$$\vec{v}_c = \dot{x}\vec{n}_1 + \dot{\theta}_c L_c \vec{c}_1 \tag{7.164}$$

and acceleration is

$$\vec{a}_c = \ddot{x}\vec{n}_1 + \ddot{\theta}_c L\vec{c}_1 + \dot{\theta}_c^2 L_c \vec{c}_2 \tag{7.165}$$

Newton's law for the point mass of the second pendulum gives

$$m_c\left(\ddot{x} + \ddot{\theta}_c L_c \cos\theta_c - \dot{\theta}_c^2 L_c \sin\theta_c\right) = \vec{F}_c \cdot \vec{n}_1 \tag{7.166}$$

Newton's law for the cart is modified by one additional term, which is due to the contact force from the second pendulum. This gives

$$m_v\ddot{x} = F - \vec{F}_b \cdot \vec{n}_1 - \vec{F}_c \cdot \vec{n}_1 \tag{7.167}$$

The torque law for the second pendulum about the connection point is

$$m_c L_c g \sin\theta_c = m_c L_c \ddot{x}\cos\theta_c + m_c L_c^2 \ddot{\theta}_c \tag{7.168}$$

The model for a cart and two pendulums then is found to be

$$\begin{aligned}
(m_v + m_b + m_c)\ddot{x} + m_b L_b \ddot{\theta}_b \cos\theta_b + m_c L_c \ddot{\theta}_c \cos\theta_c & \\
-m_b \dot{\theta}_b^2 L_b \sin\theta_b - m_c \dot{\theta}_c^2 L_c \sin\theta_c &= F && (7.169)\\
m_b L^2 \ddot{\theta}_b + m_b L \ddot{x}\cos\theta_b &= m_b L_b g \sin\theta_b && (7.170)\\
m_c L_c^2 \ddot{\theta}_c + m_c L_c \ddot{x}\cos\theta_c &= m_c L_c g \sin\theta_c && (7.171)
\end{aligned}$$

The motor model is included by inserting

$$F = \frac{K_T}{r}u - \frac{J_m}{r^2}\ddot{x} \tag{7.172}$$

7.6 Example: The Furuta pendulum

The Furuta pendulum is a laboratory example where a rotational joint with vertical axis of rotation is used to balance an inverted pendulum (Aström and Furuta 2000). The inertial frame n is defined with the $\vec{n}_3$ axis vertically upwards. The frame b is obtained by a rotation θ_1 about the $\vec{n}_3$ vector, and the frame c is obtained by a rotation θ_2 about the $\vec{b}_2$ axis (Figure 7.8). According to (6.103) the frames n and b have direction cosines

$$\begin{aligned}
\vec{n}_1 \cdot \vec{b}_1 &= \cos\theta_1, \quad \vec{n}_1 \cdot \vec{b}_2 = -\sin\theta_1 && (7.173)\\
\vec{n}_2 \cdot \vec{b}_1 &= \sin\theta_1, \quad \vec{n}_2 \cdot \vec{b}_2 = \cos\theta_1 && (7.174)\\
\vec{n}_1 \cdot \vec{b}_3 &= \vec{n}_2 \cdot \vec{b}_3 = \vec{n}_3 \cdot \vec{b}_1 = \vec{n}_3 \cdot \vec{b}_2 = 0, \quad \vec{n}_3 \cdot \vec{b}_3 = 1 && (7.175)
\end{aligned}$$

and the unit vectors of frame b and frame c have direction cosines

$$\begin{aligned}
\vec{b}_1 \cdot \vec{c}_1 &= \cos\theta_2, \quad \vec{b}_1 \cdot \vec{c}_3 = \sin\theta_2 && (7.176)\\
\vec{b}_1 \cdot \vec{c}_2 &= \vec{b}_3 \cdot \vec{c}_2 = \vec{b}_2 \cdot \vec{c}_1 = \vec{b}_2 \cdot \vec{c}_3 = 0, \quad \vec{b}_2 \cdot \vec{c}_2 = 1 && (7.177)\\
\vec{b}_3 \cdot \vec{c}_1 &= -\sin\theta_2, \quad \vec{b}_3 \cdot \vec{c}_3 = \cos\theta_2 && (7.178)
\end{aligned}$$

It is noted that

$$\begin{aligned}
\vec{b}_1 &= \cos\theta_2\vec{c}_1 + \sin\theta_2\vec{c}_3 && (7.179)\\
\vec{b}_3 &= -\sin\theta_2\vec{c}_1 + \cos\theta_2\vec{c}_3 && (7.180)
\end{aligned}$$

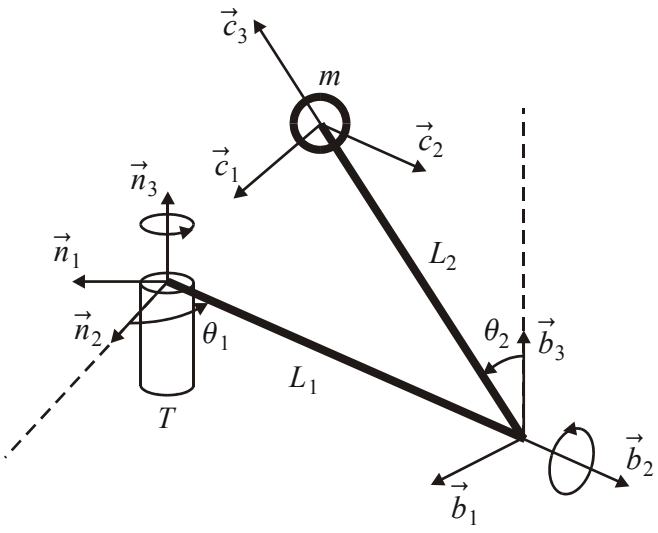

Figure 7.8: Coordinate frames n, b and c used in the description of the Furuta pendulum.

and that

$$b_1 \times \vec{c}_3 = -\cos\theta_2 \vec{c}_2 \tag{7.181}$$

$$\vec{b}_3 \times \vec{c}_1 = \cos\theta_2 \vec{c}_2, \quad \vec{b}_3 \times \vec{c}_2 = -\cos\theta_2 \vec{c}_1 - \sin\theta_2 \vec{c}_3, \quad \vec{b}_3 \times \vec{c}_3 = \sin\theta_2 \vec{c}_2 \tag{7.182}$$

The acceleration of gravity is $\vec{g} = -g\vec{n}_3$. The first link is rotated with angular velocity

$$\vec{\omega}_1 = \dot{\theta}_1 \vec{b}_3 \tag{7.183}$$

and the second link is rotated with angular velocity

$$\vec{\omega}_2 = \dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2 \tag{7.184}$$

The position of the mass

$$\vec{r} = L_1 \vec{b}_2 + L_2 \vec{c}_3$$

The velocity is

$$\begin{aligned} \vec{v} &= L_1 \dot{\theta}_1 \vec{b}_3 \times \vec{b}_2 + L_2 \left(\dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2 \right) \times \vec{c}_3 \\ &= -L_1 \dot{\theta}_1 \vec{b}_1 + L_2 \dot{\theta}_1 \sin\theta_2 \vec{c}_2 + L_2 \dot{\theta}_2 \vec{c}_1 \end{aligned} \tag{7.185}$$

and the acceleration is

$$\begin{aligned} \vec{a} &= -L_1 \ddot{\theta}_1 \vec{b}_1 + L_2 \ddot{\theta}_1 \sin\theta_2 \vec{c}_2 + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos\theta_2 \vec{c}_2 + L_2 \ddot{\theta}_2 \vec{c}_1 \\ &\quad -L_1 \dot{\theta}_1 \dot{\theta}_1 \vec{b}_3 \times \vec{b}_1 + \left(\dot{\theta}_1 \vec{b}_3 + \dot{\theta}_2 \vec{c}_2 \right) \times \left(L_2 \dot{\theta}_1 \sin\theta_2 \vec{c}_2 + L_2 \dot{\theta}_2 \vec{c}_1 \right) \\ &= -L_1 \ddot{\theta}_1 \vec{b}_1 + L_2 \ddot{\theta}_1 \sin\theta_2 \vec{c}_2 + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos\theta_2 \vec{c}_2 + L_2 \ddot{\theta}_2 \vec{c}_1 \\ &\quad -L_1 \dot{\theta}_1^2 \vec{b}_2 - L_2 \dot{\theta}_1^2 \sin\theta_2 \left(\cos\theta_2 \vec{c}_1 + \sin\theta_2 \vec{c}_3 \right) + L_2 \dot{\theta}_1 \dot{\theta}_2 \cos\theta_2 \vec{c}_2 - L_2 \dot{\theta}_2^2 \vec{c}_3 \end{aligned}$$

which gives

$$\begin{aligned}\vec{a} &= -L_1\ddot{\theta}_1\vec{b}_1 + \left(L_2\ddot{\theta}_1\sin\theta_2 + 2L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_2 - L_1\dot{\theta}_1^2\right)\vec{b}_2 \\ &\quad + \left(L_2\ddot{\theta}_2 - L_2\dot{\theta}_1^2\sin\theta_2\cos\theta_2\right)\vec{c}_1 - \left(L_2\dot{\theta}_1^2\sin^2\theta_2 + L_2\dot{\theta}_2^2\right)\vec{c}_3 \end{aligned} \tag{7.186}$$

The kinematic equations have now been established, and we will develop the equations of motion. Newton's law for the mass gives

$$\vec{F} = m\vec{a}$$

where F is the force on link 2 from link 1. The the torque law for the first angle is

$$T = J_1\ddot{\theta}_1 + (\vec{r}\times m\vec{a})\cdot\vec{b}_3 \tag{7.187}$$

After some relative extensive vector calculations we may find that

$$\begin{aligned}(\vec{r}\times m\vec{a})\cdot\vec{b}_3 &= \left[\left(L_1\vec{b}_2 + L_2\vec{c}_3\right)\times m\vec{a}\right]\cdot\vec{b}_3 \\ &= mL_1^2\ddot{\theta}_1 - mL_1L_2\cos\theta_2\ddot{\theta}_2 + mL_1L_2\dot{\theta}_2^2\sin\theta_2 \\ &\quad + mL_2^2\ddot{\theta}_1\sin^2\theta_2 + 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2 \end{aligned} \tag{7.188}$$

This gives the equation of motion

$$\begin{aligned}& J_1\ddot{\theta}_1 + mL_1^2\ddot{\theta}_1 + mL_2^2\sin^2\theta_2\ddot{\theta}_1 - mL_1L_2\cos\theta_2\ddot{\theta}_2 \\ =\ & T - mL_1L_2\dot{\theta}_2^2\sin\theta_2 - 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2 \end{aligned} \tag{7.189}$$

The equation of motion for the second link is found from

$$L_2\vec{c}_3\times(-mg\vec{b}_3) = L_2\vec{c}_3\times m\vec{a} \tag{7.190}$$

where $\vec{r}_g = L_2\vec{c}_3$ in the notation of (7.95). The component of (7.190) in the $\vec{c}_2$ direction is

$$L_2mg\sin\theta_2 = mL_2\vec{c}_3\times\left(-L_1\ddot{\theta}_1\vec{b}_1 + L_2\ddot{\theta}_2\vec{c}_1 - L_2\dot{\theta}_1^2\sin\theta_2\cos\theta_2\vec{c}_1\right)\cdot\vec{c}_2 \tag{7.191}$$

which is simplified to

$$L_2mg\sin\theta_2 = mL_2^2\ddot{\theta}_2 - mL_1L_2\cos\theta_2\ddot{\theta}_1 - mL_2^2\dot{\theta}_1^2\sin\theta_2\cos\theta_2 \tag{7.192}$$

We may then conclude as follows:

The dynamic model of the Furuta pendulum is

$$\begin{aligned}&(J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\ddot{\theta}_1 - mL_1L_2\cos\theta_2\ddot{\theta}_2 \\ &= T - mL_1L_2\dot{\theta}_2^2\sin\theta_2 - 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2 \end{aligned} \tag{7.193}$$

$$-mL_1L_2\cos\theta_2\ddot{\theta}_1 + mL_2^2\ddot{\theta}_2 = mL_2^2\dot{\theta}_1^2\sin\theta_2\cos\theta_2 + mL_2g\sin\theta_2 \tag{7.194}$$

Example 128 *The derivation of the dynamic model of the Furuta pendulum is quite complicated, and it is important to check for errors in the model. This can be done by investigating if the model satisfies the energy flow requirement that the time derivative of the total energy is equal to the power $\dot{\theta}_1 T$ supplied by the motor torque T. The total energy is the kinetic energy of the first rotational link, the kinetic energy of the mass, and the potential energy due to gravity. This gives*

$$\begin{aligned} V &= \frac{1}{2}J_1\dot{\theta}_1^2 + \frac{1}{2}m\vec{v}\cdot\vec{v} + mgL_2\cos\theta_2 \\ &= \frac{1}{2}J_1\dot{\theta}_1^2 + \frac{1}{2}m\left(L_1^2\dot{\theta}_1^2 + L_2^2\sin^2\theta_2\dot{\theta}_1^2 + L_2^2\dot{\theta}_2^2 - 2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_2\right) + mgL_2\cos\theta_2 \\ &= \frac{1}{2}(J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\dot{\theta}_1^2 + \frac{1}{2}mL_2^2\dot{\theta}_2^2 \\ &\quad -mL_1L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_2 + mgL_2\cos\theta_2 \end{aligned} \tag{7.195}$$

The time derivative for the solutions of the system is

$$\begin{aligned} \dot{V} &= \dot{\theta}_1\left[(J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\ddot{\theta}_1 - mL_1L_2\ddot{\theta}_2\cos\theta_2\right] \\ &\quad +\dot{\theta}_2\left(mL_2^2\ddot{\theta}_2 - mL_1L_2\ddot{\theta}_1\cos\theta_2\right) \\ &\quad +mL_1L_2\dot{\theta}_1\dot{\theta}_2^2\sin\theta_2 + mL_2^2\dot{\theta}_1^2\dot{\theta}_2\sin\theta_2\cos\theta_2 - \dot{\theta}_2 mgL_2\sin\theta_2 \\ &= \dot{\theta}_1\left(T - mL_1L_2\dot{\theta}_2^2\sin\theta_2 - 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2\right) \\ &\quad +\dot{\theta}_2\left(mL_2^2\dot{\theta}_1^2\sin\theta_2\cos\theta_2 + mL_2 g\sin\theta_2\right) \\ &\quad +mL_1L_2\dot{\theta}_1\dot{\theta}_2^2\sin\theta_2 + mL_2^2\dot{\theta}_1^2\dot{\theta}_2\sin\theta_2\cos\theta_2 - \dot{\theta}_2 mgL_2\sin\theta_2 \\ &= \dot{\theta}_1 T \end{aligned} \tag{7.196}$$

This shows that the model is consistent with the energy flow dynamics.

7.7 Principle of virtual work

7.7.1 Introduction

The equations of motion give the relation between the forces and torques acting on the system and the resulting accelerations. There are two classes of forces that are important in this connection: The active forces, which are also termed actuator forces, and the forces of constraint. In the design of a control system we are mainly concerned with the actuator forces, which can be command to achieve a specified motion. In contrast to this, the main concern in a mechanical design will be the forces of constraint, which are forces that ensure that the mechanical system is not damaged, and which ensure the system does not break into parts. The following examples illustrate the contrast between the two classes of forces:

- In the design of a robot control system we are interested in the motor torques required for a desired acceleration. In the mechanical design of a robot it is different, then it is important that the forces of constraint that appear in the bearings of the joints are within acceptable limits so that the joint is not damaged. Note that as long as the robot joints are intact, the forces of constraint are not relevant in the control systems design.

- In speed control of a train the control problem is to set up an engine force that give a desired acceleration. The mechanical design problem is to ensure that the tracks and the wheels can support the forces of constraint, which in this case are the forces required to keep the train on the track.
- For a football player the motion control problem is to use the muscles of the leg to set up active forces that result in a desired motion. For the knee the muscles will provide the active forces that rotate the knee about its axis of rotation. The forces of constraint will keep the knee join together so that is not damaged. As long as the joint is strong enough, the football player need not be concerned about the forces of constraint.

From this we get the idea that in the design of control systems we need not know the forces of constraint to get the solutions we are seeking. It turns out that it may be quite complicated to derive the forces of constraint, and therefore it seems to be attractive to find a way to eliminate the forces of constraint from the equations of motion. The principle of virtual work is a tool that allows us to do this, but first we have to introduce generalized coordinates and the concept of virtual displacements.

7.7.2 Generalized coordinates

Consider N particles numbered by $k = 1, \ldots, N$. Each particle is of mass m_k and has position

$$\vec{r}_k = x_k \vec{i}_1 + y_k \vec{i}_2 + z_k \vec{i}_3 \tag{7.197}$$

in an Newtonian coordinate frame i with orthogonal unit vectors $\vec{i}_1, \vec{i}_2, \vec{i}_3$ along the axes. The position vectors $\vec{r}_k$ define the *configuration* of the system. The resultant force on each particle is $\vec{F}_k^{(r)}$. Newton's law for each particle is given by

$$m_k \frac{d^2}{dt^2} \vec{r}_k = \vec{F}_k^{(r)} \tag{7.198}$$

Note that all differentiations of vectors are done in the Newtonian frame i in this section. Adding over all particles gives

$$\sum_{k=1}^{N} m_k \frac{d^2}{dt^2} \vec{r}_k = \sum_{k=1}^{N} \vec{F}_k^{(r)} \tag{7.199}$$

Suppose that there is an n-dimensional column vector $\mathbf{q} = (q_1, \ldots, q_n)^T$ so that the position $\vec{r}_k$ of all particles are given as functions of $\mathbf{q}$ and t, that is,

$$\vec{r}_k = \vec{r}_k \left[\mathbf{q}(t), t \right] \tag{7.200}$$

Then the variables $q_1, \ldots, q_n$ are called the *generalized coordinates* of the system. If n is the minimum number of generalized coordinates that define the configuration of the system, then $q_1, \ldots, q_n$ will in addition be termed the minimal coordinates. The n-dimensional space described by the generalized coordinates is called the *configuration space* of the system.

The velocity of particle k can be expressed in terms of the generalized coordinates according to

$$\vec{v}_k = \frac{d}{dt} \vec{r}_k = \sum_{i=1}^{n} \frac{\partial \vec{r}_k}{\partial q_i} \dot{q}_i + \frac{\partial \vec{r}_k}{\partial t} \tag{7.201}$$

Example 129 *For later use we note that (7.201) implies that partial differentiation of* $\vec{v}_k$ *with respect to* $\dot{q}_i$ *gives*

$$\frac{\partial \vec{v}_k}{\partial \dot{q}_i} = \frac{\partial \vec{r}_k}{\partial q_i} \tag{7.202}$$

Moreover, we find that by interchanging derivation with respect to q_i *and* t *that*

$$\frac{\partial \vec{v}_k}{\partial q_i} = \frac{\partial}{\partial q_i}\frac{d\vec{r}_k}{dt} = \frac{d}{dt}\frac{\partial \vec{r}_k}{\partial q_i} \tag{7.203}$$

7.7.3 Virtual displacements

We now introduce the concept of virtual displacements which is very important in dynamics. The *virtual displacement* $\delta\vec{r}_k$ of particle k is defined by

$$\delta\vec{r}_k = \sum_{i=1}^{n} \frac{\partial \vec{r}_k}{\partial q_i}\delta q_i \tag{7.204}$$

where δq_i is the virtual displacement in the generalized coordinate q_i. If the time derivatives $\dot{q}_i$ of the generalized coordinates are independent, then the virtual displacements δq_i are linearly independent, and there are n independent virtual displacements $\delta\vec{r}_k$, and the system is said to have n *degrees of freedom*.

If there is a linear constraint on the generalized velocities $\dot{q}_i$ given by

$$\mathbf{A}(\mathbf{q})\dot{\mathbf{q}} = \mathbf{0} \tag{7.205}$$

then the virtual displacements of the generalized coordinates will satisfy

$$\mathbf{A}(\mathbf{q})\delta\mathbf{q} = \mathbf{0} \tag{7.206}$$

where $\delta\mathbf{q} = (\delta q_1, \ldots, \delta q_n)^T$. If the null-space of $\mathbf{A}(\mathbf{q})$ has dimension $n_{dof} \leq n$, which means that there are n_{dof} independent generalized velocities $\dot{q}_i$, then there are n_{dof} independent virtual displacements $\delta\vec{r}_k$ and the system is said to have n_{dof} degrees of freedom.

7.7.4 d'Alembert's principle

From the outset there are N particles, each with three coordinates, hence, if the particles are moving independently of each other, then a system of N particles will have $3N$ degrees of freedom. However, to satisfy constraints of the form $\vec{r}_k = \vec{r}_k[\mathbf{q}(t), t]$ where the velocities $\dot{q}_i$ are independent, the system will have only n degrees of freedom. To make these constraints hold there must be certain forces acting on the particles. Such forces can be characterized in a number of ways, but it turns out to appropriate to define *forces of constraints* $\vec{F}_k^{(c)}$ that satisfy *the principle of virtual work* which is given by

$$\sum_{k=1}^{N} \delta\vec{r}_k \cdot \vec{F}_k^{(c)} = 0 \tag{7.207}$$

Here $\vec{F}_k^{(c)}$ is the force of constraint acting on particle k. Then the resultant force on particle k is given by

$$\vec{F}_k^{(r)} = \vec{F}_k^{(c)} + \vec{F}_k \tag{7.208}$$

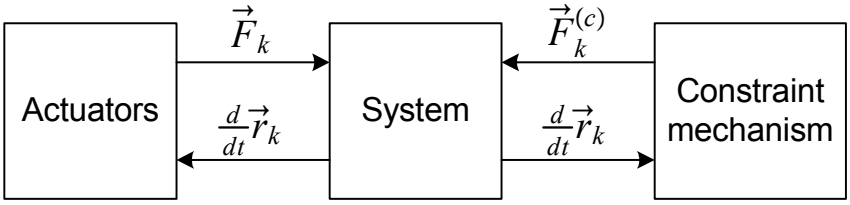

Figure 7.9: A mechanical system is driven by actuators that set up active forces $\vec{F}_k$ acting on the system. In addition there is a constraint mechanism that set up constraint forces $\vec{F}_k^{(c)}$ that ensures that the system does not break in parts. It simplifies the equation of motion greatly if the forces of constraint are formulated so that they do not influence on the action of the active forces, which is achieved by the principle of virtual work. The forces of constraint can then be eliminated from the equations of motion.

where $\vec{F}_k$ is the active force on particle k.

The principle of virtual work can now be used to eliminate the forces of constraint $\vec{F}_k^{(c)}$ from the equation of motion. This is done by taking the scalar product between the equation of motion for particle k and the virtual displacement $\delta\vec{r}_k$, and then summing over all particles. This gives

$$\begin{aligned}\sum_{k=1}^{N} \delta\vec{r}_k \cdot m_k \frac{d^2\vec{r}_k}{dt^2} &= \sum_{k=1}^{N} \delta\vec{r}_k \cdot \vec{F}_k^{(c)} + \sum_{k=1}^{N} \delta\vec{r}_k \cdot \vec{F}_k \\ &= \sum_{k=1}^{N} \delta\vec{r}_k \cdot \vec{F}_k \end{aligned} \tag{7.209}$$

and we arrive at the following formulation of the equation of motion

$$\sum_{k=1}^{N} \delta\vec{r}_k \cdot \left(m_k \frac{d^2\vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \tag{7.210}$$

which is called *d'Alembert's principle*. Note that the only forces appearing in this formulation are the externally applied forces $\vec{F}_k$.

If we insert the expression for $\delta\vec{r}_k$ from (7.204) and change the order of the summation, we find that

$$\sum_{i=1}^{n} \delta q_i \sum_{k=1}^{N} \frac{\partial\vec{r}_k}{\partial q_i} \cdot \left(m_k \frac{d^2\vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \tag{7.211}$$

If the virtual displacements δq_i in the generalized coordinates are independent, then d'Alembert's principle can be written

$$\sum_{k=1}^{N} \frac{\partial\vec{r}_k}{\partial q_i} \cdot \left(m_k \frac{d^2\vec{r}_k}{dt^2} - \vec{F}_k \right) = 0 \tag{7.212}$$

Example 130 *A train is running along a railway. The generalized coordinate of the*

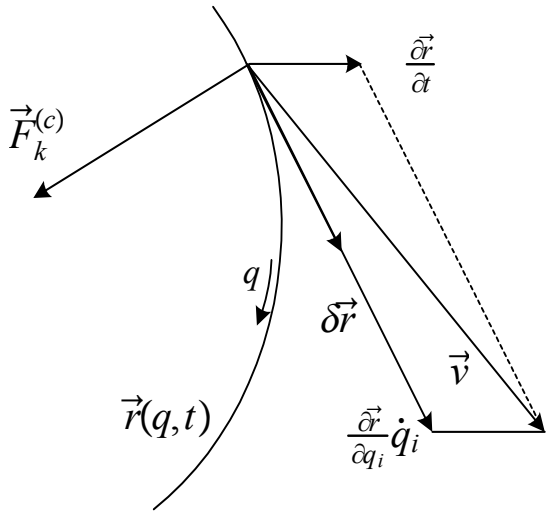

Figure 7.10: Train running along a track $\mathbf{r}(q,t)$ where the track has a velocity $\partial\mathbf{r}/\partial t$ due to the rotation of the earth. The virtual displacement $\delta\mathbf{r}$ is along the track, and the force of constraint $\mathbf{F}^{(c)}$ is perpedicular to the track in accordance with the principle of virtual work.

train is the position coordinate along the railway track, which is denoted q. The position of the train in a Newtonian coordinate frame is $\vec{r}(q,t)$, and the velocity of the train is

$$\vec{v}=\frac{d\vec{r}}{dt}=\frac{\partial\vec{r}}{\partial q}\dot{q}+\frac{\partial\vec{r}}{\partial t} \tag{7.213}$$

Here the first term on the right side is due to the motion along the track, and the other term is due to the rotation of the earth. The virtual displacement of the train is defined as

$$\delta\vec{r}=\frac{\partial\vec{r}}{\partial q}\delta q \tag{7.214}$$

which is a vector tangent to the track. The train is subjected to the resultant force $\vec{F}^{(r)}=\vec{F}+\vec{F}^{(c)}$ where $\vec{F}^{(c)}$ is the constraint force that keeps the train on the track, and $\vec{F}$ is the motor and braking force that controls the velocity of the train. The principle of virtual work then simply states that

$$\delta\vec{r}\cdot\vec{F}^{(c)}=0 \tag{7.215}$$

The physical interpretation of this is that $\vec{F}^{(c)}$ is normal to the track because $\delta\vec{r}$ is tangent to the track. This is illustrated in Figure 7.10.

Example 131 *In coordinate form we have*

$$\delta\mathbf{r}_k=\mathbf{J}_k\delta\mathbf{q} \tag{7.216}$$

The principle of virtual work states that for all $\delta\mathbf{q}$ we have

$$0=\sum_{k=1}^{N}\delta\mathbf{r}_k^T\mathbf{F}_k^{(c)}=\delta\mathbf{q}^T\sum_{k=1}^{N}\mathbf{J}_k^T\mathbf{F}_k^{(c)} \tag{7.217}$$

This means that the constraint force $\mathbf{F}_k^{(c)}$ *satisfies*

$$\sum_{k=1}^{N} \mathbf{J}_k^T \mathbf{F}_k^{(c)} = \mathbf{0} \tag{7.218}$$

which shows that $\mathbf{F}_k^{(c)}$ *is in the null-space of* $\mathbf{J}^T$*, while the active force* $\mathbf{F}_k$ *is the part of the resultant force* $\mathbf{F}_k^{(r)} = \mathbf{F}_k + \mathbf{F}_k^{(c)}$ *that is in the range space of* $\mathbf{J}_k$*. Extensive treatment of null-spaces and range space is found in (Strang 1988).*

7.8 Principle of virtual work for a rigid body

7.8.1 Virtual displacements for a rigid body

The configuration of a rigid body b can be given by a rotation matrix $\mathbf{R}_b^i$ and a position $\vec{r}_c$ of the center of the mass. The velocity is given by the velocity $\vec{v}_c$ of the center of mass and the angular velocity $\vec{\omega}_{ib}$ of frame b relative to frame i. The forces and moments acting on the rigid body are represented by a force $\vec{F}_{bc}$ with line of action through the mass center, and a torque $\vec{T}_{bc}$. The force $\vec{F}_{bc}$ can be split into an active force $\vec{F}_{bc}^{(a)}$ and a constraint force $\vec{F}_{bc}^{(c)}$. In the same way the $\vec{T}_{bc}$ can be described as a sum of an active torque $\vec{T}_{bc}^{(a)}$ and a constraint torque $\vec{T}_{bc}^{(c)}$ so that

$$\vec{F}_{bc} = \vec{F}_{bc}^{(a)} + \vec{F}_{bc}^{(c)}, \quad \vec{T}_{bc} = \vec{T}_{bc}^{(a)} + \vec{T}_{bc}^{(c)} \tag{7.219}$$

The constraint force $\vec{F}_b^{(c)}$ and the constraint torque $\vec{T}_b^{(c)}$ can be eliminated with the principle of virtual work. To do this we will have to define virtual displacements corresponding to the velocity $\vec{v}_c$ and the angular velocity $\vec{\omega}_{ib}$.

It is assumed that the configuration of the body is described by $n \leq 6$ generalized coordinates q_j. Then the velocity and angular velocity are given by

$$\vec{v}_c = \sum_{j=1}^{n} \vec{v}_{c,j} \dot{q}_j + \vec{v}_t \tag{7.220}$$

$$\vec{\omega}_{ib} = \sum_{j=1}^{n} \vec{\omega}_{ib,j} \dot{q}_j + \vec{\omega}_t \tag{7.221}$$

where

$$\vec{v}_{c,j} = \frac{\partial \vec{r}_c}{\partial q_j} = \frac{\partial \vec{v}_c}{\partial \dot{q}_j}, \quad \vec{\omega}_{ib,j} = \frac{\partial \vec{\omega}_{ib}}{\partial \dot{q}_j} \tag{7.222}$$

Following the terminology of (Kane and Levinson 1985), $\vec{v}_{c,j}$ is called partial velocity j and $\vec{\omega}_{ib,j}$ is called partial angular velocity j. The virtual displacements may then be defined by

$$\delta\vec{r}_c = \sum_{j=1}^{n} \vec{v}_{c,j} \delta q_j \tag{7.223}$$

$$\vec{\sigma}_{ib} = \sum_{j=1}^{n} \vec{\omega}_{ib,j} \delta q_j \tag{7.224}$$

Chapter 8

Analytical mechanics

8.1 Introduction

The term analytical mechanics was introduced by Lagrange with his work *Mécanique Analytique* which was published in 1788. In this work Lagrange emphasized the use of algebraic operations in the derivation and analysis of equations of motion as opposed to the earlier works of Newton and Euler which relied on vector operations. In our presentation of analytical mechanics we will first explore Lagrangian dynamics, which is based on the use of generalized coordinates, generalized forces and energy functions. Then we will present a related formulation based on the Euler-Poincaré equation, where dynamics on $SO(3)$ and $SE(3)$ can be described using energy functions without the reliance on generalized coordinates. Finally the extended Hamilton's principle and Hamilton's equations of motion will be presented. These methods are energy-based, and quite useful as they provide a systematic way of deriving energy functions that are potential Lyapunov function candidates. Moreover, Hamilton's principle and Hamilton's equations of motion provide the basis for the Hamilton-Jacobi equation which is important in optimal control theory. The material in this chapter is based on classical texts on dynamics like (Goldstein 1980) and (Lovelock and Rund 1989), more recent text on dynamics like (Arnold 1989) and (Marsden and Ratiu 1994), and robotics books like (Spong and Vidyasagar 1989), (Sciavicco and Siciliano 2000) and (Murray et al. 1994). The results that will be presented in this chapter are well established in the dynamics literature. However, a control engineer will have to consult a great number of books, some of which are quite advanced, to find the selection of analysis tools that will be presented here. Note that although some of the material may seem to be abstract at a first reading, the methods are of great use in practical controller design and analysis, and in the development of simulation systems.

8.2 Lagrangian dynamics

8.2.1 Introduction

The equations of motion for a mechanical system can be derived in the Newton-Euler formulation, which is based on Newton's second law in a vector formulation. It has been documented in robotics that the Newton-Euler equations lead to an efficient formulation suited for computations in real-time control and simulation (Luh et al. 1980). An alter-

native way of deriving the equations of motion is to use Lagrange's formulation which is based on algebraic operations on energy expressions using generalized coordinates and generalized forces. Lagrange's formulation may be better suited to derive results related to energy conservation and passivity, as it is based on the expressions for kinetic and potential energy. This is becoming even more important in control theory as many new controller designs are energy-based using Lyapunov designs or passivity (Slotine 1991), (Krstić, Kanellakopoulos and Kokotović 1995), (Khalil 1996), (Arimoto 1996), (Sepulchre, Janković and Kokotović 1997), (Lozano et al. 2000). Well-known examples in robotics is the independent-joint controller (Takegaki and Arimoto 1981), and the adaptive tracking controller (Slotine and Li 1988), and related results have appeared in other applications like attitude control (Wen and Kreutz-Delgado 1991) and vibration damping (Kelkar and Joshi 1996). It is therefore of great interest to study Lagrange's equation of motion and related concepts of analytical dynamics for use in controller design and analysis.

8.2.2 Lagrange's equation of motion

Lagrange's equations of motion for a mechanical system are equivalent to the Newton-Euler equations of motion, although the methods derive the equations of motion in two different ways. We have already presented Newton-Euler formulations, and we will now show how to derive Lagrange's equation of motion from d'Alembert's principle as presented in Section 7.7 for a system of particles (Goldstein 1980). We consider N particles, where particle k has mass m_k and position $\vec{r}_k(q_1, \ldots, q_n, t)$, where $q_1, \ldots, q_n$ are the generalized coordinates of the system. The velocity of particle k is $\vec{v}_k = d\vec{r}_k/dt$, and the acceleration is $\vec{a}_k = d\vec{v}_k/dt$. Time differentiation and partial differentiation of vectors are in a Newtonian frame in this section.

The starting point for our derivation of Lagrange's equation of motion is d'Alembert's principle in the form (7.211)

$$\sum_{i=1}^{n} \left[\sum_{k=1}^{N} \frac{\partial \vec{r}_k}{\partial q_i} \cdot \left(m_k \vec{a}_k - \vec{F}_k \right) \right] \delta q_i = 0 \tag{8.1}$$

To proceed we introduce the kinetic energy T of the system, which is

$$T = \sum_{k=1}^{N} \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \tag{8.2}$$

We find that

$$\frac{\partial T}{\partial \dot{q}_i} = \frac{\partial}{\partial \dot{q}_i} \left(\sum_{k=1}^{N} \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \right) = \sum_{k=1}^{N} \frac{\partial \vec{v}_k}{\partial \dot{q}_i} \cdot m_k \vec{v}_k = \sum_{k=1}^{N} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \tag{8.3}$$

$$\frac{\partial T}{\partial q_i} = \frac{\partial}{\partial q_i} \left(\sum_{k=1}^{N} \frac{1}{2} m_k \vec{v}_k \cdot \vec{v}_k \right) = \sum_{k=1}^{N} \frac{\partial \vec{v}_k}{\partial q_i} \cdot m_k \vec{v}_k = \sum_{k=1}^{N} \frac{d}{dt} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \tag{8.4}$$

where (7.202) and (7.203) are used. The following calculation can then be done:

$$\begin{aligned} \frac{d}{dt} \frac{\partial T}{\partial \dot{q}_i} &= \sum_{k=1}^{N} \frac{d}{dt} \left(\frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k \right) = \sum_{k=1}^{N} \left(\frac{d}{dt} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{v}_k + \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{a}_k \right) \\ &= \frac{\partial T}{\partial q_i} + \sum_{k=1}^{N} \frac{\partial \vec{r}_k}{\partial q_i} \cdot m_k \vec{a}_k \end{aligned} \tag{8.5}$$

This result combined with (8.1) leads to

$$\sum_{i=1}^{n}\left[\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right)-\frac{\partial T}{\partial q_i}-\sum_{k=1}^{N}\frac{\partial \vec{r}_k}{\partial q_i}\cdot\vec{F}_k\right]\delta q_i=0 \tag{8.6}$$

The third term in the bracket is defined to be the *generalized force*

$$Q_i:=\sum_{k=1}^{N}\frac{\partial \vec{r}_k}{\partial q_i}\cdot\vec{F}_k \tag{8.7}$$

associated with the generalized coordinate q_i. This gives

$$\sum_{i=1}^{n}\left[\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right)-\frac{\partial T}{\partial q_i}-Q_i\right]\delta q_i=0 \tag{8.8}$$

Then, under the assumption that the time derivatives $\dot{q}_i$ of the generalized coordinates are independent, the virtual displacements δq_i are arbitrary, and it follows that

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right)-\frac{\partial T}{\partial q_i}=Q_i \tag{8.9}$$

The generalized force Q_i is assumed to be given by a conservative force $-\partial U/\partial q_i$ due to a potential $U=U(\mathbf{q})$ plus the generalized actuator force τ_i. This is written

$$Q_i=-\frac{\partial U}{\partial q_i}+\tau_i \tag{8.10}$$

Then the equation of motion becomes

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right)-\frac{\partial T}{\partial q_i}+\frac{\partial U}{\partial q_i}=\tau_i \tag{8.11}$$

From this result, Lagrange's equation of motion is found:

Lagrange's equation of motion is formulated using the *Lagrangian*

$$L\left(\mathbf{q},\dot{\mathbf{q}},t\right)=T\left(\mathbf{q},\dot{\mathbf{q}},t\right)-U\left(\mathbf{q}\right) \tag{8.12}$$

The equation of motion is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right)-\frac{\partial L}{\partial q_i}=\tau_i \tag{8.13}$$

Example 136 *For use in the part on Hamiltonian dynamics we derive the following result: Time differentiation of the Lagrangian gives*

$$\begin{aligned}\frac{dL\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{dt} &= \frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial \mathbf{q}}\dot{\mathbf{q}}+\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial \dot{\mathbf{q}}}\ddot{\mathbf{q}}+\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial t}\\ &= \left(\frac{d}{dt}\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial \dot{\mathbf{q}}}-\boldsymbol{\tau}^T\right)\dot{\mathbf{q}}+\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial \dot{\mathbf{q}}}\ddot{\mathbf{q}}+\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial t}\end{aligned} \tag{8.14}$$

where Lagrange's equation of motion (8.13) has been inserted. This gives

$$\frac{dL\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{dt}=\frac{d}{dt}\left(\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial \dot{\mathbf{q}}}\dot{\mathbf{q}}\right)+\frac{\partial L\left(\mathbf{q},\dot{\mathbf{q}},t\right)}{\partial t}-\boldsymbol{\tau}^T\dot{\mathbf{q}} \tag{8.15}$$

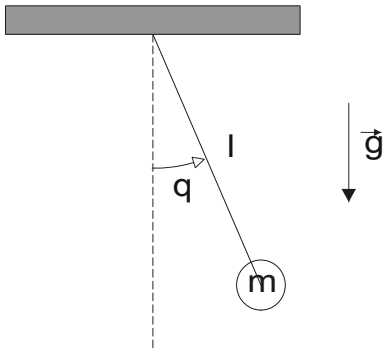

Figure 8.1: Mathematical pendulum.

8.2.3 Generalized coordinates and generalized forces

The power supplied to the system with $\vec{r}_k = \vec{r}_k(\mathbf{q})$ from the forces $\vec{F}_k$ is

$$\begin{aligned}\sum_{k=1}^{N} \frac{d\vec{r}_k}{dt} \cdot \vec{F}_k &= \sum_{k=1}^{N} \left(\sum_{i=1}^{n} \frac{\partial \vec{r}_k}{\partial q_i} \dot{q}_i \cdot \vec{F}_k \right) = \sum_{i=1}^{n} \left(\sum_{k=1}^{N} \frac{\partial \vec{r}_k}{\partial q_i} \cdot \vec{F}_k \right) \dot{q}_i \\ &= \sum_{i=1}^{n} Q_i \dot{q}_i \end{aligned} \tag{8.16}$$

This shows that the product $Q_i \dot{q}_i$ between the generalized force Q_i and the generalized speed $\dot{q}_i$ has dimension power. This means that a system with two degrees of freedom with $q_1 = x$ is a position and $q_2 = \theta$ is an angle, then Q_1 must be a force and Q_2 must be a torque.

8.2.4 Pendulum

A mathematical pendulum is a mass point of mass m in the gravity field which is connected by a massless rod of length L to a frictionless joint with angle q. The pendulum is shown in Figure 8.1. The kinetic energy is

$$T = \frac{1}{2} m v^2 = \frac{1}{2} m \ell^2 \dot{q}^2 \tag{8.17}$$

The potential energy is

$$U = mg\ell \left(1 - \cos q\right) \tag{8.18}$$

The resulting Lagrangian is

$$L = \frac{1}{2} m \ell^2 \dot{q}^2 - mg\ell \left(1 - \cos q\right) \tag{8.19}$$

and the equation of motion is

$$\frac{d}{dt} \left(m \ell^2 \dot{q} \right) + mg\ell \sin q = 0 \tag{8.20}$$

which gives

$$\ddot{q} + \omega_0^2 \sin q = 0, \quad \omega_0 = \sqrt{\frac{g}{\ell}} \tag{8.21}$$

8.2.5 Mass-spring system

A mass-spring system with mass m and spring stiffness k will have Lagrangian

$$L = T - U = \frac{1}{2}m\dot{q}^2 - \frac{1}{2}kq^2 \tag{8.22}$$

Lagrange's equation of motion is then found to be

$$\frac{d}{dt}\left(m\dot{q}\right) + kq = \tau \tag{8.23}$$

which can be written in the familiar form

$$m\ddot{q} + kq = \tau \tag{8.24}$$

8.2.6 Ball and beam

The ball and beam system presented in Section 7.4 has kinetic energy

$$\begin{aligned} T &= \frac{1}{2}J_1\dot{\theta}^2 + \frac{1}{2}J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right)^2 + \frac{1}{2}m\left[\left(\dot{x} + \dot{\theta}R\right)^2 + \left(\dot{\theta}x\right)^2\right] \\ &= \frac{1}{2}\begin{pmatrix} \dot{\theta} \\ \dot{x} \end{pmatrix}^T \begin{pmatrix} J_1 + J_2 + m\left(x^2 + R^2\right) & \frac{1}{R}\left(J_2 + mR^2\right) \\ \frac{1}{R}\left(J_2 + mR^2\right) & m + \frac{J_2}{R^2} \end{pmatrix} \begin{pmatrix} \dot{\theta} \\ \dot{x} \end{pmatrix} \end{aligned} \tag{8.25}$$

and potential energy

$$U = mg\left(R\cos\theta - x\sin\theta\right) \tag{8.26}$$

The generalized coordinates are selected as

$$q_1 = \theta \quad \text{and} \quad q_2 = x \tag{8.27}$$

Then, with $L = T - U$, we have the following partial derivatives

$$\frac{\partial L}{\partial \dot{\theta}} = J_1\dot{\theta} + J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right) + m\left[\left(\dot{x} + \dot{\theta}R\right)R + \left(\dot{\theta}x\right)x\right] \tag{8.28}$$

$$\frac{\partial L}{\partial \dot{x}} = J_2\left(\dot{\theta} + \frac{\dot{x}}{R}\right)\frac{1}{R} + m\left(\dot{x} + \dot{\theta}R\right) \tag{8.29}$$

$$\frac{\partial L}{\partial \theta} = mg\left(R\sin\theta + x\cos\theta\right) \tag{8.30}$$

$$\frac{\partial L}{\partial x} = m\dot{\theta}^2 x + mg\sin\theta \tag{8.31}$$

and the equations of motion can be written

$$\begin{aligned} &\left[J_1 + J_2 + m\left(x^2 + R^2\right)\right]\ddot{\theta} \\ &+\frac{1}{R}\left(J_2 + mR^2\right)\ddot{x} + 2mx\dot{x}\dot{\theta} = mg\left(R\sin\theta + x\cos\theta\right) + \tau \end{aligned} \tag{8.32}$$

$$\frac{1}{R}\left(J_2 + mR^2\right)\ddot{\theta} + \left(m + \frac{J_2}{R^2}\right)\ddot{x} - m\dot{\theta}^2 x = mg\sin\theta \tag{8.33}$$

We note that these equations of motion have the same form as (7.271, 7.272) which were found using the formulation of Kane. We note that the matrix formulation

$$\mathbf{M}\begin{pmatrix} \ddot{\theta} \\ \ddot{x} \end{pmatrix} = \begin{pmatrix} -2mx\dot{x}\dot{\theta} + mg\left(R\sin\theta + x\cos\theta\right) \\ m\dot{\theta}^2 x + mg\sin\theta \end{pmatrix} + \begin{pmatrix} \tau \\ 0 \end{pmatrix} \tag{8.34}$$

has a positive definite and symmetric mass matrix

$$\mathbf{M} = \begin{pmatrix} J_1 + J_2 + m\left(x^2 + R^2\right) & \frac{1}{R}\left(J_2 + mR^2\right) \\ \frac{1}{R}\left(J_2 + mR^2\right) & m + \frac{J_2}{R^2} \end{pmatrix} \tag{8.35}$$

8.2.7 Furuta pendulum

The kinetic energy T and the potential energy U of the Furuta pendulum are given by

$$\begin{aligned} T &= \frac{1}{2}(J_1 + mL_1^2 + mL_2^2 \sin^2\theta_2)\dot{\theta}_1^2 + \frac{1}{2}mL_2^2\dot{\theta}_2^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_2 & (8.36)\\ U &= mgL_2\cos\theta_2 & (8.37) \end{aligned}$$

which give the Lagrangian

$$L = \frac{1}{2}(J_1 + mL_1^2 + mL_2^2 \sin^2\theta_2)\dot{\theta}_1^2 + \frac{1}{2}mL_2^2\dot{\theta}_2^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2\cos\theta_2 - mgL_2\cos\theta_2 \tag{8.38}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial L}{\partial \dot{\theta}_1} &= (J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\dot{\theta}_1 - mL_1L_2\dot{\theta}_2\cos\theta_2 & (8.39)\\ \frac{\partial L}{\partial \dot{\theta}_2} &= mL_2^2\dot{\theta}_2 - mL_1L_2\dot{\theta}_1\cos\theta_2 & (8.40)\\ \frac{\partial L}{\partial \theta_1} &= 0 & (8.41)\\ \frac{\partial L}{\partial \theta_2} &= mL_2^2\sin\theta_2\cos\theta_2\dot{\theta}_1^2 + mL_1L_2\dot{\theta}_1\dot{\theta}_2\sin\theta_2 + mgL_2\sin\theta_2 & (8.42) \end{aligned}$$

and the equations of motion are found by evaluation

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} &= (J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\ddot{\theta}_1 - mL_1L_2\ddot{\theta}_2\cos\theta_2 \\ &\quad + 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2 + mL_1L_2\dot{\theta}_2^2\sin\theta_2 & (8.43) \end{aligned}$$

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_2}\right) - \frac{\partial L}{\partial \theta_2} &= mL_2^2\ddot{\theta}_2 - mL_1L_2\ddot{\theta}_1\cos\theta_2 + mL_1L_2\dot{\theta}_1\dot{\theta}_2\sin\theta_2 \\ &\quad - mL_2^2\sin\theta_2\cos\theta_2\dot{\theta}_1^2 - mL_1L_2\dot{\theta}_1\dot{\theta}_2\sin\theta_2 \\ &\quad - mgL_2\sin\theta_2 & (8.44) \end{aligned}$$

The equations of motion of the Furuta pendulum are

$$\begin{aligned} (J_1 + mL_1^2 + mL_2^2\sin^2\theta_2)\ddot{\theta}_1 - mL_1L_2\ddot{\theta}_2\cos\theta_2 & \\ + 2mL_2^2\dot{\theta}_1\dot{\theta}_2\sin\theta_2\cos\theta_2 + mL_1L_2\dot{\theta}_2^2\sin\theta_2 &= \tau & (8.45)\\ mL_2^2\ddot{\theta}_2 - mL_1L_2\ddot{\theta}_1\cos\theta_2 - mL_2^2\sin\theta_2\cos\theta_2\dot{\theta}_1^2 - mgL_2\sin\theta_2 &= 0 & (8.46) \end{aligned}$$

This result is in agreement with the result derived with the Newton-Euler approach. The Lagrange derivation is much simpler for this system.

8.2.8 Manipulator

In this section we will derive the Lagrangian equations of motion for a manipulator (Spong and Vidyasagar 1989), (Sciavicco and Siciliano 2000). The manipulator has n links which are rigid bodies. The links are assumed to be connected with rotary joints of one degree of freedom. The joint angle of joint i is denoted q_i. The joint angles are the generalized coordinates of the manipulator. The vector of generalized coordinates is denoted $\mathbf{q} = (q_1 \dots q_n)^T$. At each joint there is a motor torque τ_i which are the input generalized forces. The vector of generalized forces is denoted $\boldsymbol{\tau} = (\tau_1 \dots \tau_n)^T$.

The kinetic energy of link i is

$$T_i = \frac{1}{2} m_i (\mathbf{v}_{ci}^i)^T (\mathbf{v}_{ci}^i) + \frac{1}{2} (\boldsymbol{\omega}_{0i}^i)^T \mathbf{M}_{ci}^i \boldsymbol{\omega}_{0i}^i \tag{8.47}$$

where m_i is the mass, $\mathbf{v}_{ci}^i$ is the velocity of the center of mass, $\boldsymbol{\omega}_{0i}^i$ is the angular velocity, and $\mathbf{M}_{ci}^i$ is the inertia matrix around the center of mass. The velocity $\mathbf{v}_{ci}^i$ and the angular velocity $\boldsymbol{\omega}_{0i}^i$ are linear combinations of the time derivatives of the generalized coordinates, and we may write

$$\mathbf{v}_{ci}^i = \sum_{j=1}^{i} \mathbf{v}_{ci,j}^i (\mathbf{q}) \dot{q}_j = \mathbf{J}_{v_{ci}} (\mathbf{q}) \dot{\mathbf{q}} \tag{8.48}$$

$$\boldsymbol{\omega}_{0i}^i = \sum_{j=1}^{i} \boldsymbol{\omega}_{0i,j}^i (\mathbf{q}) \dot{q}_j = \mathbf{J}_{\omega_{0i}} (\mathbf{q}) \dot{\mathbf{q}} \tag{8.49}$$

Then the kinetic energy of link i can be written

$$T_i = \frac{1}{2} m_i \dot{\mathbf{q}}^T \mathbf{J}_{v_{ci}}^T (\mathbf{q}) \mathbf{J}_{v_{ci}} (\mathbf{q}) \dot{\mathbf{q}} + \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{J}_{\omega_{0i}}^T (\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}} (\mathbf{q}) \dot{\mathbf{q}} \tag{8.50}$$

and the total kinetic energy for the manipulator is

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \sum_{i=1}^{n} \left[m_i \mathbf{J}_{v_{ci}}^T (\mathbf{q}) \mathbf{J}_{v_{ci}} (\mathbf{q}) + \mathbf{J}_{\omega_i}^T (\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}} (\mathbf{q}) \right] \tag{8.51}$$

This shows that the kinetic energy of the manipulator can be written as the quadratic form

$$T = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M} (\mathbf{q}) \dot{\mathbf{q}} \tag{8.52}$$

where the $n \times n$ *mass matrix* $\mathbf{M}(\mathbf{q})$ given by

$$\mathbf{M} (\mathbf{q}) = \sum_{i=1}^{n} \left[m_i \mathbf{J}_{v_{ci}}^T (\mathbf{q}) \mathbf{J}_{v_{ci}} (\mathbf{q}) + \mathbf{J}_{\omega_{0i}}^T (\mathbf{q}) \mathbf{M}_{ci}^i \mathbf{J}_{\omega_{0i}} (\mathbf{q}) \right] \tag{8.53}$$

is symmetric. Moreover, the kinetic energy is nonnegative, which implies that $\mathbf{M}(\mathbf{q})$ is positive definite. The potential energy is due to the gravity potential, and is written

$$U (\mathbf{q}) = \sum_{i=1}^{n} U_i (\mathbf{q}) = \sum_{i=1}^{n} m_i \mathbf{g}^T \mathbf{r}_{ci} (\mathbf{q}) \tag{8.54}$$

The Lagrangian of the manipulator is therefore

$$L = \frac{1}{2} \dot{\mathbf{q}}^T \mathbf{M} (\mathbf{q}) \dot{\mathbf{q}} - U(\mathbf{q}) \tag{8.55}$$

The derivation of the Lagrangian equation of motion is a relatively complicated exercise, and we therefore state the main results first and present derivation afterwards.

The equations of motion for a manipulator can be written

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau} \tag{8.56}$$

where $\mathbf{M}(\mathbf{q}) = \mathbf{M}^T(\mathbf{q})$ is positive definite and $\mathbf{g}(\mathbf{q})$ is the gradient of the gravity potential. The matrix $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$ can be selected to be

$$\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) = \{c_{kj}\} = \left\{ \sum_{i=1}^{n} c_{ijk}\dot{q}_i \right\} \tag{8.57}$$

where

$$c_{ijk} := \frac{1}{2}\left(\frac{\partial m_{kj}}{\partial q_i} + \frac{\partial m_{ik}}{\partial q_j} - \frac{\partial m_{ij}}{\partial q_k} \right) \tag{8.58}$$

are the Christoffel symbols of the first kind. In this case the matrix $\dot{\mathbf{M}} - 2\mathbf{C}$ is skew symmetric.

To derive Lagrange's equation of motion it is convenient to use the component form

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}(\mathbf{q})\dot{q}_i\dot{q}_j \tag{8.59}$$

for the kinetic energy, which gives the Lagrangian

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}(\mathbf{q})\dot{q}_i\dot{q}_j - U(\mathbf{q}) \tag{8.60}$$

We find that

$$\begin{aligned} \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} &= \frac{d}{dt}\left(\frac{1}{2}\sum_{j=1}^{n} m_{kj}\dot{q}_j + \frac{1}{2}\sum_{i=1}^{n} m_{ik}\dot{q}_i \right) \\ &= \sum_{j=1}^{n} m_{kj}(\mathbf{q})\ddot{q}_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial m_{kj}}{\partial q_i} + \frac{\partial m_{ik}}{\partial q_j} \right)\dot{q}_i\dot{q}_j \end{aligned} \tag{8.61}$$

and that

$$\frac{\partial L}{\partial q_k} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial m_{ij}}{\partial q_k}\dot{q}_i\dot{q}_j - \frac{\partial U}{\partial q_k} \tag{8.62}$$

The resulting equation of motion is

$$\sum_{j=1}^{n} m_{kj}(\mathbf{q})\ddot{q}_j + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ijk}(\mathbf{q})\dot{q}_i\dot{q}_j + g_k(\mathbf{q}) = \tau_k \tag{8.63}$$

where c_{ijk} are the Christoffel symbols of the first kind as defined by (8.58), and

$$g_k := \frac{\partial U}{\partial q_k} \tag{8.64}$$

Then the equation of motion (8.56) appears by defining the matrix

$$\mathbf{C}\left(\mathbf{q},\dot{\mathbf{q}}\right)=\left\{c_{kj}\left(\mathbf{q},\dot{\mathbf{q}}\right)\right\},\quad c_{kj}\left(\mathbf{q},\dot{\mathbf{q}}\right)=\sum_{i=1}^{n}c_{ijk}\left(\mathbf{q}\right)\dot{q}_i \tag{8.65}$$

and the gravity vector

$$\mathbf{g}\left(\mathbf{q}\right)=\frac{\partial U}{\partial \mathbf{q}}^T \tag{8.66}$$

Finally, we will show that the matrix

$$\mathbf{N}=\dot{\mathbf{M}}-2\mathbf{C} \tag{8.67}$$

is skew symmetric. This is shown by considering element

$$n_{kj}=\dot{m}_{kj}-2c_{kj} \tag{8.68}$$

of the matrix. We find that

$$\begin{aligned}\dot{m}_{kj}-2c_{kj} &= \sum_{i=1}^{n}\left(\frac{\partial m_{kj}}{\partial q_i}-\frac{\partial m_{kj}}{\partial q_i}-\frac{\partial m_{ik}}{\partial q_j}+\frac{\partial m_{ij}}{\partial q_k}\right)\dot{q}_i \\ &= \sum_{i=1}^{n}\left(\frac{\partial m_{ij}}{\partial q_k}-\frac{\partial m_{ik}}{\partial q_j}\right)\dot{q}_i\end{aligned} \tag{8.69}$$

This implies

$$n_{kj}=-n_{jk} \tag{8.70}$$

which shows that $\mathbf{N}$ is skew symmetric.

8.2.9 Passivity of the manipulator dynamics

The time derivative of the energy $E=T+U$ is found by the chain rule to be

$$\begin{aligned}\dot{E}(\mathbf{q},\dot{\mathbf{q}}) &= \frac{d}{dt}\left(\frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\left(\mathbf{q}\right)\dot{\mathbf{q}}\right)+\frac{\partial U}{\partial \mathbf{q}}\dot{\mathbf{q}} \\ &= \dot{\mathbf{q}}^T\mathbf{M}\left(\mathbf{q}\right)\ddot{\mathbf{q}}+\frac{1}{2}\dot{\mathbf{q}}^T\dot{\mathbf{M}}\left(\mathbf{q}\right)\dot{\mathbf{q}}+\frac{\partial U}{\partial \mathbf{q}}\dot{\mathbf{q}}\end{aligned} \tag{8.71}$$

The time derivative along the solutions of the system is found by inserting the equation of motion (8.56) and (8.66). This gives

$$\begin{aligned}\dot{E}(\mathbf{q}) &= \dot{\mathbf{q}}^T\left[-\mathbf{C}\left(\mathbf{q},\dot{\mathbf{q}}\right)\dot{\mathbf{q}}-\mathbf{g}\left(\mathbf{q}\right)+\boldsymbol{\tau}\right]+\frac{1}{2}\dot{\mathbf{q}}^T\dot{\mathbf{M}}\left(\mathbf{q}\right)\dot{\mathbf{q}}+\mathbf{g}\left(\mathbf{q}\right)^T\dot{\mathbf{q}} \\ &= \dot{\mathbf{q}}^T\boldsymbol{\tau}+\frac{1}{2}\dot{\mathbf{q}}^T\left[{}^T\dot{\mathbf{M}}\left(\mathbf{q}\right)-2\mathbf{C}\left(\mathbf{q},\dot{\mathbf{q}}\right)\right]\dot{\mathbf{q}}\end{aligned} \tag{8.72}$$

Finally, the skew symmetry of $\dot{\mathbf{M}}-2\mathbf{C}$ gives the result

$$\dot{E}(\mathbf{q})=\dot{\mathbf{q}}^T\boldsymbol{\tau} \tag{8.73}$$

The kinetic energy is always nonnegative.

If there is a constant $U_{\min}$ so that the potential energy is lower bounded according to $U \geq U_{\min}$, then the storage function $V = T + U - U_{\min} \geq 0$ will have time derivative

$$\dot{V} = \dot{\mathbf{q}}^T \boldsymbol{\tau} \tag{8.74}$$

along the solutions of the system. This implies that the manipulator dynamics (8.56) with input $\boldsymbol{\tau}$ and output $\dot{\mathbf{q}}$ is passive.

8.2.10 Example: Planar two-link manipulator 1

The planar manipulator from Section 7.9.7 has kinetic energy

$$T = \frac{1}{2} m_1 \vec{v}_{c1} \cdot \vec{v}_{c1} + \frac{1}{2} m_2 \vec{v}_{c2} \cdot \vec{v}_{c2} + \frac{1}{2} \vec{\omega}_1 \cdot \vec{M}_{1/c} \cdot \vec{\omega}_1 + \frac{1}{2} \vec{\omega}_2 \cdot \vec{M}_{2/c} \cdot \vec{\omega}_2 \tag{8.75}$$

This can be written

$$T = \frac{1}{2} m_{11} \dot{q}_1^2 + m_{12} \dot{q}_1 \dot{q}_2 + \frac{1}{2} m_{22} \dot{q}_2^2 \tag{8.76}$$

where

$$\begin{aligned} m_{11} &= I_{1z} + I_{2z} + m_1 L_{c1}^2 + m_2 \left(L_1^2 + L_{c2}^2 + 2 L_1 L_{c2} \cos q_2 \right) && (8.77) \\ m_{12} &= m_{21} = I_{2z} + m_2 L_{c2}^2 + m_2 L_1 L_{c2} \cos q_2 && (8.78) \\ m_{22} &= I_{2z} + m_2 L_{c2}^2 && (8.79) \end{aligned}$$

are the elements of the inertia matrix. The potential energy is

$$U = (m_1 g L_{c1} + m_2 g L_1) \sin q_1 + m_2 g L_{c2} \sin(q_1 + q_2) \tag{8.80}$$

Then, from $L = T - U$ the partial derivatives are found to be

$$\frac{\partial L}{\partial \dot{q}_1} = \frac{\partial T}{\partial \dot{q}_1} = m_{11} \dot{q}_1 + m_{12} \dot{q}_2 \tag{8.81}$$

$$\frac{\partial L}{\partial \dot{q}_2} = \frac{\partial T}{\partial \dot{q}_2} = m_{21} \dot{q}_1 + m_{22} \dot{q}_2 \tag{8.82}$$

$$\begin{aligned} \frac{\partial L}{\partial q_1} &= \frac{\partial T}{\partial q_1} - \frac{\partial U}{\partial q_1} = -\left(m_1 L_{c1} + m_2 L_1 \right) g \cos q_1 - m_2 L_{c2} g \cos(q_1 + q_2) && (8.83) \\ \frac{\partial L}{\partial q_2} &= \frac{\partial T}{\partial q_2} - \frac{\partial U}{\partial q_2} = \frac{1}{2} \frac{\partial m_{11}}{\partial q_2} \dot{q}_1^2 + \frac{\partial m_{21}}{\partial q_2} \dot{q}_1 \dot{q}_2 - m_2 g L_{c2} g \cos(q_1 + q_2) && (8.84) \end{aligned}$$

The equations of motion are then found from (8.13) to be

$$\begin{aligned} m_{11} \ddot{q}_1 + m_{12} \ddot{q}_2 + \left(\frac{\partial m_{11}}{\partial q_2} \dot{q}_2 \right) \dot{q}_1 + \left(\frac{\partial m_{12}}{\partial q_2} \dot{q}_2 \right) \dot{q}_2 + \frac{\partial U}{\partial q_1} &= \tau_1 && (8.85) \\ m_{21} \ddot{q}_1 + m_{22} \ddot{q}_2 + \left(\frac{\partial m_{21}}{\partial q_2} \dot{q}_2 \right) \dot{q}_1 - \frac{1}{2} \frac{\partial m_{11}}{\partial q_2} \dot{q}_1^2 - \frac{\partial m_{21}}{\partial q_2} \dot{q}_1 \dot{q}_2 + \frac{\partial U}{\partial q_2} &= \tau_2 && (8.86) \end{aligned}$$

which gives the equations of motion in the form

$$\begin{aligned} \left(I_{1z} + I_{2z} + m_1 L_{c1}^2 + m_2 \left(L_1^2 + L_{c2}^2 + 2 L_1 L_{c2} \cos q_2 \right) \right) \ddot{q}_1 & \\ + \left(I_{2z} + m_2 L_{c2}^2 + m_2 L_1 L_{c2} \cos q_2 \right) \ddot{q}_2 & \\ - m_2 L_1 L_{c2} \sin q_2 \left(2 \dot{q}_1 \dot{q}_2 + \dot{q}_2^2 \right) & \\ + \left(m_1 L_{c1} + m_2 L_1 \right) g \cos q_1 + m_2 L_{c2} g \cos(q_1 + q_2) &= \tau_1 \end{aligned} \tag{8.87}$$

$$\begin{aligned}\left(I_{2z} + m_2 L_{c2}^2 + m_2 L_1 L_{c2} \cos q_2\right) \ddot{q}_1 + \left(I_{2z} + m_2 L_{c2}^2\right) \ddot{q}_2 & \\ + m_2 L_1 L_{c2} \dot{q}_1^2 \sin q_2 + m_2 L_{c2} g \cos(q_1 + q_2) &= \tau_2 \end{aligned} \tag{8.88}$$

8.2.11 Example: Planar two-link manipulator 2

In this section we will see that the equations of motion will be simplified by introducing a following change of generalized coordinates to

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \mathbf{A}\mathbf{q} \tag{8.89}$$

with associated generalized forces

$$\mathbf{K} = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \mathbf{A}^{-T}\boldsymbol{\tau} \tag{8.90}$$

as this gives

$$\mathbf{K}^T \dot{\boldsymbol{\phi}} = \boldsymbol{\tau}^T \mathbf{A}^{-1} \mathbf{A} \dot{\mathbf{q}} = \boldsymbol{\tau}^T \dot{\mathbf{q}} \tag{8.91}$$

Note that $\dot{\phi}_1 = \omega_1$ and $\dot{\phi}_2 = \omega_2$.

With the new set of generalized coordinates the kinetic energy is

$$T = \frac{1}{2}\dot{\mathbf{q}}^T \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} = \frac{1}{2}\dot{\boldsymbol{\phi}}^T \mathbf{A}^{-T} \mathbf{M}(\mathbf{q}) \mathbf{A}^{-1} \dot{\boldsymbol{\phi}} = \frac{1}{2}\dot{\boldsymbol{\phi}}^T \mathbf{D}(\boldsymbol{\phi}) \dot{\boldsymbol{\phi}} \tag{8.92}$$

where the mass matrix $\mathbf{D}(\boldsymbol{\phi}) = \{d_{ij}(\boldsymbol{\phi})\}$ corresponding to the new coordinates $\boldsymbol{\phi}$ is found to be

$$\begin{aligned} \mathbf{D}(\boldsymbol{\phi}) &= \mathbf{A}^{-T} \mathbf{M}(\mathbf{q}) \mathbf{A}^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} m_{11} - 2m_{12} + m_{22} & m_{12} - m_{22} \\ m_{12} - m_{22} & m_{22} \end{pmatrix} \end{aligned} \tag{8.93}$$

which gives

$$\mathbf{D}(\boldsymbol{\phi}) = \begin{pmatrix} I_{1z} + m_1 L_{c1}^2 + m_2 L_1^2 & m_2 L_1 L_{c2} \cos\left(\phi_2 - \phi_1\right) \\ m_2 L_1 L_{c2} \cos\left(\phi_2 - \phi_1\right) & I_{2z} + m_2 L_{c2}^2 \end{pmatrix} \tag{8.94}$$

The equations of motion are then found from (8.13) to be

$$\begin{aligned} d_{11}\ddot{\phi}_1 + d_{22}\ddot{\phi}_2 + \left(\frac{\partial d_{12}}{\partial \phi_1}\dot{\phi}_1 + \frac{\partial d_{12}}{\partial \phi_2}\dot{\phi}_2\right)\dot{\phi}_2 - \left(\frac{\partial d_{12}}{\partial \phi_1}\right)\dot{\phi}_1\dot{\phi}_2 + \frac{\partial U}{\partial \phi_1} &= K_1 \\ d_{21}\ddot{\phi}_1 + d_{22}\ddot{\phi}_2 + \left(\frac{\partial d_{21}}{\partial \phi_1}\dot{\phi}_1 + \frac{\partial d_{21}}{\partial \phi_2}\dot{\phi}_2\right)\dot{\phi}_1 - \left(\frac{\partial d_{21}}{\partial \phi_2}\right)\dot{\phi}_1\dot{\phi}_2 + \frac{\partial U}{\partial \phi_2} &= K_2 \end{aligned}$$

to be

$$\begin{aligned} \left(I_{1z} + m_1 L_{c1}^2 + m_2 L_1^2\right)\ddot{\phi}_1 + m_2 L_1 L_{c2} \cos\left(\phi_2 - \phi_1\right)\ddot{\phi}_2 & \\ - m_2 L_1 L_{c2} \sin\left(\phi_2 - \phi_1\right)\dot{\phi}_2^2 + \left(m_1 L_{c1} + m_2 L_1\right) g \cos\phi_1 &= K_1 \end{aligned} \tag{8.95}$$

$$\begin{aligned} m_2 L_1 L_{c2} \cos\left(\phi_2 - \phi_1\right)\ddot{\phi}_1 + \left(I_{2z} + m_2 L_{c2}^2\right)\ddot{\phi}_2 & \\ + m_2 L_1 L_{c2} \sin\left(\phi_2 - \phi_1\right)\dot{\phi}_1^2 + m_2 L_{c2} g \cos(\phi_2) &= K_2 \end{aligned} \tag{8.96}$$

8.2.12 Limitations of Lagrange's equation of motion

Lagrange's equation of motion is based on the use of a set of generalized coordinates. For many systems the use of generalized coordinates is convenient. Typically, this is the case for robotic manipulators where the joint angles are suitable candidates for the use as generalized coordinates. However, there are other systems which are more efficiently described in terms of the rotation matrix and the angular velocity, and for such systems the use of generalized coordinates may introduce complicated expressions.

To illustrate this we use the rotational dynamics of a rigid body as an example. The kinetic energy is

$$T = \frac{1}{2}\boldsymbol{\omega}^T \mathbf{M}\boldsymbol{\omega} \tag{8.97}$$

where $\boldsymbol{\omega}$ is the angular velocity in body-fixed coordinates, and $\mathbf{M}$ is the constant inertia matrix in body coordinates. The configuration of the rotational dynamics is given by the rotation matrix $\mathbf{R}$. To derive Lagrange's equation of motion for this system we have to select a set of generalized coordinates. The usual set of generalized coordinates for this system is the roll-pitch-yaw angles ψ, θ and ϕ, that is, $\mathbf{q} = (\phi, \theta, \psi)^T$. Then the kinetic energy is found to be

$$T = \frac{1}{2}\dot{\mathbf{q}}^T \mathbf{E}_d^T(\mathbf{q})\mathbf{M}\mathbf{E}_d(\mathbf{q})\dot{\mathbf{q}} \tag{8.98}$$

where

$$\mathbf{E}_d(\mathbf{q}) = \begin{pmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\phi & \sin\phi\cos\theta \\ 0 & -\sin\phi & \cos\phi\cos\theta \end{pmatrix} \tag{8.99}$$

Lagrange's equation of motion is

$$\frac{d}{dt}\left(\mathbf{E}_d^T(\mathbf{q})\mathbf{M}\mathbf{E}_d(\mathbf{q})\dot{\mathbf{q}}\right) - \frac{\partial}{\partial \mathbf{q}}\left(\frac{1}{2}\dot{\mathbf{q}}^T \mathbf{E}_d^T(\mathbf{q})\mathbf{M}\mathbf{E}_d(\mathbf{q})\dot{\mathbf{q}}\right) = \mathbf{0} \tag{8.100}$$

Here, we clearly see that the use of the generalized coordinate vector $\mathbf{q}$ has introduced kinematic terms in the form of the matrix $\mathbf{E}_d(\mathbf{q})$ in the equation of motion. This causes an unnecessary complication of the expressions, and moreover, the matrix $\mathbf{E}_d(\mathbf{q})$ is singular for $\cos\theta = 0$, which introduces a singularity in the mathematical model which is due to the mathematical representation. A great deal of patience is required to arrive at the result

$$\mathbf{M}\dot{\boldsymbol{\omega}} + \boldsymbol{\omega}^{\times}\mathbf{M}\boldsymbol{\omega} = \mathbf{0} \tag{8.101}$$

which is straightforward to derive in the Newton-Euler formulation.

Still, it would be useful if there were some energy-based formulation that resembled Lagrange's equation, but where the definition of generalized coordinates was not required. The form that we will use is the Euler-Poincaré equation of motion, but before we can present it, we need some background material on the calculus of variations. This includes a quite interesting development of the variation of the rotation matrix, and of the homogeneous transformation matrix.

8.3 Calculus of variations

8.3.1 Introduction

In the following the concept of variations in dynamics is discussed, and standard results on the variation of a function is presented. In addition, variational tools on $SO(3)$ and $SE(3)$ are presented.

This leads to the following expression for the Laplacian of the velocity in cylindrical coordinates:

$$\begin{aligned}\vec{\nabla}^2\vec{v} &= \left(\frac{\partial^2 v_r}{\partial r^2}+\frac{1}{r^2}\frac{\partial^2 v_r}{\partial \theta^2}+\frac{\partial^2 v_r}{\partial z^2}+\frac{1}{r}\frac{\partial v_r}{\partial r}-\frac{2}{r^2}\frac{\partial v_\theta}{\partial \theta}-\frac{v_r}{r^2}\right)\vec{j}_r \\ &+\left(\frac{\partial^2 v_\theta}{\partial r^2}+\frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial \theta^2}+\frac{\partial^2 v_\theta}{\partial z^2}+\frac{1}{r}\frac{\partial v_\theta}{\partial r}+\frac{2}{r^2}\frac{\partial v_r}{\partial \theta}-\frac{v_\theta}{r^2}\right)\vec{j}_\theta \\ &+\left(\frac{\partial^2 v_z}{\partial r^2}+\frac{1}{r^2}\frac{\partial^2 v_z}{\partial \theta^2}+\frac{\partial^2 v_z}{\partial z^2}+\frac{1}{r}\frac{\partial v_z}{\partial r}\right)\vec{j}_z \end{aligned} \tag{10.80}$$

10.4 Reynolds' transport theorem

10.4.1 Introduction

In the derivation of balance equations we will typically define a control volume that will depend on the geometry of the specific problem. The modeling procedure will then typically involve the calculation of the rate of change of mass, momentum or energy in the control volume. In connection with this calculation Reynolds' transport theorem is of great use. In the following we will present the theorem and show how it can be adapted to the case where the control volume is a material volume, a spatially fixed volume, or a general control volume with a moving boundary.

10.4.2 Basic transport theorem

The concept of a control volume V_c is used in the derivation of models based on conservation laws. In this section we will present an important kinematic result called Reynolds' transport theorem (Aris 1989), (White 1999). Reynolds' transport theorem shows the relation between the time derivative of the volume integral

$$\iiint_{V_c(t)} \phi(\mathbf{x},t)dV \tag{10.81}$$

and the time derivative of $\phi(\mathbf{x},t)$. The boundary of $V_c(t)$ is denoted $\partial V_c(t)$, and the velocity of a point on the boundary $\partial V_c(t)$ is denoted $\vec{v}_c$. We recall the following standard result from calculus:

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)dx = \int_{a(t)}^{b(t)} \frac{\partial f(x,t)}{\partial t}dx + f\left(b,t\right)\frac{db}{dt} - f\left(a,t\right)\frac{da}{dt}. \tag{10.82}$$

In analogy with this, the time derivative of the integral in (10.81) is equal to the volume integral of the time derivative of the integrand, and one term due to the changing boundary of the volume $V_c(t)$.

For a general time-varying control volume V_c the transport theorem is given by

$$\frac{d}{dt}\iiint_{V_c(t)} \phi(\mathbf{x},t)dV = \iiint_{V_c} \frac{\partial \phi(\mathbf{x},t)}{\partial t}dV + \iint_{\partial V_c} \phi\vec{v}_c\cdot\vec{n}dA \tag{10.83}$$

The last term can be explained as follows: The volume element dA on the surface $\partial V_c(t)$ has velocity $\vec{v}_c$**,** and the rate of change of the integral due to this is $\phi\vec{v}_c\cdot\vec{n}dA$ where $\vec{n}$ is the outwards unit normal of the surface. Integration over the whole surface gives the total rate of change due to the change in the volume $V_c(t)$.

10.4.3 The transport theorem for a material volume

Of particular interest is Reynolds' transport theorem for a *material volume* $V_m(t)$. The reason for this is that balance laws will typically be formulated for material volumes. By material volume it is meant a volume containing a specific set of particles. It is assumed that initially, say at $t = t_0$, these particles filled the volume $V_m(t_0) = V_0$, while at time t the same particles fill the volume $V_m(t) = V$. If we apply Reynolds transport theorem with $V_c = V_m$, then $v_c(t)$ is equal to the particle velocity $v(t)$ and Reynolds transport theorem gives

$$\frac{d}{dt}\iiint_{V_m(t)} \phi(\mathbf{x},t)dV = \iiint_{V_m(t)} \frac{\partial \phi(\mathbf{x},t)}{\partial t} dV + \iint_{\partial V_m(t)} \phi \vec{v}\cdot\vec{n} dA \tag{10.84}$$

To avoid the need to explain whether the volume is material or not, we define the notation

$$\frac{D}{Dt}\iiint_{V} \phi(\mathbf{x},t)dV := \iiint_{V} \frac{\partial \phi(\mathbf{x},t)}{\partial t} dV + \iint_{\partial V} \phi \vec{v}\cdot\vec{n} dA \tag{10.85}$$

Note that in this notation, the volume V need not be a material volume, it is merely assumed that some material volume $V_m(t)$ coincides with V at time t.

The result can be further developed by applying the divergence theorem to the last term on the right side of (10.85), and by using (10.10) and (10.15):

Reynolds' transport theorem for a material volume coinciding with V at time t is given in material form as

$$\frac{D}{Dt}\iiint_{V} \phi(\mathbf{x},t)dV = \iiint_{V} \left[\frac{D\phi(\mathbf{x},t)}{Dt} + \phi\left(\vec{\nabla}\cdot\vec{v}\right)\right] dV \tag{10.86}$$

and in divergence form as

$$\frac{D}{Dt}\iiint_{V} \phi(\mathbf{x},t)dV = \iiint_{V} \left[\frac{\partial \phi(\mathbf{x},t)}{\partial t} + \vec{\nabla}(\phi\vec{v})\right] dV \tag{10.87}$$

10.4.4 The transport theorem and balance laws

As we will see in the following there are important physical laws that can be formulated in terms of the material derivative given by (10.85). In particular, this is the case for the mass balance, the momentum balance, the angular momentum balance, and the energy balance. In the derivation of a model, however, we will often use a control volume that is not a material volume, but rather a volume that is determined from the geometry of the problem. From (10.83) and (10.85) we have the following result

For a general control volume $V_c(t)$ where a point on the surface has velocity $\vec{v}_c$ the transport theorem gives

$$\frac{d}{dt}\iiint_{V_c} \phi(\mathbf{x},t)dV = \frac{D}{Dt}\iiint_{V_c} \phi(\mathbf{x},t)dV - \iint_{\partial V_c} \phi\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n} dA \tag{10.88}$$

If the volume V_f is fixed in spatial coordinates, we get

$$\frac{d}{dt}\iiint_{V_f}\phi(\mathbf{x},t)dV = \iiint_{V_f}\frac{\partial\phi(\mathbf{x},t)}{\partial t}dV, \quad V_f \text{ is fixed.} \tag{10.89}$$

as there is no term due to a changing boundary. Combining this with (10.85) we find:

For a fixed volume V_f the transport theorem gives

$$\iiint_{V_f}\frac{\partial\phi(\mathbf{x},t)}{\partial t}dV = \frac{D}{Dt}\iiint_{V_f}\phi(\mathbf{x},t)dV - \iint_{\partial V_f}\phi\vec{v}\cdot\vec{n}dA \tag{10.90}$$

.

Chapter 11

Mass, momentum and energy balances

11.1 The mass balance

11.1.1 Differential form

We will now derive the continuity equation using the principle of mass conservation, which states that the mass of a material volume must be a constant. The mass of a material volume $V_m(t)$ is

$$m = \iiint_{V_m(t)} \rho dV \tag{11.1}$$

where ρ is the fluid density. This means that principle of mass conservation can be expressed in the form

$$\frac{D}{Dt} \iiint_V \rho dV = 0 \tag{11.2}$$

Then Reynolds' transport theorem with $\phi = \rho$ leads to

$$\iiint_V \left[\frac{D\rho}{Dt} + \rho(\vec{\nabla} \cdot \vec{v}) \right] dV = 0 \tag{11.3}$$

in material form and

$$\iiint_V \left[\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) \right] dV = 0 \tag{11.4}$$

in divergence form.

As the volume V is arbitrary, the integrand in both integral forms (11.3) and (11.4) of the continuity equation must be identically zero, and this leads to the differential formulation of the continuity in material form

$$\underbrace{\frac{D\rho}{Dt}}_{\substack{\text{rate of change of} \\ \text{density in material} \\ \text{volume element}}} + \underbrace{\rho\left(\vec{\nabla} \cdot \vec{v}\right)}_{\substack{\text{rate of change of density} \\ \text{due to divergence of} \\ \text{material volume element}}} = 0 \tag{11.5}$$

and in divergence form,

$$\underbrace{\frac{\partial \rho}{\partial t}}_{\substack{\text{rate of change of}\\ \text{density in spatial}\\ \text{volume element}}} + \underbrace{\vec{\nabla}\cdot(\rho\vec{v})}_{\substack{\text{rate of change of density}\\ \text{due to convection out of}\\ \text{spatial volume element}}} = 0 \tag{11.6}$$

Example 157 *It is possible to derive the divergence form of the continuity equation from the material form using*

$$\vec{\nabla}\cdot(\rho\vec{v}) = \left(\vec{\nabla}\rho\right)\cdot\vec{v} + \rho(\vec{\nabla}\cdot\vec{v}) \tag{11.7}$$

and the definition of the material derivative.

11.1.2 Integral form

For a fixed volume V_f we find from (10.90) that

$$\underbrace{\frac{d}{dt}\iiint_{V_f}\rho dV}_{\substack{\text{rate of change}\\ \text{of mass in}\\ \text{fixed volume}}} = -\underbrace{\iint_{\partial V_c}\rho\vec{v}\cdot\vec{n}dA}_{\substack{\text{net increase of}\\ \text{mass by}\\ \text{convection}}}, \quad V_f \text{ is constant} \tag{11.8}$$

where $\vec{n}$ is a unit normal pointing out of the volume V_f.

From (10.88) we have the following equation for a control volume V_c where $\vec{v}_c$ is the velocity of the surface ∂V_c of the volume:

$$\frac{d}{dt}\iiint_{V_c}\rho dV = \frac{D}{Dt}\iiint_{V_c}\rho dV - \iint_{\partial V_c}\rho\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n}dA \tag{11.9}$$

Using the principle of mass conservation as expressed in (11.2) we arrive at the result

$$\frac{d}{dt}\iiint_{V_c}\rho dV = -\iint_{\partial V_c}\rho\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n}dA \tag{11.10}$$

11.1.3 Control volume with compressible fluid

Consider a control volume V_c which may be time varying, and which is filled with a compressible fluid. Moreover, assume that the density ρ is the same all over the control volume. The fluid is assumed to be compressible with bulk modulus β so that

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \tag{11.11}$$

Then from (10.88) we have the mass balance in the form

$$\underbrace{\frac{d}{dt}\iiint_{V_c}\rho dV}_{\substack{\text{rate of change}\\ \text{of mass in}\\ \text{control volume}}} = \underbrace{\frac{D}{Dt}\iiint_{V_c}\rho dV}_{\substack{\text{This term equals}\\ \text{zero in view of}\\ (11.2)}} \underbrace{-\iint_{\partial V_c}\rho\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n}dA}_{\substack{\text{net mass}\\ \text{flow into}\\ \text{control volume}}} \tag{11.12}$$

This equation states that the time derivative of the mass in V_c is equal to the net mass flow into the control volume, which makes sense. We denote the mass flow into the volume by $w_1 = \rho q_1$, and the mass flow out of the volume by $w_2 = \rho q_2$, where q_1 and q_2 are the corresponding volumetric flows. Then the mass balance can be written

$$\frac{d}{dt}(\rho V_c) = w_1 - w_2 = \rho(q_1 - q_2) \tag{11.13}$$

Moreover, assume that the density ρ is the same all over the control volume. The fluid is assumed to be compressible with bulk modulus β so that

$$\frac{d\rho}{\rho} = \frac{dp}{\beta} \Rightarrow \dot{\rho} = \frac{\rho}{\beta}\dot{p} \tag{11.14}$$

Then the mass balance of a control volume V_c with compressible fluid with bulk modulus β can be written

$$\frac{V_c}{\beta}\dot{p} + \dot{V}_c = q_1 - q_2 \tag{11.15}$$

11.1.4 Mass flow through a pipe

A fluid of constant density is flowing through a pipe of cross section A with velocity $\vec{v}$ along the direction of the pipe, which is the x direction with unit vector $\vec{i}$. It is assumed that the velocity is constant over the cross section, and given by $\vec{v} = v\vec{i}$. If the flow is into the volume, then the outwards-pointing surface normal is $\vec{n} = -\vec{i}$, and the mass flow is

$$w = -\iint_A \rho\vec{v}\cdot\vec{n}dA = -\iint_A \rho\vec{v}\cdot\vec{n}dA = \rho vA \tag{11.16}$$

If the velocity varies over the cross section of the pipe, then the mass flow is

$$w = -\iint_{\partial V_c} \rho\vec{v}\cdot\vec{n}dA = \rho\int vdA = \rho\bar{v}A \tag{11.17}$$

where $\bar{v}$ is the average velocity.

Let the control volume be $V_c = AL$, which is the fixed volume from $x_1 = 0$ to $x_2 = L$ of the pipe. Then the boundary ∂V_c of the volume V_c is the wall of the pipe plus the cross sections at x_1 and at x_2. The outwards-pointing normal vector is $\vec{n} = -\vec{i}$ at x_1 and $\vec{n} = \vec{i}$ at x_2. Then the mass balance is

$$\frac{d}{dt}m_c = \rho_1\bar{v}_1A - \rho_2\bar{v}_2A = w_1 - w_2 \tag{11.18}$$

where

$$m_c = \iiint_{V_c} \rho dV \tag{11.19}$$

is the mass inside the control volume.

Example 158 *We consider gas of density ρ in a fixed volume V shown in Figure 11.1 with inlet through a pipe of cross section A_1 and outlet through a pipe of cross section A_2. We suppose that ρ is constant over the fixed volume V, while the density is ρ_1 at the inlet. We assume that the velocity in the inlet pipe is in the x direction and of magnitude*

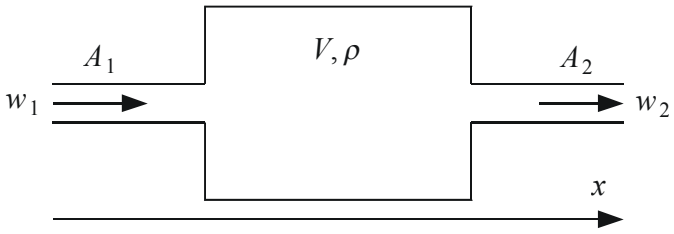

Figure 11.1: Volume V with mass flow w_1 into the wolume and w_2 out of the volume.

v_1. In the same way, the velocity in the outlet pipe is assumed to be in the x direction and of magnitude v_2. The velocity is assumed to be constant over the cross section of the pipe. Then the balance equation of mass is

$$V\frac{d\rho}{dt} = A_1\rho_1\bar{v}_1 - A_2\rho\bar{v}_2 \tag{11.20}$$

which may also be written

$$\frac{dm}{dt} = w_1 - w_2. \tag{11.21}$$

Here $m = \rho V$ is the mass contained in the volume, $w_1 = A_1\rho_1 v_1$ is the mass flow into the volume, and $w_2 = A_2\rho_2 v_2$ is the mass flow out of the volume.

Example 159 *Water of constant density ρ is flowing into a tank of cross section A with mass flow w_1 and flows out with mass flow w_2. The water level is h. The mass balance is*

$$\frac{d}{dt}(\rho Ah) = w_1 - w_2 \tag{11.22}$$

which can be written

$$\dot{h} = \frac{1}{\rho A}(w_1 - w_2) \tag{11.23}$$

11.1.5 Continuity equation and Reynolds' transport theorem

It turns out that by combining the continuity equation with Reynold's transport theorem we can derive alternative forms of Reynold's transport theorem. This is useful in the development of the momentum balance in Section 11.2.1.

First it is noted that the divergence form (10.87) of Reynolds' transport theorem for the function $\rho\phi$ gives

$$\frac{D}{Dt}\iiint_V \rho\phi dV = \iiint_V \left[\frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla}\cdot(\rho\phi\vec{v})\right] dV \tag{11.24}$$

Then it is used that the material form (10.86) of Reynolds' transport theorem gives

$$\begin{aligned}\frac{D}{Dt}\iiint_V \rho\phi dV &= \iiint_V \left[\frac{D(\rho\phi)}{Dt} + \rho\phi(\vec{\nabla}\cdot\vec{v})\right] dV \\ &= \iiint_V \left\{\rho\frac{D\phi}{Dt} + \phi\left[\frac{D\rho}{Dt} + \rho(\vec{\nabla}\cdot\vec{v})\right]\right\} dV\end{aligned} \tag{11.25}$$

The last two terms of the integrand cancel, which can be seen from the continuity equation (11.5). This gives the following important result for a volume V.

The continuity equation in combination with the transport theorem gives the result

$$\frac{D}{Dt}\iiint_V \rho\phi dV = \iiint_V \rho\frac{D\phi}{Dt}dV \tag{11.26}$$

By comparing this with (11.24) and accounting for the fact that the volume V is arbitrary, it is found that

$$\rho\frac{D\phi}{Dt} = \frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla}\cdot(\rho\phi\vec{v}) \tag{11.27}$$

Note that the last term on the right hand side of (11.27) is a divergence term. The importance of this is made clear by integrating the equation over a volume V and using the divergence theorem. This gives

$$\iiint_V \rho\frac{D\phi}{Dt}dV = \iiint_V \frac{\partial(\rho\phi)}{\partial t}dV + \iint_{\partial V}\rho\phi(\vec{v}\cdot\vec{n})\,dA \tag{11.28}$$

where V can be any volume, and $\vec{n}$ is the outwards pointing surface normal. We see that the first term on the right side is the rate of change of the quantity of $\rho\phi$ in the volume, while the second term on the right side is the flow of $\rho\phi$ into the volume over the volume boundary.

We note that ϕ may be the component of a vector $\mathbf{u}$, that is, $\phi = u_i$ which leads to the following vector equations

The continuity equation and the transport theorem for a vector $\mathbf{u}$ gives the results

$$\frac{D}{Dt}\iiint_V \rho\vec{u}dV = \iiint_V \rho\frac{D\vec{u}}{Dt}dV \tag{11.29}$$

and

$$\rho\frac{D\vec{u}}{Dt} = \frac{\partial(\rho\vec{u})}{\partial t} + \vec{\nabla}\cdot(\rho\vec{v}\vec{u}) \tag{11.30}$$

The last term in (11.30) is verified in a Cartesian coordinate system with orthogonal unit vectors $\vec{a}_i$ with the following computation:

$$\vec{\nabla}\cdot(\rho\vec{v}\vec{u}) = \sum_k \frac{\partial}{\partial x_k}\vec{a}_k\cdot\left(\rho\sum_j v_j\vec{a}_j\sum_i u_i\vec{a}_i\right) = \sum_i \frac{\partial}{\partial x_j}(\rho v_j u_i)\,\vec{a}_i \tag{11.31}$$

The integral form of (11.30) is found to be

$$\iiint_V \rho\frac{D\vec{u}}{Dt}dV = \iiint_V \frac{\partial(\rho\vec{u})}{\partial t}dV + \iint_{\partial V}\rho\vec{u}(\vec{v}\cdot\vec{n})\,dA \tag{11.32}$$

This result has a nice structure, and the terms on the right side has the same physical interpretation as in the scalar case.

Finally we note that from the expressions in (10.10) of the material derivative, the following alternative expressions are obtained

$$\frac{\partial(\rho\phi)}{\partial t} + \vec{\nabla}\cdot(\rho\phi\vec{v}) = \rho\frac{D\phi}{Dt} = \rho\left(\frac{\partial\phi}{\partial t} + \vec{v}\cdot\vec{\nabla}\phi\right) \tag{11.33}$$

$$\frac{\partial(\rho\vec{u})}{\partial t} + \vec{\nabla}\cdot(\rho\vec{v}\vec{u}) = \rho\frac{D\vec{u}}{Dt} = \rho\left(\frac{\partial\vec{u}}{\partial t} + \vec{v}\cdot\vec{\nabla}\vec{u}\right) \tag{11.34}$$

11.1.6 Multi-component systems

To describe systems with chemical reactions we may need the continuity equation for a volume with several components, and where different mass components are generated or used in the chemical reactions. The presentation is adopted from the introductory part of (de Groot and Mazur 1984). We consider a fluid with n components where there may be r chemical reactions between the components. The mass m_k of component k in a material volume V satisfies

$$\frac{d}{dt}m_k = \sum_{j=1}^{r}\iiint_V \nu_{kj}J_j dV \tag{11.35}$$

where $\nu_{kj}J_j$ is the *rate of production* of component k per unit volume in reaction j. Using the transport theorem, this gives

$$\iiint_V \left[\frac{\partial\rho_k}{\partial t} + \vec{\nabla}\cdot(\rho_k\vec{v}_k)\right]dV = \sum_{j=i}^{r}\iiint_V \nu_{kj}J_j dV \tag{11.36}$$

where ρ_k is the density of component k, and v_k is the velocity of component k. As the volume V is arbitrary, it follows that the *continuity equation of component* k is

$$\frac{\partial\rho_k}{\partial t} + \vec{\nabla}\cdot(\rho_k\vec{v}_k) = \sum_{j=i}^{r}\nu_{kj}J_j \tag{11.37}$$

Since mass is conserved in each of the chemical reactions it follows that

$$\sum_{k=i}^{n}\nu_{kj}J_j = 0 \tag{11.38}$$

Then, by adding the continuity equations of all components the continuity equation

$$\frac{\partial\rho}{\partial t} + \vec{\nabla}\cdot(\rho\vec{v}) = 0 \tag{11.39}$$

where ρ is the *total density*

$$\rho = \sum_{k=1}^{n}\rho_k \tag{11.40}$$

and $\mathbf{v}$ is the *barycentric velocity*, which is defined as the velocity of the center of mass

$$\vec{v} := \sum_{k=1}^{n}\frac{\rho_k\vec{v}_k}{\rho} \tag{11.41}$$

Define the *barycentric material derivative* by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \vec{v} \cdot \vec{\nabla} \tag{11.42}$$

where **v** is the barycentric velocity. Insertion into the continuity equation (11.37) for component k gives

$$\frac{D\rho_k}{Dt} - \vec{v} \cdot \vec{\nabla}\rho_k + \vec{\nabla} \cdot (\rho_k \vec{v}_k) = \sum_{j=i}^{r} \nu_{kj} J_j \tag{11.43}$$

The last term on the left side is expanded to give

$$\frac{D\rho_k}{Dt} - \vec{v} \cdot \vec{\nabla}\rho_k + \vec{\nabla} \cdot (\rho_k \vec{v}) + \vec{\nabla} \cdot [\rho_k (\vec{v}_k - \vec{v})] = \sum_{j=i}^{r} \nu_{kj} J_j \tag{11.44}$$

By defining the *diffusion flow* of component k as

$$\vec{j}_k = \rho_k (\vec{v}_k - \vec{v}) \tag{11.45}$$

and, accounting for (10.15), we find the following result:

The continuity equation for component k is found to be

$$\frac{D\rho_k}{Dt} = -\rho_k \left(\vec{\nabla} \cdot \vec{v}\right) - \vec{\nabla} \cdot \vec{j}_k + \sum_{j=i}^{r} \nu_{kj} J_j \tag{11.46}$$

while from (11.39) and (11.42) the continuity equation for the total density is

$$\frac{D\rho}{Dt} + \rho \left(\vec{\nabla} \cdot \vec{v}\right) = 0 \tag{11.47}$$

Example 160 *In terms of* mass fractions

$$c_k = \frac{\rho_k}{\rho} \tag{11.48}$$

the continuity equation for component k becomes

$$\rho \frac{Dc_k}{Dt} = -\vec{\nabla} \cdot \vec{j}_k + \sum_{j=i}^{r} \nu_{kj} J_j \tag{11.49}$$

which is found by inserting $\rho_k = \rho c_k$ in the continuity equation (11.46) for component k and then use (11.47).

11.2 The momentum balance

11.2.1 Euler's equation of motion

We consider a material volume element dV of a fluid with density ρ and velocity $\vec{v}$. The momentum of the volume element is $\rho\vec{v}dV$. Newton's law for this set of particles is

$$\frac{D}{Dt}(\rho\vec{v}dV) = d\vec{F}. \tag{11.50}$$

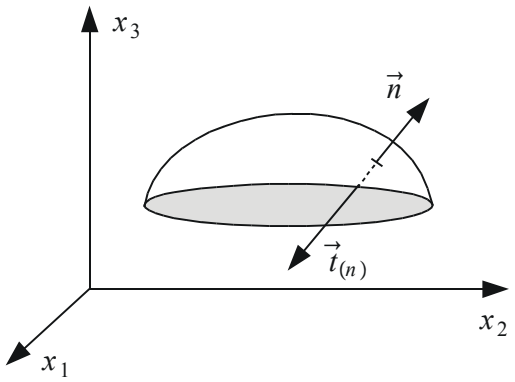

Figure 11.2: The stress vector in an inviscid fluid is parallel to the surface normal.

where $d\vec{F}$ is the force acting on the differential volume dV. Note that the material derivative is used, as Newton's law applies to a material volume element. The mass ρdV of the particles in the material volume element is constant, and it follows that

$$\frac{D}{Dt}\left(\rho\vec{v}dV\right) = \frac{D}{Dt}\left(\rho dV\right)\vec{v} + \rho\frac{D\vec{v}}{Dt}dV = \rho\frac{D\vec{v}}{Dt}dV \tag{11.51}$$

We may therefore write Newton's law in the form

$$\rho\frac{D\vec{v}}{Dt}dV = d\vec{F} \tag{11.52}$$

The force $d\vec{F}$ denotes the total force on the volume element, which is the mass force plus the surface force. When this is integrated over the material volume V we get

$$\iiint_V \rho\frac{D\vec{v}}{Dt}dV = \iiint_V d\vec{F} = \vec{F}^{(r)} \tag{11.53}$$

where $\vec{F}^{(r)}$ is the resultant force acting on the volume V. The surface forces cancel out inside the volume due to Newton's third law of action and reaction. This is referred to as *the principle of local equilibrium of the stresses*. Because of this the total force $\vec{F}^{(r)}$ is given by the sum of surface forces acting on ∂V plus the mass force on the volume. Assume that the fluid is *inviscid* in which case the only surface forces are the pressure forces. This gives

$$\vec{F}^{(r)} = -\iint_{\partial V} p\vec{n}dA + \iiint_V \rho\vec{f}dV \tag{11.54}$$

where $\rho\vec{f}$ is the mass force, and $-p\vec{n}dA$ is the surface force in the form of pressure forces. The divergence theorem and (10.13) then gives

$$\iiint_V \rho\frac{D\vec{v}}{Dt}dV = \iiint_V \left(-\vec{\nabla}p + \rho\vec{f}\right)dV \tag{11.55}$$

The volume V is arbitrary, and this leads to *Euler's equation of motion*

Euler's equation of motion for an inviscid fluid is given by

$$\rho\frac{D\vec{v}}{Dt} = -\vec{\nabla}p + \rho\vec{f} \tag{11.56}$$

Alternative formulations of Euler's equation found from (11.34) are the divergence form

$$\frac{\partial\left(\rho\vec{v}\right)}{\partial t} + \vec{\nabla}\cdot\left(\rho\vec{v}\vec{v}\right) = -\vec{\nabla}p + \rho\vec{f} \tag{11.57}$$

and the formulation

$$\rho\frac{\partial\vec{v}}{\partial t} + \rho\left(\vec{v}\cdot\vec{\nabla}\right)\vec{v} = -\vec{\nabla}p + \rho\vec{f} \tag{11.58}$$

Example 161 *We consider the one-dimensional case where the velocity is v in the x direction. Then, if the pressure gradient is zero, the mass forces are zero, and ρ is a constant, Euler's equation as given by (11.58) gives*

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = 0 \tag{11.59}$$

which is known as Burger's equation (Evans 1998). This simple equation is interesting as it may have analytical solutions that can be used to check the accuracy of numerical solution techniques, and it may exhibit shocks where the velocity gradient approaches infinity.

11.2.2 The momentum equation for a control volume

From (11.29) we have the following expression

$$\frac{D}{Dt}\iiint_V \rho\vec{v}dV = \iiint_V \rho\frac{D\vec{v}}{Dt}dV \tag{11.60}$$

From (10.88) we have the following equation for a control volume V_c

$$\frac{d}{dt}\iiint_{V_c} \rho\vec{v}dV = \frac{D}{Dt}\iiint_{V_c} \rho\vec{v}dV - \iint_{\partial V_c} \rho\vec{v}\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n}dA \tag{11.61}$$

where $\vec{v}$ is the velocity of the fluid and $\vec{v}_c$ is the velocity of the surface ∂V_c of the control volume. Combining these two equations with (11.53) we get

$$\underbrace{\frac{d}{dt}\iiint_{V_c} \rho\vec{v}dV}_{\substack{\text{rate of change}\\ \text{of momentum}\\ \text{in control}\\ \text{volume}}} = \underbrace{\vec{F}^{(r)}}_{\substack{\text{resultant force}\\ \text{on fluid in}\\ \text{control}\\ \text{volume}}} \underbrace{- \iint_{\partial V_c} \rho\vec{v}\left(\vec{v}-\vec{v}_c\right)\cdot\vec{n}dA}_{\substack{\text{net increase of}\\ \text{momentum}\\ \text{by convection}}} \tag{11.62}$$

Example 162 *For the system in Example 158 the momentum conservation in the x direction gives*

$$\frac{d}{dt}\iiint_V v\rho dV = F + p_1A_1 - p_2A_2 + v_1w_1 - v_2w_2 \tag{11.63}$$

where F is the force in the x direction acting on the gas from the tank, p_1A_1 is the force due to pressure on the inlet, and p_2A_2 is the force due to pressure at the outlet. It is assumed that the velocity is constant over the cross section.

11.2.3 Example: Waterjet

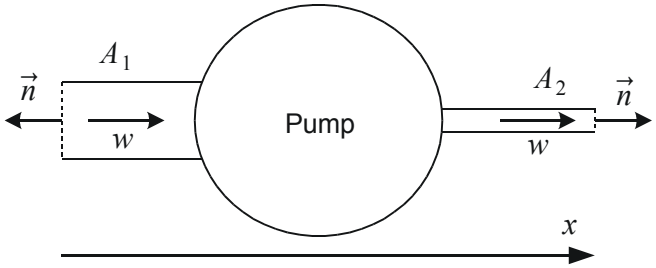

Figure 11.3: Schematic diagram of a waterjet.

We consider a waterjet (Figure 11.3) where water enter through the intake which is a pipe with cross section A_1, and flows out through an outlet pipe of cross section A_2. A pump is used to force the water through the waterjet. The water flows axially in the pipes with velocity $\mathbf{v} = v_1\mathbf{i} = -v_1\mathbf{n}$ at the inlet and $\mathbf{v} = v_2\mathbf{i} = v_2\mathbf{n}$ at the outlet where $\mathbf{i}$ is the unit vector in the x direction. Stationary conditions are assumed. Moreover, the water is assumed to be incompressible, so that the mass flow in is equal to the mass flow out. Then the continuity equation gives

$$A_1\rho v_1 = A_2\rho v_2 = w \tag{11.64}$$

where w is the mass flow. We assume that the pressure forces over the cross sections A_1 and A_2 of the pipes can be left out. Then the momentum equation in the x direction gives

$$F + v_1A_1\rho v_1 - v_2\rho A_2 v_2 = 0. \tag{11.65}$$

We define the thrust T of the waterjet as the force from the fluid on the casing. The thrust is given by $T = -F$, and we get the result

$$T = -\left(1 - \frac{A_2}{A_1}\right) wv_2 \approx -wv_2 \tag{11.66}$$

where it is assumed that $A_2 \ll A_1$. We see that if the outlet area is much smaller than the inlet area, then the thrust is equal to mass flow times outlet velocity, and that the thrust is directed in the opposite direction of the flow through the waterjet. Suppose that the outlet cross section is reduced. Then if the pump is sufficiently powerful so that the mass flow w is unchanged, then $v_2 = w/(A_2\rho)$ will increase, and the thrust $T \approx -wv_2$ will increase in magnitude.

11.2.4 Example: Sand dispenser and conveyor

Sand is dispensed from a container with mass flow w down on a conveyor belt as shown in Figure 11.4. The conveyor belt is driven by a motor torque T acting on a shaft of radius r with angular velocity ω_m. The velocity of the conveyor belt is therefore $v = \omega_m r$.

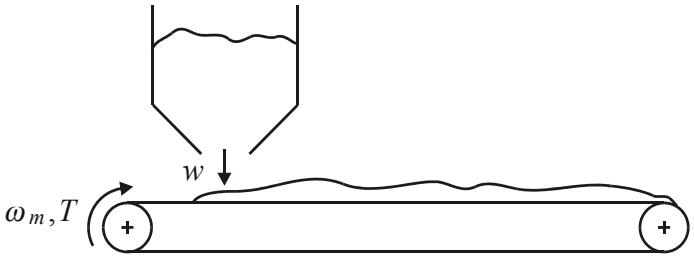

Figure 11.4: Sand of mass flow w falling down on a conveyor belt.

Here the mass m and the momentum $p = mv$ of the sand are conserved quantities. The balance equation for the mass is

$$\frac{d}{dt}m = w - w_e, \tag{11.67}$$

while the balance equation for the momentum is

$$\frac{d}{dt}(mv) = -vw_e + F \tag{11.68}$$

Here F is the force from the conveyor belt on the sand. The equation of motion for the belt is

$$J\dot{\omega}_m = T - Fr \tag{11.69}$$

where J is the inertia experienced by the motor. The equation of motion can be expressed in terms of the velocity to give

$$\frac{J}{r^2}\dot{v} = \frac{1}{r}T - F. \tag{11.70}$$

The momentum equation gives

$$\dot{m}v + m\dot{v} = -vw_e + F \tag{11.71}$$

and insertion of the mass balance and the equation of motion gives

$$\left(m + \frac{J}{r^2}\right)\dot{v} = \frac{1}{r}T - vw \tag{11.72}$$

The results seem reasonable as the belt is slowed down when sand with zero horizontal velocity falls down on the belt.

11.2.5 Irrotational Bernoulli equation

The convective term $\rho(\vec{v}\cdot\vec{\nabla})\vec{v}$ in (11.58) can be written

$$\left(\vec{v}\cdot\vec{\nabla}\right)\vec{v} = \vec{\nabla}\left(\frac{1}{2}\vec{v}^2\right) - \vec{v}\times\left(\vec{\nabla}\times\vec{v}\right) \tag{11.73}$$

which can be verified by evaluation the components on both sides. It follows that for irrotational flow, which occurs for $\vec{\nabla}\times\vec{v} = \vec{0}$, the Euler equation can be written

$$\frac{\partial\vec{v}}{\partial t} + \vec{\nabla}\left(\frac{1}{2}\vec{v}^2\right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p = \vec{0} \tag{11.74}$$

Suppose that the fluid is incompressible so that ρ is a constant. Moreover, assume that the mass force is the gradient $\vec{f} = -\vec{\nabla}(gz)$ of the gravitational potential gz, where z is the coordinate in the vertical upwards direction. As $\vec{\nabla} \times \vec{v} = \vec{0}$ there will be a velocity potential ϕ so that $\vec{v} = \vec{\nabla}\phi$. Then Euler's equation can be written as the gradient equation

$$\vec{\nabla}\left[\frac{\partial \phi}{\partial t} + \frac{1}{2}\vec{v}^2 + gz + \frac{p}{\rho}\right] = 0 \tag{11.75}$$

where it is used that ρ is a constant for incompressible fluids. This implies that

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\vec{v}^2 + \frac{p}{\rho} + gz = \text{constant} \tag{11.76}$$

which is the *irrotational Bernoulli equation*. In the stationary case we then have

$$\frac{1}{2}\left(\vec{v}_2^2 - \vec{v}_1^2\right) + \frac{(p_2 - p_1)}{\rho} + (z_2 - z_1)\, g = 0 \tag{11.77}$$

for irrotational flow of an inviscid and incompressible fluid.

Example 163 *The velocity term can be expressed using the gradient of the velocity potential, which gives*

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\left(\vec{\nabla}\phi\right) \cdot \vec{\nabla}\phi + \frac{p}{\rho} + gz = \text{constant} \tag{11.78}$$

11.2.6 Bernoulli's equation along a streamline

It is seen from (11.58) and (11.73) that the Euler equation can be written

$$\frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left(\frac{1}{2}\vec{v}^2\right) - \vec{v} \times \left(\vec{\nabla} \times \vec{v}\right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p = \vec{0} \tag{11.79}$$

To proceed we need to eliminate the term $\vec{v} \times (\vec{\nabla} \times \vec{v})$. There are two ways to do this that give interesting results (White 1999). The first approach, which was discussed in the previous section, is to require that $\vec{\nabla} \times \vec{v}{=}\vec{0}$, which is the case for irrotational flow. The second approach, which will be investigated here, is to integrate the expression on the left hand side of (11.79) along a streamline.

Consider the following integral form of the Euler equation (11.79):

$$\int \left[\frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left(\frac{1}{2}\vec{v}^2\right) - \vec{v} \times \left(\vec{\nabla} \times \vec{v}\right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p\right] \cdot d\vec{x}{=}0 \tag{11.80}$$

where the differential $d\vec{x}$ is parallel to the velocity and satisfies $d\vec{x}/dt = \vec{v}$. Then

$$\vec{v} \times \left(\vec{\nabla} \times \vec{v}\right) \cdot d\vec{x}{=}\, \mathbf{0} \tag{11.81}$$

and the integral expression becomes

$$\int \left[\frac{\partial \vec{v}}{\partial t} + \vec{\nabla} \cdot \left(\frac{1}{2}\vec{v}^2\right) - \vec{f} + \frac{1}{\rho}\vec{\nabla}p\right] \cdot d\vec{x}{=}0 \tag{11.82}$$

We assume that $\vec{f} = -g\vec{a}_3$, and denote the vertical coordinate $z = x_3$, and write $|d\vec{x}| = ds$. This gives

$$\int_1^2 \frac{\partial\, |\vec{v}|}{\partial t} ds + \int_1^2 d\left(\frac{1}{2}\vec{v}^2\right) + \int_1^2 g dz + \int_1^2 \frac{dp}{\rho}{=}0 \tag{11.83}$$

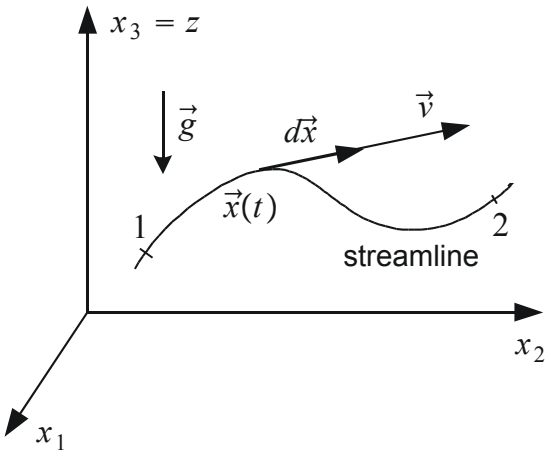

Figure 11.5: Streamline $\mathbf{x}(t)$ with two points 1 and 2 on the streamline.

where 1 and 2 denotes two points on the same streamline. Two of the integrals are exact, and we find that

$$\int_1^2 \frac{\partial |\vec{v}|}{\partial t} ds + \frac{1}{2}\left(\vec{v}_2^2 - \vec{v}_1^2\right) + \int_1^2 \frac{dp}{\rho} + g\left(z_2 - z_1\right) = 0 \tag{11.84}$$

which is *Bernoulli's equation for frictionless flow along a streamline*. Under stationary conditions $\partial |\vec{v}| / \partial t = 0$, and

$$\frac{1}{2}\left(\vec{v}_2^2 - \vec{v}_1^2\right) + \int_1^2 \frac{dp}{\rho} + \left(z_2 - z_1\right) g = 0 \tag{11.85}$$

For incompressible flow ρ is a constant and

$$\frac{1}{2}\left(\vec{v}_2^2 - \vec{v}_1^2\right) + \frac{\left(p_2 - p_1\right)}{\rho} + \left(z_2 - z_1\right) g = 0, \quad \text{1 and 2 on a streamline} \tag{11.86}$$

which is the Bernoulli equation for stationary frictionless flow along a streamline for an incompressible fluid. We see that if $z_1 = z_2$, then the pressure along a streamline will decrease when the velocity increases.

The additional assumption that was made for the irrotational Bernoulli's equation was that the flow is irrotational. The equation (11.77) is valid for arbitrary points 1 and 2 in the fluid, whereas Bernoulli's equation (11.86) along a streamline is only valid if the points 1 and 2 are on a streamline.

11.2.7 Example: Transmission line

A hydraulic transmission line is a pipe of cross section A and length L with a compressible fluid. The dynamic model for a hydraulic transmission line is developed from the mass balance and momentum balance of a differential control volume Adx where A is the cross sectional area of the pipe and x is the length coordinate along the pipe. It is assumed that the density of the fluid is not varying over the cross section, so that $\rho = \rho(x,t)$. The mass flow is

$$w(x,t) = \int_A \rho v dA = \rho \bar{v} A \tag{11.87}$$

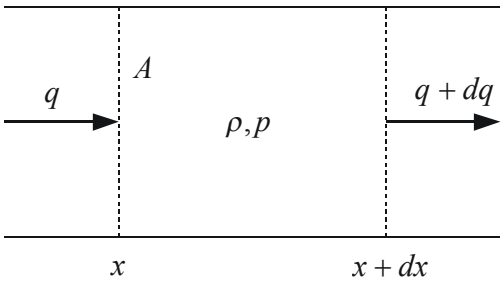

Figure 11.6: Volume element for hydraulic transmission line.

where $\bar{v}$ is the average velocity. The mass balance is taken for the fixed differential control volume Adx from x to $x + dx$. The mass flow into the volume is w at x, while the mass flow out of the volume is $w + dw$ at $x + dx$. The mass balance is then found from (11.18) to be

$$Adx\frac{\partial \rho}{\partial t} = w - (w + dw) = -dw$$

Dividing by Adx we get

$$\frac{\partial \rho}{\partial t} = -\frac{1}{A}\frac{\partial w}{\partial x} \tag{11.88}$$

A change of variables from density ρ to pressure p is achieved in the mass balance using the constitutive equation $dp = (\beta/\rho)d\rho$ where β is the bulk modulus of the fluid. This gives

$$\frac{\partial p}{\partial t} = -\frac{\beta}{\rho A}\frac{\partial w}{\partial x}$$

The momentum equation is found from (11.63) to be

$$\frac{\partial}{\partial t}\left(\rho\bar{v}\right)Adx = Ap - A(p + dp) + \int_A \rho v^2 dA - \int_A \left[\rho v^2 + d\left(\rho v^2\right)\right] dA - Fdx \tag{11.89}$$

where Fdx is the friction force. This gives

$$\frac{\partial w}{\partial t} = -A\frac{\partial p}{\partial x} - A\frac{\partial}{\partial x}\int_A \rho v^2 dA - F \tag{11.90}$$

We will assume that the average velocity $\bar{v}$ is close to zero, so that the second term on the right side can be set to zero. The model becomes

$$\frac{\partial p}{\partial t} = -\frac{\beta}{\rho A}\frac{\partial w}{\partial x} \tag{11.91}$$

$$\frac{\partial w}{\partial t} = -A\frac{\partial p}{\partial x} - F \tag{11.92}$$

These equations are usually formulated in terms of the pressure p and the volumetric flow q by treating the density as a constant ρ_0 so that $w = \rho_0 q$.

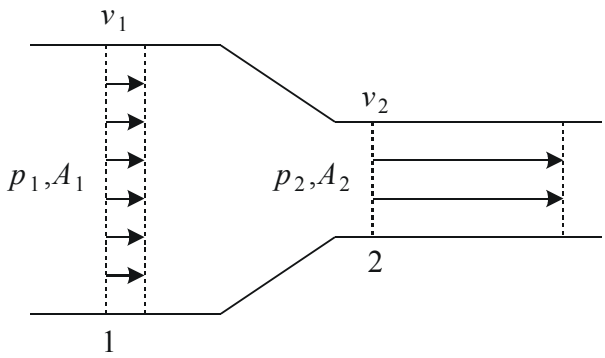

Figure 11.7: Incompressible fluid flowing through a pipe of cross section A_1 with a restriction with cross section A_2.

The transmission line model linearized around $q = 0$ and $\rho = \rho_0$ is given by

$$\frac{\partial p}{\partial t} = -\frac{\beta}{A}\frac{\partial q}{\partial x} \tag{11.93}$$

$$\frac{\partial q}{\partial t} = -\frac{A}{\rho_0}\frac{\partial p}{\partial x} - \frac{F}{\rho_0} \tag{11.94}$$

11.2.8 Liquid mass flow through a restriction

We consider a liquid, that is an incompressible fluid, which flows through a pipe with cross sectional area A_1 with a restriction with cross sectional area A as shown in Figure 11.7. The continuity equation implies that the mass flow $w_1 = \rho q_2$ at the inlet is the same as the mass flow $w_2 = \rho q_2$ at the outlet. As the fluid is incompressible, this implies that also the volumetric flow is the same at the inlet and the outlet, so that the volumetric flow q is given by

$$q = v_1 A_1 = v_2 A_2 \tag{11.95}$$

Bernoulli's equation (11.86) gives

$$\frac{1}{2}v_1^2 + \frac{p_1}{\rho} = \frac{1}{2}v_2^2 + \frac{p_2}{\rho} \tag{11.96}$$

which gives

$$\begin{aligned} p_1 - p_2 &= \frac{\rho}{2}\left(v_2^2 - v_1^2\right) = \frac{\rho}{2}\left[1 - \left(\frac{A_2}{A_1}\right)^2\right] v_2^2 \\ &= \left[1 - \left(\frac{A_2}{A_1}\right)^2\right] \frac{\rho q_2^2}{2A_2} \end{aligned} \tag{11.97}$$

This gives the following result:

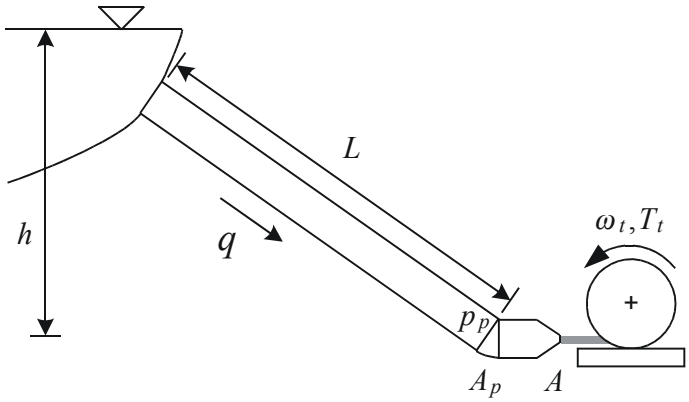

Figure 11.8:

Frictionless and incompressible flow through a restriction A_2 in a pipe with cross section A_1 is given by

$$q = A_2 \sqrt{\frac{2}{\rho} \frac{(p_1 - p_2)}{1 - \left(\frac{A_2}{A_1}\right)^2}} \tag{11.98}$$

If the flow is from a volume, then $A_1 \to \infty$, and the expression becomes

$$q = A_2 \sqrt{\frac{2}{\rho} (p_1 - p_2)} \tag{11.99}$$

This expression (11.98) is adjusted with the *discharge coefficient* C_d to account for the effect that the cross section of the flow will be somewhat smaller than the cross section A_2 of the restriction. This gives

$$q = C_d A_2 \sqrt{\frac{2}{\rho} \frac{(p_1 - p_2)}{1 - \left(\frac{A_2}{A_1}\right)^2}} \tag{11.100}$$

At very low flow rates the friction will be the dominating physical phenomenon. Then Bernoulli's equation is no longer valid, and the flow becomes linear in the pressure difference. This is discussed in Section 4.2.2.

11.2.9 Example: Water turbine

Model

In this section we will study the dynamics of a hydroelectric power system consisting of a pipe that transports water from a reservoir with water level h to an impulse turbine with a Pelton wheel (White 1999) at water level 0. Between the outlet of the pipe and the turbine there is a control device that sets the cross section A of the water flowing into the turbine. The cross section A is the input control variable of the system, while the turbine torque T_t is the output. The turbine torque is of interest as the equation of

motion for the turbine shaft is

$$J_t\dot{\omega}_t = T_t - T_L \tag{11.101}$$

where J_t is the moment of inertia of the turbine shaft, ω_t is the turbine shaft speed and T_L is the load torque which will typically be the driving torque for an electrical generator. The model will be developed by deriving the model for the pipe, the model for the control device, and the model for the turbine. Then the complete model is obtained by connecting the three component models. This approach makes it easy at a later stage to change the pipe model from an incompressible flow model to a compressible flow model. Also this approach will hopefully give some structure to the presentation so that the reader will not get lost in the many equations.

The pipe is of length L, and has inlet at the elevation h where the inlet pressure is zero. The outlet of the pipe has pressure p_p and volumetric flow q. We treat the pipe as a two-port with pressure and volumetric flow as port variables. The pressure at the line ends are the inputs to the model of the pipe. The inlet pressure is supposed to be the constant ambient pressure $p_a = 0$. Therefore, the flow q at the outlet of the pipe will depend on the outlet pressure p_p. To get a result that will be valid for different pipe models, we will at this stage assume that the linearized dynamics of the pipe are given by the transfer function

$$H_{pq}(s) := \frac{-\Delta p_p}{\Delta q}(s) \tag{11.102}$$

where $\Delta q = q - q_0$ and $\Delta p_p = p_p - p_{p0}$ are deviations from a constant solution (q_0, p_{p0}). Note that the negative pressure change $-\Delta p_p$ is used in the definition of the transfer function to ensure that the $H_{pq}(s)$ has positive gain.

The inlet of the control device has a constant cross section A_p, and the inlet pressure is p_p. At the outlet of the control device the cross section is controlled to A, the pressure is p, and water velocity is $v = q/A$. It is assumed that the outlet pressure p is small and constant so that $p = 0$ can be used. It is assumed that the mass of the water in the control device is small so that Bernoulli's equation applies to describe the relation between the inlet pressure and velocity and the outlet pressure and velocity. According to (11.96) this gives

$$p_p = \frac{\rho}{2}\left(\frac{q^2}{A^2} - \frac{q^2}{A_p^2}\right) \tag{11.103}$$

Linearization of the control device equation (11.103) around the nominal area A_0 and a corresponding nominal flow q_0 gives

$$\Delta p_p = \frac{\rho\alpha q_0}{A_0^2}\Delta q - \frac{\rho q_0^2}{A_0^3}\Delta A$$

where $\alpha = 1 - A_0^2/A_p^2$. Dividing by Δq and rearranging we find that

$$\frac{\rho q_0^2}{A_0^3}\frac{\Delta A}{\Delta q}(s) = \frac{\rho\alpha q_0}{A_0^2} + \frac{-\Delta p_p}{\Delta q}(s) \tag{11.104}$$

The transfer function from the control input A to the flow q is then found to be given by

$$H_{qA}(s) := \frac{\Delta q}{\Delta A}(s) = \frac{q_0}{\alpha A_0}\frac{1}{1 + \frac{A_0^2}{\alpha\rho q_0}H_{pq}(s)} \tag{11.105}$$

where we have used (11.102). We note that the control device can be connected to a particular pipe by inserting the transfer function $H_{pq}(s)$ of the pipe.

The shaft torque T_t for an impulse turbine with a Pelton wheel is given by (White 1999)

$$T_t = 2r_t\rho q(v - r_t\omega_t) = 2r_t\rho\left(\frac{q^2}{A} - qr_t\omega_t\right) \tag{11.106}$$

where r_t is the radius of the wheel. We will treat the shaft speed ω_t as a constant in the linearization of the shaft torque. Linearization of the turbine torque equation (11.106) will then give

$$\Delta T = 2r_t\rho\frac{q_0^2}{A_0^2}\left(\beta\frac{A_0}{q_0}\Delta q - \Delta A\right) \tag{11.107}$$

where

$$\beta = 2 - \frac{r_t\omega_{t0}A_0}{q_0}$$

is a constant of linearization. The transfer function from the control input A to the turbine torque is then found to be

$$\frac{\Delta T_t}{\Delta A}(s) = 2r_t\rho\frac{q_0^2}{A_0^2}\left(\beta\frac{A_0}{q_0}H_{qA}(s) - 1\right) \tag{11.108}$$

Insertion of $H_{qA}(s)$ from (11.105) gives

$$\frac{\Delta T_t}{\Delta A}(s) = 2r_t\rho\frac{q_0^2}{A_0^2}\left(\frac{\beta}{\alpha}\frac{1}{1 + \frac{A_0^2}{\alpha\rho q_0}H_{pq}(s)} - 1\right)$$

and some algebra leads to the transfer function in the form (Hutarew 1969), (Ervik 1971)

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2r_t\rho}{\gamma}\frac{q_0^2}{A_0^2}\frac{1 - \gamma\frac{A_0^2}{\alpha\rho q_0}H_{pq}}{1 + \frac{A_0^2}{\alpha\rho q_0}H_{pq}} \tag{11.109}$$

where the constant γ is given by

$$\gamma = \frac{1}{\frac{\beta}{\alpha} - 1} \approx \frac{1}{1 - \frac{r_{pt}\omega_{t0}A_0}{q_0}}$$

The power on the turbine shaft is $P = T\omega_t$, and if we assume $\alpha = 1$, then it is a straightforward exercise to show that $\beta = 1.5$ and $\gamma = \alpha/(1.5 - \alpha) \approx 2$ at full load where the power is maximized.

Water turbine with incompressible water supply

The pipe is of length L and cross section A_p, and the reservoir has water level h. The water is assumed to be incompressible with density ρ. The volumetric flow is q, and the velocity of the water is $v_p = q/A_p$. The equation of motion for water in the pipe is

$$L\rho\dot{q} = mgh + A_p(p_0 - p_p) \tag{11.110}$$

where mgh is the constant gravity force in the flow direction that acts on the water in the pipe, p_0 is the constant ambient pressure, and p_p is the pressure at the end of the pipe. Laplace transformation leads to the pipe transfer function

$$H_{pq}(s) := \frac{-\Delta p_p}{\Delta q}(s) = \frac{\rho Ls}{A_p} \tag{11.111}$$

Note that the negative pressure is used in the definition of the transfer function $H_{pq}(s)$ to achieve a transfer function with a positive gain. The transfer function $H_{qA}(s)$ can then be found from (11.105) to be

$$H_{qA}(s) = \frac{q_0}{\alpha A_0} \frac{1}{1 + \mu \frac{T_r}{2} s}$$

where we have defined the time constant T_r and the nondimensional flow constant μ by

$$T_r = 2\frac{LA_0^2 q_{\max}}{\alpha q_0^2 A_p}, \quad \mu = \frac{q_0}{q_{\max}} \tag{11.112}$$

The transfer functions for the complete system is found from (11.109) to be

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2}{\gamma} \frac{r_t \rho q_0^2}{A_0^2} \frac{\left(1 - \gamma\mu \frac{T_r}{2} s\right)}{\left(1 + \mu \frac{T_r}{2} s\right)} \tag{11.113}$$

At full load with $\mu = 1$ and $\gamma = 2$, the transfer function is

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{r_t \rho q_0^2}{A_0^2} \frac{(1 - T_r s)}{\left(1 + \frac{T_r}{2} s\right)} \tag{11.114}$$

Example 164 *Francis or Kaplan type turbines are reaction turbines that are driven by power transfer from the water flow. The shaft torque is*

$$T_{ft} = \frac{P}{\omega_t} \tag{11.115}$$

where

$$P = q\rho \left(\frac{1}{2} v^2\right) = \frac{\rho}{2} \frac{q^3}{A^2} \tag{11.116}$$

is the power supplied to the turbine. Linearization of the power expression gives

$$\Delta P = \frac{\rho}{2} \frac{q_0^2}{A_0^2} \left(3\Delta q - 2\frac{q_0}{A_0} \Delta A\right) \tag{11.117}$$

Then the transfer function from A to P can be found from

$$\frac{\Delta P}{\Delta A}(s) = \frac{\rho}{2} \frac{q_0^2}{A_0^2} \left(3H_{qA}(s) - 2\frac{q_0}{A_0}\right)$$

by inserting (11.105). This gives

$$\left(1 + \frac{A_0^2}{\rho\alpha q_0} H_{pq}(s)\right) \frac{\Delta P}{\Delta A}(s) = \frac{\rho q_0^3}{A_0^3} \left(\frac{3}{\alpha} - 2 - 2\frac{A_0^2}{\rho\alpha q_0} H_{pq}(s)\right)$$

and, using the reasonable approximation $\alpha = 1$, we arrive at the well-known power transfer function (Hutarew 1969)

$$\frac{\Delta P}{\Delta A}(s) = \frac{\rho q_0^3}{A_0^3} \frac{(1 - \mu T_r s)}{\left(1 + \mu \frac{T_r}{2} s\right)} \tag{11.118}$$

where T_r and μ are given in (11.112).

Water turbine with compressible water supply

We now include compressibility effects in the supplying pipe. The inlet of the pipe is open, so the transfer function from the volumetric flow $w = \rho q$ at the lower end of the pipe to the pressure p at the same place is given by (4.180) and (4.195) as

$$H_{pq} = \frac{-\Delta p}{\Delta q}(s) = \frac{\rho c}{A_p} \tanh \frac{L}{c} s \tag{11.119}$$

Note that $H_{pq}(s)$ tends to the incompressible solution $\rho L s/A_p$ when $c \to \infty$, which corresponds to the incompressible case where $\beta \to \infty$. We find that when the compressibility effects of the water in the pipe is included the transfer functions to torque and power becomes

$$\frac{\Delta T_t}{\Delta A}(s) = \frac{2}{\gamma} \frac{r_t \rho q_0^2}{A_0^2} \frac{\left(1 - \gamma \frac{A_0^2}{\alpha q_0} \frac{c}{A_p} \tanh \frac{L}{c} s\right)}{\left(1 + \frac{A_0^2}{\alpha q_0} \frac{c}{A_p} \tanh \frac{L}{c} s\right)} \tag{11.120}$$

11.2.10 Example: Waterhammer

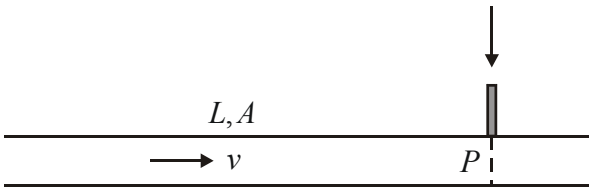

Figure 11.9: The waterhammer effect occurs when the pipe is suddenly closed at P.

The waterhammer effect (Merritt 1967), (Holmboe and Rouleau 1967) occurs when fluid is flowing through a pipe, and the pipe is suddenly closed for example by a valve (Figure 11.9). A fluid with velocity v and density ρ flowing in a pipe of length L and cross section A will have a kinetic energy

$$K = \frac{1}{2} \rho V v^2 \tag{11.121}$$

where $V = LA$ is the volume of the fluid. We note that for a material volume V of a set of particle with mass m and density $\rho = m/V$ the volume differential is

$$dV = d\left(\frac{m}{\rho}\right) = -\frac{m}{\rho^2} d\rho = -\frac{m}{\rho^2} \frac{\rho}{\beta} dp = -\frac{V}{\beta} dp \tag{11.122}$$

Then it follows that if the fluid is instantaneously stopped the kinetic energy K will give an increase ΔP due to compression, which is given by

$$\Delta P = -\int_1^2 p dV = \int_1^2 p \frac{V}{\beta} dp = \frac{1}{2} \frac{V}{\beta} \left(p_2^2 - p_1^2\right) \tag{11.123}$$

where p_1 is the pressure just before the pipe is closed, and p_2 is the pressure just after the pipe is closed. From $K = \Delta P$ the pressure increase is seen to be

$$\sqrt{p_2^2 - p_1^2} = \rho c v \tag{11.124}$$

where $c = \sqrt{\beta/\rho}$ is the sonic speed. In the case that the initial pressure p_1 is small, this is approximated by

$$p_2 = \rho c v \tag{11.125}$$

Example 165 *For water $c = 1500$ m/s and $\rho = 10^3$ kg/m^3, and $p_2 = 1.5 \cdot 10^6 \frac{\text{Pa}}{\text{m/s}} \cdot v$, or $p_2 = 15 \frac{\text{atm}}{\text{m/s}} \cdot v$, so that 5 m/s gives a pressure rise of 75 bar. For hydraulic fluids $c = 1250$ m/s and $\rho = 800$ kg/m^3 which gives $p_2 = 10 \frac{\text{atm}}{\text{m/s}} \cdot v$ so that 5 m/s gives a pressure rise of 50 bar.*

11.3 Angular momentum balance

11.3.1 General expression

The angular momentum equation is important in the modeling of compressors and turbines. Whereas the momentum equation is derived from Newton's law for an infinitesimal material volume, the angular momentum equation is derived from Euler's law of angular momentum

$$\rho \frac{D}{Dt} (\vec{r} \times \vec{v}) \, dV = \vec{r} \times d\vec{F} \tag{11.126}$$

for a material volume element dV. Here $\vec{r}$ is the position vector of the volume element from a specified point o. The force $d\vec{F}$ denotes the resultant force on the volume element, which is the mass force plus the surface force. When this is integrated over the material volume V we get

$$\iiint_V \rho \frac{D}{Dt} (\vec{r} \times \vec{v}) \, dV = \frac{D}{Dt} \iiint_V \vec{r} \times \rho \vec{v} dV = \vec{N}_o \tag{11.127}$$

where

$$\vec{N}_o = \iiint_V \vec{r} \times d\vec{F} \tag{11.128}$$

is the moment about the point o.

The angular momentum equation is given by

$$\frac{D}{Dt} \iiint_V \vec{r} \times \rho \vec{v} dV = \vec{N}_o \tag{11.129}$$

For a general control volume V_c the angular momentum equation is written

$$\frac{d}{dt} \iiint_{V_c} \vec{r} \times \rho \vec{v} dV + \iint_{\partial V_c} (\vec{r} \times \rho \vec{v}) (\vec{v} - \vec{v}_c) \cdot \vec{n} dA = \vec{N}_o \tag{11.130}$$

where $\vec{v}_c$ is the velocity of a point on the surface of V_c.

11.3.2 Centrifugal pump with radial blades

A pump is a device where power is supplied from the pump axis to a fluid to make the fluid flow with a mass flow w. The pump axis may be driven by an electrical motor, an engine or a turbine. We will first consider a centrifugal pump with radial blades acting on an incompressible fluid (Figure 11.10). The pump has angular shaft velocity ω. The

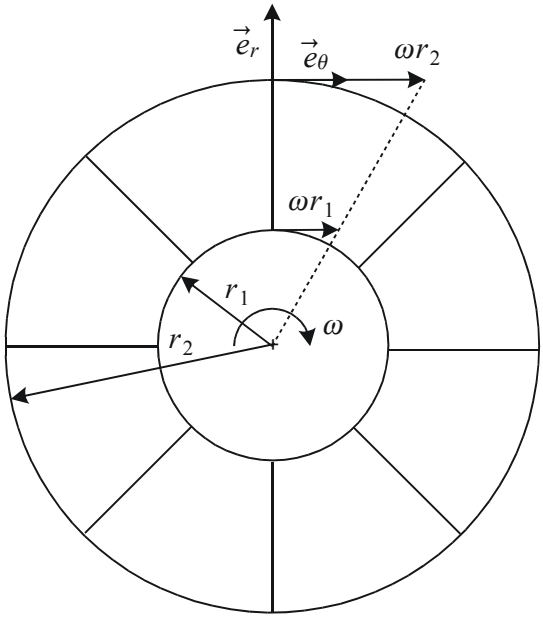

Figure 11.10: Centrifugal pump with radial blades.

fluid enters in the center, and flows through an arrangement of radial blades with inner blade tips at a radius r_1 and outer blade tips at a radius r_2. We define a frame with orthogonal unit vectors $\vec{e}_r, \vec{e}_\theta, \vec{e}_z$ where $\vec{e}_r$ is in the radial direction, $\vec{e}_\theta$ in the tangential direction, and $\vec{e}_z$ is along the pump axis. We note that the inner tip speed of the blades is $\vec{U}_1 = r_1\omega\vec{e}_\theta$ while the outer tip speed of the blades is $\vec{U}_2 = r_2\omega\vec{e}_\theta$. We will assume that the fluid flow is constant and with a mass flow

$$w = 2\pi r_1 b v_{1r} = 2\pi r_2 b v_{2r} \tag{11.131}$$

where b is the width of the pump, v_{1r} is the radial fluid velocity at the blade inlet, and v_{2r} is the radial fluid velocity at the blade outlet. We will consider the moment about the pump axis, which means that the point o is in the center of the pump, so that we have $\vec{r}_1 = r_1\vec{e}_r$ at the blade inlet and $\vec{r}_2 = r_2\vec{e}_r$ at the outlet. The fluid velocity at the blade inlet is denoted $\vec{v}_1$ and the velocity at the blade outlet is denoted $\vec{v}_2$ where

$$\vec{v}_1 = \frac{w}{2\pi r_1 b}\vec{e}_r + r_1\omega\vec{e}_\theta, \quad \vec{v}_2 = \frac{w}{2\pi r_2 b}\vec{e}_r + r_2\omega\vec{e}_\theta \tag{11.132}$$

This gives

$$\vec{r}_1 \times \vec{v}_1 = r_1^2\omega\vec{e}_z, \quad \vec{r}_2 \times \vec{v}_2 = r_2^2\omega\vec{e}_z \tag{11.133}$$

The control volume V_c is taken to be the volume between the blade inlet and the blade outlet. This is a volume that is fixed in space, so that $\vec{v}_c = \vec{0}$. The outwards pointing surface normal is $\vec{n} = -\vec{e}_\theta$ at the inlet and $\vec{n} = \vec{e}_\theta$ at the outlet. The angular momentum balance (11.130) gives

$$\iiint_{V_c} \frac{\partial}{\partial t}\left(\rho r^2\omega\vec{e}_z\right) dV + w\left(r_2^2\omega - r_1^2\omega\right)\vec{e}_z = \vec{T}_p \tag{11.134}$$

where $\vec{T}_p = T_p \vec{e}_z$ is the load torque on the shaft. This gives

$$J_f \dot{\omega} + w\omega \left(r_2^2 - r_1^2\right) = T_p \tag{11.135}$$

where

$$J_f = \frac{\pi b \rho}{2} \left(r_2^4 - r_1^4\right) = \frac{m_f}{2} \left(r_1^2 + r_2^2\right) \tag{11.136}$$

is the moment of inertia due to the fluid, and $m_f = \pi b \rho \left(r_2^2 - r_1^2\right)$ is the mass of the fluid in V_c. We see that the stationary shaft torque needed to pump a mass flow of w is

$$T_p = w\omega \left(r_2^2 - r_1^2\right) \tag{11.137}$$

The shaft power is

$$P_p = T_p \omega = w\omega^2 \left(r_2^2 - r_1^2\right) \tag{11.138}$$

11.3.3 Euler's turbomachinery equation

In a more well-designed centrifugal pump the blades will be curved, and the blades will have an inlet angle β_1 and outlet angle β_2. The velocity $\vec{v}_1$ at the inlet and the velocity $\vec{v}_2$ at the blade outlet is written

$$\vec{v}_1 = v_{1r}\vec{e}_r + v_{1t}\vec{e}_\theta, \quad \vec{v}_2 = v_{2r}\vec{e}_r + v_{2t}\vec{e}_\theta \tag{11.139}$$

We get

$$\vec{r}_1 \times \vec{v}_1 = r_1 v_{1t} \vec{e}_z, \quad \vec{r}_2 \times \vec{v}_2 = r_2 v_{2t} \vec{e}_z \tag{11.140}$$

Then, proceeding as in the previous section, we get the shaft power

$$T_p = w \left(r_2 v_{2t} - r_1 v_{1t}\right) \tag{11.141}$$

The shaft power is found to be

$$P_p = T\omega = w\omega \left(r_2 v_{2t} - r_1 v_{1t}\right) \tag{11.142}$$

A turbine is a device where a fluid delivers power to the turbine shaft by changing the momentum of the fluid. This means that a turbine converts kinetic energy in a fluid to mechanical energy in the form of rotational energy of the shaft. We note that for the centrifugal pump the shaft torque T is zero when the shaft speed ω is zero. This shows that the centrifugal pump with radial blades cannot be used as a turbine.

11.3.4 Pump instability

The direction of the velocity vectors $\vec{v}_1$ and $\vec{v}_2$ are described by the flow angles

$$\tan \alpha_1 = \frac{v_{1t}}{v_{1r}}, \quad \tan \alpha_2 = \frac{v_{2t}}{v_{2r}} \tag{11.143}$$

We define $\vec{W}_1$ and $\vec{W}_2$ by

$$\vec{v}_1 = \vec{U}_1 + \vec{W}_1, \quad \vec{v}_2 = \vec{U}_2 + \vec{W}_2 \tag{11.144}$$

$$\vec{W}_1 = W_{1r}\vec{e}_r + W_{1t}\vec{e}_\theta, \quad \vec{W}_2 = W_{1r}\vec{e}_r + W_{1t}\vec{e}_\theta \tag{11.145}$$

At the blade outlet the fluid flow will be along the blade, so that the velocity $\vec{W}_2$ will have direction given by the blade outlet angle β_2. At design speed a design rule is to

select the inlet blade angle β_1 so that the inlet flow will be along the blade at the inlet, so that $\vec{W}_1$ will have direction given by β_1. Then

$$W_{1t} = -v_{1r}\text{cotan}\beta_1, \quad W_{2t} = -v_{2r}\text{cotan}\beta_2 \tag{11.146}$$

will be the tangential fluid velocities relative to the blades. We will consider the situation when there is no *pre-whirl*, which means that the tangential speed at the blade inlet is zero. Then

$$v_{1t} = 0 \quad \Rightarrow \quad \tan\beta_1 = \frac{v_{1r}}{U_1} \tag{11.147}$$

and the torque is found to be

$$\begin{aligned} T &= wr_2\left(U_2 - v_{2r}\text{cotan}\beta_2\right) \\ &= wr_2\left(\omega r_2 - \frac{w}{2\pi r_2 b\rho}\text{cotan}\beta_2\right) \end{aligned} \tag{11.148}$$

Suppose that the pump is delivering an incompressible fluid to a pipe of cross section A and length L. The velocity at the inlet of the pipe is denoted v, and it is assumed that the mass flow is

$$w = \rho A v \tag{11.149}$$

The equation of motion for the fluid is

$$\rho AL\dot{v} = F - F_{\text{out}} \tag{11.150}$$

where F_{out} is the force acting at the pipe outlet. We assume that the shaft power $T\omega$ is converted to kinetic power Fv for the fluid in the pump so that $T\omega = Fv$. Then the force F at the inlet of the pipe is found to be

$$F = \frac{\omega}{v}T = \rho A\frac{\omega}{w}T = \rho A\omega r_2\left(\omega r_2 - \frac{Av}{2\pi r_2 b}\text{cotan}\beta_2\right) \tag{11.151}$$

and the equation of motion becomes

$$\rho AL\dot{v} = \rho A\omega^2 r_2^2 - v\frac{\rho A^2\omega}{2\pi b}\text{cotan}\beta_2 - F_{\text{out}} \tag{11.152}$$

The force consist of a term that is proportional to ω^2 which can be considered as the forcing term. In addition there is the second term on the right side of (11.152) which is proportional to the outlet fluid velocity v. If $\beta_2 > 90°$, which is the case if the blade have a backsweep at the outlet, then the velocity term will have the same effect as viscous friction, and has a stabilizing effect. However, if the blades are swept forward, then $\beta_2 > 90°$, and the second term on the right side of (11.152) will give the same effect as a positive velocity feedback, which may cause the system to be unstable.

Example 166 *The pump delivers an incompressible fluid through a pipe of cross section A to a basin. The fluid level in the basin is denoted h. Water flows out of the basin through at throttle with mass flow $w_t(h) = C\sqrt{h}$. The model for the system is*

$$\dot{v} = -\frac{A}{2\pi bL}\omega\text{cotan}\beta_2 v - \frac{g}{L}h + \frac{r_2^2\omega^2}{L} \tag{11.153}$$

$$\dot{h} = \frac{A}{A_b}v - \frac{1}{A_b\rho}w_t\left(h\right) \tag{11.154}$$

where the pump velocity ω^2 is considered to be the control input. This can be achieved by velocity control of the motor driving the pump. Linearization gives

$$\dot{v} = a_{11}v + a_{12}h + b\omega^2 \tag{11.155}$$
$$\dot{h} = a_{21}v + a_{22}h \tag{11.156}$$

and the characteristic equation of the linearized system is found to be

$$\lambda^2 - (a_{11} + a_{22})\,\lambda - a_{12}a_{21} = 0 \tag{11.157}$$

Stability results whenever

$$a_{11} + a_{22} = -\frac{A}{2\pi bL}\omega \mathrm{cotan}\beta_2 - \frac{1}{A_b\rho}\frac{dw_t}{dh} < 0 \tag{11.158}$$

which is the case if

$$\mathrm{cotan}\beta_2 < -\frac{2\pi bL\omega}{\rho A A_b}\frac{dw_t}{dh} \tag{11.159}$$

This means that if the blade outlets are backswept so that $\mathrm{cotan}\beta_2 \geq 0$*, then the system will be stable. Forward swept blade outlets may cause instability depending on the system parameters.*

Example 167 *Under stationary conditions is may be assumed that the mechanical power $T\omega$ from the shaft is converted to power Fv supplied to the fluid, so that*

$$F = \frac{\omega}{v}T \tag{11.160}$$

In transients there will be energy loss until the stationary flow pattern is established. It is reasonable to assume that these transient flow will last for at least the time time it takes a fluid particle to flow through the pump, and in some cases up to 5 times of this time. Then a reasonable model for the transients in the shaft torque is

$$\dot{F} = \frac{1}{\alpha T_{\mathrm{trans}}}\left(\frac{\omega}{v}T - F\right) \tag{11.161}$$

where T_{trans} can be taken to be the transport time of a fluid particle through the pump, and α is in the range from 1 to 5.

11.4 The energy balance

11.4.1 Material volume

A material volume has a fixed set of particles. Therefore the total energy of a material volume is conserved. This means that the rate of change of the total energy of a material volume is equal to the net rate of energy supplied to the volume. We assume here that the total energy in a volume element dV is $\rho e dV$ where

$$e = u + \frac{1}{2}\vec{v}^2 + \phi \tag{11.162}$$

is the specific energy, u is the specific internal energy, $(1/2)\vec{v}^2$ is the specific kinetic energy, and ϕ is the specific potential energy. We assume that the body forces are derived from the potential ϕ in the sense that

$$\vec{\nabla}\phi = -\vec{f} \;\Rightarrow\; \frac{D\phi}{Dt} = \left(\vec{\nabla}\phi\right)\cdot\vec{v} = -\vec{f}\cdot\vec{v} \tag{11.163}$$

The material time derivative of the total energy in a volume V is equal to the net rate of energy supplied to the volume. Suppose that the net supplied energy is the sum of the heat flow into the volume due to the heat flux density $\vec{j}_Q$ plus the power added from the pressure force $-p\vec{n}$ acting on the surface. This is written

$$\frac{D}{Dt}\iiint_V \rho e dV = -\iint_{\partial V} p\vec{v}\cdot\vec{n}dA - \iint_{\partial V}\vec{j}_Q\cdot\vec{n}dA \tag{11.164}$$

The volume V is arbitrary, and it follows from the divergence theorem that

$$\underbrace{\rho\frac{D}{Dt}\left(u+\frac{1}{2}\vec{v}^2+\phi\right)}_{\substack{\text{rate of change}\\ \text{in internal, kinetic}\\ \text{and potential energy}\\ \text{for material}\\ \text{volume element}}} = - \underbrace{\vec{\nabla}\cdot(p\vec{v})}_{\substack{\text{pressure work}\\ \text{on the surface of}\\ \text{the volume element}}} - \underbrace{\vec{\nabla}\cdot\vec{j}_Q}_{\substack{\text{heat}\\ \text{conduction}}} \tag{11.165}$$

The divergence form is found by changing the left hand side as follows:

$$\rho\frac{De}{Dt} = \frac{\partial}{\partial t}(\rho e) + \vec{\nabla}\cdot(\rho e\vec{v}) \tag{11.166}$$

If we leave out the potential energy, then (11.163) can be used to express the energy equation in the form

$$\underbrace{\rho\frac{D}{Dt}\left(\frac{1}{2}\vec{v}^2+u\right)}_{\substack{\text{rate of change}\\ \text{in internal and}\\ \text{kinetic energy}\\ \text{for material}\\ \text{volume element}}} = - \underbrace{\vec{\nabla}\cdot(p\vec{v})}_{\substack{\text{pressure work}\\ \text{on the surface of}\\ \text{the volume element}}} - \underbrace{\vec{\nabla}\cdot\vec{j}_Q}_{\substack{\text{heat}\\ \text{conduction}}} + \underbrace{\rho\vec{v}\cdot\vec{f}}_{\substack{\text{work of body}\\ \text{forces on volume}\\ \text{element}}} \tag{11.167}$$

Example 168 *If the pressure is constant over the volume, then the pressure work can be written*

$$\iint_{\partial V} p\vec{v}\cdot\vec{n}dA = p\iint_{\partial V}\vec{v}\cdot\vec{n}dA = p\frac{DV}{Dt} \tag{11.168}$$

which means that the pressure work is equal to the pressure times the time derivative of the material volume.

11.4.2 Fixed volume

If the volume V is fixed then the energy balance can be written in the material form using

$$\frac{D}{Dt}\iiint_V \rho e dV = \frac{d}{dt}\iiint_V \rho e dV + \iint_{\partial V}\rho e\vec{v}\cdot\vec{n}dA \tag{11.169}$$

Insertion of (11.164) gives the result

$$\frac{d}{dt}\iiint_V \rho e dV = -\iint_{\partial V} \rho\left(e+\frac{p}{\rho}\right)\vec{v}\cdot\vec{n}dA - \iint_{\partial V}\vec{j}_Q\cdot\vec{n}dA \tag{11.170}$$

where the first term on the right side is the convected energy plus the pressure work on the volume. At this stage it is useful to introduce the *specific enthalpy* h which is defined by

$$h = u + \frac{p}{\rho} \tag{11.171}$$

Then the energy balance can be written

$$\underbrace{\frac{d}{dt}\iiint_V \rho\left(u+\frac{1}{2}\vec{v}^2+\phi\right)dV}_{\substack{\text{rate of change}\\ \text{of energy}\\ \text{in fixed volume}}} = \underbrace{-\iint_{\partial V}\rho\left(h+\frac{1}{2}\vec{v}^2+\phi\right)\vec{v}\cdot\vec{n}dA}_{\substack{\text{convected enthalpy,}\\ \text{kinetic energy and}\\ \text{potential energy}}} \underbrace{-\iint_{\partial V}\vec{j}_Q\cdot\vec{n}dA}_{\substack{\text{heat}\\ \text{conduction}}} \tag{11.172}$$

Note that in the convection term the enthalpy h enters in place of the internal energy u as the pressure work is included in the convection term.

Example 169 *Suppose that the specific energy of the system in Example 158 is simply $e = u$, which means that the kinetic and potential energy can be neglected. Moreover, suppose that there is no heat flow into the volume, that is, $\vec{j}_Q = \vec{0}$, that there is no mass or energy generation in the volume, and that ρ and u are constants over the volume. Then the energy balance is*

$$V\frac{d}{dt}(u\rho) = u_1A_1\rho_1v_1 - uA_2\rho v_2 + p_1A_1v_1 - pA_2v_2 \tag{11.173}$$

which gives

$$\frac{d}{dt}E = h_1w_1 - hw_2 \tag{11.174}$$

where we used $E = mu$ where $m = \rho V$, and where we have used the enthalpy $h = u+p/\rho$. We may obtain an equation for the specific internal energy u by expanding the left side. This gives

$$m\dot{u} + \dot{m}u = h_1w_1 - hw_2. \tag{11.175}$$

Combining this with the mass balance

$$\dot{m} = w_1 - w_2 \tag{11.176}$$

we get a differential equation for the specific internal energy in the form

$$m\dot{u} = h_1w_1 - hw_2 - u(w_1 - w_2) \tag{11.177}$$

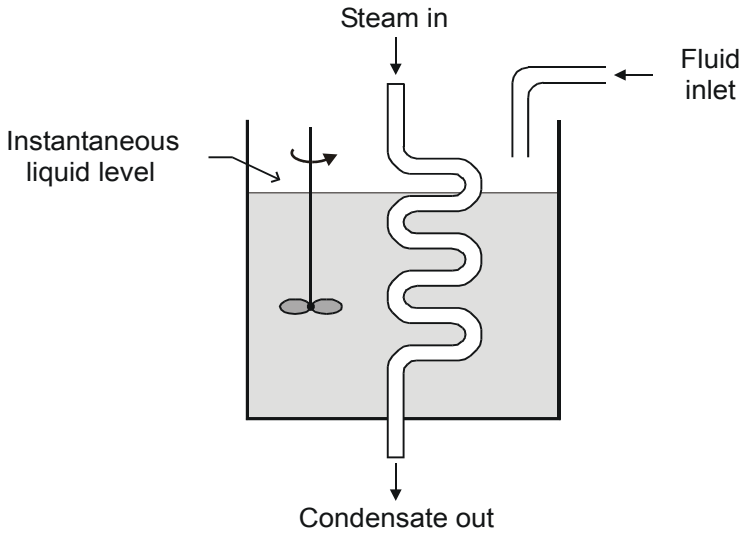

Figure 11.11: Water tank heated by a coil filled with steam.

which gives

$$\dot{u} = \frac{w_1}{m}(h_1 - u) - \frac{w_2}{m}\frac{p}{\rho} \tag{11.178}$$

We will later see that this leads to a differential equation for the temperature by using $h = c_p T$ *and* $u = c_v T$.

Example 170 *This example and the next example are adopted from (Bird, Stewart and Lightfoot 1960, p. 473). A cylindrical tank with cross section* A *is filled with a liquid with a mass flow* w *(Figure 11.11). The volume of the liquid in the tank is* $V = Az$ *where* z *is the height of the liquid surface. The liquid in the tank is heated with a coil filled with steam of temperature* T_s*. The heat transfer coefficient per length unit of the coil from the coil to the liquid is* G*. The tank is stirred so that the temperature of the liquid in the tank is uniform. The energy of the liquid is supposed to be* $u = c_p T$*. The mass and energy balances are*

$$\frac{d}{dt}(\rho V) = w \tag{11.179}$$

$$\frac{d}{dt}(\rho u V) = w u_1 + Gz(T_s - T). \tag{11.180}$$

The first term on the right side of the energy balance is the convected internal energy, while the second term is a heat conduction term as in the general expression (11.172). The energy balance can be written out as

$$\left(\frac{d}{dt}\rho V\right) c_p T + \rho V c_p \frac{dT}{dt} = w c_p T_1 + Gz(T_s - T). \tag{11.181}$$

Insertion of the mass balance in the energy balance gives

$$\rho V c_p \frac{dT}{dt} = w c_p T_1 - w c_p T + Gz(T_s - T) \tag{11.182}$$

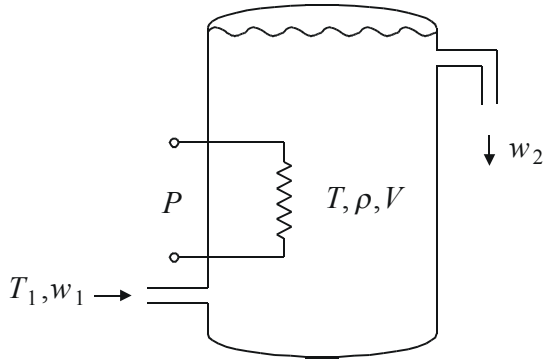

Figure 11.12: Heated tank.

and some straightforward manipulations lead to the model

$$\dot{z} = \frac{w}{\rho A} \tag{11.183}$$

$$\dot{T} = \frac{w}{\rho A z}(T_1 - T) + \frac{G}{\rho A c_p}(T_s - T). \tag{11.184}$$

Example 171 *A liquid is heated by pumping it through a tank with an electrical heating element supplying the power P as shown in Figure 11.12. The temperature of the liquid in the tank is T, the density ρ is constant, and the volume of the tank is V. The inlet has mass flow w_1 and temperature T_1, while the outlet has mass flow w_2. The liquid flowing through the outlet has the temperature T of the liquid in the tank. The energy of the liquid in the tank is mc_pT where $m = \rho V$ is the mass of the liquid in the tank. The mass balance implies that*

$$w_1 = w_2 = w \tag{11.185}$$

The energy balance is then

$$\frac{d}{dt}(mc_pT) = c_p w (T_1 - T) + P. \tag{11.186}$$

From the mass balance we have $\dot{m} = 0$, and we get

$$\dot{T} = \frac{w}{m}(T_1 - T) + \frac{P}{c_p m}. \tag{11.187}$$

11.4.3 General control volume

For a general control volume V_c Reynolds' transport theorem (10.88) gives

$$\frac{d}{dt}\iiint_{V_c} \rho e dV = \frac{D}{Dt}\iiint_{V_c} \rho e dV - \iint_{\partial V_c} \rho e (\vec{v} - \vec{v}_c) \cdot \vec{n} dA \tag{11.188}$$

From equation (11.164) we have the following expression for the material derivative of the energy:

$$\frac{D}{Dt}\iiint_{V_c} \rho e dV = -\iint_{\partial V_c} p\vec{v} \cdot \vec{n} dA - \iint_{\partial V_c} \vec{j}_Q \cdot \vec{n} dA \tag{11.189}$$

Combining the two equations we find that

$$\begin{aligned}\frac{d}{dt}\iiint_{V_c}\rho e dV &= -\iint_{\partial V_c}\rho\left(e+\frac{p}{\rho}\right)(\vec{v}-\vec{v}_c)\cdot\vec{n}dA \\ &\quad -\iint_{\partial V_c}p\vec{v}_c\cdot\vec{n}dA-\iint_{\partial V_c}\vec{j}_Q\cdot\vec{n}dA \end{aligned} \tag{11.190}$$

The first term on the right side is the convected energy plus the pressure work on the volume. The specific enthalpy $h = u + p/\rho$ is inserted. Then the energy balance can be written

$$\begin{aligned}\underbrace{\frac{d}{dt}\iiint_{V_c}\rho\left(u+\frac{1}{2}\vec{v}^2+\phi\right)dV}_{\substack{\text{rate of change}\\ \text{of energy}\\ \text{in control volume}}} &= \underbrace{-\iint_{\partial V_c}\rho\left(h+\frac{1}{2}\vec{v}^2+\phi\right)(\vec{v}-\vec{v}_c)\cdot\vec{n}dA}_{\substack{\text{convected enthalpy,}\\ \text{kinetic energy and}\\ \text{potential energy}}} \\ &\quad -\underbrace{\iint_{\partial V_c}p\vec{v}_c\cdot\vec{n}dA}_{\substack{\text{pressure work}\\ \text{due to change in}\\ \text{control volume}}}-\underbrace{\iint_{\partial V_c}\vec{j}_Q\cdot\vec{n}dA}_{\substack{\text{heat}\\ \text{conduction}}}\end{aligned} \tag{11.191}$$

Note that the velocity in the convection term is $\vec{v} - \vec{v}_c$ which is the particle velocity relative to the surface of the control volume V_c.

Example 172 *If the pressure is constant over the volume then the pressure work on the surface of the control volume can be written*

$$\iint_{\partial V_c}p\vec{v}_c\cdot\vec{n}dA = p\iint_{\partial V_c}\vec{v}_c\cdot\vec{n}dA = p\dot{V}_c \tag{11.192}$$

11.4.4 The heat equation

Heat conduction in a solid is described by the heat equation. The energy appears in the form of internal energy $u = c_pT$, and energy flow is due to heat conduction according to the constitutive equation in the form of Fourier's law

$$\vec{j}_Q = -\alpha\vec{\nabla}(\rho c_p T) \tag{11.193}$$

where the α is the thermal diffusivity in m^2/s. The energy balance is simply

$$\rho\frac{\partial u}{\partial t} = -\vec{\nabla}\cdot\vec{j}_Q \tag{11.194}$$

which in combination with Fourier's law with constant α and ρ gives

$$\frac{\partial T}{\partial t}-\alpha\nabla^2 T = 0 \tag{11.195}$$

where $\nabla^2 = \vec{\nabla}\cdot\vec{\nabla}$ is the Laplacian operator.

11.4.5 Transfer function for the heat equation

The heat equation in one dimension for the temperature $T(x,t)$ in a bar from $x=0$ to $x=L$ is given by

$$\frac{\partial T(x,t)}{\partial t} - \alpha \frac{\partial^2 T(x,t)}{\partial x^2} = 0 \tag{11.196}$$

Suppose that the bar is insulated at $x=0$, and that the heat flux j_Q is controlled at $x=L$ according to $j_Q = -\alpha\rho c_p u$ where u is the control variable. Then the boundary conditions are

$$\frac{\partial T(0,t)}{\partial x} = 0, \quad \frac{\partial T(L,t)}{\partial x} = u \tag{11.197}$$

Laplace transformation of (11.196) gives

$$\frac{\partial^2 T(x,s)}{\partial x^2} - \frac{s}{\alpha} T(x,s) = 0 \tag{11.198}$$

which has the solution

$$T(x,s) = A\cosh\left(\sqrt{\frac{s}{\alpha}}x\right) + B\sinh\left(\sqrt{\frac{s}{\alpha}}x\right) \tag{11.199}$$

with derivative

$$\frac{\partial T(x,s)}{\partial x} = A\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}x\right) + B\sqrt{\frac{s}{\alpha}}\cosh\left(\sqrt{\frac{s}{\alpha}}x\right) \tag{11.200}$$

The boundary condition at $x=0$ gives $B=0$, and the boundary condition at $x=L$ gives

$$A\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}L\right) = u \tag{11.201}$$

so that the temperature is given by

$$T(x,s) = \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}x\right)}{\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}L\right)} u(s) \tag{11.202}$$

The transfer function from the heat flux to the temperature at $x=L$ is found to be

$$\frac{T(L,s)}{u(s)} = \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}{\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}L\right)} \tag{11.203}$$

The zeros of the transfer function are found from

$$L\sqrt{\frac{s}{\alpha}} = j\left(k\pi + \frac{\pi}{2}\right) \Rightarrow \frac{s}{\alpha} = -\frac{1}{L^2}\left(k\pi + \frac{\pi}{2}\right)^2 \tag{11.204}$$

while the singularities are given by

$$L\sqrt{\frac{s}{\alpha}} = jk\pi \Rightarrow \frac{s}{\alpha} = -\frac{1}{L^2}(k\pi)^2 \tag{11.205}$$

Numerical values are given in Table 11.1.

Zeros		Singularities	
$L\sqrt{\frac{s}{\alpha}}$	$L^2\frac{s}{\alpha}$	$L\sqrt{\frac{s}{\alpha}}$	$L^2\frac{s}{\alpha}$
1.570 8	−2.467 4	0	0
4.712 4	−22.207	3.141 6	−9.869 6
7.854 0	−61.685	6.283 2	−39.478
10.995541	−120.9019	9.424 8	−88.826

(11.206)

Table 11.1: Singularites for the one-dimensional heat equation when the beam is insulated at $x = 0$, and the heat flux is controlled at $x = L$.

Example 173 *The heat equation is studied for the bar of the previous example, but the boundary condition at $x = L$ is changed so that the bar is in contact with a reservoir with temperature u, which is the control input. The heat-transfer coefficient is β. Then the boundary conditions are changed to*

$$\frac{\partial T(0,t)}{\partial x} = 0, \quad \frac{\partial T(L,t)}{\partial x} = \beta[u - T(L,t)] \tag{11.207}$$

The boundary condition at $x = 0$ gives $B = 0$, while the boundary condition at $x = L$ in combination with (11.199) gives

$$\sqrt{\frac{s}{\alpha}} A \sinh\left(\sqrt{\frac{s}{\alpha}} L\right) = \beta \left[u(s) - A\cosh\left(\sqrt{\frac{s}{\alpha}} L\right)\right]$$

This implies that

$$A = \frac{\beta}{\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}L\right) + \beta\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)} u(s) \tag{11.208}$$

and insertion in (11.199) gives the transfer function

$$\frac{T(L,s)}{u(s)} = \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}{\frac{1}{\beta}\sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}L\right) + \cosh\left(\sqrt{\frac{s}{\alpha}}L\right)} \tag{11.209}$$

11.5 Viscous flow

11.5.1 Introduction

So far the balance equations for momentum and energy have been developed for inviscid fluids, that is, for fluids without viscosity. In some problems, viscous effects may be important, and in the following balance equations for the viscous case will be developed. The mathematical level is somewhat more advanced than for the inviscid case. The main reason for this is the appearance of the viscous stress tensor which necessitates the introduction of tensor notation.

11.5.2 Tensor notation

The derivation of certain important results in fluid mechanics are best done in tensor notation (Aris 1989), (Lovelock and Rund 1989). This involves a systematic notation

Part V

Simulation

Chapter 14

Simulation

14.1 Introduction

14.1.1 The use of simulation in automatic control

Simulation of dynamic processes involves the numerical solution of differential equations which are normally in the form of initial value problems. The numerical schemes that are used for this are called *numerical integrators*. There is a large literature on numerical integrators (Hairer, Nørsett and Wanner 1993), (Hairer and Wanner 1996), (Lambert 1991), (Shampine 1994), and a wide range of methods are available. These methods have different properties and the selection of which method to use depends on the properties of the system to be simulated. In this chapter a range of numerical integrators are presented and analyzed, and it is attempted to give some advice on how a suitable method can be selected for important dynamic systems.

Simulation plays an important part in the design, maintenance and upgrading of control systems. Dynamic systems to be controlled are usually described by differential equations or transfer functions, and simulation is used to check the qualitative behavior of the system for typical parameter values and for expected modes of operation. When a controller is designed for a system it is usual practice to test the controller in simulations before implementing it. This allows for rapid changes and correction of errors before the system is designed. Also it is important that procedures for handling of discrete events and errors can be tested. For systems where a controller has already been developed, quantitative aspects of simulation is important for the fine tuning of controller parameters and the redesign of the system to be controlled.

An example where this is useful is in the development of industrial robots. In applications like spot welding in car production lines there are ever-increasing demands on the robot to finish spot welding tasks faster while maintaining the weld quality specifications. Then, simulation must be used to improve controller parameters, to try out friction compensation, and to improve the mechanical construction so that elastic deformations can be reduced. The alternative to using simulation would be iterative mechanical redesign which is costly and time consuming.

In car engines new regulations of emissions are enforced, and there is a demand for engines that are lighter, that use less fuel and that pollute less. The introduction of new electronic control systems is necessary to achieve this. Car manufacturers use simulation systems to reduce mechanical vibrations, to shape the combustion chamber for efficient combustion of the fuel, to reduce the formation of pollutants, to optimize electronic

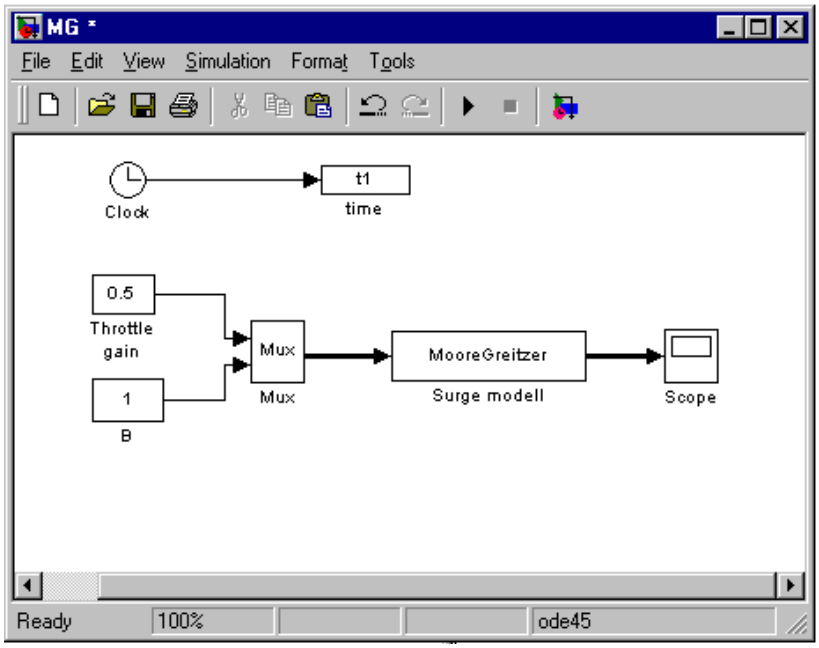

Figure 14.1: The model implemented in Simulink.

controls for components like fuel injectors, turbochargers, and valves for exhaust gas recycling. Also in the design and testing of control systems for ABS brakes simulation is an indispensable tool.

For ship control systems there are large costs involved in commissioning of control systems, which involves installation and calibration. Moreover, a typical situation is that there is very little time available for the control engineer to commission the controllers before the ship is to be set into commercial operation. By use of simulation the time for commissioning can be reduced significantly, and this may be a decisive factor to make control systems commercially attractive for the marine industry.

The last few years new and powerful tools have been made available for simulation which makes it much easier to run simulations than what have been the case. Also simulation tools and control systems development tool have been integrated, and the role of simulation in automatic control is becoming even more important than it used to be. Still, it is important to know the properties of the numerical schemes that are used so that the results can be interpreted in the right way.

In the following, three examples are presented where the dynamics of systems without controllers are presented. Simulation of the dynamics of these systems reveals the qualitative properties of the systems, and this is useful a starting point for designing controllers. MATLAB code is included for two of the examples to make it easy for the reader to simulate the systems with MATLAB or SIMULINK.

14.1.2 The Moore Greitzer model

A jet engine consists basically of a compressor, a combustion chamber, a turbine and connecting ducts. The compressor delivered compressed air to the combustion chamber

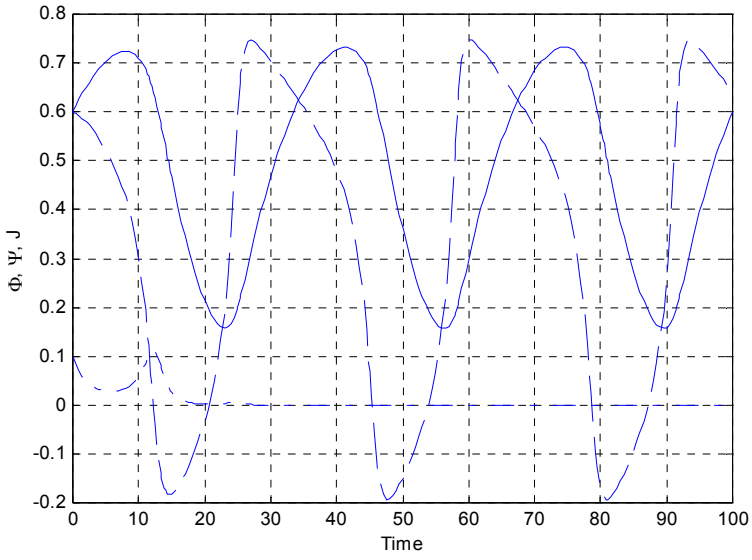

Figure 14.2: Simulation of rotating stall. Φ, Ψ and J are plotted with dashed solid and dash-dotted lines respectively.

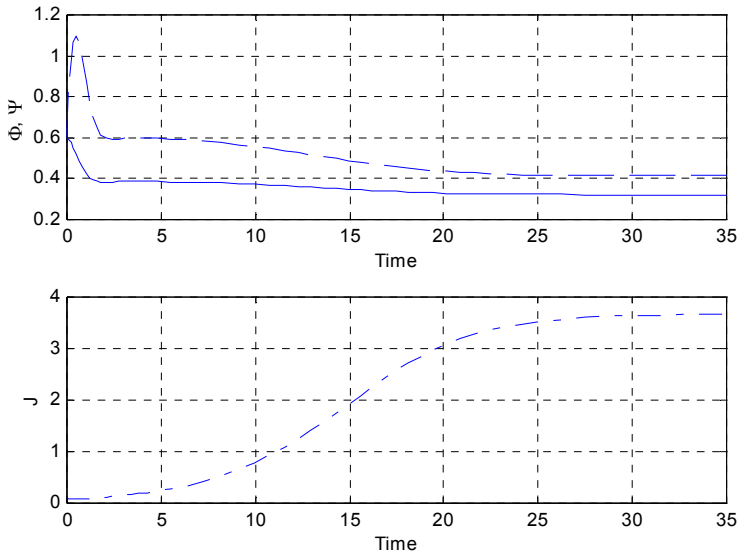

Figure 14.3: Simulation of surge. Φ, Ψ and J are plotted with dashed solid and dash-dotted lines respectively.

where fuel is added. The gas expands and drives the turbine which delivers the required power to the compressor over a shaft. The thrust force comes from the mass flow through the restriction behind the turbine, and in modern jet from a fan that is driven by the turbine If the jet-stream of another aeroplane meets the intake, or if the aspect angle of the aeroplane becomes too large there will be a severe disturbance in the fluid flow at the compressor inlet. This may cause the mass flow through the compressor to become smaller than a critical value given by the surge line, and the engine will enter one of two unstable operating modes known as surge and rotating stall. Surge is an axisymmetric pulsation of the flow through the compressor, while rotating stall is an instability where the circumferential flow pattern is disturbed. If the engine enters rotating stall it will be necessary to shut down the engine, which may lead to the plane falling out of the sky. These physical phenomena are described by the Moore-Greitzer model (Moore and Greitzer 1986) that describes the transients in an axial compression system like an aircraft jet engine. Based on this model a number of control systems have been designed to stabilize surge and stall. This is discussed in (Gravdahl and Egeland 1999). In the model the turbine is modeled as a throttle and the combustion chamber is called the plenum. The model consists of three nonlinear differential equations, and the states are plenum pressure, mass flow and rotating stall amplitude. The rotating stall amplitude is a measure of the unstable non-axisymmetric flow disturbance. All states have been made normalized. The Moore-Greitzer model is written

$$\dot{\Psi} = \frac{1}{4l_c B^2}\left(\Phi - \gamma_T\sqrt{\Psi}\right) \tag{14.1}$$

$$\dot{\Phi} = \frac{H}{l_c}\left(-\frac{\Psi - \psi_{c0}}{H} + 1 + \frac{3}{2}\left(\frac{\Phi}{W} - 1\right)\left(1 - \frac{J}{2}\right) - \frac{1}{2}\left(\frac{\Phi}{W} - 1\right)^3\right) \tag{14.2}$$

$$\dot{J} = J\left(1 - \left(\frac{\Phi}{W} - 1\right)^2 - \frac{J}{4}\right)\sigma \tag{14.3}$$

where Ψ is the nondimensional plenum pressure (pressure divided by density and the square of compressor rotational speed), Φ is the average mass flow coefficient (axial flow velocity divided by compressor rotational speed), and J is the squared amplitude of rotating stall amplitude. The constant l_c is the total length of the compressor and duct, A_c is the cross sectional flow area, γ_T is a parameter proportional to the throttle opening, and H, W, ψ_{c0} and σ are constants describing the compressor. Finally,

$$B = \frac{U}{2c}\sqrt{\frac{V_p}{A_c l_c}} \tag{14.4}$$

is Greitzer's B-parameter, where U is the tangential speed of the compressor, c is the speed of sound, V_p is the plenum volume, and A_c is the cross sectional flow area.

Numerical values for a laboratory compression system in unstable operation is $H = 0.18$, $W = 0.25$, $\psi_{c0} = 0.30$, $\sigma = 0.38$, $\gamma_T = 0.5$ and $l_c = 2$. Initial conditions that correspond to a stable operating point are given by $\Psi(0) = \Phi(0) = 0.6, J(0) = 0.1$. By setting the B-parameter at $B = 0.1$, the engine will go into rotating stall, and by setting the B-parameter at $B = 1$, the engine will start to surge.

The model (14.1)-(14.3) can be simulated in SIMULINK under MATLAB by implementing the following SIMULINK s-function:

```
function [sys,x0] = MooreGreitzer(t,x,u,flag)
H=0.18;
```

```
W=0.25;
l_c=2;
psi_co=0.30;
s=0.38;

if flag == 1,
  %return state derivatives
  gamma_T=u(1);
  B=u(2);
  sys(1)=1/(4*l_c*B^2)*(x(2)-gamma_T*sqrt(x(1)));
  sys(2)=H/l_c*(-(x(1)-psi_co)/H+1+1.5*(x(2)/W-1)*(1-0.5*x(3))
         -0.5*(x(2)/W-1)^3);
  sys(3)=x(3)*(1-(x(2)/W-1)^2-x(3)/4)*s;
elseif flag == 0,
  % return initial conditions
  sys=[3;0;3;2;0;0];
  x0=[0.6;0.6;0.1];
elseif flag == 3,
  % return outputs
  sys=[x(1) x(2) x(3)];
else
  sys = [];
end
```

The model may now be simulated in SIMULINK by making a block diagram as shown in Figure 14.1. The simulation result for both rotating stall and surge is shown in Figures 14.2 and 14.3.

14.1.3 The restricted three-body problem

The restricted three-body problem describes the motion of a satellite moving in the combined gravitational field of the moon and the earth. There are three bodies in the problem, the satellite, the moon and the earth. The mass of the spacecraft is assumed to be so small that it does not influence the motion of the moon or the earth. The normalized model is derived in Section 8.9.3, and is given by

$$\dot{x} = v_x \tag{14.5}$$

$$\dot{y} = v_y \tag{14.6}$$

$$\dot{v}_x = 2v_y + x - \frac{m_1\left(x+m_2\right)}{r_1^3} - \frac{m_2\left(x-m_1\right)}{r_2^3} \tag{14.7}$$

$$\dot{v}_y = -2v_x + y - \frac{m_1 y}{r_1^3} - \frac{m_2 y}{r_2^3} \tag{14.8}$$

where

$$r_1 = \sqrt{\left(x+m_2\right)^2 + y^2} \tag{14.9}$$

$$r_2 = \sqrt{\left(x-m_1\right)^2 + y^2} \tag{14.10}$$

$$m_1 + m_2 = 1 \tag{14.11}$$

Orbit	Numerical values	
1	x_0	0.994
	y_0	0
	v_{x0}	0
	v_{y0}	−2.00158510637908252240537862224
	T	17.06521656015796255889172 09
	m_2	0.012277471
2	x_0	0.994
	y_0	0
	v_{x0}	0
	v_{y0}	−2.031732629557336835730205 7924
	T	11.124340337266085134999734047
	m_2	0.012277471
3	x_0	1.2
	y_0	0
	v_{x0}	0
	v_{y0}	−1.04935750983031990726
	T	6.192169333131963970674
	m_2	$(82.45)^{-1}$

Table 14.1: Initial conditions and periode T for three periodic orbit of the resticted three-body problem.

Here x and y are the position coordinates of the satellite, v_x is the velocity in the x direction and v_y is the velocity in the y direction. The mass of the earth is m_1 and the mass of the moon is m_2. The acceleration terms are due to the gravitational field, and Coriolis and centrifugal effects due to the rotation of the earth-moon system. The energy function of the system is given by

$$h = \frac{1}{2}\left(v_x^2 + v_y^2 - x^2 - y^2\right) - \frac{m_1}{r_1} - \frac{m_2}{r_2} \tag{14.12}$$

and the conservation of energy implies that h is a constant during the motion of the system. This can be used to check the accuracy of computed solutions.

We would like to compute the solution of the differential equation using a numerical scheme. Several periodic orbits have been found for this system that can be used to check the accuracy of numerical integrators (Hairer and Wanner 1996), (Shampine et al. 1997). It turns out that the solution is very sensitive close to the moon at $(x, y) = (1, 0)$, and close to the earth at $(x, y) = (0, 0)$. As a consequence of this, a standard fixed step integrator will be useless for the integration of this system. The widely used Euler's method give large errors even with 24000 time steps per orbit, and even the fourth order Runge-Kutta method RK4 gives significant errors with 6000 time steps. The parameters describing three periodic orbits are given in Table 14.1.

The solutions were computed with the ode45 function in MATLAB with a relative tolerance of 10^{-6}. The computation of the first orbit took 697 steps for Orbit 1, 621 steps for Orbit 2, and 601 steps for orbit 3. The results for the computation of two orbits are shown in Figures 14.4–14.6. Note that the integration is sufficiently accurate for the two orbits to coincide.The code for generating the plots is the MATLAB script

```
tf1=17.065216560157962558891720 6249; %Periode
```

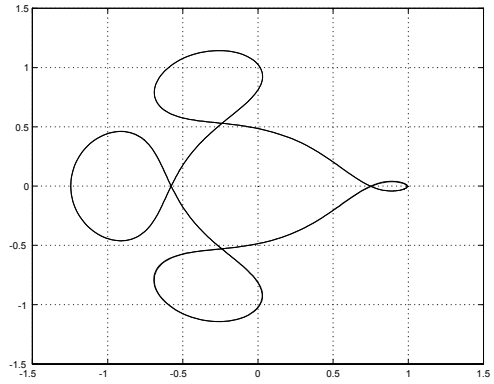

Figure 14.4: Periodic orbit 1 of the restricted three-body problem

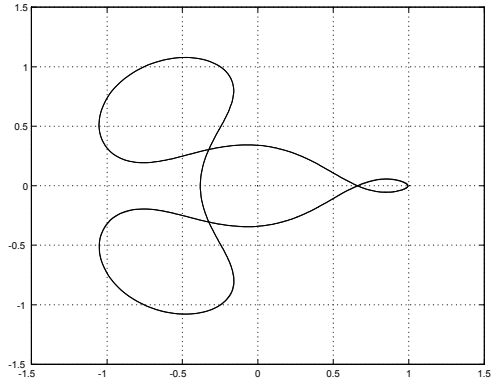

Figure 14.5: Periodic orbit 2 of the restricted three-body problem

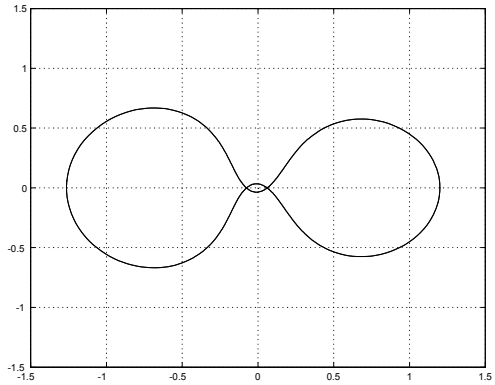

Figure 14.6: Periodic orbit 3 of the restricted three-body problem

```
tf2=11.124340337266085134999734047;
tf3=6.192169333131963970674;
x0=0.994; x03=1.2; y0=0; vx0=0.0; %Initial conditions
vy01=-2.001585106379082522405378622 24;
vy02=-2.031732629557336835730205792 4;
vy03=-1.04935750983031990726;
N=1; %Number of orbits
options = odeset('RelTol',1e-6);
[t,y] = ode45('OrbitODEEq',[0 N*tf1],[x0 0 0 vy01],options);
plot(y(:,1),y(:,2),0,0,'.',1,0,'.');
axis([-1.5 1.5 -1.5 1.5]); grid; size(t) %number of steps
options = odeset('RelTol',1e-6);
[t,y] = ode45('OrbitODEEq',[0 N*tf2],[x0 0 0 vy02],options);
figure; plot(y(:,1),y(:,2),0,0,'.',1,0,'.');
axis([-1.5 1.5 -1.5 1.5]);grid; size(t) %number of steps
options = odeset('RelTol',1e-6);
[t,y] = ode45('OrbitODEEq2',[0 N*tf3],[x03; 0; 0; vy03],options);
figure; plot(y(:,1),y(:,2),0,0,'.',1,0,'.');
axis([-1.5 1.5 -1.5 1.5]); grid; size(t) %number of steps
```

and the function

```
function dydt = OrbitODEEq(t,y)

m2 = 0.012277471;
m1 = 1 - m2;
r13 = (((y(1) + m2)^2 + y(2)^2) ^1.5);
r23 = (((y(1) - m1)^2 + y(2)^2) ^1.5);
dydt = [ y(3)
         y(4)
         (2*y(4) + y(1) - m1*((y(1)+m2)/r13)...
            - m2*((y(1)-m1)/r23))
         (-2*y(3) + y(2) - m1*(y(2)/r13)...
            - m2*(y(2)/r23)) ];
```

The function OrbitODEEq2 is identical to OrbitODEEq except for the numerical value of m_2.

14.1.4 Mass balance of chemical reactor

A chemical reaction

$$\begin{array}{rcll} A & \rightarrow & B & \text{(slow)} \\ B+B & \rightarrow & B+C & \text{(very fast)} \\ B+C & \rightarrow & A+C & \text{(fast)} \end{array}$$

in a closed tank has the mass balance equations

$$\begin{array}{rrrl} \dot{y}_1 = & -0.04y_1 + 10^4 y_2 y_3 & & y_1(0) = 1 \\ \dot{y}_2 = & 0.04y_1 - 10^4 y_2 y_3 & -3 \cdot 10^7 y_2^2 & y_2(0) = 0 \\ \dot{y}_3 = & & 3 \cdot 10^7 y_2^2 & y_3(0) = 0 \end{array} \qquad (14.13)$$

The solution of these equations can be computed numerically. This is a difficult system, however, to integrate, as it has both very fast and slow dynamics. Because of this, the process has been used as a benchmark for testing the performance of numerical integrators.

14.2 Preliminaries

14.2.1 Notation

We will investigate the problem of computing a numerical solution to the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y},t), \quad \mathbf{y}(t_0) = \mathbf{y}_0 \tag{14.14}$$

The system has the exact solution $\mathbf{y}(t)$, and we would like to compute a numeric solution which approximates the exact solution with satisfactory accuracy. This will be done with a time step h so that the solution is computed for $(t_0, t_1, \ldots, t_n, \ldots, t_N)$ where $t_{n+1} - t_n = h$. The numerical solution at time t_n is denoted $\mathbf{y}_n$, while the exact solution at time t_n is denoted $\mathbf{y}(t_n)$.

14.2.2 Computation error

To analyze the accuracy of a computed solution it is useful to have a measure of how much the error increases in one time-step. To do this we introduce the concept of a *local solution* $\mathbf{y}_L(t_n;t)$, which is the exact solution of (14.14) with initial condition $\mathbf{y}_n$ at t_n, that is,

$$\dot{\mathbf{y}}_L(t_n;t) = \mathbf{f}[\mathbf{y}_L(t_n;t)], \quad \mathbf{y}_L(t_n;t_n) = \mathbf{y}_n \tag{14.15}$$

In particular we will be concerned with the local solution at the next time-step, which is $\mathbf{y}_L(t_n;t_{n+1})$. The deviation of the computed solution $\mathbf{y}_{n+1}$ from the local solution $\mathbf{y}_L(t_n;t_{n+1})$ will then be the error introduced by the numerical scheme from time t_n to time t_{n+1}.

The local error $\mathbf{e}_{n+1}$ is the is the difference of the computed solution $\mathbf{y}_{n+1}$ from the local solution $\mathbf{y}_L(t_n;t_{n+1})$ at time t_{n+1}:

$$\mathbf{e}_{n+1} = \mathbf{y}_{n+1} - \mathbf{y}_L(t_n;t_{n+1}) \tag{14.16}$$

The global error $\mathbf{E}_{n+1}$ is the error in the computed solution $\mathbf{y}_{n+1}$ relative to the exact solution $\mathbf{y}(t_{n+1})$ at time t_{n+1}:

$$\mathbf{E}_{n+1} = \mathbf{y}_{n+1} - \mathbf{y}(t_{n+1}) \tag{14.17}$$

The local error $\mathbf{e}_{n+1}$ is the error in the solution resulting from the computation from time t_n to t_{n+1}.The global error $\mathbf{E}_{n+1}$ is the error in the solution resulting from the computation from initial time t_0 to t_{n+1}.

14.2.3 The order of a one-step method

A one-step method is a numerical scheme which computes $\mathbf{y}_{n+1}$ as a function of $\mathbf{y}_n$ according to

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\boldsymbol{\phi}(\mathbf{y}_n, t_n) \tag{14.18}$$

where $\boldsymbol{\phi}(\cdot)$ is given by the particular numerical method that is used. We would like our method to give a small error in some sense when the time step is small. One way of characterizing different methods is the concept of the order of the method. We say that the method is of order p if p is the smallest integer so that

$$\mathbf{e}_{n+1} = O(h^{p+1}) \tag{14.19}$$

Here we have used the order notation $O(\cdot)$ (Lin and Segel 1974). The function $\phi(x)$ satisfies

$$\phi(x) = O[\psi(x)] \tag{14.20}$$

if there exists a constant $C > 0$ so that

$$|\phi(x)| \leq C\,|\psi(x)| \tag{14.21}$$

when x is close to zero.

Example 210 *The expression $\phi(x) = O(x^m)$ implies that there exists a $C > 0$ so that $|\phi(x)| \leq C\,|x^m|$. Moreover, if $C > 0$, then $Ch^m = O(h^m)$, which is implied by $|Ch^m| \leq C\,|h^m|$.*

To investigate the order of a method it is useful to develop the Taylor series expansion of $\mathbf{y}_L(t_n; t_{n+1})$ around $\mathbf{y}_n$. The Taylor series is given by

$$\mathbf{y}_L(t_n; t_{n+1}) = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \ldots + \frac{h^p}{p!}\frac{d^{p-1}\mathbf{f}(\mathbf{y}_n, t_n)}{dt^{p-1}} + \frac{h^{p+1}}{(p+1)!}\frac{d^p\mathbf{f}\left[\mathbf{y}_L(\tau), \tau\right]}{dt^p} \tag{14.22}$$

where $t_n \leq \tau \leq t_{n+1}$. As the local error is $\mathbf{e}_{n+1} = \mathbf{y}_{n+1} - \mathbf{y}_L(t_n; t_{n+1})$, and we arrive at the following result

A one-step method is of order p if p is the smallest integer so that $\mathbf{e}_{n+1} = O(h^{p+1})$. If the numerical solution $\mathbf{y}_{n+1}$ satisfies the equation

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t) + \ldots + \frac{h^p}{p!}\frac{d^{p-1}\mathbf{f}(\mathbf{y}_n, t_n)}{dt^{p-1}} + O(h^{p+1}) \tag{14.23}$$

then $\mathbf{e}_{n+1} = O(h^{p+1})$, and it follows that the method is of order p.

Analysis of the global error is somewhat more complicated. However, we state without further analysis that for one-step methods the global error is $\mathbf{E}_{n+1} = O(h^p)$.

14.2.4 Linearization

The stability and performance of a one-step method for the numerical integration of (14.14) can be investigated in terms of the linearized system equations, and in this section we will see how this can be done. The basic idea is to apply a one-step method to the linearized system. From basic systems theory it is known that the dynamics of a linearized system is to a large extent determined by the location of the eigenvalues of the Jacobian matrix. In the same way, the performance of a one-step method applied to a linear system can be described by the eigenvalues of the Jacobian in terms of the *stability function* of the method. We will first establish the necessary mathematical background for this.

Suppose that $\mathbf{y}^*(t)$ is a solution of the differential equation

$$\dot{\mathbf{y}}^* = \mathbf{f}(\mathbf{y}^*,t), \quad \mathbf{y}^*(t_0) = \mathbf{y}_0^* \tag{14.24}$$

Linearization of the differential equation around the solution $\mathbf{y}^*(t)$ is based on writing $\mathbf{y} = \mathbf{y}^* + \Delta\mathbf{y}$ and using the Taylor series

$$\dot{\mathbf{y}}^* + \Delta\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}^*,t) + \left.\frac{\partial \mathbf{f}(\mathbf{y},t)}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{y}^*} \Delta\mathbf{y} \tag{14.25}$$

We define the *Jacobian* $\mathbf{J}$ of the system to be

$$\mathbf{J} = \left.\frac{\partial \mathbf{f}(\mathbf{y},t)}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{y}^*} = \left\{ \left.\frac{\partial f_i(\mathbf{y},t)}{\partial y_j}\right|_{\mathbf{y}=\mathbf{y}^*} \right\} \tag{14.26}$$

and obtain the *linearization* of (14.14) which is

$$\Delta\dot{\mathbf{y}} = \mathbf{J}\Delta\mathbf{y} \tag{14.27}$$

The solution $\Delta\mathbf{y}(t)$ of (14.27) can be expressed as a linear combination

$$\Delta\mathbf{y} = \sum_{i=1}^{d} q_i(t)\mathbf{m}_i \tag{14.28}$$

of solutions $q_i(t)$ of the scalar differential equations

$$\dot{q}_i = \lambda_i q_i, \quad i = 1,\ldots,d \tag{14.29}$$

where $\mathbf{q} = (q_1,\ldots,q_n)^T$, λ_i are the eigenvalues of $\mathbf{J}$, and $\mathbf{m}_i$ are the eigenvectors of $\mathbf{J}$. This means that we can study the dynamics of the linearized system (14.27) by finding the eigenvalues of $\mathbf{J}$. In particular, if we apply a one-step method to (14.27), then the solution $\Delta\mathbf{y}_{n+1}$ will be the same as if we apply the method to (14.29) and compute

$$\Delta\mathbf{y}_{n+1} = \sum_{i=1}^{d} (q_i)_{n+1}\mathbf{m}_i \tag{14.30}$$

Suppose that there is a function $R(s)$, which will be called the *stability function,* so that the one-step method gives the numerical solution

$$(q_i)_{n+1} = R(h\lambda_i)(q_i)_n \tag{14.31}$$

when applied to (14.29). Then

$$\Delta\mathbf{y}_n = \sum_{i=1}^{d} R^n(h\lambda_i)(q_i)_0\mathbf{m}_i \tag{14.32}$$

and the following conclusion may be drawn:

The numerical solution $\Delta\mathbf{y}_n$ of the linearized system (14.27) is stable if the magnitude of the stability function is less than or equal to unity for all the eigenvalues, that is, if

$$|R(h\lambda_i)| \le 1, \quad i = 1,\ldots,d \tag{14.33}$$

where h is the step-length and λ_i is an eigenvalue of $\mathbf{J}$.

Example 211 *Consider the system*

$$\dot{y}_1 = y_2 \tag{14.34}$$
$$\dot{y}_2 = -\gamma y_1^3 - c y_2 \tag{14.35}$$

The linearization around $y_1 = 0$, $y_2 = 0$ *is*

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -c \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{14.36}$$

which has eigenvalues $\lambda_1 = 0$ *and* $\lambda_2 = -c$. *The linearization around a solution* $y_1^*(t)$, $y_2^*(t)$ *is*

$$\begin{pmatrix} \Delta\dot{y}_1 \\ \Delta\dot{y}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -3\gamma\left(y_1^*\right)^2 & -c \end{pmatrix} \begin{pmatrix} \Delta y_1 \\ \Delta y_2 \end{pmatrix} \tag{14.37}$$

The eigenvalues are given by

$$\lambda^2 + c\lambda + 3\gamma\left(y_1^*\right)^2 = 0 \tag{14.38}$$

which gives

$$\lambda = -\frac{c}{2} \pm \sqrt{\left(\frac{c}{2}\right)^2 - 3\gamma\left(y_1^*\right)^2} \tag{14.39}$$

which implies that $\mathrm{Re}[\lambda] \leq 0$. *We see that for large* $|y_1^*|$, *that is when* $3\gamma\left(y_1^*\right)^2 \gg \left(\frac{c}{2}\right)^2$, *then the system becomes oscillatory, while for small* $|y_1^*|$ *the system is overdamped.*

14.2.5 The linear test function

Important insight on the properties of a numerical integration scheme is gained by analyzing the performance of the method for the linearization of the system. From the previous section it is clear that the performance of a numerical integrator for linear systems can be investigated by applying the method to the *linear test system*

$$\dot{y} = \lambda y \tag{14.40}$$

The numerical solution for this system is

$$y_{n+1} = R(h\lambda) y_n \tag{14.41}$$

where $R(h\lambda)$ is the stability function for the method. Stability of the numerical scheme is ensured if the difference equation satisfies

$$|y_{n+1}| \leq |y_n| \tag{14.42}$$

and we see that this is ensured if

$$|R(h\lambda)| \leq 1 \tag{14.43}$$

This gives conditions on the time-step h and the location of the eigenvalue λ for the numerical solution to be stable.

14.3 Euler methods

14.3.1 Euler's method

A simple but important numerical integration scheme is *Euler's method*, where the numerical solution is computed from

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) \tag{14.44}$$

Comparison with (14.23) shows that the method is of order 1.

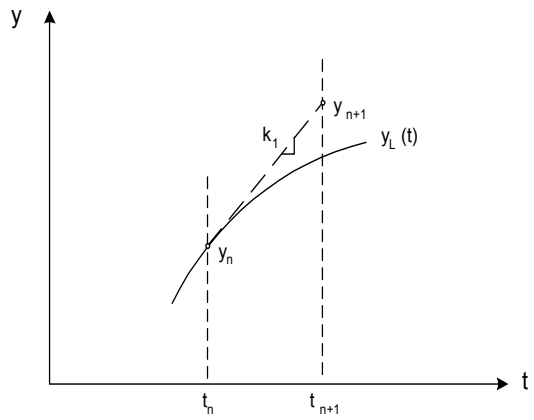

Figure 14.7: Euler's method

The linear stability of Euler's method can be investigated with the scalar test system

$$\dot{y} = \lambda y \tag{14.45}$$

Euler's method gives

$$y_{n+1} = y_n + h\lambda y_n = (1 + h\lambda)y_n \tag{14.46}$$

which shows that the stability function is

$$R(h\lambda) = 1 + h\lambda \tag{14.47}$$

Stability is ensured whenever

$$|R(h\lambda)| = |1 + h\lambda| \leq 1 \tag{14.48}$$

This is the case if $h\lambda$ is inside the circle of radius one around -1. For real eigenvalues λ stability is ensured when

$$-\frac{2}{h} \leq \lambda \leq 0 \tag{14.49}$$

or, equivalently,

$$h \leq -\frac{2}{\lambda} \tag{14.50}$$

The region of stability is shown in Figure 14.14.

Example 212 *The system*

$$\dot{y} = -y, \quad y(0) = 1 \tag{14.51}$$

was integrated for $0 \leq t \leq 8$ *with Euler's method. The stability limit for the time step is* $h = 2$*, as* $\lambda = -1$ *for this system. First a solution was calculated with* $h = 0.5$*, then with* $h = 1.5$*, then with the stability limit* $h = 2.0$*, and finally with the unstable value* $h = 2.2$*. The results are shown in Figure 14.8. It is clear from the results that the time step should be less than* $h = 0.5$ *to achieve a reasonably accurate solution.*

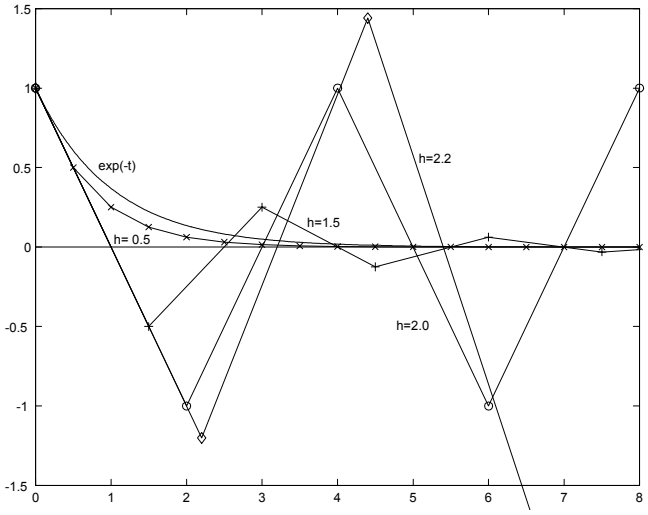

Figure 14.8: Calculated solutions for the system $\dot{y} = -y$ with Euler's method for four different time steps h. The exact solution $\exp(-t)$ is shown for comparison.

Example 213 *The first order system*

$$\dot{y} = -\beta y^3 \tag{14.52}$$

in Euler's method gives the algorithm

$$y_{n+1} = y_n - h\beta y_n^3 \tag{14.53}$$

The linearization of the differential equation around zero gives

$$\dot{y} = 0 \tag{14.54}$$

that is, the test system $\dot{y} = \lambda y$ *with* $\lambda = 0$*. This is stable for all time steps* h*.*

Example 214 *The system*

$$\dot{y} = -\alpha y - \beta y^3 \tag{14.55}$$

in Euler's method gives

$$y_{n+1} = y_n - h\left(\alpha y_n + \beta y_n^3\right) \tag{14.56}$$

The linearization of the differential equation around zero gives

$$\dot{y} = -\alpha y \tag{14.57}$$

while the linearization around y^* gives

$$\Delta\dot{y} = -\left[\alpha + 3\beta\left(y^*\right)^2\right]\Delta y \tag{14.58}$$

A large $|y^*|$ requires a small h for the stability condition to hold. Here, the eigenvalue is $\lambda = -\left[\alpha + 3\beta\left(y^*\right)^2\right]$, and the stability condition for the linearized system is

$$h \leq \frac{2}{\alpha + 3\beta\left(y^*\right)^2} \tag{14.59}$$

Example 215 *Consider the second order system*

$$\ddot{x} = F(x, \dot{x}) \tag{14.60}$$

To apply Euler's method, the system must first be brought into the form (14.14). This can be done by defining $y_1 = x$ and $y_2 = \dot{x}$. This gives

$$\dot{y}_1 = y_2 \tag{14.61}$$
$$\dot{y}_2 = F(y_1, y_2) \tag{14.62}$$

and Euler's method gives the integration algorithm

$$y_{1,n+1} = y_{1,n} + hy_{2,n} \tag{14.63}$$
$$y_{2,n+1} = y_{2,n} + hF(y_1, y_2) \tag{14.64}$$

14.3.2 The improved Euler method

The *improved Euler method* includes an evaluation $\widehat{\mathbf{y}}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n)$ according to Euler's method. Then an approximation of $\mathbf{f}\left(\widehat{\mathbf{y}}_{n+1}, t_{n+1}\right)$ at the time t_{n+1} is computed using $\widehat{\mathbf{y}}_{n+1}$. This value is used to improve the accuracy of the numerical solution $\mathbf{y}_{n+1}$. The method is given by

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \tag{14.65}$$
$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n + h\mathbf{k}_1, t_n + h) \tag{14.66}$$
$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}\left(\mathbf{k}_1 + \mathbf{k}_2\right) \tag{14.67}$$

To find the order of this method a Taylor series expansion around $(\mathbf{y}_n, t_n)$ is used. The Taylor series of $\mathbf{k}_2$ is

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n, t_n) + h\frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^2}{2}\frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^3) \tag{14.68}$$

This gives the Taylor series

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \frac{h^2}{2}\frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^3}{4}\frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^4) \tag{14.69}$$

The first two terms coincide with the Taylor series expansion of the local solution $\mathbf{y}_L(t_n; t_{n+1})$, and the remaining terms are $O(h^3)$. Comparison with (14.23) leads to the conclusion that the improved Euler's method is of order $p = 2$.

To investigate stability of the method, we apply the method to the test equation

$$\dot{y} = \lambda y$$

This results in

$$\begin{aligned} k_1 &= \lambda y_n \\ k_2 &= \lambda (1 + h\lambda) y_n \\ y_{n+1} &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_n \end{aligned}$$

which is stable whenever

$$\left|1 + h\lambda + \frac{(h\lambda)^2}{2}\right| \leq 1$$

On the real axis this corresponds to $-2/h \leq \lambda \leq 0$. The region of stability is shown in Figure 14.14

Example 216 *The first order system*

$$\dot{y} = -\alpha y^3 \tag{14.70}$$

in the improved Euler method gives the algorithm

$$k_1 = -\alpha y_n^3 \tag{14.71}$$

$$k_2 = -\alpha (y_n + h k_1)^3 \tag{14.72}$$

$$y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2) \tag{14.73}$$

Example 217 *Consider the second order system*

$$\ddot{x} + c\dot{x} + \gamma(x)x = 0 \tag{14.74}$$

which is set in standard form by using $y_1 = x$ and $y_2 = \dot{x}$. Then the differential equation is written

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\gamma(y_1)y_1 - cy_2 \end{pmatrix} \tag{14.75}$$

The improved Euler's method gives the integration algorithm

$$\begin{pmatrix} k_{1,1} \\ k_{1,2} \end{pmatrix} = \begin{pmatrix} y_{2,n} \\ -\gamma(y_{1,n})y_{1,n} - cy_{2,n} \end{pmatrix} \tag{14.76}$$

$$\begin{pmatrix} k_{2,1} \\ k_{2,2} \end{pmatrix} = \begin{pmatrix} y_{2,n} + hk_{1,2} \\ -\gamma(y_{1,n} + hk_{1,1})(y_{1,n} + hk_{1,1}) - c\,(y_{2,n} + hk_{1,2}) \end{pmatrix} \tag{14.77}$$

$$y_{1,n+1} = y_{1,n} + \frac{h}{2}(k_{1,1} + k_{2,1}) \tag{14.78}$$

$$y_{2,n+1} = y_{2,n} + \frac{h}{2}(k_{1,2} + k_{2,2}) \tag{14.79}$$

14.3.3 The modified Euler method

The *modified Euler method*, also called the *explicit midpoint rule*, is derived in a similar way as the improved Euler method. In the modified Euler method an approximation of $\mathbf{f}$ at $\left(\mathbf{y}(t+\frac{h}{2}), t+\frac{h}{2}\right)$ is used to find the solution. This approximation is computed using Euler's method to find an estimate of $\mathbf{y}(t+\frac{h}{2})$. The method is illustrated in Figure 14.9 and is given by

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \tag{14.80}$$

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n + \frac{h}{2}\mathbf{k}_1, t_n + \frac{h}{2}) \tag{14.81}$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{k}_2 \tag{14.82}$$

A Taylor series expansion of $\mathbf{k}_2$ gives

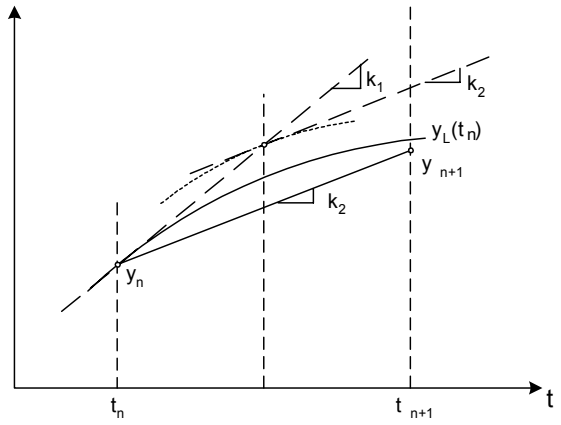

Figure 14.9: The modified Euler method

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n, t_n) + \frac{h}{2}\frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{\left(\frac{h}{2}\right)^2}{2}\frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^3) \tag{14.83}$$

which gives

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n, t_n) + \frac{h^2}{2}\frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + \frac{h^3}{8}\frac{d^2\mathbf{f}}{dt^2}(\mathbf{y}_n, t_n) + O(h^4) \tag{14.84}$$

The method is seen to be of order $p = 2$.

Application of the method to the test systems $\dot{y} = \lambda y$ gives

$$k_2 = \lambda\left(1 + \frac{h}{2}\lambda\right) y_n$$

$$y_{n+1} = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y_n$$

which leads to the same stability conditions as for the improved Euler method.

14.4 Explicit Runge-Kutta methods

14.4.1 Introduction

It was demonstrated above that Euler's method, which is of order $p = 1$, can be modified to a method of order $p = 2$ by computing $\mathbf{y}_{n+1}$ as a linear combination of $\mathbf{f}(\mathbf{y}_n, t_n)$ and an approximation of $\mathbf{f}\left[\mathbf{y}\left(t_n + ch\right), t_n + ch\right]$ where $0 < c \leq 1$. This result can be extended to higher order methods by computing more approximations of $\mathbf{f}$ over the interval, and then compute $\mathbf{y}_{n+1}$ as a linear combination of these approximations. This is done in the explicit Runge-Kutta methods. A Runge-Kutta method is said to have σ stages if σ approximations, or stages, of the function derivative $\mathbf{f}$ is used.

14.4.2 Numerical scheme

An explicit Runge-Kutta method with σ *stages* for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{14.85}$$

is given by

$$\begin{aligned} \mathbf{k}_i &= \mathbf{f}(\mathbf{y}_n + h\sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j, t_n + c_i h),\ i = 1, \ldots, \sigma \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{j=1}^{\sigma} b_j \mathbf{k}_j \end{aligned}$$

The explicit Runge-Kutta method can be written out as

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n, t_n) && (14.86)\\ \mathbf{k}_2 &= \mathbf{f}(\mathbf{y}_n + ha_{21}\mathbf{k}_1, t_n + c_2 h) && (14.87)\\ \mathbf{k}_3 &= \mathbf{f}\left(\mathbf{y}_n + h\left(a_{31}\mathbf{k}_1 + a_{32}\mathbf{k}_2\right), t_n + c_3 h\right) && (14.88)\\ &\vdots && (14.89)\\ \mathbf{k}_\sigma &= \mathbf{f}\left(\mathbf{y}_n + h\left(a_{\sigma 1}\mathbf{k}_1 + \ldots + a_{\sigma,\sigma-1}\mathbf{k}_{\sigma-1}\right), t_n + c_\sigma h\right) && (14.90)\\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h(b_1\mathbf{k}_1 + \ldots + b_\sigma \mathbf{k}_\sigma) && (14.91) \end{aligned}$$

The equations for $\mathbf{k}_1, \ldots, \mathbf{k}_\sigma$ are called the stage computations. The interpolation parameters $c_i,\ i \in \{2, \ldots, \sigma\}$ are in the range $0 \leq c_i \leq 1$ and form an increasing sequence, that is, $0 \leq c_1 \leq \ldots \leq c_\sigma \leq 1$. The weighting parameters at stage i are denoted $a_{ij},\ i \in \{2, \ldots, \sigma\},\ j \in \{1, \ldots, i-1\}$, and satisfy the normalization condition

$$\sum_{j=1}^{i-1} a_{ij} = c_i \leq 1 \tag{14.92}$$

The weighting parameters b_i of the solution $\mathbf{y}_{n+1}$ are required to satisfy the normalization condition $\sum_{i=1}^{\sigma} b_i = 1$. Each explicit Runge-Kutta method is described by its parameters

a_{ij}, b_i and c_i, which can be arranged in a *Butcher array* of the form

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_\sigma & a_{\sigma 1} & a_{\sigma 2} & \dots & a_{\sigma,\sigma-1} & \\ \hline & b_1 & b_2 & \dots & b_{\sigma-1} & b_\sigma \end{array} \tag{14.93}$$

Alternatively, the parameters can be expressed by the matrix $\mathbf{A}$ and the vectors $\mathbf{b}$ and $\mathbf{c}$ defined by

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & \dots & 0 & 0 \\ a_{31} & a_{32} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{\sigma 1} & a_{\sigma 2} & \dots & a_{\sigma,\sigma-1} & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_\sigma \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0 \\ c_2 \\ c_3 \\ \vdots \\ c_\sigma \end{pmatrix}$$

The Butcher array is then written

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array}$$

We note that the matrix $\mathbf{A}$ is singular for explicit Runge-Kutta methods.

14.4.3 Order conditions

The parameters of an explicit Runge-Kutta method of σ stages must satisfy certain conditions to be of order p. Here, a derivation of the conditions for a method with $\sigma = 2$ stages to be of order $p = 2$ will be done. The Butcher array is

$$\begin{array}{c|cc} 0 & & \\ c_2 & a_{21} & \\ \hline & b_1 & b_2 \end{array}$$

A Taylor series expansion of $\mathbf{k}_2$ gives

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_n, t_n) + a_{21} h \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + O(h^2)$$

The condition $a_{21} = c_2$ from (14.92) then gives

$$\mathbf{y}_{n+1} = \mathbf{y}_n + (b_1 + b_2)\, h\mathbf{f}(\mathbf{y}_n, t_n) + b_2 c_2 h^2 \frac{d\mathbf{f}}{dt}(\mathbf{y}_n, t_n) + O(h^3) \tag{14.94}$$

and it is seen that the right hand side is equal to the Taylor series expansion of $\mathbf{y}_{n+1}$ for terms up to h^2 if the parameters satisfy

$$b_1 + b_2 = 1 \tag{14.95}$$

$$b_2 c_2 = \frac{1}{2} \tag{14.96}$$

Example 218 *The improved Euler method has $b_1 = b_2 = \frac{1}{2}$ and $c_2 = 1$, which satisfies the conditions in (14.95) and (14.96). The modified Euler method has $b_1 = 0$, $b_2 = 1$ and $c_2 = \frac{1}{2}$, which also agrees with the conditions (14.95) and (14.96). The is in agreement with the result that both of these methods have $\sigma = 2$ stages, and are of order $p = 2$.*

In the same way 4 conditions can be found for $\sigma = p = 3$, while 8 conditions can be found for $\sigma = p = 4$.

For higher order methods there are certain lower bounds for how many stages that are needed (Hairer et al. 1993). For order $5 \leq p \leq 6$, an explicit Runge-Kutta method must have $\sigma \geq p+1$ stages. For order $p = 7$, an explicit Runge-Kutta method must have $\sigma \geq p + 2$ stages, while to achieve order $p \geq 8$, a method with at least $\sigma \geq p + 3$ stages.

14.4.4 Some explicit Runge-Kutta methods

The following explicit Runge-Kutta methods are of order $p = \sigma$. Euler's method, which is of order 1, has the Butcher array

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}$$

The improved Euler method is an explicit Runge-Kutta method with array

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

The modified Euler method has the array

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

Heun's method has the following array

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}$$

The region of stability is shown in Figure 14.14.

The famous fourth order Runge-Kutta method RK4 is of order 4 and has the array

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

The region of stability is shown in Figure 14.14.

14.4.5 Case study: Pneumatic spring

Consider the pneumatic spring system in Figure 14.10. The cylinder has cross section $A = 0.01$ m^2 and a vertical center axis pointing upwards with coordinate x. The cylinder

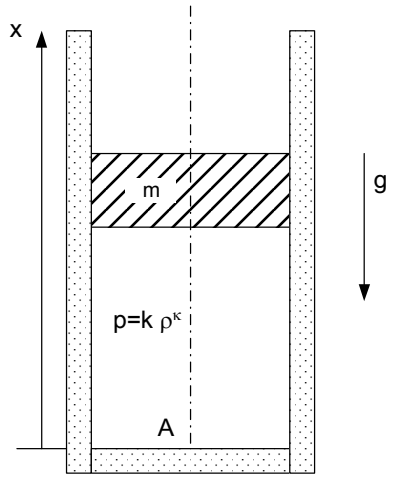

Figure 14.10: Pneumatic spring with gravity acting on the position.

is filled with air and has a piston of mass $m = 200$ kg that compresses the air. The density of the air inside the cylinder is

$$\rho = \frac{m_a}{V_a} = \frac{m_a}{Ax} \tag{14.97}$$

where m_a is the mass and $V_a = Ax$ is the volume of the air. The air is assumed to be isentropic which implies that the pressure inside the cylinder is

$$p = p_0 \left(\frac{\rho}{\rho_0}\right)^{\kappa} = p_0 \left(\frac{x_0}{x}\right)^{\kappa} \tag{14.98}$$

where $\kappa = 1.4$ and $p_0 = 2 \cdot 10^5$ N/m^2 is the pressure corresponding to a piston position $x_0 = 1$ m, and the density $\rho_0 := m_a/(Ax_0)$. The total force acting on the piston is gravity and pressure forces:

$$F = -mg + Ap = -mg + Ap_0 \left(\frac{x_0}{x}\right)^{\kappa} \tag{14.99}$$

where $g = 10$ m/s^2. Inserting the numerical values we see that $mg = Ap_0$, which implies that when $x = x_0$ the force is $F = 0$ and the system is at an equilibrium at $x = x_0$. The equation of motion can then be written

$$\ddot{x} + g\left[1 - x^{-\kappa}\right] = 0 \tag{14.100}$$

The standard form is $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is obtained by setting

$$\mathbf{y} = \begin{pmatrix} x \\ v \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} v \\ -g\left[1 - x^{-\kappa}\right] \end{pmatrix} \tag{14.101}$$

where $v = \dot{x}$ is the velocity of the piston. We see that the system has an equilibrium at $x = 1$, $v = 0$, where $\ddot{x} = 0$.

Linearization around $x^* = 1$ gives

$$\Delta\dot{\mathbf{y}} = \mathbf{J}\Delta\mathbf{y} \tag{14.102}$$

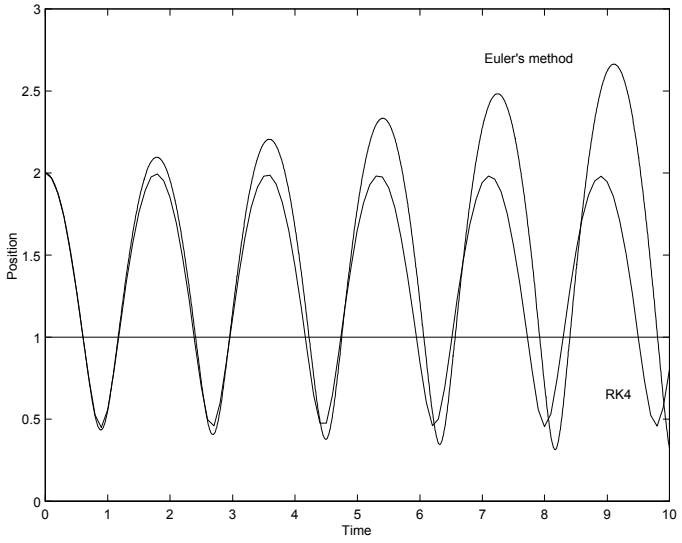

Figure 14.11: Position of piston integrated with Euler's method with $h = 0.005$ s and with RK4 with $h = 0.1$ s.

where

$$\Delta\mathbf{y} = \begin{pmatrix} x-1 \\ v \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 & 1 \\ -g\kappa\left(x^*\right)^{-(\kappa+1)} & 0 \end{pmatrix} \tag{14.103}$$

The eigenvalues of the linearization are found to be

$$\lambda_{1,2} = \pm j\omega_0, \quad \omega_0 = \sqrt{g\kappa\left(x^*\right)^{-(\kappa+1)}} \tag{14.104}$$

Numerical values are

x^*	0.5	1	2
ω_0	4.3	3.7	3.3

(14.105)

The total energy E is the sum of the internal energy $U = pV/(\kappa - 1)$, the gravity potential mgx, and the kinetic energy $\frac{1}{2}mv^2$:

$$E = \frac{1}{\kappa - 1}p_0 A x^{-(\kappa-1)} + mgx + \frac{1}{2}mv^2 \tag{14.106}$$

The total energy has its minimum value at the equilibrium state were the energy is

$$E_{\min} = \frac{1}{\kappa - 1}p_0 A + mg = 7000 \text{ J} \tag{14.107}$$

The system was simulated with Euler's method with time step $h = 0.005$, and with the fourth order RK4 method with time step $h = 0.1$. The result is shown in Figure 14.11.The solution computed with Euler's method was unstable even with the very short time step of 0.005, while the solution with RK4 was stable with a time step that was 20 times larger than for the Euler solution. To check the accuracy of the solutions the total energy was computed for the numerical solutions. For the exact solution the total energy will be constant as there is no energy loss terms in the equation of motion. The solution from

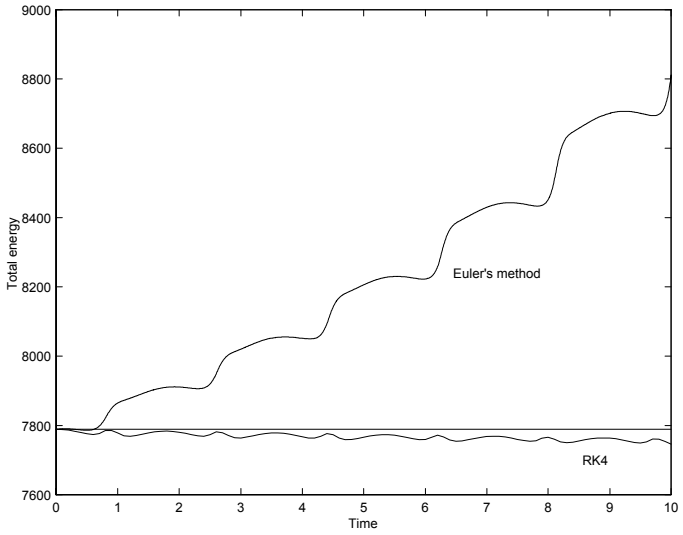

Figure 14.12: Total energy of solutions computed with Euler's method and RK4. It is seen that Euler's method increases the energy in the system, while RK4 gives a slight decrease in energy.

Euler's method gave a steady increase in energy which is not in agreement with the physics of the system is it has no energy source. The RK4 solution gave a slight decrease in energy, which means that the RK4 introduced some damping in the system. The results are shown in Figure 14.12.

The system has eigenvalues $\pm j\omega_0$ on the imaginary axis, and the stability limit for RK4 is then $h = 2.83/\omega_0$, which can be seen from Figure 14.14. As the largest eigenvalue occurs for $\omega_0 = 4.3$ this indicates that the stability limit would be $h_{\min} = 0.65$ s. In simulations it turned out that a slightly smaller value, $h = 0.52$ s was the stability limit for this trajectory. This is demonstrated in Figure 14.13. The difference between the theoretical value and the value found in simulations should be due to the system being nonlinear.

14.4.6 Stability function

A general formula for the stability function of an explicit Runge-Kutta method in terms of $\mathbf{c}$, $\mathbf{A}$ and $\mathbf{b}$ is found as follows: Application of a general explicit Runge-Kutta method to the linear time-invariant test system

$$\dot{y} = \lambda y$$

gives

$$\begin{aligned} k_1 &= \lambda y_n \\ &\vdots \\ k_\sigma &= \lambda \left[y_n + h \left(a_{\sigma 1} k_1 + \ldots + a_{\sigma,\sigma-1} k_{\sigma-1} \right) \right] \\ y_{n+1} &= y_n + h \left(b_1 k_1 + \ldots + b_\sigma k_\sigma \right) \end{aligned}$$

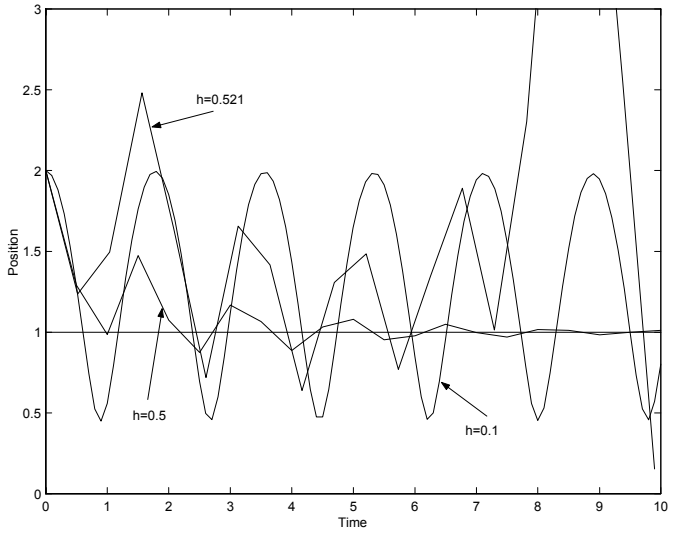

Figure 14.13: Simulation of peumatic spring usning RK4 for three different step lengths.

In vector notation with $\boldsymbol{\kappa} = (k_1, k_2, \ldots, k_\sigma)^T$ and $\mathbf{1} = (1, 1, \ldots, 1)^T$ this can be written

$$\boldsymbol{\kappa} = \lambda \left(\mathbf{1} y_n + h \mathbf{A} \boldsymbol{\kappa} \right) \tag{14.108}$$

$$y_{n+1} = y_n + h \mathbf{b}^T \boldsymbol{\kappa} \tag{14.109}$$

Here $\boldsymbol{\kappa}$ can be solved from (14.108) and inserted into (14.109), which gives

$$R(h\lambda) = 1 + \lambda h \mathbf{b}^T \left(\mathbf{I} - h\lambda \mathbf{A} \right)^{-1} \mathbf{1} \tag{14.110}$$

Alternatively, the system (14.108, 14.109) can be written

$$\begin{pmatrix} \mathbf{I} - h\lambda \mathbf{A} & \mathbf{0} \\ -h\mathbf{b}^T & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\kappa} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} \lambda \mathbf{1} \\ 1 \end{pmatrix} y_n$$

From Cramer's rule it is seen that the stability function can be written

$$R(h\lambda) = \frac{\det \left[\mathbf{I} - \lambda h \left(\mathbf{A} - \mathbf{1}\mathbf{b}^T \right) \right]}{\det \left(\mathbf{I} - \lambda h \mathbf{A} \right)} \tag{14.111}$$

This formula has the advantage that it clearly shows how the numerator and denominator depend on $h\lambda$, $\mathbf{A}$ and $\mathbf{b}$. For an explicit Runge-Kutta method the $\mathbf{A}$ matrix have nonzero elements only below the diagonal, and it follows that $\det \left(\mathbf{I} - \lambda h \mathbf{A} \right) = 1$. Using (14.111) we find:

For an explicit Runge-Kutta method the stability function can be written

$$R_E(h\lambda) = \det \left[\mathbf{I} - \lambda h \left(\mathbf{A} - \mathbf{1}\mathbf{b}^T \right) \right] \tag{14.112}$$

This expression shows that for explicit Runge-Kutta methods

1. $|R_E(h\lambda)|$ will tend to infinity when $|\lambda|$ goes to infinity
2. $R_E(h\lambda)$ is a polynomial in $h\lambda$ of order less than or equal to σ.

Example 219 *Consider the improved Euler method where*

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

Then

$$R_E(h\lambda) = \det \begin{pmatrix} 1+\frac{\lambda h}{2} & \frac{\lambda h}{2} \\ -\frac{\lambda h}{2} & 1+\frac{\lambda h}{2} \end{pmatrix} = 1 + \lambda h + \frac{(\lambda h)^2}{2} \tag{14.113}$$

We see that $R_E(h\lambda)$ is a polynomial in $h\lambda$ of order 2 which is equal to the number of stages.

Next we will comment on explicit Runge-Kutta methods where the number of stages equals the order of the method. The local solution $y_L(t_n; t_{n+1})$ starting from $y_L(t_n; t_n) = y_n$ is given by

$$y_L(t_n; t_{n+1}) = e^{\lambda h} y_n$$

A Taylor series expansion of the local solution is therefore

$$y_L(t_n; t_{n+1}) = \left[1 + h\lambda + \frac{(h\lambda)^2}{2!} + \frac{(h\lambda)^3}{3!} + \ldots\right] y_n \tag{14.114}$$

Therefore, if an explicit Runge-Kutta method of order p is used, then the numerical solution y_{n+1} for a linear test system with will have the Taylor series expansion

$$y_{n+1} = \left[1 + h\lambda + \frac{(h\lambda)^2}{2!} + \ldots + \frac{(h\lambda)^p}{p!} + O(h^{p+1})\right] y_n \tag{14.115}$$

It follows that the stability function for a explicit method of order p satisfies

$$R_E(\lambda h) = 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \ldots + \frac{(h\lambda)^p}{p!} + O(h^{p+1}) \tag{14.116}$$

The stability function of an explicit Runge-Kutta method with σ stages is a polynomial in λh of degree less than or equal to the number of stages σ. If the method is of $\sigma = p \leq 4$ stages the stability function must have exactly p terms, and this is only possible if

$$R_E(\lambda h) = 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \ldots + \frac{(h\lambda)^p}{p!} \quad \text{when } p = \sigma$$

Example 220 *The improved Euler method has stability function*

$$R(\lambda h) = 1 + \lambda h + \frac{(\lambda h)^2}{2}$$

which coincides with the Taylor series expansion with two terms. This agrees with the fact that the method has 2 stages and is of order 2.

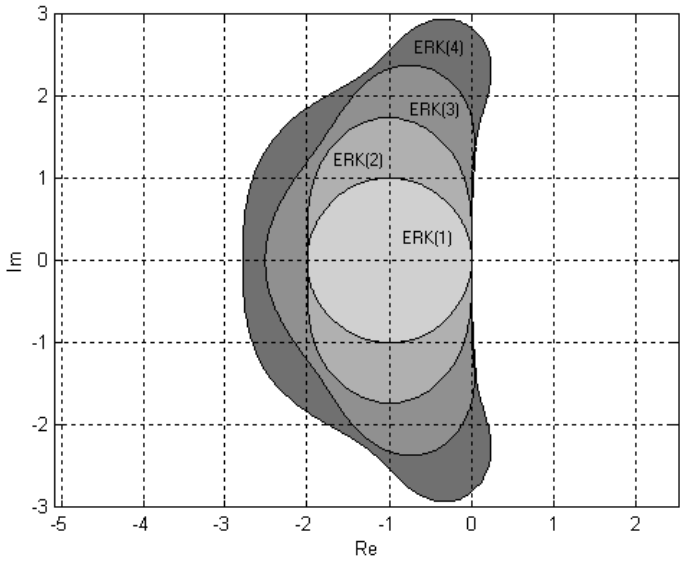

Figure 14.14: Regions of stability in $s = h\lambda$ for the test system $\dot{y} = \lambda y$ for the explicit Runge-Kutta methods. ERK(1): Euler's method, ERK(2): The modified and the improved Euler method, ERK(3): Heun's third order method, and ERK(4): The fourth order Rung-Kutta method RK4.

14.4.7 FSAL methods

We will here take a closer look at explicit Runge-Kutta methods of the FSAL type.

An explicit Runge-Kutta method is said to be an FSAL method if

$$\mathbf{k}_\sigma = \mathbf{f}\left(\mathbf{y}_{n+1}, t_{n+1}\right) \tag{14.117}$$

From the definition we see that in an FSAL method gives some savings in computations as

$$\mathbf{k}_\sigma^n = \mathbf{k}_1^{n+1} \tag{14.118}$$

where $\mathbf{k}_\sigma^n$ denotes the last stage in the calculation of $\mathbf{y}_{n+1}$, and $\mathbf{k}_1^{n+1}$ denotes the first stage in the computation of $\mathbf{y}_{n+2}$. This is the reason for calling such methods *First Same As Last*, which is abbreviated to FSAL. In an FSAL method the weighting vector $\mathbf{b}$ is equal to the last row in the stage matrix $\mathbf{A}$.

14.5 Implicit Runge-Kutta methods

14.5.1 Stiff systems

When a system $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t)$ is integrated with an explicit Runge-Kutta method the time step h cannot be selected so that $h\left|\lambda_{\max}\right|$ is significantly larger than unity, where $\lambda_{\max}$

is the largest eigenvalue of the Jacobian $\mathbf{J} = \partial \mathbf{f}(\mathbf{y},t)/\partial \mathbf{y}$. As an example of this, $h\,|\lambda_{\max}|$ must be less than 2 for Euler's method, and it is seen from Figure 14.14 that approximately the same hold for e.g. RK4. Some systems have a large spread in eigenvalues, and as the time-step of an explicit method must be selected to ensure stability, it follows that very many time steps are required to compute the dynamics corresponding to the small eigenvalues. This gives problems with simulation time and accuracy. Systems that have a large spread in eigenvalues of the Jacobian are referred to as stiff systems. Stiff systems are difficult to solve with explicit methods. This has lead to a recent and more pragmatic definition of stiff systems as systems that are difficult to solve with explicit methods. Examples of stiff systems are the restricted three-body problem in Section 14.1.3, and the mass balances in Section 14.1.4. We will see that stiff problems can be solved efficiently by implicit Runge-Kutta method, that are presented in the following.

14.5.2 Implicit Runge-Kutta methods

An implicit Runge-Kutta method with σ stages for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{14.119}$$

is given by

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n + h\left(a_{11}\mathbf{k}_1 + \ldots + a_{1\sigma}\mathbf{k}_\sigma\right), t_n + c_1 h) && (14.120)\\
&\vdots && (14.121)\\
\mathbf{k}_\sigma &= \mathbf{f}(\mathbf{y}_n + h\left(a_{\sigma 1}\mathbf{k}_1 + \ldots + a_{\sigma\sigma}\mathbf{k}_\sigma\right), t_n + c_\sigma h) && (14.122)\\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h(b_1\mathbf{k}_1 + \ldots + b_\sigma\mathbf{k}_\sigma) && (14.123)
\end{aligned}$$

As for explicit Runge-Kutta methods, the interpolation parameters $c_i,\ i \in \{1,\ldots,\sigma\}$ are in the range $0 \le c_i \le 1$. The weighting factors satisfy the normalization equation $\sum_{i=1}^{\sigma} b_i = 1$, and usually the weighting factors at each stage satisfy $\sum_{j=1}^{\sigma} a_{ij} = c_i$.

14.5.3 Implicit Euler method

The *implicit Euler method* is an implicit Runge-Kutta method with one stage described by the following array:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

This gives

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n + h\mathbf{k}_1, t_{n+1})\\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\mathbf{k}_1
\end{aligned}$$

This method is said to be a Radau IIA method.

The stability function is found by applying the method to the linear test system $\dot{y} = \lambda y$. Then $k_1 = \lambda y_n + \lambda h k_1$ can be solved for k_1, and inserting this into the equation for y_{n+1} we get

$$y_{n+1} = y_n + \frac{h\lambda}{1 - h\lambda} y_n = \frac{1}{1 - h\lambda} y_n \tag{14.124}$$

The stability function is seen to be

$$R(h\lambda) = \frac{1}{1 - h\lambda} \tag{14.125}$$

The region in the complex plane where the method is stable is given by $|h\lambda - 1| \geq 1$ and shown as the shaded region in Figure 14.15.

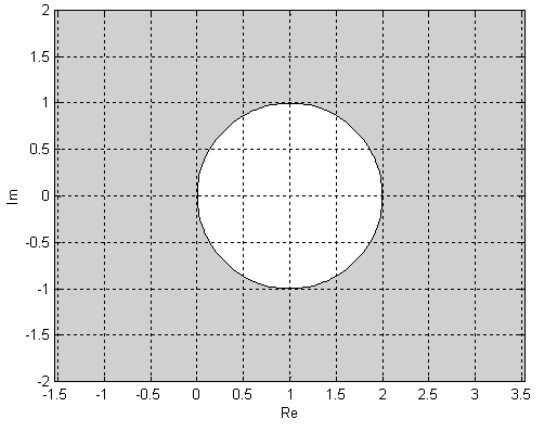

Figure 14.15: The shaded area shows where the implicit Euler method is stable as a function of the complex variable $s = \lambda h$.

14.5.4 Trapezoidal rule

Consider the implicit Runge-Kutta method

$$\mathbf{k}_1 = \mathbf{f}(\mathbf{y}_n, t_n) \tag{14.126}$$

$$\mathbf{k}_2 = \mathbf{f}\left[\mathbf{y}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2), t_n + h\right] \tag{14.127}$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2) \tag{14.128}$$

which is a Lobatto IIIA method of order 2. The Butcher array is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

A closer look reveals that

$$\mathbf{k}_2 = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1}) \tag{14.129}$$

as the last row in $\mathbf{A}$ is equal to $\mathbf{b}^T$. This implies that the expression for $\mathbf{y}_{n+1}$ can be rewritten in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}\left[\mathbf{f}(\mathbf{y}_n, t_n) + \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})\right] \tag{14.130}$$

which is known as the *trapezoidal rule*.

The stability function is found from (14.130) which for the test equation gives

$$y_{n+1} = y_n + \frac{h\lambda}{2}\left(y_n + y_{n+1}\right) \tag{14.131}$$

and it follows that

$$R(\lambda h) = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \tag{14.132}$$

We see that

$$|R(\lambda h)|^2 = \frac{\left(1 + \mathrm{Re}\left[\frac{h\lambda}{2}\right]\right)^2 + \left(\mathrm{Im}\left[\frac{h\lambda}{2}\right]\right)^2}{\left(1 - \mathrm{Re}\left[\frac{h\lambda}{2}\right]\right)^2 + \left(\mathrm{Im}\left[\frac{h\lambda}{2}\right]\right)^2} \tag{14.133}$$

and it follows that $|R(\lambda h)| \leq 1$ and the method is stable for all λ that have negative real part. The area in the complex plane where the trapezoidal rule is stable is therefore the left half plane as shown in Figure 14.16.

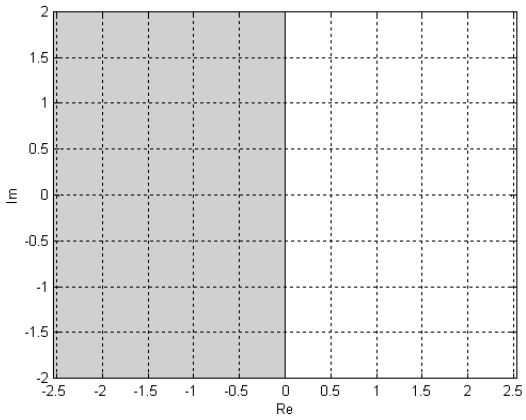

Figure 14.16: The shaded area shows where the trapezoidal rule is stable as a function of the complex variable $s = \lambda h$.

14.5.5 Implicit midpoint rule

We consider the implicit Runge-Kutta method

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}\left(\mathbf{y}_n + \frac{h}{2}\mathbf{k}_1, t_n + \frac{h}{2}\right) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + h\mathbf{k}_1 \end{aligned}$$

with Butcher array

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

This method is a Gauss method of order 2. This implicit Runge-Kutta method can be reformulated as a scheme known as the *implicit mid-point rule*. To do this we first

note that the equation for $\mathbf{y}_{n+1}$ gives $h\mathbf{k}_1 = \mathbf{y}_{n+1} - \mathbf{y}_n$. Inserting this into the stage computation gives

$$\mathbf{y}_{n+1} - \mathbf{y}_n = h\mathbf{f}\left[\mathbf{y}_n + \frac{1}{2}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right), t_n + \frac{h}{2}\right]$$

which is simplifies to the following scheme

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}\left(\frac{\mathbf{y}_n + \mathbf{y}_{n+1}}{2}, t_n + \frac{h}{2}\right) \tag{14.134}$$

which is called the implicit mid-point rule.

From (14.134) we find that the stability function for the implicit mid-point rule is

$$R(\lambda h) = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} \tag{14.135}$$

which is identical to the stability function for the trapezoidal rule. Therefore the stability properties of the two methods are the same for linear time-invariant systems. However it turns out that for nonlinear systems the implicit mid-point rule has much better stability properties, as will be seen in the following sections.

14.5.6 The theta method

Consider the implicit Runge-Kutta method

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n, t_n) && (14.136)\\
\mathbf{k}_2 &= \mathbf{f}\left[\mathbf{y}_n + h\left[\theta\mathbf{k}_1 + (1-\theta)\,\mathbf{k}_2\right], t_n + h\right] && (14.137)\\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left[\theta\mathbf{k}_1 + (1-\theta)\,\mathbf{k}_2\right] && (14.138)
\end{aligned}$$

where $\theta \in [0, 1]$ is a parameter. The Butcher array is

$$\begin{array}{c|cc}
0 & 0 & 0 \\
1 & \theta & 1-\theta \\
\hline
 & \theta & 1-\theta
\end{array}$$

As for the trapezoidal rule, the second stage can be written $\mathbf{k}_2 = \mathbf{f}\left(\mathbf{y}_{n+1}, t_{n+1}\right)$. Then the expression for $\mathbf{y}_{n+1}$ can be rewritten in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\left[\theta\mathbf{f}(\mathbf{y}_n, t_n) + (1-\theta)\,\mathbf{f}\left(\mathbf{y}_{n+1}, t_{n+1}\right)\right]$$

which is known as the *theta method*. We see that for $\theta = 1$ the method is Euler's method, for $\theta = \frac{1}{2}$ the method is the trapezoidal rule, and for $\theta = 0$ it is the implicit Euler method. The stability function is

$$R(h\lambda) = \frac{1 + h\lambda\theta}{1 - h\lambda\left(1-\theta\right)} \tag{14.139}$$

14.5.7 Stability function

The application of an implicit Runge-Kutta method to a linear test system gives

$$\begin{aligned}
\boldsymbol{\kappa} &= \lambda\left(\mathbf{1}y_n + h\mathbf{A}\boldsymbol{\kappa}\right) && (14.140)\\
y_{n+1} &= y_n + h\mathbf{b}^T\boldsymbol{\kappa} && (14.141)
\end{aligned}$$

as for explicit methods, where the notation is defined in connection with equations (14.108) and (14.109). From (14.110) and (14.111) we may conclude as follows:

The stability function for an implicit Runge-Kutta method is given by the two alternative expressions

$$R(h\lambda) = \left[1 + \lambda h \mathbf{b}^T \left(\mathbf{I} - h\lambda \mathbf{A}\right)^{-1} \mathbf{1}\right] \tag{14.142}$$

$$R(h\lambda) = \frac{\det\left[\mathbf{I} - \lambda h \left(\mathbf{A} - \mathbf{1}\mathbf{b}^T\right)\right]}{\det\left(\mathbf{I} - \lambda h \mathbf{A}\right)} \tag{14.143}$$

From (14.143) it is seen that the stability function for an implicit Runge-Kutta method is a rational expression in $s = \lambda h$. We will see in the following that certain properties of the implicit methods will depend on the degree of the numerator and denominator polynomials in the stability function. In particular it will be shown that the most important implicit methods have stability functions $R(s)$ given by Padé approximations of e^s, and that interesting conclusions can be drawn from this fact. However, first we will present some implicit methods and a case study.

14.5.8 Some implicit Runge-Kutta methods

Gauss order 2, which is the implicit mid-point rule

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Gauss order 4, which is the Hammer and Hollingsworth method of order 4

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Radau IA, order 3

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}$$

Radau IIA, order 3

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

Lobatto IIIA, order 2, which is the trapezoidal rule

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Lobatto IIIB order 2

$$\begin{array}{c|cc} 0 & \frac{1}{2} & 0 \\ 1 & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Lobatto IIIC order 2

$$\begin{array}{c|cc} 0 & \frac{1}{2} & -\frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

14.5.9 Case study: Pneumatic spring revisited

The pneumatic spring from Section 14.4.5 was simulated with implicit Runge-Kutta methods with a time step $h = 0.5$ s, which was found to be the stability limit for this system when the explicit RK4 was used (Figure 14.13). The methods that were used was the Gauss method of order 2 (the implicit mid-point rule), Radau IIA of order 3, Lobatto IIIC of order 2 and the implicit Euler method. The results are shown in Figure 14.17. The Gauss method gave no damping, while the Radau method gave some damping, the Lobatto method gave more damping than the Radau method, and the implicit Euler method gave the most damping. This is clearly seen in Figure 14.18 where the total energy corresponding to the numerical solutions is plotted. It is seen that the energy of the solution from the Gauss method fluctuates around the correct value, while the other methods introduce what can be termed numerical dissipation of energy. In particular it is seen that the implicit Euler method gave a solution where the total energy quickly converged to the energy of the equilibrium state.

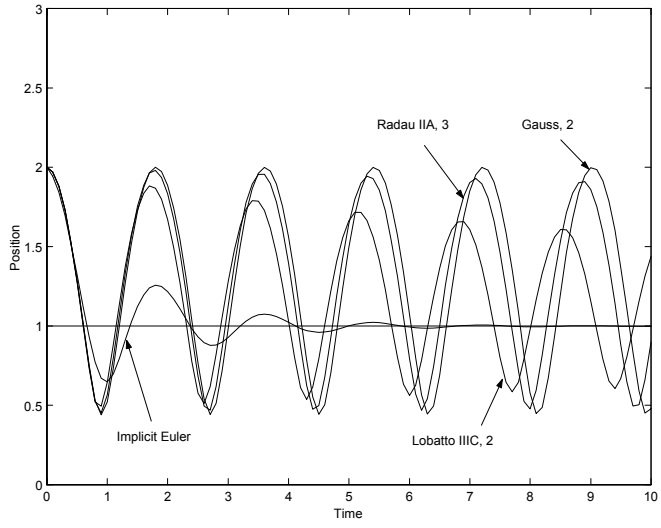

Figure 14.17: Position of piston computed with the implicit Runge-Kutta methods Gauss of order 2, Radau IIA of order 3, Lobatto IIIC of order 2 and the implicit Euler method.

To study how Runge-Kutta methods work for stiff oscillatory systems the pneumatic spring system was modified to include a mechanical resonance in the mass as shown

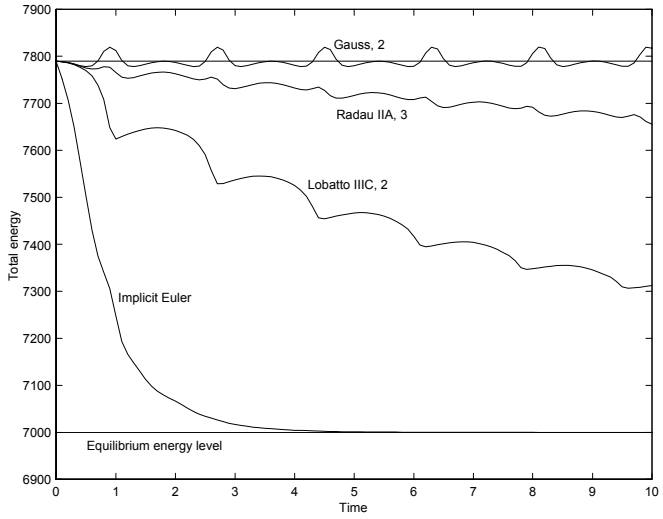

Figure 14.18: Total energy for the pneumatic system when the solution is computed with the implicit Runge-Kutta methods. The energy oscillates around the correct value for the Gauss of order 2, while the energy is numerically dissipated with the methods Radau IIA of order 3, Lobatto IIIC of order 2 and the implicit Euler method.

in Figure 14.19. This was done by splitting the mass $m = 200$ kg into two masses $m_1 = m_2 = 100$ kg, which are connected by a spring with stiffness $K = m_1\omega_2^2/2$ with axis along the vertical axis. The position of m_1 is denoted x_1 and the position of m_2 is denoted x_2. The coordinate x_2 is given an offset so that $x_1 = x_2$ when the spring force is zero. The equilibrium energy of this system with two degrees of freedom is the same as for the one degree of freedom system studied above. The equations of motion are

$$\ddot{x}_1 + g\left(1 - \frac{m}{m_1}x_1^{-\kappa}\right) + \frac{\omega_2^2}{2}(x_1 - x_2) = 0 \tag{14.144}$$

$$\ddot{x}_2 + g + \frac{\omega_2^2}{2}(x_2 - x_1) = 0 \tag{14.145}$$

where $\omega_2 = 1000$ rad/s, $m_1 = m_2 = 100$ kg and $m = m_1 + m_2 = 200$ kg. The standard form is $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$ is obtained by setting

$$\mathbf{y} = \begin{pmatrix} x_1 \\ v_1 \\ x_2 \\ v_2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} v_1 \\ -g(1 - \frac{m}{m_1}x^{-\kappa}) - \frac{\omega_2^2}{2}(x_1 - x_2) \\ v_2 \\ -g - \frac{\omega_2^2}{2}(x_2 - x_1) \end{pmatrix} \tag{14.146}$$

where $v_i = \dot{x}_i$ is the velocity of the piston. We see that the system has an equilibrium at $x_1^* = 1$, $v_1^* = 0$, $x_2^* = 1 - \frac{m_2}{K}$ where $\ddot{x}_i = 0$. Linearization around x_1^*, x_2^* gives

$$\Delta\dot{\mathbf{y}} = \mathbf{J}\Delta\mathbf{y} \tag{14.147}$$

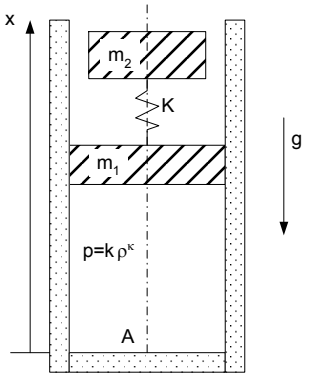

Figure 14.19: Pneumatic spring with mechanical resonance in load. The x coordinate of mass m_2 is given an offset so that $x_1 = x_2$ when the spring is unloaded.

where

$$\Delta\mathbf{y} = \begin{pmatrix} x_1 - x_1^* \\ v_1 \\ x_2 - x_2^* \\ v_2 \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -g\frac{m}{m_1}\kappa\left(x^*\right)^{-(\kappa-1)} - \frac{\omega_2^2}{2} & 0 & \frac{\omega_2^2}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\omega_2^2}{2} & 0 & -\frac{\omega_2^2}{2} & 0 \end{pmatrix} \tag{14.148}$$

The eigenvalues of the linearization around the equilibrium are found to be

$$\lambda_{1,2} = \pm j\omega_1, \quad \omega_1 = 3.7 \text{ rad/s} \tag{14.149}$$

$$\lambda_{3,4} = \pm j\omega_2, \quad \omega_2 = 1000 \text{ rad/s} \tag{14.150}$$

This means that the system has the eigenvalues at $\pm j3.7$ as the pneumatic spring, and in addition a new set of eigenvalues have been introduced at $\pm j1000$.

The system was simulated with RK4 with a time step $h = 0.0005$ s, where the step size was selected so that $h\omega_2 = 0.5$. The simulation result as shown in Figure 14.20 is fairly accurate, but it is seen from Figure 14.21 that the high frequency motion is damped out even though there is no damping in the system. The phenomenon is clearly seen from the plot of the total energy in Figure 14.22 where it is seen that the energy converges to the energy level of the slow dynamics corresponding to $\lambda_{1,2} = \pm j3.7$, which is the energy that is obtained if $x_1 = x_2$.

The system was then simulated with the implicit methods Gauss of order 2 and Lobatto IIIC of order 2 with a time step $h = 0.05$. This gives Nyquist frequency $\omega_N = \pi/0.05 = 62.8$ rad/s, so that the resonance at 1000 rad/s is well above the Nyquist frequency. The solution of the Gauss method, which is shown in Figure 14.23 gave no damping, but the aliasing effect moved the energy of the fast dynamics associated with the eigenvalues $\pm j1000$ to oscillations with frequency lower than the Nyquist frequency. This gave a beat phenomenon which is clearly seen in Figure 14.24, while it is seen from Figure 14.25 that the total energy is constant for the Gauss solution. The Lobatto IIIC solution gave quick damping of the fast dynamics, and a slight damping of the slow dynamics. It is seen from Figure 14.25 that the energy associated with the fast dynamics is dissipated in one step, and the total energy remains on the level of the energy associated with the slow dynamics.

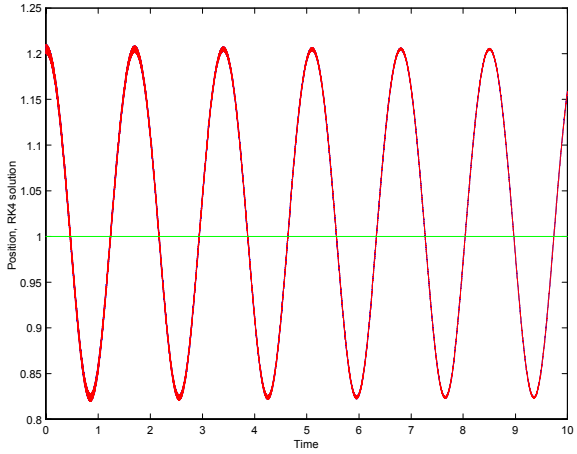

Figure 14.20: Position of the two masses computed with RK4 with time step $h = 0.0005$ s.

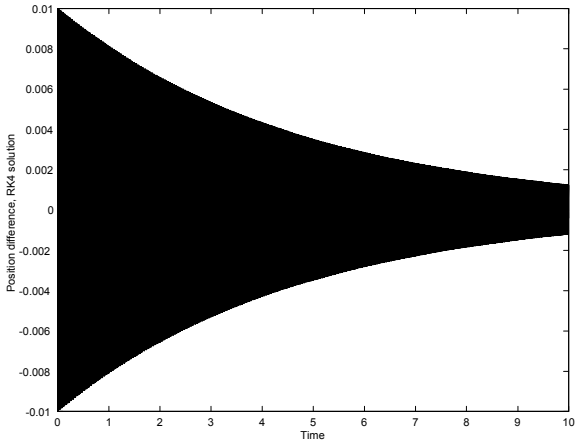

Figure 14.21: Offset equilibrium for the spring between masses one and two computed with RK4 with $h = 0.0005$ s. The oscillation is seen to be lightly damped by the integration method.

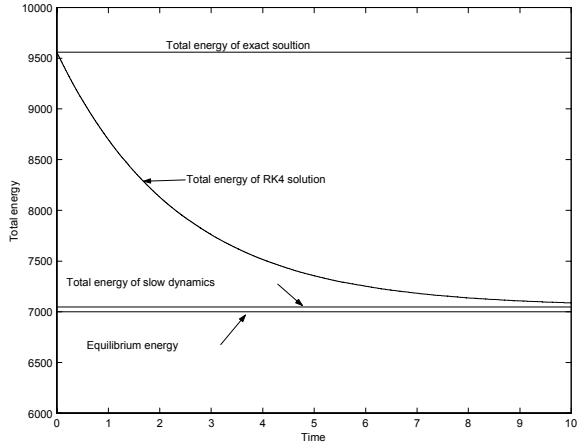

Figure 14.22: Total energy corresponding to the numerical solution computed with RK4 with $h = 0.0005$ s. It is seen that the energy related to the fast dynamics is slowly damped out.

14.6 Stability of Runge-Kutta methods

14.6.1 Aliasing

We consider the test equation $\dot{y} = \lambda y$, and write the eigenvalue in the form

$$\lambda = \sigma + j\omega \tag{14.151}$$

where σ is the real part and $j\omega$ is the imaginary part. It is assumed that $\omega < \pi/h$, that is, ω is assumed to be less that the Nyquist frequency π/h. The local solution of the test system is

$$y_L\left(t_n; t_{n+1}\right) = e^{\lambda h} y_n$$

Consider a system $\dot{y} = \mu y$, which has the local solution

$$y_L\left(t_n; t_{n+1}\right) = e^{\mu h} y_n \tag{14.152}$$

The two systems will give the same local solutions at t_{n+1} whenever $e^{\lambda h} = e^{\mu h}$ which is implied by

$$\mu = \lambda + j2k\frac{\pi}{h} = \sigma + j\left(\omega + 2k\frac{\pi}{h}\right),\ k = 0, \pm 1, \pm 2, \ldots \tag{14.153}$$

If

$$\mu = \lambda + 2kj\frac{\pi}{h},\ k = \pm 1, \pm 2, \ldots \tag{14.154}$$

then the system $\dot{y} = \mu y$ where $\mathrm{Im}\left(\lambda\right) > \pi/h$ will have the same solution as the system $\dot{y} = \lambda y$ where $\mathrm{Im}\left(\lambda\right) < \pi/h$. This phenomenon is called aliasing.

14.6.2 A-stability, L-stability

The test system $\dot{y} = \lambda y$ is stable when $\mathrm{Re}\,\lambda \leq 0$. We consider an integration method which gives $y_{n+1} = R(\lambda h) y_n$ when applied to the test system. It would seem to be a

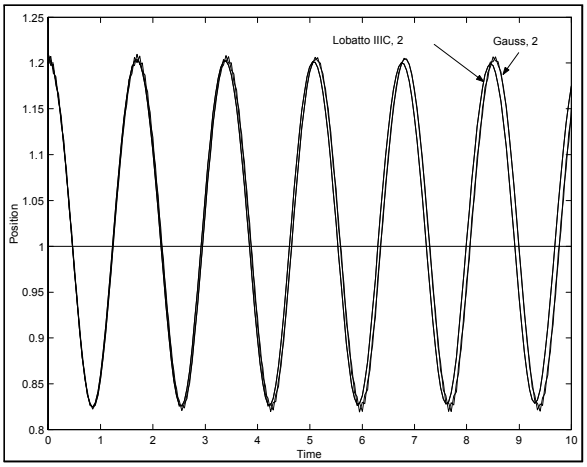

Figure 14.23: Position of the two masses computed with a Gauss method of order 2 and a Lobatto IIIC method of order 3. The time step was $h = 0.05$ with both methods.

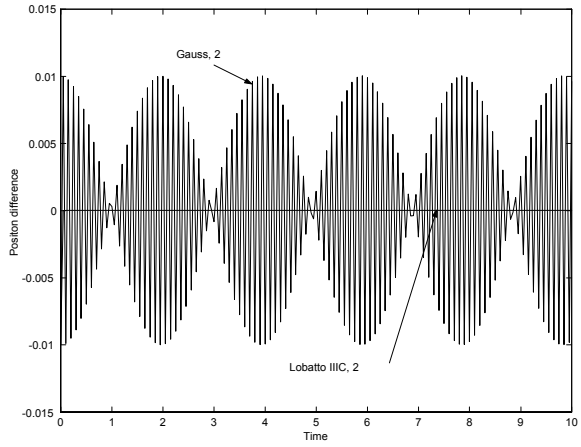

Figure 14.24: Offset in position between the two masses calculated with Gauss order 2 and Lobatto IIIC order 2 with $h = 0.05$ s. The Gauss method gave no damping, but the energy of the fast dynamics was shifted to frequencies below the Nyquist frequency $\omega_N = 62.8$ rad/s. The Lobatto method damped out the fast dynamics in one step.

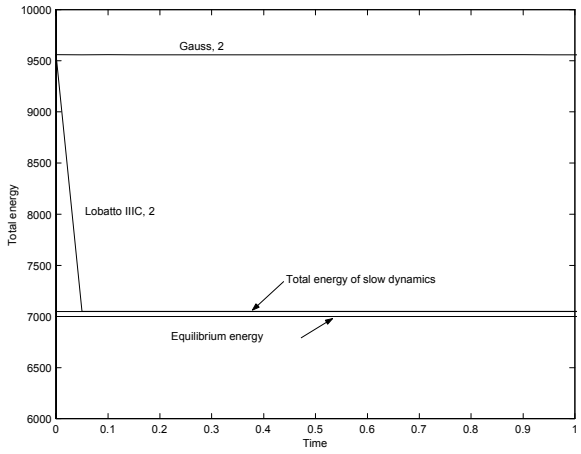

Figure 14.25: Total energy corresponding to the numerical solution of Gauss order 2 and Lobatto IIIC order 2. The Gauss method gave a constant total energy in agreement with the exact solution, while the Lobatto method damped out the energy associated witht the fast dynamics.

useful property for an integration method if the method was stable for all stable test systems. This property is called A-stability.

An method is A-stable if $|R(\lambda h)| \leq 1$ for all $\mathrm{Re}\,\lambda \leq 0$.

Integration methods that are A-stable will be stable also for systems with very fast dynamics which in this context are systems that have dynamics which is significantly faster that the time step h of the integration method. In particular, aliasing can be problematic for A-stable methods, as high frequency oscillations will appear in the computed solution as an oscillation with frequency below the Nyquist frequency π/h. As the integration cannot give an accurate computation of such fast dynamics, it may be useful that the method damp out the fast dynamics. This is the case for L-stable integration methods.

A method is L-stable if it is A-stable and, in addition, if $|R(j\omega h)| \to 0$ when $\omega \to \infty$ for all systems $\dot{y} = \lambda y$ where $\lambda = j\omega$.

Example 221 *We note that explicit Runge-Kutta methods have stability functions*

$$R_E(\lambda h) = \det\left[\mathbf{I} - \lambda h\left(\mathbf{A} - \mathbf{1}\mathbf{b}^T\right)\right] \tag{14.155}$$

It is clear that $|R_E(\lambda h)| \to \infty$ whenever $|\lambda| \to \infty$, and it follows that an explicit Runge-Kutta method cannot be A-stable.

14.6.3 Stiffly accurate methods

We will here take a closer look at implicit Runge-Kutta methods that are *stiffly accurate*. The stability function of an implicit Runge-Kutta method is found in the same way as

for explicit Runge-Kutta methods. Therefore, the expression

$$R(h\lambda) = \left[1 + \lambda h \mathbf{b}^T \left(\mathbf{I} - h\lambda \mathbf{A}\right)^{-1} \mathbf{1}\right]$$

(14.110) can be used also for implicit Runge-Kutta methods. Suppose that $\mathbf{A}$ is nonsingular, and consider the case where λh tends to infinity. Then if the limit $R(\infty) := \lim_{s\to\infty} R(s)$ exists, it is given by

$$\begin{aligned} R(\infty) &= \lim_{s\to\infty} \left[1 + s\mathbf{b}^T \left(\mathbf{I} - s\mathbf{A}\right)^{-1} \mathbf{1}\right] \\ &= \lim_{s\to\infty} \left[1 - s\mathbf{b}^T \left(s\mathbf{A}\right)^{-1} \mathbf{1}\right] \\ &= 1 - \mathbf{b}^T \mathbf{A}^{-1} \mathbf{1} \end{aligned} \tag{14.156}$$

Moreover, it is noted that for an implicit Runge-Kutta method where

$$\mathbf{k}_\sigma = \mathbf{f}\left(\mathbf{y}_{n+1}, t_{n+1}\right) \tag{14.157}$$

then the weighting vector $\mathbf{b}$ is equal to the last row in the stage matrix $\mathbf{A}$. This gives

$$\mathbf{b} = \mathbf{A}^T \mathbf{e}_\sigma \tag{14.158}$$

where $\mathbf{e}_\sigma = (0, 0, \ldots, 1)^T$ is a σ-dimensional unit vector. Insertion of (14.158) into (14.156) gives

$$R(\infty) = 1 - \lambda h \mathbf{e}_\sigma^T \mathbf{A} (\lambda h \mathbf{A})^{-1} \mathbf{1} = 1 - \mathbf{e}_\sigma^T \mathbf{1} = 0 \tag{14.159}$$

An implicit Runge-Kutta method is said to be stiffly accurate if the stage matrix $\mathbf{A}$ is nonsingular and in addition $\mathbf{b} = \mathbf{A}^T \mathbf{e}_\sigma$.

From (14.159) we find that

An A-stable Runge-Kutta method that is stiffly accurate will be L-stable.

Moreover, from (14.159) we may conclude that a stiffly accurate method will damp out dynamics corresponding to eigenvalues λ_i that are large in the sense that $|\lambda_i h|$ are much larger that unity. Consider the case where a stiffly accurate method is applied to a stiff system, and the time step h is selected in the dynamic range of the slow dynamics. Then the fast dynamics will have eigenvalues so that $|\lambda_i h| \gg 1$. The fast dynamics will therefore be damped out, and the solution will mainly correspond to the slow dynamics. In particular, if there is an eigenvalue λ_j so that $|\lambda_j h| \to \infty$, then the dynamics associated with this eigenvalue will be damped to zero.

Example 222 *It is clear that a Gauss method cannot be stiffly accurate as for these methods* $|R(j\omega)| = 1$ *for all* ω.

Example 223 *The Trapezoidal rule is a Lobatto IIIA method with*

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

The last row in $\mathbf{A}$ *is equal to* $\mathbf{b}^T$*, but the matrix* $\mathbf{A}$ *is singular. Thus the method is not stiffly accurate. This agrees with the fact that the stability function is*

$$\begin{aligned} R(s) &= 1 + s\mathbf{b}^T(\mathbf{I} - s\mathbf{A})^{-1}\mathbf{1} \\ &= 1 + \frac{s}{2}\left[\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{\frac{s}{2}}{1-\frac{s}{2}} & \frac{\frac{1}{2}}{1-\frac{s}{2}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] \\ &= \frac{1+\frac{s}{2}}{1-\frac{s}{2}} \end{aligned}$$

which means that

$$|R(j\omega)| = 1$$

for all ω*.*

14.6.4 Padé approximations

The stability function $R(s)$ of a Runge-Kutta method, which may be implicit or explicit, is given by a rational expression in s, which is seen from (14.143) with $s = \lambda h$. Properties like A-stability and L-stability depend on the stability function. Also the region of stability for a method is found from the stability function. Instead of checking such properties for each method, it is possible to have more general results. This is done by introducing a classification of Runge-Kutta methods based on a special characterization of the stability functions. This will be done in the following using the Padé approximations of the exponential function e^s.

First it is noted that the local solution of the test equation $\dot{y} = \lambda y$ over the time step from t_n to t_{n+1} is

$$y_L(t_n; t_{n+1}) = e^{\lambda h} y_n \tag{14.160}$$

while the numerical solution is

$$y_{n+1} = R(\lambda h) y_n \tag{14.161}$$

From these two equations it is seen that the accuracy of the numerical solution y_n will depend on to what extent the stability function approximates the exponential function, that is, the accuracy of the numerical solution y_n depends on the difference

$$e^s - R(s) \tag{14.162}$$

between the exponential function in the exact solution (14.160), and stability the numerical solution (14.161). Here we have used $s = \lambda h$ to simplify the notation. An explicit Runge-Kutta method with σ stages approximates the exponential function e^s by the polynomial approximation

$$R(s) = 1 + \beta_1 s + \ldots + \beta_\sigma s^\sigma$$

In the special case that the method is of order $p = \sigma \leq 4$ the stability function is given by the Taylor series expansion of e^s given by

$$R(s) = 1 + s + \ldots + \frac{s^p}{p!}$$

An implicit Runge-Kutta method of σ stages approximates e^s by the rational approximation

$$R(s) = \frac{1 + \beta_1 s + \ldots + \beta_k s^k}{1 + \gamma_1 s + \ldots + \gamma_m s^m}$$

		k			
		0	1	2	3
m	0	1	$1+s$	$1+s+\frac{s^2}{2!}$	$1+s+\frac{s^2}{2!}+\frac{s^3}{3!}$
	1	$\frac{1}{1-s}$	$\frac{1+\frac{1}{2}s}{1-\frac{1}{2}s}$	$\frac{1+\frac{2}{3}s+\frac{1}{6}s^2}{1-\frac{1}{3}s}$	$\frac{1+\frac{3}{4}s+\frac{1}{4}s^2+\frac{1}{24}s^3}{1-\frac{1}{4}s}$
	2	$\frac{1}{1-s+\frac{s^2}{2!}}$	$\frac{1+\frac{1}{3}s}{1-\frac{2}{3}s+\frac{1}{6}s^2}$	$\frac{1+\frac{1}{2}s+\frac{1}{12}s^2}{1-\frac{1}{2}s+\frac{1}{12}s^2}$	$\frac{1+\frac{3}{5}s+\frac{3}{20}s^2+\frac{1}{60}s^3}{1-\frac{2}{5}s+\frac{1}{20}s^2}$
	3	$\frac{1}{1-s+\frac{s^2}{2!}-\frac{s^3}{3!}}$	$\frac{1+\frac{1}{4}s}{1-\frac{3}{4}s+\frac{1}{4}s^2-\frac{1}{24}s^3}$	$\frac{1+\frac{2}{5}s+\frac{1}{20}s^2}{1-\frac{3}{5}s+\frac{3}{20}s^2-\frac{1}{60}s^3}$	$\frac{1+\frac{s}{2}+\frac{s^2}{10}+\frac{s^3}{120}}{1-\frac{s}{2}+\frac{s^2}{10}-\frac{s^3}{120}}$

Table 14.2: The Padé approximations $P_m^k(s)$ for $m, n = 0, 1, 2, 3$

		k			
		0	1	2	3
m	0		Euler's method	Mod. Euler	Heun's, 3
	1	Radau, 1	Gauss, 2, Trapez.		
	2	Lobatto IIIC, 2	Radau, 3	Gauss, 4	
	3		Lobatto IIIC, 4	Radau, 5	Gauss, 6

Table 14.3: Methods that have Pade approximations $P_m^k(s)$ as stability functions.

where $m \leq \sigma$ and $k \leq \sigma$.

One particular rational approximation of the exponential function is the *Padé approximation* (Golub and van Loan 1989).

The Padé approximation $P_m^k(s)$ of the exponential function e^s is a rational function of s with a numerator of degree k and a denominator of degree m. The Padé approximation $P_m^k(s)$ of e^s is the rational approximation of e^s which has the highest order in s when the numerator is of order k and the denominator is of order m.

The Padé approximation $P_m^k(s)$ is given by

$$P_m^k(s) = \frac{Q_{mk}(s)}{Q_{km}(-s)} \tag{14.163}$$

where

$$Q_{mk}(s) = 1 + \sum_{i=1}^{k} \frac{k!\,(m+k-i)!}{(k-i)!\,(m+k)!} \frac{s^i}{i!} \tag{14.164}$$

The error in the approximation is given by

$$e^s - P_m^k(s) = \frac{(-1)^k\, m!k!}{(k+m)!} \frac{s^{k+m+1}}{(k+m+1)!} + O(s^{k+m+2})$$

which shows that the approximation is of order $k+m$. The Padé approximations $P_m^k(s)$ for $m, k = 0, 1, 2, 3$ are shown in Table 14.2.

Example 224 *Long division of* $P_1^1(s)$ *gives*

$$\frac{1+\frac{1}{2}s}{1-\frac{1}{2}s} = 1 + s + \frac{s^2}{2} + \frac{s^3}{4} + O(s^4)$$

where the error is $\frac{s^3}{6} + O(s^4)$.

Example 225 *We note that an explicit Runge-Kutta method with* $p = \sigma \leq 4$ *have stability functions* $R_E(s) = P_0^p(s)$

14.6.5 Stability for Padé approximations

An important result related to A-stability of methods is the following:

$$P_m^k(s) \leq 1, \quad \text{when } \operatorname{Re}[s] \leq 0 \text{ for } k \leq m \leq k+2 \tag{14.165}$$

This is derived using order stars in (Hairer and Wanner 1996). Moreover, in relation to L-stability it is interesting to study $P_m^k(j\omega)$. From Table 14.2 it is seen that the Padé approximations where the degree of the numerator polynomial equals the denominator polynomials satisfy

$$|P_m^m(j\omega)| = 1, \quad \text{for all } \omega \tag{14.166}$$

whereas for Padé approximations where the degree of the numerator polynomial is less than the degree of the denominator polynomials we have

$$|P_m^k(j\omega)| \to 0, \quad \text{when } \omega \to \infty \text{ for } m > k \tag{14.167}$$

Combining these results we arrive at the following result:

A one-step method with stability function

$$R(s) = P_m^m(s) \tag{14.168}$$

is A stable. A one-step method with stability function

$$R(s) = P_m^k(s) \quad \text{where } m = k+1 \text{ or } m = k+2 \tag{14.169}$$

is L-stable.

Example 226 *The Gauss methods, including the implicit mid-point rule, the Lobatto IIIA, including the trapezoidal rule, and the Lobatto IIIB have stability functions*

$$R(s) = P_m^m(s) \tag{14.170}$$

which implies that the methods are A-stable.

Example 227 *Radau methods and Lobatto IIIC methods have stability functions*

$$R(s) = P_m^k(s), \quad k < m \leq k+2 \tag{14.171}$$

and this implies that the methods are L-stable.

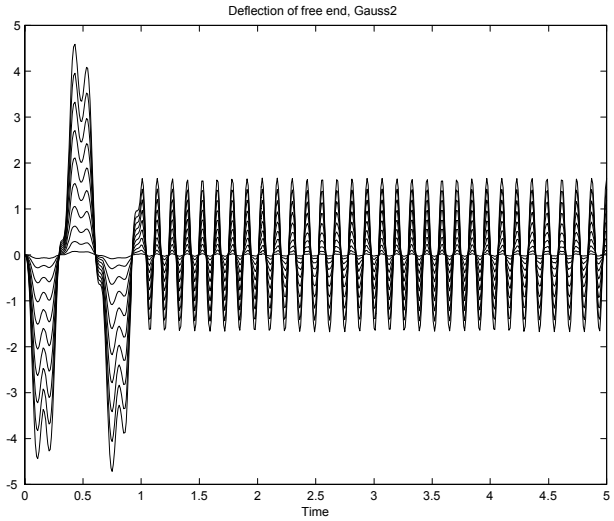

Figure 14.26: Simulation of vibrations in an Euler-Bernoulli beam with a Gauss method. There is an initial excitation that is switched off after 1 s. After this the vibrational energy of the system is a constant. The simulation reflects this.

14.6.6 Example: Mechanical vibrations

An Euler-Bernoulli beam was simulated using finite elements spatial discretization. The beam was modelled as having no damping, and this means that the vibrational energy will be constant when the beam is not excited from external forces. The vibration of the beam will occur at resonance frequencies ω_i, $i = 1, 2, \ldots$. The number of resonant frequencies in the model used for simulation depends on the way the model is implemented. A simulation was done with a discretization using 10 finite elements leading to 10 resonance frequencies. The resulting system is stiff as the fastest resonances have very large eigenvalues $\lambda_i = j\omega_i h$, so that $|\lambda_i| = \omega_i h \gg 1$. The system was simulated with a Gauss method and a Lobatto IIIC method (Kristiansen 2000). The results are shown in Figures 14.26 and 14.27.

14.6.7 Frequency response

The Runge-Kutta methods have stability functions

$$R(s) = 1 + s\mathbf{b}^T(\mathbf{I} - s\mathbf{A})^{-1}\mathbf{1}$$

where $R(s)$ appear in $y_{n+1} = R(\lambda h)y_n$ when the method is applied to the test equation $\dot{y} = \lambda y$. To study the performance of Runge-Kutta methods it is of interest to plot $|R(s)|$ as a function of the complex variable s. One approach to this is to plot the order stars of a method (Hairer and Wanner 1996), which are contour plots of $|R(s)/e^s$ in the complex plane. We will follow a different approach in the following where we plot the magnitude of $R(s)$ for the imaginary axis $s = j\omega$, and for $s = \sigma$ for $-\infty < \sigma \leq 0$ which is the negative part of the real axis. We note that for $s = \lambda h$ the Nyquist frequency is found at $s = \pm j\pi$. The absolute value $|R(j\omega)|$ of the stability function evaluated on the imaginary axis is shown for explicit Runge-Kutta methods in Figure 14.28, and for implicit Runge-Kutta

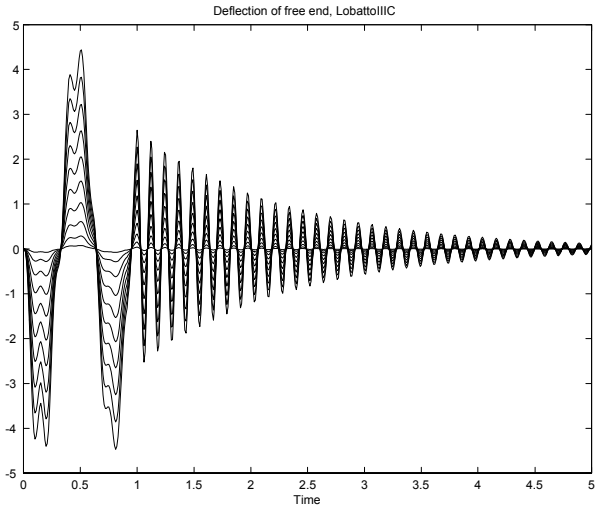

Figure 14.27: Simulation of vibrations in an Euler-Bernoulli beam with a Lobatto IIIC method. There is an initial excitation that is switched off after 1 s. After this the vibrational energy of the system is a constant. The simulation method is seen to introduce damping of the response, and the high frequency components are seen to be more damped than the low frequency components.

methods in Figure 14.29. The absolute value of the stability function evaluated on the negative part of the real axis is shown in Figures 14.30 and 14.31.

A special case occurs when $R(\lambda h)$ has a zero, that is, when there is a $\lambda_z(h)$ so that $R(\lambda_z h) = 0$. This results in a dead-beat response

$$y_{n+1} = 0 \quad \text{when} \quad R(\lambda h) = 0 \tag{14.172}$$

In this section we will take a closer look at the stability functions for the Runge-Kutta methods that have the Padé approximations as their stability functions. To simplify notation we use $s = \lambda h$. The explicit methods of order $p \leq 4$ with $\sigma = p$ stages have stability functions

$$R(s) = P_p^0(s) = \begin{cases} 1 + s & \text{when } p = 1 \\ 1 + s + \frac{s^2}{2} & \text{when } p = 2 \\ 1 + s + \frac{s^2}{2} + \frac{s^3}{6} & \text{when } p = 3 \\ 1 + s + \frac{s^2}{2} + \frac{s^3}{6} + \frac{s^4}{24} & \text{when } p = 4 \end{cases}$$

We see that $R(s) = 0$ occurs for

$$\begin{aligned} s_1 &= -1 && \text{when } p = 1 \\ s_{1,2} &= -1 \pm j && \text{when } p = 2 \\ s_{1,2} &= -0.7020 \pm j1.8073, \quad s_3 = -1.5961 && \text{when } p = 3 \\ s_{1,2} &= -0.2706 \pm j2.5048, \quad s_{3,4} = -1.7294 \pm j0.8890 && \text{when } p = 4 \end{aligned}$$

To study the performance of the method when the test equation has a pole $\lambda = j\omega$ on the imaginary axis we insert $s = j\omega$ in the stability function and get $R(j\omega)$. It is seen that for all the explicit Runge-Kutta methods, $|R(j\omega)| \to \infty$ when $\omega \to \infty$.

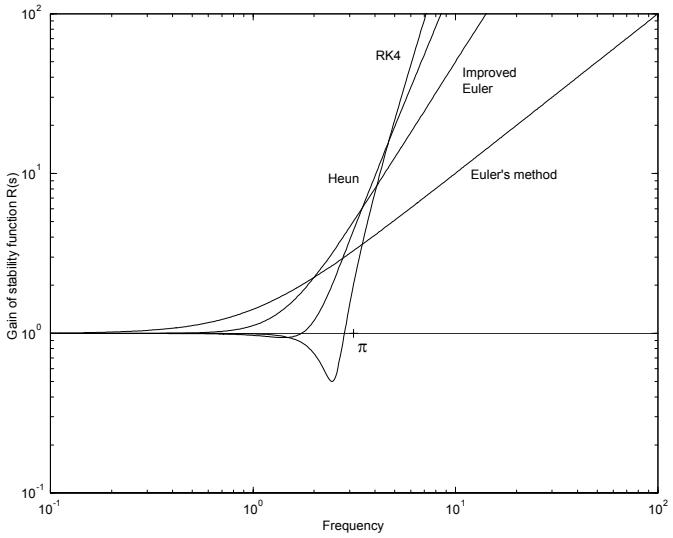

Figure 14.28: Stability function $R(j\omega)$ of explicit Runge-Kutta methods evaluated for $\lambda h = j\omega$. The Nyquist frequency ω_N is plotted at $\omega_N h = \pi$.

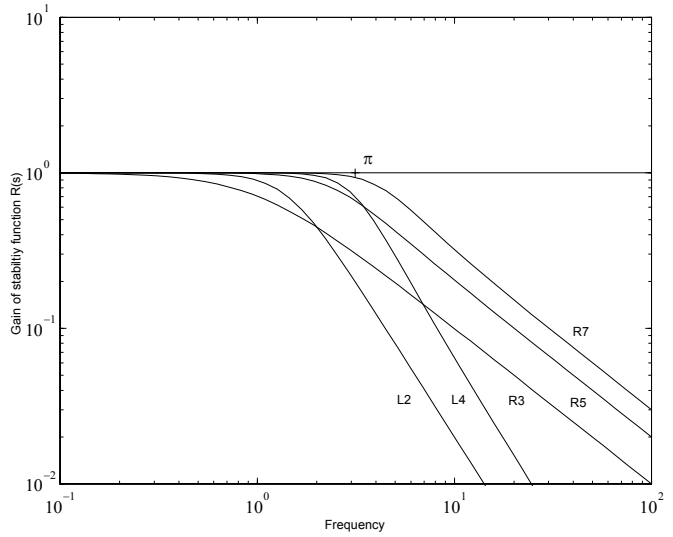

Figure 14.29: Stability function $R(j\omega)$ of implicit Runge-Kutta methods evaluated for $\lambda h = j\omega$. The Nyquist frequency ω_N is plotted at $\omega_N h = \pi$. The methods are seen to damp out frequency components over the Nyquist frequency. The Radau methods have a roll-off of -1, and the Lobatto IIIC methods have a roll-off of -2.

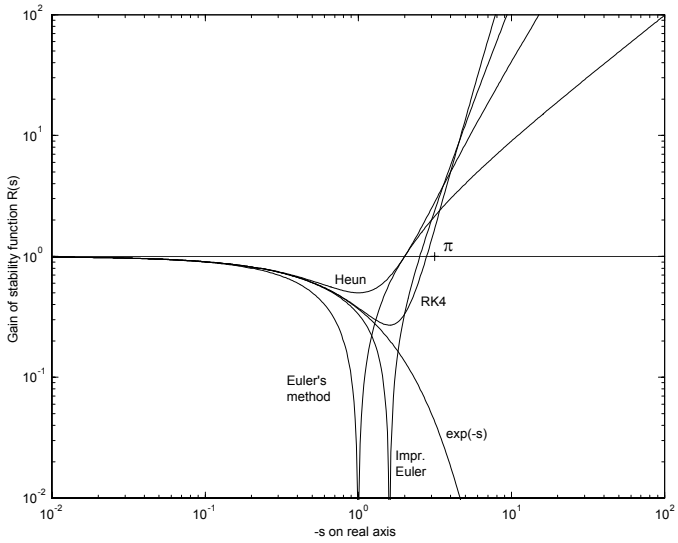

Figure 14.30: Absolute value of stability function $|R(-s)|$ of explicit Runge-Kutta methods evaluated for $\lambda h = -s$. The exact value $\exp(-s)$ is plotted for comparison.

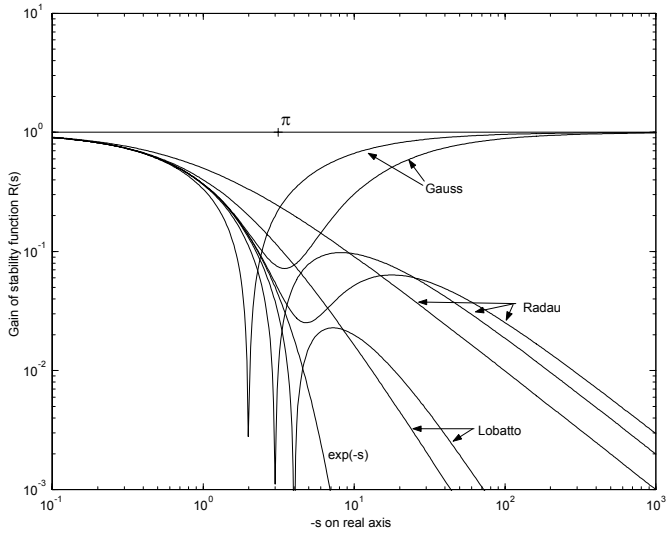

Figure 14.31: Absolute value of stability function $|R(-s)|$ of implicit Runge-Kutta methods evaluated for $\lambda h = -s$. The exact value $\exp(-s)$ is plotted for comparison. The Radau and Lobatto IIIC have a roll-off for $-s$ large, while $|R(-s)| \to 1$ when $-s \to \infty$ for the Gauss methods.

A Gauss method with σ stages, which is an implicit Runge-Kutta method of order 2σ, has stability function

$$R(s) = P_\sigma^\sigma(s) = \begin{cases} \frac{1+\frac{1}{2}s}{1-\frac{1}{2}s} & \text{when } \sigma = 1 \\ \frac{1+\frac{1}{2}s+\frac{1}{12}s^2}{1-\frac{1}{2}s+\frac{1}{12}s^2} & \text{when } \sigma = 2 \\ \frac{1+\frac{1}{2}s+\frac{1}{10}s^2+\frac{1}{120}s^3}{1-\frac{1}{2}s+\frac{1}{10}s^2-\frac{1}{120}s^3} & \text{when } \sigma = 3 \end{cases}$$

The stability function is zero for

$$\begin{array}{cc} s_1 = -2 & \text{when } \sigma = 1 \\ s_{1,2} = -3 \pm j1.7321 & \text{when } \sigma = 2 \\ s_{1,2} = -3.6778 \pm j3.5088, \quad s_3 = -4.6444 & \text{when } \sigma = 3 \end{array}$$

It is quite interesting to study the stability function of Gauss methods for $s = j\omega$. Then, for $p = 1$ we see that

$$|R(j\omega)| = \left| \frac{1 + \frac{1}{2}j\omega}{1 - \frac{1}{2}j\omega} \right| = 1$$

The Radau IIA methods of order $p = 2\sigma - 1$ have stability functions

$$R(s) = P_\sigma^{\sigma-1}(s) = \begin{cases} \frac{1}{1-s} & \text{when } \sigma = 1 \\ \frac{1+\frac{1}{3}s}{1-\frac{2}{3}s+\frac{1}{6}s^2} & \text{when } \sigma = 2 \\ \frac{1+\frac{2}{5}s+\frac{1}{20}s^2}{1-\frac{3}{5}s+\frac{3}{20}s^2-\frac{1}{60}s^3} & \text{when } \sigma = 3 \end{cases}$$

There is no zero in $R(s)$ for $\sigma = 1$, while there is a zero in $s = -3$ for $\sigma = 2$. For Radau IIA with $\sigma = 1$

$$|R(j\omega)| = \begin{cases} 1 & \text{when } \omega \ll \omega_{R1} \\ \frac{1}{\omega} & \text{when } \omega \gg \omega_{R2} \end{cases}$$

The Lobatto IIIC methods of order $p = 2\sigma - 2$ have stability functions

$$R(s) = P_\sigma^{\sigma-2}(s) = \begin{cases} \frac{1}{1-s+\frac{s^2}{2!}} & \text{when } \sigma = 2 \\ \frac{1}{1-s+\frac{1}{2}s^2-\frac{1}{6}s^3} & \text{when } \sigma = 3 \end{cases}$$

For $\sigma = 2$ we have

$$|R(j\omega)| = \begin{cases} 1 & \text{when } \omega \ll \omega_{L1} \\ \frac{1}{\omega^2} & \text{when } \omega \gg \omega_{L2} \end{cases}$$

14.6.8 AN-stability

Before we turn our attention to the nonlinear stability analysis of Runge-Kutta methods we will present an intermediate result on linear time-varying systems. In this connection the linear time-varying test system

$$\dot{y} = \lambda(t)y$$

will be used. The exact solution for linear time-varying test system satisfies

$$y(t_{n+1}) = y(t_n)\exp\left[\int_{t_n}^{t_{n+1}} \lambda(t)dt\right]$$

It is clear that the system is stable in the sense that $|y(t_{n+1})| \leq |y(t_n)|$ if $\mathrm{Re}\,[\lambda(t)] \leq 0$ for all $t \in [t_n, t_{n+1}]$.

An implicit Runge-Kutta method for this system can be written

$$\boldsymbol{\kappa} = \boldsymbol{\Lambda}\left(\mathbf{1}y_n + h\mathbf{A}\boldsymbol{\kappa}\right) \tag{14.173}$$

$$y_{n+1} = y_n + h\mathbf{b}^T\boldsymbol{\kappa} \tag{14.174}$$

where $\boldsymbol{\kappa} = (k_1 \ldots k_\sigma)^T$ and $\mathbf{1} = (1 \ldots 1)^T$ and

$$\boldsymbol{\Lambda} = diag\left(\lambda_1, \ldots, \lambda_\sigma\right), \lambda_i = \lambda(t_n + c_i h) \tag{14.175}$$

Equation (14.173) gives $\boldsymbol{\kappa} = (1 - h\boldsymbol{\Lambda}\mathbf{A})^{-1}\boldsymbol{\Lambda}\mathbf{1}y_n$, and insertion into (14.174) gives

$$y_{n+1} = R_{AN}(h\boldsymbol{\Lambda})y_n$$

where we have defined the stability function

$$R_{AN}(h\boldsymbol{\Lambda}) = 1 + \mathbf{b}^T(\mathbf{I} - h\boldsymbol{\Lambda}\mathbf{A})^{-1}h\boldsymbol{\Lambda}\mathbf{1} \tag{14.176}$$

An implicit Runge-Kutta method is said to be AN-stable if $\mathrm{Re}[\lambda_i] \leq 0$ implies that $|R_{AN}(h\boldsymbol{\Lambda})| \leq 1$ and that $(1 - h\boldsymbol{\Lambda}\mathbf{A})$ is nonsingular.
From this definition it is clear that

$$\boxed{\text{AN-stability}} \Rightarrow \boxed{\text{A-stability}} \tag{14.177}$$

Example 228 *The trapezoidal rule has*

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

and after some algebra it is found that

$$R_{AN}(h\boldsymbol{\Lambda}) = \frac{1 + \frac{h\lambda_1}{2}}{1 - \frac{h\lambda_2}{2}}$$

It is seen that if $\lambda_2 = 0$ and λ_1 is large, then $|R_{AN}(h\boldsymbol{\Lambda})| > 1$, and the method is not AN stable.

Example 229 *Cramer's rule can be used to find the function $R_{AN}(h\boldsymbol{\Lambda})$ defined in (14.176) in the same way as the stability function $R(h\lambda)$. This gives*

$$R_{AN}(h\boldsymbol{\Lambda}) = \frac{\det\left[\mathbf{I} - (\mathbf{A} - \mathbf{1}\mathbf{b}^T)h\boldsymbol{\Lambda}\right]}{\det\left(\mathbf{I} - \mathbf{A}h\boldsymbol{\Lambda}\right)} \tag{14.178}$$

In Lobatto IIIA the first row of $\mathbf{A}$ and the last row of $\mathbf{A} - \mathbf{1}\mathbf{b}^T$ have only zeros. This means that the numerator of $R_{AN}(h\boldsymbol{\Lambda})$ is not a function of λ_σ, and the denominator of $R_{AN}(h\boldsymbol{\Lambda})$ is not a function of λ_1. This means that if $\lambda_2 = \ldots = \lambda_\sigma = 0$, then $|R_{AN}(h\boldsymbol{\Lambda})|$ can be made arbitrarily large by selecting a large $|\lambda_1|$. This means that Lobatto IIIA is not AN-stable.

Example 230 *In Lobatto IIIB the last column of $\mathbf{A}$ and the first column of $\mathbf{A} - \mathbf{1}\mathbf{b}^T$ have only zeros. This means that the numerator of $R_{AN}(h\boldsymbol{\Lambda})$ is not a function of λ_1, and the denominator of $R_{AN}(h\boldsymbol{\Lambda})$ is not a function of λ_σ. This means that if $\lambda_1 = \ldots = \lambda_{\sigma-1} = 0$, then $|R_{AN}(h\boldsymbol{\Lambda})|$ can be made arbitrarily large by selecting a large $|\lambda_\sigma|$. This means that Lobatto IIIB is not AN-stable.*

14.6.9 B-stability

Using the concept of B-stability it is possible to analyze the stability of Runge-Kutta methods for contracting nonlinear systems. For such systems we will see that the stability of Runge-Kutta methods can be studied in terms of a simple algebraic condition on the Runge-Kutta parameters $\mathbf{A}$ and $\mathbf{b}$.

Consider the nonlinear systems

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{14.179}$$

and the scalar nonnegative function

$$V = \frac{1}{2}(\mathbf{y} - \widetilde{\mathbf{y}})^T \mathbf{P}(\mathbf{y} - \widetilde{\mathbf{y}}) \tag{14.180}$$

where $\widetilde{\mathbf{y}}$ is a solution of $\frac{d}{dt}\widetilde{\mathbf{y}} = \mathbf{f}(\widetilde{\mathbf{y}}, t)$ and $\mathbf{P}$ is a positive definite symmetric matrix. We note that the eigenvalues of $\mathbf{P}$ must be real and positive, and we denote the largest eigenvalue by $\lambda_{\max}(\mathbf{P})$ and the smallest eigenvalue by $\lambda_{\min}(\mathbf{P})$**.** The time derivative of V along solutions of the systems is

$$\dot{V} = (\mathbf{y} - \widetilde{\mathbf{y}})^T \mathbf{P}[\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\widetilde{\mathbf{y}}, t)]$$

Suppose that the system is contracting, which means that there exists a positive definite symmetric matrix $\mathbf{P}$ and a constant $\gamma \geq 0$ so that

$$(\mathbf{y} - \widetilde{\mathbf{y}})^T \mathbf{P}[\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\widetilde{\mathbf{y}}, t)] \leq -\gamma(\mathbf{y} - \widetilde{\mathbf{y}})^T \mathbf{P}(\mathbf{y} - \widetilde{\mathbf{y}}), \quad \forall \mathbf{y}, \widetilde{\mathbf{y}}$$

This implies that

$$\dot{V} \leq -2\gamma V$$

and it follows that

$$V(t) \leq V(t_0) e^{-2\gamma(t - t_0)}$$

and that

$$\|\mathbf{y}(t) - \widetilde{\mathbf{y}}(t)\| \leq \left(\frac{\lambda_{\max}(\mathbf{P})}{\lambda_{\min}(\mathbf{P})}\right)^{\frac{1}{2}} \|\mathbf{y}(t_0) - \widetilde{\mathbf{y}}(t_0)\| e^{-\gamma(t - t_0)}$$

This means that the two solutions $\mathbf{y}(t)$ and $\widetilde{\mathbf{y}}(t)$ of a contracting system will converge exponentially to each other.

Suppose that the system (14.179) is contracting with $\mathbf{P} = \mathbf{I}$ so that

$$(\mathbf{y} - \widetilde{\mathbf{y}})^T [\mathbf{f}(\mathbf{y}, t) - \mathbf{f}(\widetilde{\mathbf{y}}, t)] \leq -\gamma(\mathbf{y} - \widetilde{\mathbf{y}})^T (\mathbf{y} - \widetilde{\mathbf{y}})$$

and that a numerical solution is computed using a Runge-Kutta method. Then, if the computed solutions $\mathbf{y}_{n+1}$ starting from $\mathbf{y}_n$ and $\widetilde{\mathbf{y}}_{n+1}$ starting from $\widetilde{\mathbf{y}}_n$ satisfies the condition

$$\|\mathbf{y}_{n+1} - \widetilde{\mathbf{y}}_{n+1}\| \leq \|\mathbf{y}_n - \widetilde{\mathbf{y}}_n\|$$

then the Runge-Kutta method is said to be B-stable.

Consider the linear time-varying test system

$$\dot{y} = \lambda(t) y \tag{14.181}$$

Then, for two solutions $y(t)$ and $\tilde{y}(t)$ we have

$$V = \frac{1}{2}(y-\tilde{y})^2 \Rightarrow \dot{V} = (y-\tilde{y})\lambda(t)(y-\tilde{y}) \tag{14.182}$$

It follows that if $\mathrm{Re}\,\lambda(t) \leq 0$ for all t, then the linear time invariant test equation (14.181) is contracting. Therefore, if a B-stable method is used the numerical solutions will satisfy $|y_{n+1} - \tilde{y}_{n+1}| \leq |y_n - \tilde{y}_n|$. As $\tilde{y}(t) = 0$ is a solution it follows that a B-stable method will also be AN-stable. We may then conclude that

$$\boxed{\text{B-stability}} \Rightarrow \boxed{\text{AN-stability}} \Rightarrow \boxed{\text{A-stability}} \tag{14.183}$$

14.6.10 Algebraic stability

The property of B-stability is important as it applies to nonlinear contracting systems, and as B-stability implies A-stability. It is problematic, however, to check if a method is B-stable by working with the nonnegative function V defined in (14.180). Therefore it is better to work with algebraic stability which can be established by algebraic manipulations of the elements of the Butcher array. In this section algebraic stability is defined, and it will be shown that algebraic stability implies B-stability.

An implicit Runge-Kutta method with Butcher array

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array}$$

is said to be algebraically stable if $b_i \geq 0$ for $i = 1, \ldots, \sigma$ and

$$\mathbf{M} = \mathrm{diag}\,(\mathbf{b})\,\mathbf{A} + \mathbf{A}^T \mathrm{diag}(\mathbf{b}) - \mathbf{b}\mathbf{b}^T \geq 0$$

We note that the elements of $\mathbf{M}$ are given by

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j \tag{14.184}$$

An algebraically stable Runge-Kutta method is B-stable, that is,

$$\boxed{\text{Algebraic stability}} \Rightarrow \boxed{\text{B-stability}} \tag{14.185}$$

This is shown as follows (Hairer et al. 1993): First we make a change of variables in the Runge-Kutta methods and write

$$\mathbf{Y}_i = \mathbf{y}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{f}\left(\mathbf{Y}_j, t_n + c_j h\right),\ i = 1, \ldots, \sigma \tag{14.186}$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\sum_{i=1}^{\sigma} b_i \mathbf{f}\left(\mathbf{Y}_i, t_n + c_i h\right) \tag{14.187}$$

Then we denote the differences between the two solutions $\mathbf{y}_n$ and $\widetilde{\mathbf{y}}_n$ by

$$\begin{aligned}\Delta\mathbf{y}_n &= \mathbf{y}_n - \widetilde{\mathbf{y}}_n, \Delta\mathbf{y}_{n+1} = \mathbf{y}_{n+1} - \widetilde{\mathbf{y}}_{n+1}, \Delta\mathbf{Y}_i = \mathbf{Y}_i - \widetilde{\mathbf{Y}}_i \\ \Delta\mathbf{f}_i &= h\left[\mathbf{f}\left(\mathbf{Y}_i, t_n + c_i h\right) - \mathbf{f}\left(\widetilde{\mathbf{Y}}_i, t_n + c_i h\right)\right]\end{aligned}$$

where $\mathbf{Y}_i$ is a vector corresponding to the vector $\mathbf{y}_n$, and $\widetilde{\mathbf{Y}}_i$ is a vector corresponding to the vector $\widetilde{\mathbf{y}}_n$. Subtraction of the Runge-Kutta equations for the solution $\widetilde{\mathbf{y}}_n$ from the equations of $\mathbf{y}_n$ gives

$$\begin{aligned}\Delta\mathbf{Y}_i &= \Delta\mathbf{y}_n + \sum_{j=1}^{\sigma} a_{ij}\Delta\mathbf{f}_j \\ \Delta\mathbf{y}_{n+1} &= \Delta\mathbf{y}_n + \sum_{i=1}^{\sigma} b_i\Delta\mathbf{f}_i\end{aligned}$$

Then, we have

$$\begin{aligned}\left(\Delta\mathbf{y}_{n+1}\right)^T\Delta\mathbf{y}_{n+1} &= \left(\Delta\mathbf{y}_n\right)^T\Delta\mathbf{y}_n + 2\sum_{i=1}^{\sigma} b_i\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{y}_n \\ &\quad + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i b_j\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{f}_j \\ &= \left(\Delta\mathbf{y}_n\right)^T\Delta\mathbf{y}_n + 2\sum_{i=1}^{\sigma} b_i\left(\Delta\mathbf{f}_i\right)^T\left(\Delta\mathbf{Y}_i - \sum_{j=1}^{\sigma} a_{ij}\Delta\mathbf{f}_j\right) \\ &\quad + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i b_j\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{f}_j \\ &= \left(\Delta\mathbf{y}_n\right)^T\Delta\mathbf{y}_n + 2\sum_{i=1}^{\sigma} b_i\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{Y}_i \\ &\quad - \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} m_{ij}\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{f}_j \qquad (14.188)\end{aligned}$$

where m_{ij} is element (i,j) of matrix $\mathbf{M}$. As

$$\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{Y}_i = h\left[\mathbf{f}\left(\mathbf{Y}_i, t_n + c_i h\right) - \mathbf{f}\left(\widetilde{\mathbf{Y}}_i, t_n + c_i h\right)\right]^T\left(\mathbf{Y}_i - \widetilde{\mathbf{Y}}_i\right) \leq 0$$

by assumption, and as $\sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} m_{ij}\left(\Delta\mathbf{f}_i\right)^T\Delta\mathbf{f}_j \geq 0$ for positive semidefinite $\mathbf{M}$, it follows that

$$\|\Delta\mathbf{y}_{n+1}\| \leq \|\Delta\mathbf{y}_n\|$$

which shows that an algebraically stable method is B-stable.

Example 231 *If there exists a positive definite symmetric matrix* $\mathbf{P}$ *so that*

$$\left[\mathbf{f}\left(\mathbf{y}, t\right) - \mathbf{f}\left(\widetilde{\mathbf{y}}, t\right)\right]^T\mathbf{P}\left(\mathbf{y} - \widetilde{\mathbf{y}}\right) = 0 \qquad (14.189)$$

then it follows from the derivation above that a Runge-Kutta method that satisfies

$$\mathbf{M} = \text{diag}\left(\mathbf{b}\right)\mathbf{A} + \mathbf{A}^T\text{diag}(\mathbf{b}) - \mathbf{b}\mathbf{b}^T = \mathbf{0}$$

will give

$$\left(\Delta\mathbf{y}_{n+1}\right)^T\mathbf{P}\Delta\mathbf{y}_{n+1} = \left(\Delta\mathbf{y}_n\right)^T\mathbf{P}\Delta\mathbf{y}_n \qquad (14.190)$$

Method	Order	Stability function	Linear stability	Nonlinear stability	Stiffly Accurate
Explicit, $p = \sigma$	σ	P_0^σ	$\|h\lambda\|$ small	-	No
Gauss	2σ	P_σ^σ	A	Algebraic	No
Radau IA	$2\sigma - 1$	$P_\sigma^{\sigma-1}$	L	Algebraic	No
Radau IIA	$2\sigma - 1$	$P_\sigma^{\sigma-1}$	L	Algebraic	Yes
Lobatto IIIA	$2\sigma - 2$	$P_{\sigma-1}^{\sigma-1}$	A	not AN	No
Lobatto IIIB	$2\sigma - 2$	$P_{\sigma-1}^{\sigma-1}$	A	not AN	No
Lobatto IIIC	$2\sigma - 2$	$P_\sigma^{\sigma-2}$	L	Algebraic	Yes

Table 14.4: Order and stability properties for some important Runge-Kutta methods.

14.6.11 Properties of Runge-Kutta methods

The properties of some important Runge-Kutta methods are summarized in Table 14.4.

14.7 Automatic adjustment of step size

14.7.1 Estimation of the local error for Runge-Kutta methods

The selection of the step size h is critical for the performance of a Runge-Kutta method. The main issues in this connection is accuracy and stability. In general the accuracy of the computed solution depends on the step size. We will see in this section that it is possible to specify the desired accuracy of the computed solution, and then to have automatic selection of the step size that ensures the required accuracy. This feature is used in the standard integrators of MATLAB.

In some applications it may be desirable for simplicity to compute the solution with a constant step size. For explicit methods the step size must then be selected so that the computations are stable. For non-stiff systems that are approximately linear in the sense that the eigenvalues of the Jacobian $\mathbf{J}$ do not vary much, it will normally be possible to select a reasonable step size that ensures stability and a certain accuracy. For systems with strong nonlinearities so that the eigenvalues of $\mathbf{J}$ exhibit large variations, the step size of an explicit Runge-Kutta method may have to be very small to account for worst-case situations. For such systems the use of a constant step size is not recommended.

The automatic selection of the step size h is based on finding an estimate of the local error, and then adjusting the time step so that the local error is less than some specified tolerance. This is done by computing the numerical solution with two explicit Runge-Kutta method with different order. Assume that the solution $\mathbf{y}_{n+1}$ is computed with a method

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array}$$

of order p, and that the solution $\widehat{\mathbf{y}}_{n+1}$ of the same system is computed with a method

$$\begin{array}{c|c} \widehat{\mathbf{c}} & \widehat{\mathbf{A}} \\ \hline & \widehat{\mathbf{b}}^T \end{array}$$

of order $\widehat{p} = p + 1$. The computation of the value at t_{n+1} starts with $\mathbf{y}_n = \widehat{\mathbf{y}}_n$. Then the

local solution $\mathbf{y}_L(t_n; t_{n+1})$ satisfies

$$\mathbf{y}_L(t_n; t_{n+1}) = \mathbf{y}_{n+1} + \mathbf{e}_{n+1} = \widehat{\mathbf{y}}_{n+1} + \widehat{\mathbf{e}}_{n+1}$$

where $\mathbf{e}_{n+1} = O(h^{p+1})$ is the local error in the computation of $\mathbf{y}_{n+1}$, and $\widehat{\mathbf{e}}_{n+1} = O(h^{p+2})$ is the local error in the computation of $\widehat{\mathbf{y}}_{n+1}$. Because $\widehat{\mathbf{e}}_{n+1}$ is of higher order in h than $\mathbf{e}_{n+1}$, we can find an estimate of $\mathbf{e}_{n+1}$ from

$$\widehat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1} = \mathbf{e}_{n+1} - \widehat{\mathbf{e}}_{n+1} \approx \mathbf{e}_{n+1}$$

The step size can then be adjusted to achieve a specified accuracy in the local error $\mathbf{e}_{n+1}$.

The estimated local error $\mathbf{e}_{n+1}$ is an estimate of the local error of the lower order solution $\mathbf{y}_{n+1}$. However, the solution $\widehat{\mathbf{y}}_{n+1}$ is more accurate, so it makes more sense to use $\widehat{\mathbf{y}}_{n+1}$ as a starting point for the next time step. The use of $\widehat{\mathbf{y}}_{n+1}$ instead of $\mathbf{y}_{n+1}$ is called *local extrapolation*, and is normally used. When local extrapolation is used then $\mathbf{y}_{n+1}$ will be used to denote the high order solution, while $\widehat{\mathbf{y}}_{n+1}$ denotes the embedded low order solution.

To make the computations efficient, the two methods are usually designed so that $\mathbf{c} = \widehat{\mathbf{c}}$ and $\mathbf{A} = \widehat{\mathbf{A}}$. Then the stage computations will be the same in both methods, and need only be done once. The solution $\widehat{\mathbf{y}}$ is said to be an *embedded solution* in this case. The algorithm is

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(\mathbf{y}_n, t_n) && (14.191)\\
&\vdots && (14.192)\\
\mathbf{k}_\sigma &= \mathbf{f}(\mathbf{y}_n + h\left(a_{\sigma 1}\mathbf{k}_1 + \ldots + a_{\sigma,\sigma-1}\mathbf{k}_{\sigma-1}\right), t_n + c_\sigma h) && (14.193)\\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h(b_1\mathbf{k}_1 + \ldots + b_\sigma\mathbf{k}_\sigma) && (14.194)\\
\widehat{\mathbf{y}}_{n+1} &= \mathbf{y}_n + h(\widehat{b}_1\mathbf{k}_1 + \ldots + \widehat{b}_\sigma\mathbf{k}_\sigma) && (14.195)\\
\mathbf{e}_{n+1} &= \widehat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1} && (14.196)
\end{aligned}$$

The computation of $\mathbf{y}_{n+1}$ and $\widehat{\mathbf{y}}_{n+1}$ is described by an array as follows:

$\mathbf{c}$	$\mathbf{A}$
$\mathbf{y}$	$\mathbf{b}^T$
$\widehat{\mathbf{y}}$	$\widehat{\mathbf{b}}^T$
$\mathbf{e}$	$\mathbf{E}^T$

where $\mathbf{E} = \widehat{\mathbf{b}} - \mathbf{b}$.

Runge-Kutta-Fehlberg 4(5) is a method where $\mathbf{y}$ is computed with order $p = 4$ using five stages. The embedded solution $\widehat{\mathbf{y}}$ is of order $p = 5$ and is computed using six stages. The method is optimized for accuracy in the fourth order solution $\mathbf{y}_{n+1}$. The method is

given by the following array.

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
y	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	
$\widehat{y}$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
Δe	$\frac{1}{360}$	0	$-\frac{128}{4275}$	$-\frac{2197}{75240}$	$\frac{1}{50}$	$\frac{2}{55}$

Dormand-Prince 5(4) is a method where $\mathbf{y}_{n+1}$ is computed with order $p = 5$. This requires six stages. The embedded solution $\widehat{\mathbf{y}}$ is of order $p = 4$ and is computed using seven stages. The seventh stage is $\mathbf{k}_7 = \mathbf{y}_{n+1}$ to reduce the number of computations. This is recognized as a FSAL method. The method is optimized for accuracy in the fifth order solution $\mathbf{y}_{n+1}$. This is the standard MATLAB method for integrating initial value problems (Shampine and Reichelt 1997). The method is given by the following array.

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
y	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
$\widehat{y}$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$
Δe	$\frac{71}{57600}$	0	$-\frac{71}{16695}$	$\frac{71}{1920}$	$-\frac{17253}{339200}$	$\frac{22}{525}$	$-\frac{1}{40}$

Another variable-step explicit Runge-Kutta method used in MATLAB is the BS23 method of Bogacki and Shampine (Shampine and Reichelt 1997). This is a method where $\mathbf{y}_{n+1}$ is computed with a third order method, and the error estimate is found by comparing the result with an embedded second order method. Also here local extrapolation is used. The Butcher array is

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	0	$\frac{3}{4}$		
1	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
y	$\frac{21}{72}$	$\frac{1}{4}$	$\frac{3}{9}$	$\frac{1}{8}$
$\widehat{y}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$	
Δe	$-\frac{5}{72}$	$\frac{1}{12}$	$\frac{1}{9}$	$-\frac{1}{8}$

14.7.2 Adjustment algorithm

Suppose that the specified accuracy is specified in terms of a tolerance e_{tol} on the local error, and that the size of the local error is described by $\varepsilon_{n+1} = |e_{i,n+1}|$ where $e_{i,n+1}$ is the element of highest absolute value of the vector $\mathbf{e}_{n+1}$. Then, because the method is of order p we will have $\varepsilon_{n+1} \leq Ch^{p+1}$ for some constant C. Let h_{new} be defined by $e_{\text{tol}} = Ch_{\text{new}}^{p+1}$. Then, if $\varepsilon_{n+1} > e_{\text{tol}}$ the local error is larger than the specified tolerance. We may then expect that the tolerance can be obtained by using the new and smaller time step

$$h_{\text{new}} = h \left(\frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{\frac{1}{p+1}} \tag{14.197}$$

In practice a somewhat smaller value may be used by adjusting with a factor of about 0.8. If the tolerance is met, then the time step can be carefully increased.

An alternative adjustment algorithm that has given good results is based on a PI control method. The derivation of this algorithm is based on the resulting equation when the logarithm of the adjustment algorithm (14.197) is taken:

$$\ln h_{\text{new}} = \ln h - \frac{1}{p+1} \left(\ln \varepsilon_{n+1} - \ln e_{\text{tol}} \right) \tag{14.198}$$

This can be compared with an incremental form of a PI controller

$$u_{n+1} = u_n - K_p \left(e_n - e_{n-1} \right) - K_p \frac{h}{T_i} e_n \tag{14.199}$$

One may compare the adjustment formula with an I controller. Proportional action is included using

$$\ln h_{\text{new}} = \ln h - K_p \left(\ln \varepsilon_{n+1} - \ln \varepsilon_n \right) - K_p \frac{h}{T_i} \left(\ln \varepsilon_{n+1} - \ln e_{\text{tol}} \right) \tag{14.200}$$

which gives the adjustment formula

$$h_{\text{new}} = h \left(\frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{K_p \frac{h}{T_i}} \left(\frac{\varepsilon_n}{\varepsilon_{n+1}} \right)^{K_p} \tag{14.201}$$

which is simplified to

$$h_{\text{new}} = h \left(\frac{e_{\text{tol}}}{\varepsilon_{n+1}} \right)^{K_p \left(1 + \frac{h}{T_i} \right)} \left(\frac{\varepsilon_n}{e_{\text{tol}}} \right)^{K_p} \tag{14.202}$$

The following parameters have been suggested.

$$K_p = 0.4/(p+1), \quad T_i = 1.3h \tag{14.203}$$

14.8 Implementation aspects

14.8.1 Solution of implicit equations

The implicit Runge-Kutta methods involves the solution of a set of implicit nonlinear equations. To solve these equations it is useful to make a change of variables and write

the stage computations in the form

$$\begin{aligned}\mathbf{z}_1 &= h\left[a_{11}\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_1,t_n+c_1h\right)+\ldots+a_{1\sigma}\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_\sigma,t_n+c_\sigma h\right)\right]\\ &\vdots\\ \mathbf{z}_\sigma &= h\left[a_{\sigma 1}\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_1,t_n+c_1h\right)+\ldots+a_{\sigma\sigma}\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_\sigma,t_n+c_\sigma h\right)\right]\end{aligned}$$

and to compute the solution $\mathbf{y}_{n+1}$

$$\mathbf{y}_{n+1}=\mathbf{y}_n+h\left[b_1\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_1,t_n+c_1h\right)+\ldots+b_\sigma\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_\sigma,t_n+c_\sigma h\right)\right]$$

Example 232 *If* $\mathbf{A}$ *is nonsingular, the update can be found from*

$$\mathbf{y}_{n+1}=\mathbf{y}_n+d_1\mathbf{z}_1+\ldots+d_\sigma\mathbf{z}_\sigma \tag{14.204}$$

where

$$\left(d_1,\ldots,d_\sigma\right)=\left(b_1,\ldots,b_\sigma\right)\mathbf{A}^{-1} \tag{14.205}$$

In particular, if $a_{\sigma i}=b_i$ *then*

$$\mathbf{y}_{n+1}=\mathbf{y}_n+\mathbf{z}_\sigma \tag{14.206}$$

To solve for $\mathbf{z}_1,\ldots,\mathbf{z}_\sigma$, a Newton search method is used. The equation is written in vector form as

$$\mathbf{Z}=h\left(\mathbf{A}\otimes\mathbf{I}_\sigma\right)\mathbf{F}(\mathbf{Z})$$

where

$$\mathbf{Z}=\begin{pmatrix}\mathbf{z}_1\\ \vdots\\ \mathbf{z}_\sigma\end{pmatrix},\ \mathbf{F}(\mathbf{Z})=\begin{pmatrix}\mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_1,t_n+c_1h\right)\\ \vdots\\ \mathbf{f}\left(\mathbf{y}_n+\mathbf{z}_\sigma,t_n+c_\sigma h\right)\end{pmatrix}$$

are vectors of dimension $d\sigma$, $\mathbf{I}_\sigma$ is the $\sigma\times\sigma$ identity matrix and

$$\mathbf{A}\otimes\mathbf{I}_\sigma=\begin{pmatrix}a_{11}\mathbf{I}_\sigma & \ldots & a_{\sigma 1}\mathbf{I}_\sigma\\ \vdots & \ddots & \vdots\\ a_{1\sigma}\mathbf{I}_\sigma & \ldots & a_{\sigma\sigma}\mathbf{I}_\sigma\end{pmatrix}$$

is the Kronecker tensor product of $\mathbf{A}$ and $\mathbf{I}_\sigma$**.**

The solution is found by minimizing the function

$$L=\left[\mathbf{Z}-h\left(\mathbf{A}\otimes\mathbf{I}_\sigma\right)\mathbf{F}(\mathbf{Z})\right]^T\left[\mathbf{Z}-h\left(\mathbf{A}\otimes\mathbf{I}_\sigma\right)\mathbf{F}(\mathbf{Z})\right] \tag{14.207}$$

with respect to $\mathbf{Z}$ using a Newton search, which is done by the iteration

$$\mathbf{H}\left(\mathbf{Z}^{i+1}-\mathbf{Z}^i\right)=-\mathbf{Z}^i+h\left(\mathbf{A}\otimes\mathbf{I}_\sigma\right)\mathbf{F}(\mathbf{Z}^i) \tag{14.208}$$

which is solved for $\mathbf{Z}^{i+1}$. Here $\mathbf{Z}^i$ is iteration i of $\mathbf{Z}$**,** and

$$\mathbf{H}=\mathbf{I}-h\left(\mathbf{A}\otimes\mathbf{J}\right)=\begin{pmatrix}1-ha_{11}\mathbf{J} & \ldots & -ha_{\sigma 1}\mathbf{J}\\ \vdots & \ddots & \vdots\\ -ha_{1\sigma}\mathbf{J} & \ldots & 1-ha_{\sigma\sigma}\mathbf{J}\end{pmatrix}$$

is an approximation to the Hessian matrix of dimension $n\sigma\times n\sigma$, where

$$\mathbf{J}=\frac{\partial\mathbf{f}}{\partial\mathbf{y}}(\mathbf{y}_n,t_n)$$

is the Jacobian evaluated at $(\mathbf{y}_n, t_n)$. The initial value for the iterations is $\mathbf{Z}^0 = \mathbf{0}$. To solve for $\mathbf{Z}^{i+1}$, a Gaussian elimination is used where the LU decomposition of $\mathbf{H}$ is needed. The reason for using the approximation of a constant $\mathbf{J}$ is that this makes it possible to use only one LU decomposition at each time step. Details on how to do the Gaussian elimination is given in (Golub and van Loan 1989) where also algorithms are included.

14.8.2 Dense outputs

A Runge-Kutta method computes the numerical solution $\ldots \mathbf{y}_{n-1}, \mathbf{y}_n, \mathbf{y}_{n+1} \ldots$ at discrete time instants $\ldots t_{n-1}, t_n, t_{n+1} \ldots$ for the system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y},t), \quad \mathbf{y}(t_0) = \mathbf{y}_0 \tag{14.209}$$

In some situations it is not sufficient to have the function values only at the time-steps. The reason for this is that for some systems it is very important to detect the exact time of certain events. In particular this is important for systems with discontinuities in $\mathbf{f}(\mathbf{y},t)$. In addition it may be desirable to have function values between the timesteps for plotting. The solution to this problem is to use an interpolation scheme where an interpolation $\mathbf{y}_n(\alpha)$ is computed so that

$$\mathbf{y}_n(\alpha), \;\; \alpha \in [0,1], \;\; \mathbf{y}_n(0) = \mathbf{y}_n, \quad \mathbf{y}_n(1) = \mathbf{y}_{n+1} \tag{14.210}$$

This can be done by using the original stage computations of the Runge-Kutta method, possibly with some additional stages, and then interpolating the solution by interpolating the weighting factors b_j. The resulting scheme is called a *continuous Runge-Kutta method*.

A continuous Runge-Kutta method is a Runge-Kutta method where interpolation is used to compute *dense outputs* $\mathbf{y}_n(\alpha), \;\; \alpha \in [0,1]$ from the scheme

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_n + h\sum_{j=1}^{\sigma^*} a_{ij}\mathbf{k}_j, t_n + c_i h), \; i = 1, \ldots, \sigma^* \tag{14.211}$$

$$\mathbf{y}_n(\alpha) = \mathbf{y}_n + h\sum_{j=1}^{\sigma^*} b_j(\alpha)\mathbf{k}_j \tag{14.212}$$

The dense outputs are of order p^* if $\mathbf{y}_n(\alpha) - \mathbf{y}_L(t_n; t_n + \alpha h) = O(h^{p^*+1})$, where $\mathbf{y}_L(t_n; t_n + \alpha h)$ is the local solution starting at $\mathbf{y}_L(t_n; t_n) = \mathbf{y}_n$.

For the Dormand-Prince 5(4) method, which is the numerical integration method of the ode45 in MATLAB, a dense output with order 4 can be computed with the original stage computations using Hermite interpolation (Dormand and Prince 1986), (Hairer

et al. 1993). The weighting factors of the method are given by the Hermite polynomials.

$$b_1(\alpha) = \alpha^2(3-2\alpha)b_1 + \alpha(\alpha-1)^2 - \alpha^2(\alpha-1)^2\frac{5(2,558,722,523-31,403,016\alpha)}{11,282,082,432} \quad (14.213)$$

$$b_2(\alpha) = 0 \quad (14.214)$$

$$b_3(\alpha) = \alpha^2(3-2\alpha)b_3 + \alpha^2(\alpha-1)^2\frac{100(882,725,551-15,701,508\alpha)}{32,700,410,799} \quad (14.215)$$

$$b_4(\alpha) = \alpha^2(3-2\alpha)b_4 - \alpha^2(\alpha-1)^2\frac{25(443332067-31,403,016\alpha)}{1,880,347,072} \quad (14.216)$$

$$b_5(\alpha) = \alpha^2(3-2\alpha)b_5 + \alpha^2(\alpha-1)^2\frac{32805(23,143,187-3,489,224\alpha)}{199,316,789,632} \quad (14.217)$$

$$b_6(\alpha) = \alpha^2(3-2\alpha)b_6 - \alpha^2(\alpha-1)^2\frac{55(29,972,135-7,076,736\alpha)}{822,651,844} \quad (14.218)$$

$$b_7(\alpha) = \alpha^2(\alpha-1) + \alpha^2(\alpha-1)^2\frac{10(7,414,447-829,305\alpha)}{29,380,432} \quad (14.219)$$

Note that $b_j(0) = 0$, and that $b_j(1) = b_j$, where b_j are the coefficients of the fifth order solution in the Dormand-Prince 5(4) method. It is therefore clear that $\mathbf{y}_n(0) = \mathbf{y}_n$ and $\mathbf{y}_n(1) = \mathbf{y}_{n+1}$. In addition, it can be shown that the time derivatives of the dense solutions satisfy $\dot{\mathbf{y}}_n(0) = h\mathbf{f}(\mathbf{y}_n, t_n)$ and $\dot{\mathbf{y}}_n(1) = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})$. This means that the dense outputs and their derivatives are continuous at the time-steps t_n.

14.8.3 Event detection

Event detection can be formulated as a *zero crossing* problem by defining a function g so that the event is given by the condition

$$g(\mathbf{y},t) = 0 \quad (14.220)$$

The event can then be detected by computing the numerical solution $\mathbf{y}_n$ and for each step check if there is a change of sign from $g(\mathbf{y}_n,t_n)$ to $g(\mathbf{y}_{n+1},t_{n+1})$. If there is a change in sign, then the dense output $\mathbf{y}_n(\alpha)$ is used to find the time of event by solving

$$g[\mathbf{y}_n(\alpha),t+\alpha h] = 0 \quad (14.221)$$

numerically for α. Then the time of the event is given by $t_n + \alpha h$.

This type of event detection can be used for systems with signum terms in $\mathbf{f}(\mathbf{y},t)$, as for problems with dry friction. Then the event that the velocity becomes zero, or leaves zero, may be detected with this method.

14.8.4 Systems with inertia matrix

There are important applications where the differential equation may be in the form

$$\mathbf{M}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u}) \quad (14.222)$$

where $\mathbf{M}$ is a nonsingular matrix. Runge-Kutta methods can the be implemented with the stage computations

$$\begin{aligned}\mathbf{k}_i &= \boldsymbol{\phi}(\mathbf{u}_n + \sum_{j=1}^{i-1} a_{ij}\mathbf{k}_j, t_n + c_i h),\ i = 1, \ldots, \sigma \\ \mathbf{u}_{n+1} &= \mathbf{u}_n + \mathbf{M}^{-1}\left(h\sum_{j=1}^{\sigma} b_j \mathbf{k}_j\right)\end{aligned}$$

Example 233 *One example of this is in robotics where the equation of motion is of the form*

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \boldsymbol{\tau}$$

where $\mathbf{q}$ *is the vector of generalized coordinates and* $\boldsymbol{\tau}$ *is the vector of input generalized forces. The matrix* $\mathbf{M}$*, which is called the inertia matrix is positive definite and symmetric. The system can be written*

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}(\mathbf{q}) \end{pmatrix}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u})$$

where

$$\mathbf{u} = \begin{pmatrix} \mathbf{q} \\ \dot{\mathbf{q}} \end{pmatrix},\ \boldsymbol{\phi}(\mathbf{u}) = \begin{pmatrix} \dot{\mathbf{q}} \\ -\mathbf{C}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau} \end{pmatrix}$$

The system could have been written in the form that has been used so far, that is,

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} \dot{\mathbf{q}} \\ \mathbf{M}\left(\mathbf{q}\right)^{-1}\left[-\mathbf{C}\left(\mathbf{q}, \dot{\mathbf{q}}\right)\dot{\mathbf{q}} - \mathbf{g}(\mathbf{q}) + \boldsymbol{\tau}\right] \end{pmatrix}$$

Then the evaluation of $\mathbf{f}(\mathbf{y})$ *would involve a computationally expensive Gauss elimination. Therefore it is advantageous to leave the system in the form (14.222) and do a slight modification to the Runge-Kutta algorithm.*

14.9 Invariants

14.9.1 Introduction

The material presented on linear and quadratic invariants in this section is based on (Hairer 1999), while the section of Hamiltonian systems is based on (Sanz-Serna and Calvo 1994) and (Hairer 1999).

14.9.2 Linear invariants

Suppose that there is a function

$$L\left(\mathbf{y}\right) = \mathbf{w}^T\mathbf{y} \tag{14.223}$$

where $\mathbf{w} = \left(w_1 \ldots w_d\right)^T$ is a vector of constants, so that for all $\mathbf{y}$ the time derivative along solutions of the system $\dot{\mathbf{y}} = \mathbf{f}\left(\mathbf{y}, t\right)$ is zero, that is

$$\dot{L}\left(\mathbf{y}\right) := \mathbf{w}^T\dot{\mathbf{y}} = \mathbf{w}^T\mathbf{f}\left(\mathbf{y}, t\right) = 0 \quad \text{for all } \mathbf{y} \tag{14.224}$$

with step size control using a FSAL computation

$$\begin{aligned}\mathbf{V}\mathbf{k}_3 &= 2\mathbf{f}\left(\mathbf{y}_n, t_n\right) + \left(6+\sqrt{2}\right)\mathbf{f}\left(\mathbf{y}_n + \frac{h}{2}\mathbf{k}_1, t_n + \frac{h}{2}\right) + \mathbf{f}\left(\mathbf{y}_{n+1}, t_{n+1}\right) \\ &\quad -2\mathbf{k}_1 - \left(6+\sqrt{2}\right)\mathbf{k}_2 + h\rho\dot{\mathbf{f}}\left(\mathbf{y}_n, t_n\right) \\ \text{error} &= \frac{h}{6}\left(\mathbf{k}_1 - 2\mathbf{k}_2 + \mathbf{k}_3\right)\end{aligned}$$

This method is similar to a Rosenbrock method, but the computation of the second stage has a term of the type $-\mathbf{k}_1$ instead of $h\mathbf{J}\rho_{21}\mathbf{k}_1$. This method is used in the MATLAB function ode23s (Shampine and Reichelt 1997).

14.11 Multistep methods

14.11.1 Explicit Adams methods

The explicit Adams methods, also called Adams-Bashforth methods, has the equation

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{y}(t), t)dt$$

as a starting point. The idea is to calculate a numerical solution from the approximation

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \int_{t_n}^{t_{n+1}} \mathbf{P}(t)dt$$

where $\mathbf{P}(t)$ is a polynomial approximation of $\mathbf{f}$ of order q so that

$$\mathbf{P}(t_{n+1-i}) = \mathbf{f}\left(\mathbf{y}_{n+1-i}, t_{n+1-i}\right) =: \mathbf{f}_{n+1-i}, \quad i = 1, 2, \ldots, q. \tag{14.279}$$

This is done with the polynomial

$$\mathbf{P}(t) = \sum_{i=1}^{q} \mathbf{f}_{n+1-i} L_i(t)$$

where $L_i(t)$, $i = 1, \ldots, q$ are the fundamental Lagrange polynomials (Shampine et al. 1997)

$$L_i(t) = \prod_{j=1,\, j\neq i}^{q} \left(\frac{t - t_{n+1-j}}{t_{n+1-i} - t_{n+1-j}}\right), \quad i = 1, \ldots, q$$

These polynomials have the property that

$$L_i(t_{n+1-j}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

as shown in Figure 14.32. The polynomial $P(t)$ is shown in Figure 14.33

It is convenient to describe the methods in terms of backward differences. To do this we define the backward difference operator ∇ by

$$\nabla\mathbf{y}_n = \mathbf{y}_n - \mathbf{y}_{n-1}$$

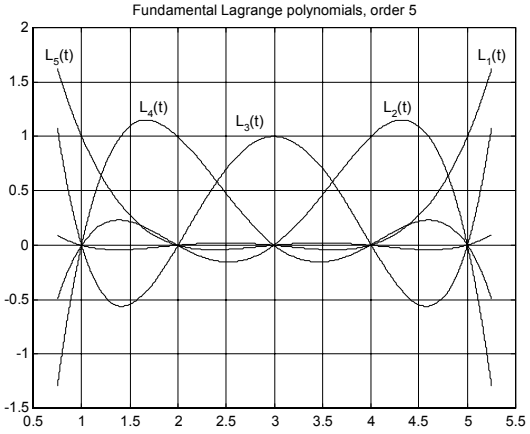

Figure 14.32: The Lagrange ploynomials $L_i(t_{6-j})$ for $i = 1, \ldots, 5$ when $h = 1$.

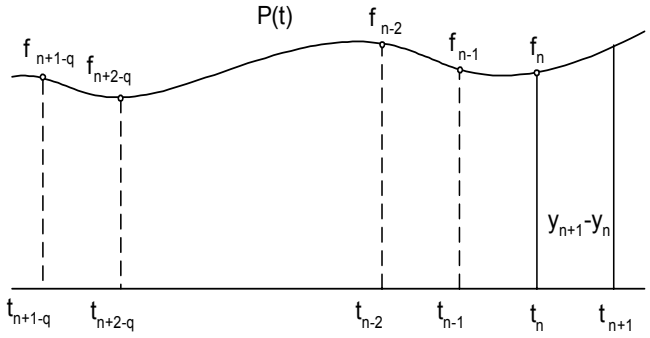

Figure 14.33: The explicit Adams method

Repeated use of the backward difference operator gives

$$\nabla^{m+1}\mathbf{y}_n = \nabla\left(\nabla^m \mathbf{y}_n\right) = \nabla^m \mathbf{y}_n - \nabla^m \mathbf{y}_{n-1}$$

for $m = 0, 1, 2, \ldots$ where

$$\nabla^0 \mathbf{y}_n = \mathbf{y}_n$$

A constant h is assumed. Then in the interval $t_n \leq t \leq t_{n+1}$ the polynomial $\mathbf{P}(t)$ can be written using a Newton interpolation formula

$$\mathbf{P}(t_n + \alpha h) = \sum_{m=0}^{q-1} \frac{\alpha(\alpha+1)\ldots(\alpha+m-1)}{m!} \nabla^m \mathbf{f}_n$$

This leads to the explicit Adams method of order q :

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{m=0}^{q-1} \gamma_m \nabla^m \mathbf{f}(\mathbf{y}_n, t_n)$$

where

$$\gamma_m = \int_0^1 \frac{\alpha(\alpha+1)\ldots(\alpha+m-1)}{m!} d\alpha$$

It can be shown that γ_m can be found recursively from the recurrence equation

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \ldots + \frac{1}{m+1}\gamma_0 = 1$$

The numerical values for γ_m are found from the recurrence equation to be

m	0	1	2	3	4
γ_m	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$

The numerical algorithms are found by inserting the expression for the backwards difference operator. The algorithms are

$$\begin{aligned}
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\mathbf{f}_n \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{3}{2}\mathbf{f}_n - \frac{1}{2}\mathbf{f}_{n-1}\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{23}{12}\mathbf{f}_n - \frac{4}{3}\mathbf{f}_{n-1} + \frac{5}{12}\mathbf{f}_{n-2}\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{55}{24}\mathbf{f}_n - \frac{59}{24}\mathbf{f}_{n-1} + \frac{37}{24}\mathbf{f}_{n-2} - \frac{9}{24}\mathbf{f}_{n-3}\right)
\end{aligned}$$

We see that the first order explicit Adams method is Euler's method.

14.11.2 Implicit Adams methods

In implicit Adams methods, which are also called Adams-Moulton methods, the approximating polynomial $\mathbf{P}(t)$ is required to satisfy

$$\mathbf{P}(t_{n+1-i}) = \mathbf{f}\left(\mathbf{y}_{n+1-i}, t_{n+1-i}\right), \quad i = 0, 1, \ldots, q-1 \tag{14.280}$$

as shown in Figure 14.34.

This is achieved with

$$\mathbf{P}^*(t_n + \alpha h) = \sum_{m=0}^{q} \frac{(\alpha-1)\alpha(\alpha+1)\ldots(\alpha+m-2)}{m!} \nabla^m \mathbf{f}_{n+1}$$

This gives the implicit Adams method of order $q+1$:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\sum_{m=0}^{q} \gamma_m^* \nabla^m \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})$$

where

$$\gamma_m^* = \int_0^1 \frac{(\alpha-1)\alpha(\alpha+1)\ldots(\alpha+m-2)}{m!} d\alpha$$

Numerical values are

m	0	1	2	3	4
γ_m^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$

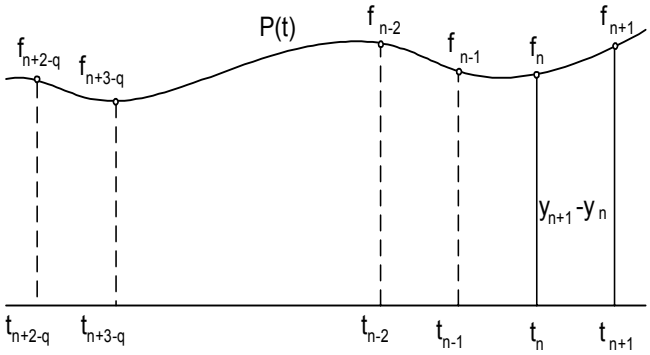

Figure 14.34: The implicit Adams method

which can also be found from the recurrence equation

$$\gamma_m^* + \frac{1}{2}\gamma_{m-1}^* + \ldots + \frac{1}{m+1}\gamma_0^* = 0,\ m \geq 0,\ \ \gamma_0^* = 1$$

The resulting algorithms are

$$\begin{aligned}
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\mathbf{f}_{n+1} \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{1}{2}\mathbf{f}_{n+1} + \frac{1}{2}\mathbf{f}_n\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{5}{12}\mathbf{f}_{n+1} + \frac{8}{12}\mathbf{f}_n - \frac{1}{12}\mathbf{f}_{n-1}\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{9}{24}\mathbf{f}_{n+1} + \frac{19}{24}\mathbf{f}_n - \frac{5}{24}\mathbf{f}_{n-1} + \frac{1}{24}\mathbf{f}_{n-2}\right)
\end{aligned}$$

It is seen that the first order implicit Adams method is the implicit Euler method, and that the second order implicit Adams method is the trapezoidal rule.

14.11.3 Predictor-Corrector implementation

An approximate implementation of the implicit Adams method is based on computing a predictor

$$\widehat{\mathbf{y}}_{n+1} = \mathbf{y}_n + h\sum_{m=0}^{q-1} \gamma_m \nabla^m \mathbf{f}(t_n, x_n)$$

with the explicit Adams method, and then use $\hat{\mathbf{f}}_{n+1} := \mathbf{f}(t_{n+1}, \widehat{\mathbf{y}}_{n+1})$ in the place of $\mathbf{f}_{n+1}$ in the implicit Adams method. This gives

$$\begin{aligned}
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\hat{\mathbf{f}}_{n+1} \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{1}{2}\hat{\mathbf{f}}_{n+1} + \frac{1}{2}\mathbf{f}_n\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{5}{12}\hat{\mathbf{f}}_{n+1} + \frac{8}{12}\mathbf{f}_n - \frac{1}{12}\mathbf{f}_{n-1}\right) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\left(\frac{9}{24}\hat{\mathbf{f}}_{n+1} + \frac{19}{24}\mathbf{f}_n - \frac{5}{24}\mathbf{f}_{n-1} + \frac{1}{24}\mathbf{f}_{n-2}\right)
\end{aligned}$$

This is called a Predictor-Corrector method, which is abbreviated to PECE.

14.11.4 Backwards differentiation methods

In the Backwards Differentiation Formula (BDF) the vector $\mathbf{P}(t)$ of polynomials of order q is required to satisfy the $q+1$ constraints

$$\mathbf{P}(t_{n-q+1}) = \mathbf{y}_{n-q+1}, \ldots, \mathbf{P}(t_n) = \mathbf{y}_n,\ \mathbf{P}(t_{n+1}) = \mathbf{y}_{n+1} \tag{14.281}$$

In this method, the numerical solution at t_{n+1} is generated by requiring the polynomial $\mathbf{P}(t)$ to satisfy

$$\dot{\mathbf{P}}(t_{n+1}) = \mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})$$

as shown in Figure 14.35. This is done with the Newton interpolating polynomial

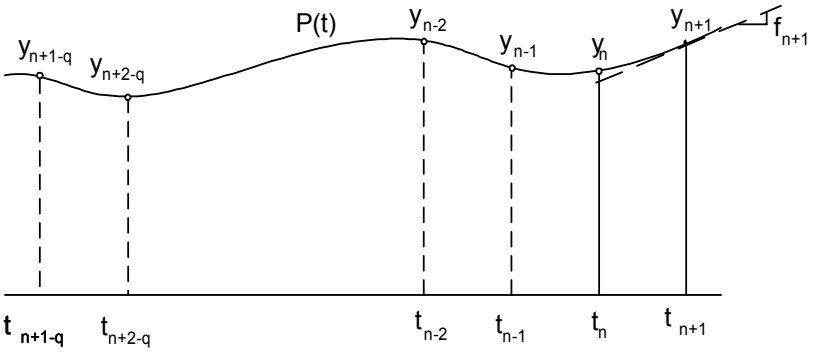

Figure 14.35: The BDF method

$$\mathbf{P}(t_n + \alpha h) = \left(1 + \sum_{m=1}^{q} \frac{(\alpha-1)\alpha(\alpha+1)\ldots(\alpha+m-2)}{m!}\nabla^m\right)\mathbf{y}_{n+1}$$

From

$$\begin{aligned}\left.\frac{d}{d\alpha}\mathbf{P}(t_n+\alpha h)\right|_{\alpha=1} &= \sum_{m=1}^{q}\left.\frac{d}{d\alpha}\frac{(\alpha-1)\alpha(\alpha+1)\ldots(\alpha+m-2)}{m!}\right|_{\alpha=1}\nabla^m\mathbf{y}_{n+1} \\ &= \sum_{m=1}^{q}\frac{1}{m}\nabla^m\mathbf{y}_{n+1}\end{aligned}$$

the BDF method of order q is found to be

$$\sum_{m=1}^{q}\frac{1}{m}\nabla^m\mathbf{y}_{n+1} = h\mathbf{f}(\mathbf{y}_{n+1}, t_{n+1})$$

This gives the following algorithms for $q = 1, \ldots, 6$:

$$\begin{aligned}
\mathbf{y}_{n+1} - \mathbf{y}_n &= h\mathbf{f}_{n+1} \\
\frac{3}{2}\mathbf{y}_{n+1} - 2\mathbf{y}_n + \frac{1}{2}\mathbf{y}_{n-1} &= h\mathbf{f}_{n+1} \\
\frac{11}{6}\mathbf{y}_{n+1} - 3\mathbf{y}_n + \frac{3}{2}\mathbf{y}_{n-1} - \frac{1}{3}\mathbf{y}_{n-2} &= h\mathbf{f}_{n+1} \\
\frac{25}{12}\mathbf{y}_{n+1} - 4\mathbf{y}_n + 3\mathbf{y}_{n-1} - \frac{4}{3}\mathbf{y}_{n-2} + \frac{1}{4}\mathbf{y}_{n-3} &= h\mathbf{f}_{n+1} \\
\frac{137}{60}\mathbf{y}_{n+1} - 5\mathbf{y}_n + 5\mathbf{y}_{n-1} - \frac{10}{3}\mathbf{y}_{n-2} + \frac{5}{4}\mathbf{y}_{n-3} - \frac{1}{5}\mathbf{y}_{n-4} &= h\mathbf{f}_{n+1} \\
\frac{147}{60}\mathbf{y}_{n+1} - 6\mathbf{y}_n + \frac{15}{2}\mathbf{y}_{n-1} - \frac{20}{3}\mathbf{y}_{n-2} + \frac{15}{4}\mathbf{y}_{n-3} - \frac{6}{5}\mathbf{y}_{n-4} + \frac{1}{6}\mathbf{y}_{n-5} &= h\mathbf{f}_{n+1}
\end{aligned}$$

In this case the first order method is the implicit Euler method.

A variant of the BDF method is the NDF method (Numerical Differentiation Formulas) (Shampine and Reichelt 1997) where an additional term is introduced as follows

$$\sum_{m=1}^{q} \frac{1}{m} \nabla^m \mathbf{y}_{n+1} = h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) + \kappa \sum_{m=1}^{q} \frac{1}{m} \left(\mathbf{y}_{n+1} - \sum_{m=0}^{q} \nabla^m \mathbf{y}_n \right)$$

14.11.5 Linear stability analysis

Multistep methods are of the form

$$\alpha_q \mathbf{y}_{n+1} + \alpha_{q-1}\mathbf{y}_n + \ldots + \alpha_0 \mathbf{y}_{n+1-q} = h\left(\beta_q \mathbf{f}_{n+1} + \beta_{q-1}\mathbf{f}_n + \ldots + \beta_0 \mathbf{f}_{n+1-q}\right)$$

which is written

$$N(z)\mathbf{y}_{n+1} = hD(z)\mathbf{f}_{n+1} \tag{14.282}$$

where

$$\begin{aligned}
N(z) &= \alpha_q z^q + \alpha_{q-1} z^{q-1} + \ldots + \alpha_0 \\
D(z) &= \beta_q z^q + \beta_{q-1} z^{q-1} + \ldots + \beta_0
\end{aligned}$$

and z is viewed as the time-shift operator defined by $z^{-1} y_{n+1} = y_n$.

Consider the linear test system

$$\dot{y} = \lambda y$$

Then the multistep method gives

$$N(z) y_{n+1} = h\lambda D(z) y_{n+1}$$

Introduction of the z transform gives

$$N(z) y(z) = h\lambda D(z) y(z)$$

where z is the complex z transform variable, and $y(z)$ is the z transform of the numerical solution y_{n+1}. Equivalently, this is written

$$\left[N(z) - h\lambda D(z)\right] y(z) = 0 \tag{14.283}$$

The stability of the method can then be investigated by studying the roots of the characteristic equation

$$N(z) - h\lambda D(z) = 0$$

which implies stability of the multistep method if the roots are inside the unit circle. Also the location of the continuous time poles λ can be found as a function of z from

$$h\lambda = \frac{N(z)}{D(z)}$$

This equation makes it possible to find the poles λ that correspond to the limit of stability for the multistep method. The stability limit occurs when $|z| = 1$, which can be parameterized by $z = e^{j\omega}$, $-\pi \leq \omega \leq \pi$. Then the limit of stability in the s plane is found by plotting

$$h\lambda = \frac{N(e^{j\theta})}{D(e^{j\theta})}, \quad -\pi \leq \theta \leq \pi \tag{14.284}$$

14.11.6 Stability of Adams methods

In terms of the z transformation the backwards differences operator ∇ is replaced by $1 - z^{-1}$. This is seen from the z transform of

$$\nabla y_n = y_n - y_{n-1}$$

which gives

$$\mathcal{Z}\{\nabla y_n\} = (1 - z^{-1})y(z)$$

For explicit Adams methods the z transform gives

$$zy(z) = y(z) + h\lambda \sum_{m=0}^{q-1} \gamma_m (1 - z^{-1})^m y(z)$$

This gives

$$h\lambda = \frac{z - 1}{\sum_{m=0}^{q-1} \gamma_m (1 - z^{-1})^m}$$

The regions of stability for methods of order 1 to 4 were computed as in (14.284), and are shown in Figure 14.36. For implicit Adams methods the z transform gives

$$zy(z) = y(z) + h\lambda \sum_{m=0}^{q} \gamma_m^* (1 - z^{-1})^m z y(z)$$

which gives

$$h\lambda = \frac{1 - z^{-1}}{\sum_{m=0}^{q} \gamma_m^* (1 - z^{-1})^m}$$

The stability regions can then be plotted as in equation (14.284). This gives the stability regions shown in Figure 14.37. In the PECE Adams method the solution $\widehat{y}_{n+1}$ of the explicit Adams method is inserted for y_{n+1} on the right hand side of the implicit method.

$$zy(z) = y(z) + h\lambda \left\{ \gamma_0^* z\widehat{y}(z) + \gamma_1^* \left[z\widehat{y}(z) - y(z) \right] + \gamma_2^* \left[z\widehat{y}(z) - 2y(z) + z^{-1}y(z) \right] + \ldots \right\}$$

After some calculation it can be established that $h\lambda$ satisfies the second order equation

$$\begin{aligned} A(h\lambda)^2 + Bh\lambda + C &= 0 \\ A &= \left(\sum_{m=0}^{q} \gamma_m^* \right) \left[\sum_{m=0}^{q-1} \gamma_m (1 - z^{-1})^m \right] \\ B &= (1 - z) \sum_{m=0}^{q} \gamma_m^* + z \sum_{m=0}^{q} \gamma_m^* (1 - z^{-1})^m \\ C &= 1 - z \end{aligned}$$

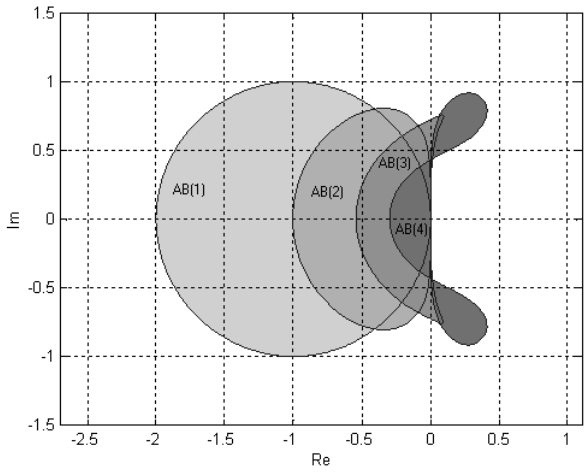

Figure 14.36: The stability regions of the explicit Adams (Adams-Bashforth) methods of order 1 to 4. We recognize the stability area of AB(1) as that of Euler's method.

14.11.7 Stability of BDF methods

For BDF, z transformation gives

$$\sum_{m=1}^{q} \frac{1}{m}(1-z^{-1})^m y(z) = h\lambda y(z)$$

and it follows that

$$h\lambda = \sum_{m=1}^{q} \frac{1}{m}(1-z^{-1})^m$$

The stability areas are found by plotting $h\lambda$ for $z = e^{j\theta}$, $-\pi \leq \theta \leq \pi$, and are shown in Figure 14.38. It is seen that both the first order and the second order BDF are stable for $\dot{y} = \lambda y$ whenever $\mathrm{Re}(\lambda) \leq 0$.

14.11.8 Frequency response

From

$$[N(z) - h\lambda D(z)]\, y(z) = 0 \tag{14.285}$$

the dynamics of the numerical solution of $\dot{y} = \lambda y$ can be analyzed in the z plane as a function of $h\lambda$. We recall that if there is a z_p so that

$$N(z_p) - h\lambda D(z_p) = 0 \tag{14.286}$$

then the dynamics of $y(z)$ have a pole in the z plane at $z = z_p$. If $z_p = 0$, then this gives a one-step reset response where $y_{n+1} = 0$. If $z_p = 1$, then we have the dynamics of an integrator where $y_{n+1} = y_n$.

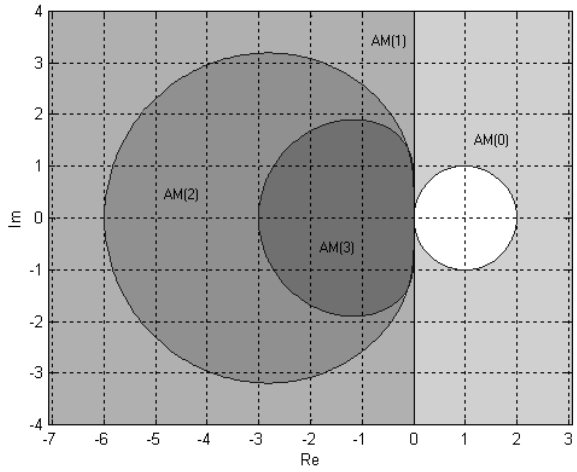

Figure 14.37: Stability areas of implicit Adams (Adams-Moulton) methods of order 1 to 4. The methods are denoted by AM(q) where $q+1$ is the order of the method. Note that AM(1) is the implicit Euler method, and AM(2) is the trapezoidal rule.

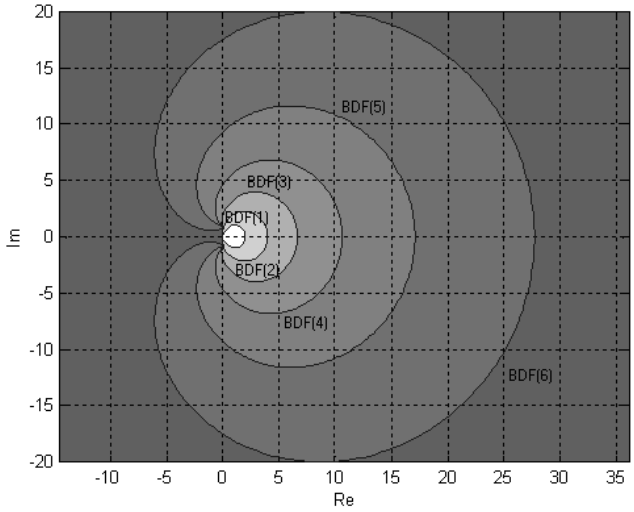

Figure 14.38: Stability areas for BDF methods of order 1 to 6. We note that BDF1 is the implicit Euler method.

14.11.9 Adams methods

Explicit Adams methods have the dynamics

$$\left[z - 1 - s\sum_{m=1}^{q} \gamma_m (1 - z^{-1})^m\right] y(z) = 0$$

while implicit Adams methods have the dynamics

$$\left[z - 1 - s\sum_{m=0}^{q} \gamma_m^* (1 - z^{-1})^m z\right] y(z)$$

It is seen that for $s = 0$ both methods have one pole which is at $z = 1$. Moreover, when $|s| \to \infty$, the explicit method has poles defined by

$$\sum_{m=1}^{q} \gamma_m (1 - z^{-1})^m = 0$$

Clearly, at least one of the poles for $|s| \to \infty$ is at $z = 1$, which is also the case for the implicit methods. This means that high frequency modes are not damped out in the Adams methods.

14.11.10 BDF methods

When a BDF method is applied to the test equation $\dot{y} = \lambda y$ we have the expression

$$\left[\sum_{m=1}^{q} \frac{1}{m}(1 - z^{-1})^m - \lambda h\right] y(z) = 0$$

which shows that when $\lambda h = 0$, there is a pole at $z = 1$.

The expression

$$\alpha_q y_{n+1} + \alpha_{q-1} y_n + \ldots + \alpha_0 = \lambda h y_{n+1} \tag{14.287}$$

leads to

$$(\alpha_q - \lambda h)\, y_{n+1} + \alpha_{q-1} y_n + \ldots + \alpha_0 = 0 \tag{14.288}$$

We see that when $\lambda h \to \infty$, then $y_{n+1} \to 0$. In the z transform the result is found from

$$\left[(\alpha_q - \lambda h)\, z^q + \alpha_{q-1} z^{q-1} + \ldots + \alpha_0\right] y(z) = 0 \tag{14.289}$$

where it is seen that when $\lambda h \to \infty$ the dynamics tend to $z^q y(z) = 0$ which is q poles at the origin of the z plane. This means that dynamics corresponding to $\lambda h \gg 1$ are damped out, and because of this the BDF methods are well suited for stiff systems. The standard MATLAB integrator for stiff systems is ode15s which is a variable order BDF solver (Shampine and Reichelt 1997).

14.12 Differential-algebraic equations

Consider the system

$$\mathbf{M}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u})$$

where $\mathbf{u} \in R^d$ and $\mathbf{M}$ is a square matrix of dimension $d \times d$. To begin with we assume that $\mathbf{M}$ has the simple form

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{d_1} & \mathbf{0} \\ \mathbf{0} & \epsilon \mathbf{I}_{d_2} \end{pmatrix} \tag{14.290}$$

where $d_1 + d_2 = d$ and ϵ is a constant. It is seen that $\mathbf{M}$ is singular whenever $\epsilon = 0$.

Let

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{u}, \; \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \boldsymbol{\phi}$$

where $\mathbf{y}, \mathbf{f} \in R^{d_1}$ and $\mathbf{z}, \mathbf{g} \in R^{d_2}$. Then the system can be written in the form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}) \tag{14.291}$$
$$\epsilon \dot{\mathbf{z}} = \mathbf{g}(\mathbf{y}, \mathbf{z}) \tag{14.292}$$

It is seen that if $\epsilon \neq 0$ then the system is of order d, and is described by the differential equations above, while for $\epsilon = 0$ the system is of order d_1 and is described by the *differential-algebraic equation*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{z}) \tag{14.293}$$
$$\mathbf{0} = \mathbf{g}(\mathbf{y}, \mathbf{z}) \tag{14.294}$$

If

$$\frac{\partial \mathbf{g}(\mathbf{y}, \mathbf{z})}{\partial \mathbf{z}} = \left\{ \frac{\partial g_i}{\partial z_j} \right\}$$

is nonsingular the differential algebraic equation is said to be of index 1. It is then possible to solve $\mathbf{z}$ from $\mathbf{0} = \mathbf{g}(\mathbf{y}, \mathbf{z})$ giving

$$\mathbf{z} = \mathbf{z}(\mathbf{y})$$

and the dynamics of the system can be written

$$\dot{\mathbf{y}} = \mathbf{f}\left[\mathbf{y}, \mathbf{z}(\mathbf{y})\right] \tag{14.295}$$

The system (14.295) can be solved with any numerical integration scheme, and the algebraic condition is automatically satisfied.

However, in some cases it is desirable to leave the system in the original form and let ϵ tend to zero. in particular, this is done if there is no explicit solution $\mathbf{z} = \mathbf{z}(\mathbf{y})$ available, or that the system is in the form $\mathbf{M}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u})$ where $\mathbf{M}$ is possibly nonsingular.

The system

$$\mathbf{M}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u})$$

is said to be a differential algebraic equation of index 1 if it can be transformed in to a index 1 system as defined above by a change of variables.

14.12.1 Implicit Runge-Kutta methods for index 1 problems

An implicit Runge-Kutta method for the system (14.291, 14.292) is given by

$$\begin{aligned}
\mathbf{Y}_i &= \mathbf{y}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{f}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\epsilon\mathbf{Z}_i &= \epsilon\mathbf{z}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{i=1}^{\sigma} b_i\mathbf{f}(\mathbf{Y}_i, \mathbf{Z}_i) \\
\epsilon\mathbf{z}_{n+1} &= \epsilon\mathbf{z}_n + h\sum_{i=1}^{\sigma} b_i\mathbf{g}(\mathbf{Y}_i, \mathbf{Z}_i)
\end{aligned}$$

We will now show how this scheme can be reformulated so that the equation for $\mathbf{z}_{n+1}$ does not include ϵ. This is done by solving $\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j)$ from the equation for $\epsilon\mathbf{Z}_i$, which gives

$$h\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) = \epsilon\sum_{j=1}^{\sigma} \omega_{ij}\left(\mathbf{Z}_i - \mathbf{z}_n\right)$$

where $\mathbf{A}^{-1} = \mathbf{\Omega} = \{\omega_{ij}\}$. This expression is inserted into the equation for $\epsilon\mathbf{z}_{n+1}$, and the result is

$$\begin{aligned}
\epsilon\mathbf{z}_{n+1} &= \epsilon\mathbf{z}_n + \epsilon\sum_{i=1}^{\sigma} b_i \sum_{j=1}^{\sigma} \omega_{ij}\left(\mathbf{Z}_i - \mathbf{z}_n\right) \\
&= \epsilon\left(1 - \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\right)\mathbf{z}_n + \epsilon\sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{Z}_i
\end{aligned}$$

We note that ϵ may be cancelled from this equation, and recall from (14.156) that

$$R(\infty) = 1 - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{1} = 1 - \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij} \tag{14.296}$$

This leads to the expression

$$\mathbf{z}_{n+1} = R(\infty)\mathbf{z}_n + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{Z}_i$$

which can be used to compute $\mathbf{z}_{n+1}$, and we get a reformulation of the Runge-Kutta method in the form

$$\begin{aligned}
\mathbf{Y}_i &= \mathbf{y}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{f}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\epsilon\mathbf{Z}_i &= \epsilon\mathbf{z}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{i=1}^{\sigma} b_i\mathbf{f}(\mathbf{Y}_i, \mathbf{Z}_i) \\
\mathbf{z}_{n+1} &= R(\infty)\mathbf{z}_n + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{Z}_i
\end{aligned}$$

Note that ϵ only appears in the equation for $\epsilon\mathbf{Z}_i$. If we let ϵ go to zero, the Runge-Kutta method becomes

$$\begin{aligned}
\mathbf{Y}_i &= \mathbf{y}_n + h\sum_{j=1}^{\sigma} a_{ij}\mathbf{f}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\mathbf{0} &= \sum_{j=1}^{\sigma} a_{ij}\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) \\
\mathbf{y}_{n+1} &= \mathbf{y}_n + h\sum_{i=1}^{\sigma} b_i\mathbf{f}(\mathbf{Y}_i, \mathbf{Z}_i) \\
\mathbf{z}_{n+1} &= R(\infty)\mathbf{z}_n + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{Z}_i
\end{aligned}$$

The following observations for the case $\epsilon = 0$ are important. At each stage the algebraic equation $\mathbf{g}(\mathbf{Y}_j, \mathbf{Z}_j) = \mathbf{0}$ is satisfied because $\mathbf{A}$ is nonsingular. The algebraic condition is not necessarily satisfied for $\mathbf{z}_{n+1}$. If $R(\infty) = 0$, we get

$$\mathbf{z}_{n+1} = \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{Z}_i$$

where $\mathbf{z}_{n+1}$ is a linear combination of stages $\mathbf{Z}_i$. Still the algebraic equations

$$\mathbf{g}(\mathbf{y}_{n+1}, \mathbf{z}_{n+1}) = \mathbf{0}$$

are not necessarily satisfied. However, if the method is stiffly accurate, that is, if it has a nonsingular $\mathbf{A}$ matrix and the last row of $\mathbf{A}$ equals $\mathbf{b}^T$, then $\mathbf{y}_{n+1} = \mathbf{Y}_\sigma$ and $\mathbf{z}_{n+1} = \mathbf{Z}_\sigma$, and as $\mathbf{g}(\mathbf{Y}_\sigma, \mathbf{Z}_\sigma) = \mathbf{0}$ it follows that $\mathbf{g}(\mathbf{y}_{n+1}, \mathbf{z}_{n+1}) = \mathbf{0}$.

To conclude: Suppose that a stiffly accurate Runge-Kutta method is used to solve (14.291,14.292) for $\epsilon = 0$. Then the computed solution will be the same as if the Runge-Kutta method was applied to the system $\dot{\mathbf{y}} = \mathbf{f}[\mathbf{y}, \mathbf{z}(\mathbf{y})]$. The same method can be used for an arbitrarily small ϵ.

For a general possibly singular $\mathbf{M}$ the method is written

$$\begin{aligned}\mathbf{M}\left(\mathbf{U}_i - \mathbf{u}_n\right) &= h\sum_{j=1}^{\sigma} a_{ij}\boldsymbol{\phi}(\mathbf{U}_j)\\ \mathbf{u}_{n+1} &= R(\infty)\mathbf{u}_n + \sum_{i=1}^{\sigma}\sum_{j=1}^{\sigma} b_i\omega_{ij}\mathbf{U}_i\end{aligned}$$

Also in this case the algebraic condition is satisfied for the stages, and also for $\mathbf{y}_{n+1}$ if a stiffly accurate method is used.

14.12.2 Multistep methods for index 1 problems

The BDF and NDF methods are of the form

$$\sum_{m=1}^{q} \alpha_q \mathbf{y}_{n+m-q} = h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$$

when applied to systems of the form

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$$

For index 1 systems in the form

$$\mathbf{M}\dot{\mathbf{u}} = \boldsymbol{\phi}(\mathbf{u})$$

the BDF and NDF method are given by

$$\sum_{m=1}^{q} \alpha_q \left(\mathbf{M}\mathbf{u}\right)_{n+m-q} = h\boldsymbol{\phi}(\mathbf{u}_{n+1})$$

This method works also for singular $\mathbf{M}$. In the case

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

the system can be written

$$\begin{aligned}\dot{\mathbf{y}} &= \mathbf{f}(\mathbf{y}, \mathbf{z})\\ \mathbf{0} &= \mathbf{g}(\mathbf{y}, \mathbf{z})\end{aligned}$$

where

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{u},\ \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \boldsymbol{\phi}$$

Then, BDF and NDF gives

$$\begin{aligned}\sum_{m=1}^{q} \alpha_q \mathbf{y}_{n+m-q} &= h\mathbf{f}(\mathbf{y}_{n+1}, \mathbf{z}_{n+1})\\ \mathbf{0} &= h\mathbf{g}(\mathbf{y}_{n+1}, \mathbf{z}_{n+1})\end{aligned}$$

It is seen that the algebraic condition is satisfied at each time step.